MODELING AND ANALYZING MULTIVARIATE LONGITUDINAL LEFT-CENSORED BIOMARKER DATA

by

Ghideon Solomon Ghebregiorgis

B.S. Mathematics, University of Asmara, Asmara, Eritrea 1998

M.S. Mathematical Statistics, Southern Illinois University,

Carbondale, IL 2003

M.A. Statistics, University of Pittsburgh, Pittsburgh, PA 2005

Submitted to the Graduate Faculty of

Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

ARTS AND SCIENCES

This dissertation was presented

by

Ghideon Solomon Ghebregiorgis

It was defended on

May 9, 2008

and approved by

Lisa Weissfeld, Ph.D., Professor, Department of Biostatistics

Henry W. Block, Ph.D., Professor, Department of Statistics

Leon J. Gleser, Ph.D., Professor, Department of Statistics

Wesly Thompson, Ph.D., Asst. Professor, Department of Statistics

Lan Kong, Ph.D., Asst. Professor, Department of Biostatistics

Dissertation Director: Lisa Weissfeld, Ph.D., Professor, Department of Biostatistics

Copyright © by Ghideon Solomon Ghebregiorgis
2008

MODELING AND ANALYZING MULTIVARIATE LONGITUDINAL LEFT-CENSORED BIOMARKER DATA

Ghideon Solomon Ghebregiorgis, PhD

University of Pittsburgh, 2008

Many medical studies collect biomarker data to gain insight into the biological mechanisms underlying both acute and chronic diseases. These markers may be obtained at a single point in time to aid in the diagnosis of an illness or may be collected longitudinally to provide information on the relationship between changes in a given biomarker as it relates to the course of the illness. While there are many different biomarkers presented in the medical literature there are very few studies that examine the relationship between multiple biomarkers, measured longitudinally, and predictors of interest.

The first part of this dissertation addresses the analysis of multiple biomarkers subject to left-censoring over time. Imputation methods and methods that account for censoring are extended to handle multiple outcomes and are compared and evaluated for both accuracy and efficiency through a simulation study. Estimation is based on a parametric multivariate linear mixed model for longitudinally measured biomarkers. For left censored biomarkers an extension of this method based on MLE is used.

The linear mixed effects model based on a full likelihood is one of the few methods available to model longitudinal data subject to left-censoring. However, a full likelihood approach is complicated algebraically due to the large dimension of the numeric computations, and maximum likelihood estimation can be computationally prohibitive when the data are heavily censored. Moreover, the complexity of the computation increases as the dimension of the random effects in the model increases. The second part of the dissertation focuses on developing a method that addresses these problems. We propose a method based on a

pseudo likelihood function that simplifies the computational complexities, allows all possible multivariate models, and that can be used for any data structure including settings where the level of censoring is high. A robust variance-covariance estimator is used to adjust and correct the variance-covariance estimate. A simulation study is conducted to evaluate and compare the performance of the proposed method for efficiency, simplicity and convergence with existing methods. The proposed methodology is illustrated in the analysis of Genetic and Inflammatory Markers for Sepsis (GenIMS) study conducted at the University of Pittsburgh.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

To God all is possible. I thank God Almighty for granting me strength, wisdom, courage and perseverance to carry out this long and demanding study.

I would like to express my deepest gratitude to my advisor Prof. Lisa Weissfeld for her invaluable insight, expertise, patience and motivation. She has been a great help and an incredible mentor, which I am very thankful for. I am so grateful to have the opportunity to work with Dr. Weissfeld. I would also like to thank my dissertation committee members' professors Henry W. Block, Leon J. Gleser, Wesley K. Thompson and Lan Kong for their valuable suggestions and time.

I would like to extend a special thanks to my family my mother Tibleth, my sisters Luul, Bisrat, Netsanet and Rahwa and my brothers Biniam, Yonas and Tedros for their unfailing love, support and encouragement for many years. Thank you for being a pillar of my strength I would not have done it without your support and encouragement. I love you all dearly and I am very much in debt to you.

I would also like to thank all my friends and my relatives both in Eritrea and in the United States for their friendship, unending motivation, encouragement and support.

Finally I would like to dedicate this dissertation to my late father Solomon Ghebregiorgis, who thought me the value of education right from the start of my early childhood , my late brother Samson Solomon Ghebregiorgis and to all my family and friends for their endless love, continue support and their determination to my success.

1.0 BACKGROUND AND MOTIVATION

1.1 INTRODUCTION

1.1.1 Longitudinal Data Analysis

Longitudinal studies can be defined broadly as studies in which the response of each individual is observed on two or more occasions. The defining feature of longitudinal studies is that measurements of the same individuals are taken repeatedly through time, thereby allowing the direct study of change over time. The primary goal of a longitudinal study is to characterize the change in response over time and the factors that influence change. Longitudinal studies are in contrast to cross-sectional studies, in which a single outcome is measured for each individual. With repeated measures on individuals, one can capture within-individual change. Indeed, the assessment of within-subject changes in the response over time can only be achieved within a longitudinal study design. For example, in a cross-sectional study one can only obtain estimates of between-individual differences in the response. That is, a cross-sectional study may allow comparisons among sub-populations that happen to differ in age, but it does not provide any information about how individuals change during the corresponding period.

In the health sciences, longitudinal studies play an important role in enhancing our understanding of the development and persistence of disease. Longitudinal studies represent one of the principle research strategies employed in medical and social science research (Goldstein 1979; Nesselroade and Baltes 1979). There is much natural heterogeneity among individuals in terms of how diseases develop and progress. This heterogeneity is due to genetic, environmental, social and behavioral factors. A longitudinal study design permits the discovery

of individual characteristics that can explain these inter-individual difference in changes in health outcomes over time. In summary, the fundamental objective of a longitudinal analysis is the assessment of within-individual changes in the response and the explanation of systematic differences among individuals in their changes.

1.1.2 Overview of Longitudinal Data Models

There are many challenges to the development of statistical methods for the analysis of longitudinal data, which include the computational burden of estimation when multiple outcomes are considered, the complexity of specifying multivariate longitudinal models and the specification of the covariance structure. Parameter estimation can be computationally intensive due to the need for numerical or Monte Carlo simulation methods to evaluate the likelihood of mixed effects regression models. Observations are not, by definition, independent and we must account for the dependence in the data using more sophisticated statistical methods, especially for more sophisticated models that permit more general forms of correlation among the repeated measurements.

The most commonly used techniques to analyze longitudinal data are generalized estimating equations (GEE) and the mixed effects model. The focus of the current study will be on the use of the mixed effects model to address the challenges of longitudinal data analysis. Methods based on the mixed effects model that can be used for both single and multiple outcomes will be proposed.

1.1.3 Mixed Effects Modeling

Traditional analysis of variance methods are of limited use for longitudinal data analysis because of restrictive assumptions concerning the variance-covariance structure of the repeated measures, and also such methods impose model assumptions that are usually not met in observational studies, since the circumstances under which the measurements are collected cannot always fully be controlled. That is, individuals can enter the study at any time, and they can also withdraw from the study at any time, for different reasons. Moreover, not only may individuals be observed for a different number of times, and for a different periods of

time, the interval between observations my be different as well. The univariate analysis of variance (mixed model) assumes that the variance and covariance of the dependent variable across time are equal (*i.e.*, compound symmetry). Alternatively, the multivariate analysis of variance for repeated measures only includes subjects with complete data across time. For these and other reasons, mixed effects regression models, such as the linear and nonlinear mixed effects models, have been used as alternatives and have become popular for modeling longitudinal data.

The basic characteristic of these models is the inclusion of random subject effects into regression models in order to account for the influence of subjects on their repeated observations. These random subject effects thus describe each person's trend across time, and explain the correlational structure of the longitudinal data. Additionally, they indicate the degree of subject variation that exists in the population of subjects.

There are several features that make mixed effects models especially useful in longitudinal research. First, subjects are not assumed to be measured on the same number of time points, thus, subjects with incomplete data across time are included in the analysis. The ability to include subjects with incomplete data across time is an important advantage relative to procedures that require complete data across time for two reasons, first by including all data, the analysis has increased statistical power, second the complete-case analysis may suffer from biases to the extent that subjects with complete data are not representative of the larger population of subjects. Because time is treated as a continuous variable in a mixed effects model, subjects do not have to be measured at the same time points. This is useful for analysis of longitudinal studies where follow-up times are not uniform across all subjects.

Linear mixed effects models are used when the relationship between a longitudinal response variable and its covariance can be expressed via a linear model. The linear mixed effects model introduced by Laird and Ware (1982) can be generally written as:

$$Y_i = X_i\beta + Z_i\gamma_i + \epsilon_i \tag{1.1}$$

$$\epsilon_i \sim N(0, \Sigma_i),\ \gamma_i \sim N(0, G), i = 1, 2, \ldots, N$$

where Y_i and ϵ_i are, respectively, the vectors of responses and measurement errors for the i^{th} subject, β and γ_i are, respectively, the vectors of fixed effects (population parameters) and

random effects (individual parameters), and the X_i and Z_i are the associated fixed effects and random effects design matrices. One can show that the mean and the variance of Y_i are given by $\mathrm{E}(Y_i) = X_i\beta$ and $\mathrm{Var}(Y_i) = Z_i G Z_i^T + \Sigma_i$, respectively.

1.2 MOTIVATION AND ORGANIZATION OF THE DISSERTATION

1.2.1 Motivation

Many medical studies collect biomarker data to gain insight into the biological mechanisms underlying disease. These markers may be obtained at a single point in time to aid in diagnosis or may be collected longitudinally to provide information on the relationship between changes in biomarkers and the course of disease. There are many challenges to the development of statistical methods for these analyses which include the handling of left truncated data due to the sensitivity of assays, the complexity of specifying multivariate longitudinal models and the computational burden of estimation when multiple biomarkers are considered.

These issues have been of importance for the analysis of data from the Genetic and Inflammatory Markers of Sepsis (GenIMS) study. This study conducted at the University of Pittsburgh enrolled 2320 subjects with community acquired pneumonia from the emergency departments of 28 hospitals in southwestern Pennsylvania, Connecticut, Michigan and Tennessee between 2001 and 2003. One major goal of this study is to understand the role of inflammation in the development of sepsis within this cohort. To this end, a battery of inflammatory markers were measured throughout the course of hospitalization. In addition, septic patients were also identified and all subjects were followed for a period of 1 year to assess mortality. These measurements are to be used to understand the relationship between pro-inflammatory and anti-inflammatory markers in sepsis creating the need for methods that can adequately model multiple longitudinal markers simultaneously.

1.2.2 Overview and Organization of the Dissertation

1.2.2.1 Overview: This research focuses on the development of methodological and theoretical methods that address the challenges discussed in the previous sections. These methods are being developed in response to methods needed for the analysis of the biomarker data collected in the GenIMS study described above.

The research can be divided in two parts. The first part focuses on evaluating and comparing existing methods for modeling and analyzing truncated longitudinal data using simulation techniques. We extend the method of maximum likelihood (ML) that accounts for the loss of information due to censoring to handle multiple outcomes studied simultaneously. These methods are compared in a simulation study in order to highlight the strengths and weaknesses of the current approaches for this problem.

With the use of the maximum likelihood methods in the simulation study it became apparent that the large number of nuisance parameters meant that the estimation problem is complicated algebraically and computationally prohibitive as it involves numerical complexities that require higher dimensional integrations. Moreover, these methods are limited by the data structure studied and the covariance structure used. The goal of the second part of this research is to develop a method that addresses these and other convergence related problems. We propose a method based on pseudo maximum likelihood estimation, which divides the estimation procedure into two steps. We evaluate and compare the performance of the proposed method with the existing methods through empirical findings and simulation study.

1.2.2.2 Organization: The proposal is organized as follows, in chapter 2 we discuss methods for modeling longitudinal data subject to left-censoring, an extensive study is carried out to evaluate and compare these methods through simulation study. Naive methods and methods that account for censoring have been evaluated and compared. In section 2.3 we extend the maximum likelihood method to a multivariate model and a simulation study is conducted to assess the performance of the extended method.

Chapter 3 discusses the proposed pseudo maximum likelihood method. In section 3.2 we

present the proposed methodology, the multivariate linear mixed model is given in section 3.2.1 and the pseudo-likelihood for left-censored data is developed in section 3.2.2. Computational details are given in section 3.2.3. A simulation study, conducted to assess the performance of the proposed model, is summarized in section 3.3 and the method is applied to the GenIMS data in section 3.4. Results of this method are compared with results obtained using existing methods. Brief discussion of the results and concluding remarks are given in section 3.5.

In chapter 4 we discuss the findings and results of the research and direction for future study is given.

2.0 MODELING MULTIVARIATE LONGITUDINAL DATA SUBJECT TO LEFT-CENSORING

2.1 INTRODUCTION

In medical sciences, studies are often designed to investigate changes in one or more variables which are measured repeatedly over time in the participating subjects. Many statistical models have been proposed for the analysis of one single outcome. The analysis of multiple outcomes, measured longitudinally, is often restricted to the analysis of each response separately. However, research questions can often only fully be addressed in a joint analysis of all outcomes simultaneously. For example, the association structure can be of direct scientific relevance. A possible question might be how the association between outcomes evolves over time or how outcome-specific evolutions are related to each other. Interest may be in the comparison of average trends for different outcomes. For example, consider testing the difference in evolution between many outcomes or joint testing of a treatment effect on a set of outcomes. All of these situations require a joint model for all outcomes. However, computational problems are likely to occur when the number of outcomes increases and the complexity of specifying multivariate longitudinal models could be a challenge. Another challenge to the development of statistical methods for this type of data include the handling of left truncated data due to the sensitivity of assays.

In this chapter we will evaluate existing methods for the analysis of one single outcome to longitudinal data subject to left truncation and we will propose an extension of the methods for a joint analysis of multiple outcomes simultaneously.

Several approaches have been proposed in the statistics literature for the analysis of longitudinal left censored data and all approaches differ in sophistication when handling the

truncated values (measures). Ad hoc (Naive) procedures to replace the censored measures have been used recently to adjust for a censored value. Replacing the censored value by the lower detection limit (Keet et.al, 1997), or by half of the detection limit (O'Brien et.al, 1998, Hornung and Reed , 1990) are the most frequently used approaches. Alternatively Paxton et al.(1997) used an iterative two-stage imputation procedure to replace the censored measures where they substitute the censored value with half of the lower detection limit in the first stage then the model is refitted again by imputing the new estimated values. These methods are convenient to use but they ignore the correlated structure of the data and did not adjust the standard errors of the parameter estimates for the loss of information due to censoring.

In addition to imputation of the value of the detection limit, half of the limit or use of a multiple imputation technique, methods that handle left-censoring directly have been proposed. These procedures differ in the methodological approaches they follow. Procedures based on fitting mixed effect linear models include maximum likelihood (ML) methods and Bayesian methods based on modes of posterior distributions (Carriquiry et al., 1987). Likelihood based methods include; Hughes (1999), Jacqmin-Gadda et al.(2000) and Lyles et al. (2000). Hughes (1999) modified the usual EM estimation procedure for the mixed effects model to account for left censoring. The method uses a Monte Carlo procedure to provide a general solution that can be used with left-censored data and since the expectation step of the EM algorithm is intractable, the Gibbs sampler is used to implement a Monte Carlo expectation step in the EM algorithm. Jacqmin-Gadda et. al (2000) proposed an approach of direct maximization of the likelihood without the EM or the Monte Carlo methods where maximization is based on the Marquardt algorithm. The Lyles et al. (2000) approach is based on a hierarchical formulation of the likelihood, their approach combines those of Hughes (1999) and Schluchter (1992) into a single likelihood, and estimation is carried out by direct maximization of the likelihood using built-in algorithms in SAS.

While these methods address the challenges of truncated longitudinal data, they fail to handle multivariate truncated data when multiple outcomes (markers) are considered. In this study we modified and extended the direct maximization method proposed by Jacqmin-Gadda et al. (2000) to model left-censored data with multiple outcomes accounting for the loss of information due to censoring. For the univariate case, we evaluated and compared

methods for modeling longitudinal data accounting for censored repeated measures, and the naive methods that use simple imputation to replace the censored value.

2.2 ANALYSIS FOR A SINGLE OUTCOME

2.2.1 Single outcome linear mixed model

Mixed effects models or random effects models have proven to be powerful tools for longitudinal data analysis. Here we used the standard linear mixed effects model proposed by Laird and Ware (1982). First we introduce the following notation which will be used throughout the chapter.

Let Y_{ij} denote the response variable for the i^{th} subject on the j^{th} measurement occasion. Subjects may not have the same number of repeated measures and may not be measured at the same set of occasions. To accommodate both of these features, we assume that there are n_i repeated measurements of the response on the i^{th} subject and that each Y_{ij} is observed at time t_{ij}. For convenience we group the response variable and the times of observation for the i^{th} subject into an $n_i \times 1$ vector

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix} \qquad t_i = \begin{pmatrix} t_{i1} \\ t_{i2} \\ \vdots \\ t_{in_i} \end{pmatrix} \qquad i = 1, 2, \ldots, N.$$

The vector Y_i is simply a time-ordered collection of the n_i response variables for the i^{th} subject. The Y_{ij}'s are often called the components, entries, or elements of Y_i.

The linear mixed effects model is defined (Laird and Ware, 1982) as:

$$Y_i = X_i\beta + Z_i\gamma_i + \epsilon_i, \tag{2.1}$$

where β is a p dimensional vector of fixed effects and X_i is the corresponding $n_i \times p$ design matrix; γ_i is a q dimensional vector of random effects, Z_i is an $n_i \times q$ design matrix of random

effects usually a sub-matrix of X_i and ϵ_i is n_i dimensional vector of measurement errors. The random effects $\gamma_1, \ldots, \gamma_N$ and errors $\epsilon_1, \ldots, \epsilon_N$ are assumed to be mutually independent with

$$\epsilon_i \sim N(0, \Sigma_i), \quad \gamma_i \sim N(0, G),$$

where G is a $(q \times q)$ covariance matrix of random effects with (i, j) element $g_{ij} = g_{ji}$ and Σ_i is $(n_i \times n_i)$ covariance matrix which depends on i only through its dimension n_i, i.e. the set of unknown parameters in Σ_i will not depend on i.

From (2.1) it follows that, conditional on the random effect γ_i, Y_i is normally distributed with mean vector $X_i\beta + Z_i\gamma_i$ and with covariance matrix Σ_i. Let $f(y_i|\gamma_i)$ and $f(\gamma_i)$ denote the corresponding density functions, then we can obtain the marginal density function of Y_i by

$$f(y_i) = \int f(y_i|\gamma_i) f(\gamma_i) d\gamma_i,$$

which can be shown to be the density function of a n_i dimensional normal distribution with mean vector $X_i\beta$ and with covariance matrix $V_i{=}Z_iGZ_i^{'} + \Sigma_i$. This marginal distribution of the response Y_i is used to estimate model parameters of the linear mixed effects model (2.1).

The marginal model

$$Y_i \sim N(X_i\beta, Z_iGZ_i^{'} + \Sigma_i), \tag{2.2}$$

does not imply that Y_i satisfies the hierarchical model (2.1). Hence, we do not explicitly assume the presence of random effects representing the natural heterogeneity between subjects, when we use the marginal distribution for inference. A simple example of the difference between marginal and hierarchical models is given in Verbeke and Molenberghs (1997). The classical approach to inference for the linear mixed effects model is based on estimators obtained from maximizing the marginal likelihood function.

Let η denote the vector of all variance and covariance parameters (simply called variance components) found in V_i, i.e. η consists of the $q(q+2)/2$ different elements in G and of all parameters in Σ_i. Let $\theta{=}\ (\beta^{'}, \eta^{'})^{'}$ denote the vector of all parameters in the marginal model (2.2) for Y_i, then the likelihood function with respect to θ, is given by

$$L(\theta) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{n_i/2}|V_i(\eta)|^{-1/2}} \times exp\left(-\frac{1}{2}(Y_i - X_i\beta)^{'}V_i^{-1}(\eta)(Y_i - X_i\beta)\right). \tag{2.3}$$

2.2.2 Likelihood function for censored data

Assume there are n_i^c censored measurements, and let c_i denote the censoring threshold for subject i (i.e. c_i=Lower detection limit(LOD)). Also let Y_i^o denote the n_i^o-vector of observed outcomes, Y_i^c the n_i^c-vector of censored outcomes for subject i, where $n_i = n_i^o + n_i^c$. Specifically, we define the observed and censored data as:

$$Y_i^o = \begin{cases} Y_i & \text{if } Y_i > c_i \\ c_i & \text{if } Y_i \leq c_i \ , \end{cases}$$

and the censored data Y_i^c= $\{Y_i \mid Y_i \leq c_i \}$. Inference is based on incorporating this into the likelihood function and different maximization techniques are used according to a method used.

2.2.2.1 Ad hoc approaches Estimation using the naive methods is carried out by simply replacing the censored value by fixed value. One of the most commonly used replacement methods is to substitute each non-detected value by half of its detection limit. Other commonly used replacement values are zeros or the detection limits. To avoid clumping of replaced values in cases where there are several non-detect values that share a common detection limit, values may be spaced evenly from 0 to the detection limit or according to some specified probability distributions.

In this study we consider three of the most frequently used approaches, the method that replaces the censored values with the lower detection limit (LOD), the approach of replacing the censored values with half of the lower detection limit (HLD), and a random imputation method where censored values are replaced by values from a random distribution. All these methods use the complete data likelihood after censored values are replaced by fixed values, and estimation is based on maximizing the complete likelihood in (2.3). We evaluated and compared these methods, with other methods that account for censoring, through empirical findings and simulation studies in sections 2.4 to 2.6.

2.2.2.2 Inference taking censoring into account One of the goals of the current study is to develop statistical techniques that may be employed to handle censored data and

account for the loss of information due to censoring. In this section we derive a likelihood function for the linear mixed model in (2.1) for a longitudinal left-censored data that will be used to estimate model parameters.

Reordering the matrix X_i, vector Y_i, and the covariance matrix $V_i(\eta)$ (written as V_i for simplicity) we can partition them into observed and censored components as

$$X_i = \begin{bmatrix} X_i^o \\ X_i^c \end{bmatrix}, \qquad Y_i = \begin{bmatrix} Y_i^o \\ Y_i^c \end{bmatrix}, \qquad V_i = \begin{bmatrix} V_i^o & V_i^{co} \\ V_i^{co'} & V_i^c \end{bmatrix}.$$

Using these partitions and model (2.3), Y_i^o has a multivariate normal probability density f_i^o with mean ${\mu_i}^o = X_i^o\beta$ and covariance V_i^o. Then the likelihood function of the censored data in terms of the vector of parameters θ is:

$$L(\theta) = \prod_{i=1}^{N} f_{Y_i^o|\theta}(Y_i^o|\theta) P(Y_i^c < c_i|Y_i^o, \theta), \tag{2.4}$$

that is the contribution to the likelihood for a censored value is the probability that the true outcome is below the threshold.

Using multivariate normal distribution properties, the conditional distribution of the censored, Y_i^c, given the observed, Y_i^o, is normally distributed with mean $\mu_i^{c|o}$ and variance, $V_i^{c|o}$, given by the following expressions respectively:

$$\begin{aligned} \mu_i^{c|o} &= X_i^c\beta + V_i^{co}{V_i^o}^{-1}[Y_i^o - \mu_i^o] \\ V_i^{c|o} &= V_i^c - V_i^{co}{V_i^o}^{-1}V_i^{co'}. \end{aligned}$$

Then the likelihood function of the left-censored data in (2.4) can be rewritten as:

$$L(\theta) = \prod_{i=1}^{N} f_{Y_i^o|\theta}(Y_i^o|\theta)\Phi_i^{c|o}(c_i|\theta), \tag{2.5}$$

where $\Phi_i^{c|o}(.)$ denote the conditional multivariate cumulative normal distribution of Y_i^c given Y_i^o. Using the density and cumulative distribution function of the normal distribution (2.5) can be further simplified to:

$$L(\theta) = \prod_{i=1}^{N} \frac{1}{(2\pi|V_i^o|)^{1/2}} exp\left[\frac{-1}{2}(Y_i^o - \mu_i^o)' V_i^{o^{-1}}(Y_i^o - \mu_i^o)\right]$$
$$\int_{-\infty}^{c_i 1}\int_{-\infty}^{c_i 2}\cdots\int_{-\infty}^{c_i n_i^c} \frac{1}{(2\pi|V_i^{c|o}|)^{1/2}} exp\left[\frac{-1}{2}(u - \mu_i^{c|o})' V_i^{c|o^{-1}}(u - \mu_i^{c|o})\right] du, \qquad (2.6)$$

where u is an n^c vector. Taking the logarithm of (2.5), we obtain the log-likelihood function as:

$$\ell(\theta) = \sum_{i=1}^{N} -log(2\pi) - \frac{1}{2}log|V_i^o| - \frac{1}{2}(Y_i^o - \mu_i^o)' V_i^{o-1}(Y_i^o - \mu_i^o)$$
$$+ log\left[\int_{-\infty}^{c_i 1}\int_{-\infty}^{c_i 2}\cdots\int_{-\infty}^{c_i n_i^c} \frac{1}{2\pi|V_i^{c|o}|^{1/2}} exp\left[\frac{-1}{2}(u - \mu_i^{c|o})' V_i^{c|o^{-1}}(u - \mu_i^{c|o})\right] du\right]. \qquad (2.7)$$

A FORTRAN program developed by Jacqmin-Gadda et al. (2000) is modified to estimate model parameters by directly maximizing the likelihood in (2.7). In maximizing this likelihood, a numerical computation of the integral of a multivariate normal density for each subject with censored measures is required. We performed this computation by combining a FORTRAN routine developed by Genz (1992) and a subregion adaptive multiple integration method by Berntsen et al. (1991). This simplifies the computation and places it into a form that allows efficient calculation using standard numerical multiple integration algorithms. The algorithm uses a sequence of three transformations that transform the original integral into an integral over a unit hypercube, reorders the integration variables and then applies a subregion adaptive multiple integration method.

The likelihood was re-parameterized in terms of the square root for the variance of the measurement error terms and the Cholesky decomposition of the covariance matrix of the random effects G ($G = C'C$ where C is an upper triangular matrix) is used in order to impose positive constraints for the covariance parameters. To reduce computation time for models with many covariance parameters, two optimization algorithms are combined. The first iterations were done using the simplex algorithm and then the Marquardt algorithm (Marquardt, 1963) was used near the optimum.

Instead of direct maximization of the likelihood in (2.7), Hughes (1999) proposed the use of the EM algorithm treating γ_i and Y_i^c as missing data. The M-step is carried out by

computing the expectation of the complete data sufficient statistics and maximization of the likelihood of complete data (i.e. the likelihood assuming that γ_i and Y_i^c are observed) replacing the sufficient statistics by their expectations. However, the E-step requires evaluation of integrals which are intractable by classical methods leading to MCEM algorithm. The MCEM algorithm has convergence problems and leads to more biased estimates, moreover, it only works for models with simple covariance structure (Jacqmin-Gadda et al., 2000). The method is restricted to single outcomes and can not be extended to a multivariate model.

2.3 ANALYSIS FOR MULTIPLE OUTCOMES

2.3.1 Multivariate linear mixed model

Mixed models are widely used for the analysis of a singleone repeatedly measured outcome. If more than one outcome are present, a mixed model can be used for each one. These separate models can be tied together into a multivariate mixed model by specifying a joint distribution for their random effects.

Let k be the number of outcomes in the model, it will be assumed that each of the k longitudinally measured outcomes can be modeled using the mixed model. For subject i and outcome k, we observe the n_{ik} vector of measurements:

$$Y_{ik} = \begin{pmatrix} y_{i1k} \\ y_{i2k} \\ \vdots \\ y_{ijk} \\ \vdots \\ y_{in_{ik}k} \end{pmatrix} ,$$

where y_{ijk} is the measure of subject i at occasion j for marker k. The number and times of measurements may be completely different for each subject and each outcome.

For simplicity we present the bivariate case (k=2) in this section. Let y_{ij} denote the j^{th} measure at time t_{ij} for subject i ($i = 1, \ldots, N$ and $j = 1, \ldots, n_i$) for a single outcome, and let

Y_i denote the vector of response of all measurements for subject i, i.e. $Y_i^{'} = (y_{i1}, \ldots, y_{in_i})$. Let $Y_i = \begin{bmatrix} Y_i^1 \\ Y_i^2 \end{bmatrix}$ denote the response vector for subject i, for i=1, ..., N and Y_i^k is the n_i^k vector of measurements of marker k (k=1,2). Let $\beta = \begin{bmatrix} \beta^1 \\ \beta^2 \end{bmatrix}$ be a $p \times 1$ vector of population parameters, known as fixed effects, and X_i be a known $n_i \times p$ design matrix of covariate variables linking β to Y_i. Let $\gamma_i = \begin{bmatrix} \gamma_i^1 \\ \gamma_i^2 \end{bmatrix}$ be a $q \times 1$ vector of subject-specific parameters, known as random effects and Z_i a known $n_i \times q$ design matrix of covariates linking γ_i to Y_i.

We extend the usual linear mixed model (Laird and Ware,1982) to a multivariate model.

$$Y_i = X_i\beta + Z_i\gamma_i + \epsilon_i \tag{2.8}$$

with the assumption that the n_i-dimensional vector Y_i satisfies

$$Y_i|\gamma_i \sim N(X_i\beta + Z_i\gamma_i, \Sigma_i)$$

γ_i and ϵ_i are assumed to be mutually independent with $\epsilon_i \sim N(0, \Sigma_i), \gamma_i \sim N(0, G)$ where Σ_i is the $n_i \times n_i$ covariance matrix of measurement errors, and G is the covariance matrix of the random effects. Σ_i is a diagonal matrix containing the two elements of the measurement error of each marker, that depends on i only through it's dimension n_i. Thus the set of unknown parameters in Σ_i will not depend on i. The covariance matrix of random effects G is the matrix $G = \begin{bmatrix} G^1 & G^{12} \\ G^{12^{'}} & G^2 \end{bmatrix}$, which is partitioned into four sub-matrices: G^1 is the covariance matrix including variance and covariance of random effects for the first marker, G^2 the covariance matrix including variance and covariance of random effects of the second marker and G^{12}= $G^{21^{'}}$ is the matrix of covariances between random effects of each marker. The correlation between the two markers can be studied through the matrix G^{12}. Marginally, the Y_i are independent normals with mean $X_i\beta$ and covariance matrix V_i=Var$(Y_i) = Z_iGZ_i^T + \Sigma_i$.

2.3.2 Likelihood function for Left-Censored data

As before, let $\theta=(\beta^{'},\eta^{'})^{'}$ denote all the unknown parameters in model (2.8), for a likelihood based inference we use the marginal distribution of Y_i. Conditional on the random effect γ_i, Y_i is normally distributed with mean $X_i\beta + Z_i\gamma_i$ and with covariance matrix Σ_i. The likelihood function arising from the marginal normal distribution for Y_i is:

$$L(\theta) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f(Y_i|\gamma_i,\beta) f(\gamma_i) d\gamma_i, \tag{2.9}$$

where $f(Y_i|\gamma_i,\beta)$ and $f(\gamma_i)$ are the normal density functions of the conditional distribution of $Y_i|\gamma_i$ and γ_i respectively.

The ad-hoc approaches replace the censored measures with fixed values (LOD, HLD etc) and thus use complete data, i.e. $n_i{=}n_i^o$. Model parameters in this case are estimated by maximizing the complete-data likelihood in (2.9).

For any of the responses Y_i, assume that there are n_i^o detectable values and thus $n_i^c{=}n_i{-}n_i^o$ non-detectable (censored) values. Using the notation and partitions for censored and observed responses from the previous section, we can show by the properties of the multivariate normal distribution, that the conditional distribution of Y_i^c given Y_i^o is normal with mean ${\mu_i}^{c|o}$ and covariance ${V_i}^{c|o}$. Then, we obtain the likelihood function using the marginal density of Y_i as

$$\begin{aligned} f(y_i) &= \int f(y_i|\gamma_i) f(\gamma_i) d\gamma_i \\ &= \int \left\{ \prod_{j=1}^{n_i^o} f(Y_i^o|\gamma_i) \right\} \left\{ \prod_{j=1}^{n_i^c} P(Y_i^c \le c_i|\gamma_i) \right\} f(\gamma_i) d\gamma_i \\ &= \int \left\{ \prod_{j=1}^{n_i^o} f(Y_i^o|\gamma_i) \right\} \left\{ \prod_{j=1}^{n_i^c} F_{Y_i^c}(c_i|\gamma_i) \right\} f(\gamma_i) d\gamma_i, \end{aligned}$$

where $F_{Y_i^c}$ is the cumulative distribution function (CDF) corresponding to the density of $f(Y_i^c|\gamma_i)$.

The the likelihood function of the vector θ of parameter is then given by

$$L(\theta) = \prod_{i=1}^{N} f(y_i)b = \prod_{i=1}^{N} f_{Y_i^o}(y_i^o|\theta)P(Y_i^c < c_i|Y_i^o = y_i^o, \theta)$$
$$= \prod_{i=1}^{N} f_{Y_i^o}(y_i^o|\theta) \int_{-\infty}^{c_i 1} \int_{-\infty}^{c_i 2} \dots \int_{-\infty}^{c_i n_i^c} f_{Y_i^c|Y_i^o}(u)du, \tag{2.10}$$

where $u = (u_1, u_2, \dots, u_{n_i^c})'$

The log likelihood is then obtained by taking the log of this likelihood function. Using the normal distribution of the observed and the censored data and their respective means and covariance matrix, the log likelihood can be simplified as:

$$\ell(\theta) = \sum_{i=1}^{N} -log(2\pi) - \frac{1}{2}log|V_i^o| - \frac{1}{2}(Y_i^o - \mu_i^o)' V_i^{o-1}(Y_i^o - \mu_i^o)$$
$$+ log\left[\int_{-\infty}^{c_i 1} \int_{-\infty}^{c_i 2} \dots \int_{-\infty}^{c_i n_i^c} \frac{1}{2\pi|V_i^{c|o}|^{1/2}} exp\left[\frac{-1}{2}(u - \mu_i^{c|o})' V_i^{c|o-1}(u - \mu_i^{c|o})\right] du\right]. \tag{2.11}$$

2.4 ESTIMATION

Parameter estimation in the linear mixed effect model typically involves maximum likelihood (ML) or variants of ML. Additionally, the solutions are usually iterative ones that can be numerically quite intensive. Here in this section we discuss estimation of parameters based on maximizing the likelihood functions specified for both univariate and multivariate models using methods that has been frequently used in the statistical literature and the proposed method.

2.4.1 Using Ad-hoc approaches

When data from an assay are left-censored, the lower detection limit (LOD) is known and may be used to substitute a value for the censored observation. An ad-hoc approach for dealing with the left-censored values is to replace them with the LOD value or with half of the LOD (HLD) value.

For our evaluation and comparison of methods we used three of the frequently used ad-hoc methods. The first method replaces the censored value by the lower detection limit,the second replaces censored value with half of the detection limit, while the third uses random imputation to replace the censored value. After substituting censored values, model parameters can be estimated using many existing softwares by maximizing the complete data likelihood given in (2.3) and (2.9) for univariate and multivariate outcomes, respectively. In this study we used the FORTRAN program we modified and extended to multivariate outcomes for parameter estimations. The computational details of the program is given in section 2.2.

2.4.2 Estimation accounting for Censoring

Several methods have been proposed for estimating model parameters handling the censored values, likelihood based methods include Hughes(1999), Jacqmin-Gadda et al. (2000), Lyles et al.(2000) and Thiebaut et. al (2004). However, most of these methods are developed and restricted to a model with a single outcome and can't be extended to a multivariate model. Thiebaut et al proposed a method to estimate parameters of a bivariate model accounting for left-censoring of one or both markers, but the method is limited by the number of random effects used in the model. Sy et al. (1997) used the Fisher scoring method to fit a bivariate linear random effects model including an integrated OrsteinUhlenbeck process (IOU), a stochastic process that includes Brownian motion as special limiting case. Their program is implemented using the IML module of SAS software, however, it is very complicated and it is not sufficiently flexible to allow large use by researchers not familiar with IML (Thiebaut et. al 2004).

We perform estimation of model parameters based on maximum likelihood by directly maximizing the log-likelihood given in (2.11). We extend the FORTRAN program described earlier to maximize the likelihood(2.11). The maximization is based on a Marquardt algorithm (Marquardt, 1963) that is a Newton-Raphson like algorithm where the diagonal of the Hessian matrix is inflated when adapted. To impose a positive constraint of covariance parameters, a new parameterization of the model was used in term of squared root of the

variance of the measurement errors and a Cholesky decomposition of the random effects covariance matrix. Multiple integrals of multivariate normal density as large as the number of censored measures (n_i^c) were numerically calculated using a subregion adaptative multiple integration method developed by Genz (1992). We compare estimates obtained using this method and other methods.

2.5 SIMULATION STUDY

The goal of this section is to evaluate and compare existing methods and proposed method through simulation study. The performance of the methods is evaluated through bias and precision of the estimators.

We generated data according to a linear mixed effects model with a random slope, random intercept and independent error for k markers given by:

$$Y_{ij}^k = \beta_0^k + \beta_1^k t_{ij} + \beta_2^k X_i + \beta_3^k t_{ij} * X_i + \gamma_{0i}^k + \gamma_{1i}^k t_{ij} + \epsilon_{ij}, \tag{2.12}$$

where Y_{ij}, β, γ and ϵ are as defined in previous sections of this chapter and X_i is a binary covariate. Data are simulated as follows, for $i = 1, \ldots, N$ we first generate all the components of the design matrix X. That is, a binary variable randomly generated using Bernoulli(0.5) is assigned as a covariate variable for each subject. For times of measurements random numbers (similar to the real data measurement times) uniformly distributed between 1 and 7 (mean 4) were selected. Censoring values were chosen independently of time and subject for each marker and parameter values were chosen to be close to those obtained from the real data analysis. For the fixed effects parameter the values are set at

$$\beta^1 = \begin{pmatrix} \beta_0^1 \\ \beta_1^1 \\ \beta_2^1 \\ \beta_3^1 \end{pmatrix} = \begin{pmatrix} 3.78 \\ -0.41 \\ -0.11 \\ 0.02 \end{pmatrix},$$

for the first marker and

$$\beta^2 = \begin{pmatrix} \beta_0^2 \\ \beta_1^2 \\ \beta_2^2 \\ \beta_3^2 \end{pmatrix} = \begin{pmatrix} 1.48 \\ -0.15 \\ -0.05 \\ 0.03 \end{pmatrix}$$

for the second marker.

The covariance matrix for the random effects was fixed at

$$G = \begin{pmatrix} 0.17 & 0.05 & -0.66 & -0.17 \\ & 0.02 & -0.16 & -0.11 \\ & & 4.53 & 0.96 \\ & & & 1.14 \end{pmatrix},$$

$\sigma^2_{\epsilon^1}$=1.55 and $\sigma^2_{\epsilon^2}$=0.66. With these specifications we simulated 500,000 measures of 1000 subjects with 30% - 40% censoring rate.

Tables 1-4 summarizes simulation results for data with 31% and 40% censored observations. All results reported are means of 500 replications. In table 1 we present the bias and the MSE of the fixed effects of the linear mixed model (2.12) using the three different ad-hoc methods considered in sections 2.2 and 2.3 and the proposed ML method that accounts for censoring. The LOD method refers to the ad-hoc procedure that replaces the censoring value by the lower detection limit, while HLD refers to the practice of substituting half of LOD for all non-detected values. RI refers to the ad-hoc approach that uses random imputation to replace non-detected values, and for this we generated random numbers from Uniform(0,LOD) to replace the censored value.

As the tables show, the proposed ML method that takes censoring into account produces estimates with significantly smaller bias and MSE compared to the other methods. The procedure that substitutes LOD performed poorly, producing estimates with significant bias and large MSE, and both bias and MSE gets larger with the increase of the censoring rate in the data. On the other hand, the HLD method improves the estimates slightly, but significant bias still exists. The RI method which replaces the censored value using a randomly selected value from Uniform (0,LOD) distribution produces estimates comparable to the HLD method, and improves the estimates slightly than the method of LOD. The ML method that takes censoring into account, given in the last column of the table, removes

Table 1: Simulation Results comparing the performance of the ML approach with the LOD, HLD and RI procedures for fixed effect of the linear mixed effect model. The bias values are obtained from the mean of estimates over 500 simulations.

			Bias			
			Naive Methods			*ML-Method*
% Censored	*Parameter*	*True Value*	*LOD*	*HLD*	*RI*	*Accounting for Cens.*
31	Slope of Time[1]	-0.41	-0.014	-0.014	-0.017	0.001
	Slope of Time[2]	-0.15	-0.120	-0.009	-0.080	-0.015
	Covariate[1]	-0.11	0.012	0.009	0.008	0.007
	Covariate[2]	-0.05	0.019	0.017	0.018	0.005
	Interaction[1]	0.02	-0.002	-0.002	-0.002	-0.001
	Interaction[2]	0.03	-0.013	-0.011	-0.012	0.010
40	Slope of Time[1]	-0.41	-0.057	-0.035	-0.025	0.004
	Slope of Time[2]	-0.15	-0.150	-0.129	-0.100	-0.045
	Covariate[1]	-0.11	-0.049	-0.022	-0.021	-0.018
	Covariate[2]	-0.05	-0.042	-0.023	-0.021	-0.012
	Interaction[1]	0.02	-0.002	-0.002	-0.002	-0.004
	Interaction[2]	0.03	0.064	0.050	0.047	0.026

Table 2: Simulation Results comparing the performance of the ML approach with the LOD, HLD and RI procedures for Variance components of the linear mixed effect model. The bias values are obtained from the mean of estimates over 500 simulations.

			Bias			
			Naive methods			*ML-Method*
% Censored	*Parameter*	*True Value*	*LOD*	*HLD*	*RI*	*Accounting for Cens.*
31	Time^1	0.02	-0.151	-0.161	0.13	-0.003
	Time^2	1.14	-0.131	-0.122	0.128	-0.053
	$\sigma^2_{\epsilon^1}$	1.55	-0.206	-0.201	0.143	-0.004
	$\sigma^2_{\epsilon^2}$	0.66	-0.092	-0.086	0.078	0.002
40	Time^1	0.02	-0.215	-0.218	0.209	0.032
	Time^2	1.14	-0.321	-0.311	0.254	0.100
	$\sigma^2_{\epsilon^1}$	1.55	-0.255	-0.198	0.117	0.132
	$\sigma^2_{\epsilon^2}$	0.66	-0.175	-0.148	0.089	0.026

much of the bias and reduces the MSE significantly. For example for the slope of time for the data with 40% censoring rate LOD produces an estimate with -0.057 bias for the first marker, whereas the ML method reduces this bias by more than 100% to 0.004. The same is true for second marker (-0.15 using LOD and -0.045 using ML)

Generally the performance of all the methods deteriorates when the amount of censoring was increased to 40%, but their relative performance did not alter substantially. Similar results were observed, the ML method performs better than all the other methods with significantly less bias and MSE.

Table 2 summarizes covariance parameter estimates. Similar trends between the different methods were also observed for the covariance parameters estimates, substitution LOD performed poorly underestimating the parameters with large bias, HLD and RI slightly improve the estimates and perform better. No appreciable difference between HLD and RI was noticed and the ML method remained to be the best producing estimates with significantly less bias and MSE than the other three methods. As it can be seen from table 2 the ML procedure not only accounts for the loss of information due to censoring but also reduced the bias by more than half for all the estimates, for example, the bias for the slope of time for the first marker is -0.151, -0.161, and 0.130 using LOD, HLD and RI respectively, while the ML method produce estimates with only -0.003 bias.

2.6 APPLICATION TO GENIMS DATA

We applied the proposed method and the other three methods to a data set that motivates our study, the Genetic and Inflammatory Markers for Sepsis study (GenIMS) data. We illustrated how inference can be affected if appropriate measures are not taken to adjust for the loss of information due to truncation. The GenIMS data is a longitudinal data set collected for the Genetic and Inflammatory Markers of Sepsis study at the University of Pittsburgh. One major aim of the study was to examine the relationship between a set of inflammatory markers and to determine if changes in these markers over time were related to mortality and/or development of sepsis.

Table 3: Descriptive statistics for cytokines.

	Raw Data		*Log Transformed Data*	
	IL-6	*IL-10*	*IL-6*	*IL-10*
N	1797	1797	1797	1797
Mean	140.38	12.23	2.65	1.92
SD	1736.85	42.83	1.69	0.71
(Min, Max)	(2, 126,000)	(5, 1519)	(0.69, 11.74)	(1.61, 7.33)

2.6.1 Data Description

Twenty eight hospital from south western Pennsylvania, Connecticut, Michigan and Tennessee participated in the study enrolling a total of 2320 patients from 2001 to 2003. Patient eligibility criteria included being at least 18 years old and having both a clinical diagnosis of pneumonia and a new pulmonary infiltrate on chest x-ray. During a patient's stay in the hospital, blood was drawn for cytokine assays at enrollment, and on days 2-7. The biomarkers of greatest interest are IL-6 (interleukin-6), which is pro-inflammatory, and IL-10 (interleukin-10), which is anti-inflammatory. Baseline inflammatory marker samples were collected for 1797 subjects. Septic patients were also identified and all subjects were followed for a period of one year to assess mortality. Descriptive statistics for the data are given in table 3.

The detectable limit for both IL-6 and IL-10 was 5, indicating that the concentration of the sample for these markers was below the detectable limit. Both of these markers are measured for 7 days with the censoring rate increasing over time. For IL-6 these rates for days 1 through 7 were as follows: 384/1797 or 21.4% for day1, 401/1738 or 23.1% for day 2, 464/1754 or 26.5% for day 3, 474/1463 or 32.4% for day 4, 364/1127 or 32.3% for day 5, 288/869 or 33.1% for day 6 and 229/696 or 32.9% for day 7. The censoring rates for IL-10 were substantially higher with the results for days 1 through 7 as follows: 1086/1797 or 60.4% for day 1, 1138/1738 or 65.5% for day 2, 1281/1754 or 73.0% for day 3, 1128/1463

or 77.1% for day 4, 844/1127 or 74.9% for day 5, 670/869 or 77.1% for day 6 and 532/696 or 76.4% for day 7. Overall 9283 (49.1%) measures of the combined bivariate cases were left-censored.

2.6.2 Application of methods

To examine the relationship between these markers and mortality over time, we fit a linear mixed model with random intercept and random slope for each biomarker (IL-6 and IL-10). Before applying the methods a normalizing transformation is considered to assure normality, and measurements are transformed using a log transformation. This results in the following model

$$Y_{ij}^1 = \beta_0^1 + \beta_1^1 t_{ij} + \beta_2^1 Mortality_i + \beta_3^1 (t_{ij} * Mortality_i) + \gamma_{0i}^1 + \gamma_{1i}^1 t_{ij} + \epsilon_{ij}$$
$$Y_{ij}^2 = \beta_0^2 + \beta_1^2 t_{ij} + \beta_2^2 Mortality_i + \beta_3^2 (t_{ij} * Mortality_i) + \gamma_{0i}^2 + \gamma_{1i}^2 t_{ij} + \epsilon_{ij},$$

where the superscript 1 and 2 represents the two biomarkers IL-6 and IL-10 respectively, and mortality is a covariate variable indicating the mortality status of patients on the 30^{th} day of the study. i.e.

$$\text{Mortality} = \begin{cases} 1 & \text{if subject is dead on day 30} \\ 0 & \text{if subject is alive on day 30.} \end{cases}$$

The three ad-hoc methods and the proposed ML method described in sections 2.2 and 2.3 are applied to these models and parameters were estimated following the techniques described in section 2.4.

Tables 4 and 5 present selected fixed effect estimates and variance components from the four different analyses of the GenIMS data for the chosen model. The first analysis replaces censored values with the detection limit. The second analysis replaces censored values with half of the detection limit, while the third analysis is the random imputation method. The fourth analysis is the ML method proposed here and uses the methodology described in sections 2 and 3 of this chapter. From these results it is clear that simply replacing censored values by the detection limit leads to bias in both the fixed effect and variance component

Table 4: Selected fixed effects estimates of Bivariate linear mixed model for IL-6 and IL-10,using several naive methods and the method that account for censoring.

Parameter	*Naive Method*	*Estimate(S.E)*	*ML Censoring Accounted Estimate(S.E)*
Slope of time for IL-6	*LOD*	-0.224(0.009)	-0.234(0.009)
	HLD	-0.235(0.01)	
	RI	-0.234(0.009)	
Slope of time for IL-10	*LOD*	-3.057(0.568)	-3.096(0.655)
	HLD	-3.088(0.656)	
	RI	-3.082(0.656)	
Mortality-IL-6	LOD	1.163(0.170)	1.194(0.176)
	HLD	1.194(0.177)	
	RI	1.196(0.179)	
Mortality-IL-10	LOD	0.364(0.079)	0.532(0.127)
	HLD	0.554(0.103)	
	RI	0.576(0.108)	
Mortality*time IL-6	LOD	0.607(0.236)	0.659(0.249)
	HLD	0.662(0.249)	
	RI	0.659(0.249)	
*Mortality*time IL-10*	*LOD*	0.685(0.097)	0.703(0.100)
	HLD	0.70(0.098)	
	RI	0.698(0.100)	

estimates. For all the methods, the mean IL-6 level exhibited a decrease during the first 7 days for alive patients and an increase for dead. The estimated mean IL-6 level at 7 days was 1.662 by LOD, 1.566 by HLD, 1.561 by RI and 1.541 by ML for the alive patients and 7.074, 7.394, 7.37 and 7.104 by LOD, HLD, RI and ML respectively for the dead. Plots of the estimated means obtained from the model is presented in figure 1.

Table 5 reports results of comparison between analyses using separate two univariate (single outcome) models vs. analysis of multiple outcomes studied simultaneously using a bivariate model. The joint modeling of the two markers using a bivariate model not only allows the study of the correlation between markers, but also produces better estimates. The estimated standard error of the slope of time for IL-6 (0.009) is much smaller than that estimated with a univariate model (0.166). This underlines that information provided by IL-10 data in the joint modeling contributes to the estimation of the IL-6. Moreover, the joint analysis of multiple outcomes simultaneously results in a different fixed effects estimates with different results, for example the mortality time interaction for IL-10 was not significant using a univariate model but found to be significant using the bivariate model.

Table 5: Fixed effects estimates of the linear mixed model using Bivariate model and two separate univariate model.

	Analysis taking left-censoring into account		*Analysis imputing threshold value**
Parameter	*Two Univariate Models*	*Bivariate Model*	*Bivariate Model*
Slope time-IL6	-0.329(0.160)	-0.234(0.009)	-0.234(0.009)
Slope time-IL10	-3.279(0.616)	-3.096(0.655)	-3.082(0.656)
Intercept-IL6	2.954(0.062)	3.179(0.043)	3.235(0.043)
Intercept-IL10	1.372(0.055)	1.667(0.021)	2.074(0.020)
Mortality-IL6	1.277(0.228)	1.194(0.176)	1.196(0.179)
Mortality-IL10	0.496(0.267)	0.532(0.127)	0.576(0.108)
*Mortality*time-IL6*	0.160(0.349)	0.659(0.249)	0.658(0.249)
*Mortality*time-IL10*	0.036((0.125)	0.703(0.100)	0.698(0.100)

*RI Method is used

3.0 PSEUDO MAXIMUM LIKELIHOOD METHOD FOR ANALYSIS OF MULTIVARIATE TRUNCATED LONGITUDINAL DATA

3.1 INTRODUCTION

Many medical studies collect biomarker data to gain insight into the biological mechanisms underlying both acute and chronic diseases. These markers may be obtained at a single point in time to aid in the diagnosis of an illness or may be collected longitudinally to provide information on the relationship between changes in a given biomarker as it relates to the course of the illness. While there are many different biomarkers presented in the medical literature there are very few studies that examine the relationship between multiple biomarkers, measured longitudinally, and predictors of interest. The exception is the HIV literature where CD4 counts and viral loads are jointly modeled over time (Jacqmin-Gadda et al 2000; Thibaut, et al 2003).

Analysis of biomarker data has been important in understanding the relationship between markers of inflammation and the development of sepsis in the Genetic and Inflammatory Markers of Sepsis (GenIMS) study. The study enrolled 2320 subjects with community acquired pneumonia through the emergency department of 28 hospitals in southwestern Pennsylvania, Connecticut, Michigan and Tennessee between 2001 and 2003. A battery of inflammatory markers was measured throughout the course of hospitalization in this cohort which was followed for a period of one year. With one goal being that of understanding the relationship between the trajectories of the pro-inflammatory and anti-inflammatory markers and the development of sepsis, there was a need for statistical methods that can accommodate multiple longitudinal biomarkers rather than relying on a series of separate longitudinal models for each biomarker. In addition, a large percentage of the biomarker data

is left censored due to the sensitivity of the assays, so that much of the currently available methodology is not applicable.

While there are methods available for the analysis of left censored outcome data in the statistical literature, there are few methods that can handle multivariate truncated longitudinal data when multiple outcomes need to be studied simultaneously. To address the issue of truncation when modeling data, researchers have proposed either the use of imputed values or the development of methods to handle the censoring directly. Imputing the lower quantification limit (Keet et al ,1997) or half of this limit (O'Brien et al, 1998) to substitute for the censored value and use of random imputation procedures (Paxton et al, 1997) are the most frequently used approaches. All of these naive approaches, as seen in the previous chapter, produce estimates with a substantial bias and they do not adjust the standard errors of the estimates for the loss of information due to censoring. Hughes (1999), Jacqmin-Gadda et al (2000), and Lyles et al (2000) proposed methods that handle left-censored measures. However all of these methods, with the exception of the method proposed by Jacqmin-Gadda et al (2000), are restricted to a longitudinal model with a single outcome. In addition, since all of these methods are based on a full likelihood, they involve numeric and algebraic complexities that require the evaluation of a series of multiple integrals and become prohibitive for data with a high rate of censoring. These computations become even more unstable for more than two random effects, leading to convergence issues when the current methodology is applied.

In this chapter we propose a method that addresses the weaknesses of the current methodology for multivariate longitudinal models with left censored outcome data. The two major weaknesses, computational complexity and model instability, are addressed by applying the method of pseudo-likelihood to this problem. Using the pseudo-likelihood the estimation problem is broken into two separate steps with estimation of the parameters associated with the covariance taking place in step 1 and then computation of the remaining parameters, based on the modified likelihood, occurring in step 2. The proposed pseudo-likelihood method significantly reduces the computational burden associated with the current methods and is much more stable while preserving the properties of the original estimators.

The chapter is organized as follows: in section 3.2 we present the proposed methodology,

the multivariate linear mixed model is given in section 3.2.1 and we develop the pseudo-likelihood for left-censored data in section 3.2.2. Computational details are given in section 3.2.3. A simulation study, conducted to assess the performance of the proposed model, is summarized in section 3.3. In section 3.4 we apply the proposed method to the GenIMS data, and results of these methods are compared with results obtained using existing methods. Finally we close the chapter by giving a brief discussion of the results and concluding remarks in section 3.5.

3.2 PSEUDO LIKELIHOOD METHODOLOGY

3.2.1 Linear Mixed Model

Let k be the number of outcomes in the model, it will be assumed that each of the k longitudinally measured outcomes can be modeled using the mixed model. Let y_{ij} be the j^{th} measure at time t_{ij} for subject i ($i = 1, \ldots, N$ and $j = 1, \ldots, n_i$) for a single outcome. Let Y_i be the response vector for subject i, i.e. $Y_i^{'} = (y_{i1}, \ldots, y_{in_i})$.

For k=2, let $Y_i = \begin{bmatrix} Y_i^1 \\ Y_i^2 \end{bmatrix}$ denote the response vector for subject i, for i=$1, \ldots, N$ and Y_i^k be the n_i^k vector of measurements of marker k (k=1,2). Let $\beta = \begin{bmatrix} \beta^1 \\ \beta^2 \end{bmatrix}$ be a $p \times 1$ vector of population parameters, known as fixed effects, and X_i be a known $n_i \times p$ design matrix of covariate variables linking β to Y_i. Let $\gamma_i = \begin{bmatrix} \gamma_i^1 \\ \gamma_i^2 \end{bmatrix}$ be a $q \times 1$ vector of subject-specific parameters, known as random effects and Z_i a known $n_i \times q$ design matrix of covariates linking γ_i to Y_i.

For multivariate normal data the linear mixed model proposed by Laird and Ware (1982) can be extended;

$$Y_i = X_i\beta + Z_i\gamma_i + \epsilon_i, \tag{3.1}$$

with the assumption that the n_i-dimensional vector Y_i satisfies

$$Y_i|\gamma_i \sim N(X_i\beta + Z_i\gamma_i, \Sigma_i),$$

where γ_i and ϵ_i are assumed to be mutually independent with $\epsilon_i \sim N(0, \Sigma_i)$, $\gamma_i \sim N(0, G)$ and Σ_i is the $n_i \times n_i$ covariance matrix of measurement errors, which is a diagonal matrix containing the two elements of the measurement error of each marker, that depends on i only through its dimension n_i, and G is the covariance matrix of the random effects. Thus the set of unknown parameters in Σ_i will not depend on i. Marginally, the Y_i are independent normals with mean $X_i\beta$ and covariance matrix V_i=Var$(Y_i) = Z_i G Z_i^T + \Sigma_i$.

3.2.2 Pseudo-likelihood for left-censored function

In general, pseudo maximum likelihood estimation consists of replacing all nuisance parameters in a model by estimates and solving a reduced system of likelihood equations. We form the pseudo-likelihood by dividing the parameter space θ into the parameter of interest and nuisance parameters, treating the covariance component parameters as nuisance parameters. Let $\theta = (\beta, \eta)$ denote the parameter space, where β is the vector of fixed effect parameters which are the parameters of interest and η are the nuisance parameters. Then the model in (1) can be rewritten as

$$Y_i = f(X, \theta) + \epsilon_i, \tag{3.2}$$

where X is the design matrix of the model.

Using the notation from section 3.2.1, and letting Y_i^o denote the n_i^o-vector of observed outcomes, Y_i^c the n_i^c-vector of censored outcomes and c_i the n_i^c-vector of censoring threshold for subject i, the pseudo likelihood function is given by

$$L(\beta, \eta) = L(\ \theta) = \prod_{i=1}^{N} f_{Y_i^o|\theta}(Y_i^o|\theta) Pr(Y_i^c < c_i | Y_i^o, \theta).$$

The matrix X_i, vector Y_i, and the covariance matrix V_i in section 3.2.1 can be partitioned into observed and censored components as

$$X_i = \begin{bmatrix} X_i^o \\ X_i^c \end{bmatrix}, Y_i = \begin{bmatrix} Y_i^o \\ Y_i^c \end{bmatrix},\ V_i = \begin{bmatrix} V_i^o & V_i^{co} \\ V_i^{co'} & V_i^c \end{bmatrix}.$$

From model (3.1), Y_i^o has a multivariate normal distribution f_i^o, using properties of the multivariate normal distribution we can show that the conditional distribution of Y_i^c given Y_i^o is normally distributed with the following mean and variance expressions respectively:

$$\begin{aligned} \mu_i^{c|o} &= X_i^c\beta + \eta_i^{co}\eta_i^{o^{-1}}[Y_i^o - \mu_i^o]. \\ V_i^{c|o} &= \eta_i^c - \eta_i^{co}\eta_i^{o^{-1}}\eta_i^{co^T}, \end{aligned}$$

where η_i denotes the covariance matrix $V_i(\eta)$.

Let the multivariate normal distribution function of the conditional distribution of Y_i^c given Y_i^o be denoted by $\Phi_i^{c|o}$, then the pseudo likelihood function can be rewritten as:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{N} f_{Y_i^o|\theta}(Y_i^o|\theta)\Phi_i^{c|o}(c_i|\theta) \\ &= \prod_{i=1}^{N} \frac{1}{2\pi|\eta_i^o|^{1/2}} e^{\left\{\frac{-1}{2}(Y_i^o - X\beta)^T \eta_i{}^{o^{-1}}(Y_i^o - X\beta)\right\}} \\ &\int_{-\infty}^{c_{i1}}\int_{-\infty}^{c_{i2}}\cdots\int_{-\infty}^{c_{in_i^c}} \frac{1}{2\pi\left|\eta_i{}^{c|o}\right|^{1/2}} e^{\left\{\frac{-1}{2}(u-\mu_i^{c|o})^T \eta_i{}^{c|o^{-1}}(u-\mu_i^{c|o})\right\}} du, \end{aligned} \tag{3.3}$$

where u is an n^c vector. The log pseudo likelihood is given by:

$$\begin{aligned} \ell(\beta,\eta) = \sum_{i=1}^{N} & -log(2\pi) - \frac{1}{2}log|\eta_i^o| - \frac{1}{2}(Y_i^o - X\beta)^T \eta_i{}^{o^{-1}}(Y_i - X\beta) \\ & + log\int_{-\infty}^{c_{i1}}\int_{-\infty}^{c_{i2}}\cdots\int_{-\infty}^{c_{in_i^c}} \frac{1}{2\pi\left|\eta_i^{c|o}\right|^{1/2}} exp\left[\frac{-1}{2}(u-\mu_i^{c|o})^T \eta_i^{c|o^{-1}}(u-\mu_i^{c|o})\right] du. \end{aligned} \tag{3.4}$$

The proposed PMLE method involves a two step estimation process. In the first a consistent estimate $\hat{\eta}$ is obtained for the nuisance parameter η by some technique or approach other than the maximum likelihood estimation (two approaches are used in this dissertation). The PMLE is then obtained by maximizing the log pseudo-likelihood $\ell(\beta,\hat{\eta})$, viewed as a function of the single parameter β.

We use and compare two different methods to obtain a consistent estimator of η in the first step of estimation. In the first method we obtain $\hat{\eta}$ by the method of moment estimators, which is known to be consistent (Casella and Berger, 2002, Ch. 7). In the second method we fixed β to be a consistent estimator $\tilde{\beta}$ obtained by an imputation method (Lubin et.al, 2004;

Paxton et al., 1997), then the resulting log-likelihood function is used to obtain a consistent estimator for η. We demonstrated the consistency of the estimator obtained by this method in a theorem given in the appendix.

The second step of estimation proceeded by updating the log-pseudo likelihood in (4) using the estimator $\hat{\eta}$ to obtain the following likelihood

$$\ell(\beta,\hat{\eta}) = \sum_{i=1}^{N} -log(2\pi) - \frac{1}{2}log|\hat{\eta}_i^o| - \frac{1}{2}(Y_i^o - X\beta)^T \hat{\eta}_i^{o^{-1}}(Y_i - X\beta)$$
$$+ log \int_{-\infty}^{c_i 1} \int_{-\infty}^{c_i 2} \dots \int_{-\infty}^{c_i n_i^c} \frac{1}{2\pi|\hat{\eta}_i^{c|o}|^{1/2}} exp\left[\frac{-1}{2}(u-\mu_i^{c|o})^T \hat{\eta}_i^{c|o-1}(u-\mu_i^{c|o})\right] du. \quad (3.5)$$

Setting the first derivative of this new log pseudo-likelihood, i.e. the pseudo-score vector,

$$S(\beta,\hat{\eta}) = \frac{\partial}{\partial(\beta,\hat{\eta})}\ell(\beta,\hat{\eta})$$

equal to 0, and solving this equation for β gives the PMLE, $\hat{\beta}$. Optimization is carried out using the dual Quasi-Newton algorithm in SAS proc nlmixed (computational details are given in next section)

The PMLE, $\hat{\beta}$ is asymptotically multivariate normal (White 1982). However, since the multiple markers are generally correlated (for example, two markers Y_{ij} and Y_{ik} for subject i, may be correlated), a robust estimator of the variance-covariance is required. To this end we carried out appropriate adjustments by replacing the asymptotic covariance matrix by a robust estimator commonly known as the sandwich estimator. For this purpose we introduce the following notation:

$$A(\theta) = E\left\{-\frac{\partial^2 \ell(Y_i, f(X,\theta))}{\partial\theta\partial\theta'}\right\}$$

$$B(\theta) = E\left\{\frac{\partial \ell(Y_i, f(X,\theta))}{\partial\theta}\frac{\partial \ell(Y_i, f(X,\theta))}{\partial\theta'}\right\}.$$

Gong and Samaiego (1981) and White (1982) showed that

$$\sqrt{N}(\hat{\theta}-\theta) \xrightarrow{d} MVN(0, A(\theta)^{-1}B(\theta)A(\theta)^{-1}). \quad (3.6)$$

Once the PMLE, $\hat{\beta}$, of β is obtained, then we obtain the first and second order partial derivatives for each subject to obtain the components of the sandwich estimator $A(\beta)$ and $B(\beta)$ as:

$$\begin{aligned} A(\beta) \quad &= E\left\{-\frac{\partial^2}{\partial\beta\partial\beta'}\left\{\sum_{i=1}^{N} -log(2\pi) - \frac{1}{2}log|\hat{\eta}_i^o| - \frac{1}{2}(Y_i^o - X\beta)^T\hat{\eta}_i^{o^{-1}}(Y_i - X\beta)\right.\right. \\ &\left.\left. +log\int_{-\infty}^{c_{i1}}\int_{-\infty}^{c_{i2}}\cdots\int_{-\infty}^{c_{in_i^c}}\frac{1}{2\pi|\hat{\eta}_i^{c|o}|^{1/2}}exp\left[\frac{-1}{2}(u-\mu_i^{c|o})^T\hat{\eta}_i^{c|o^{-1}}(u-\mu_i^{c|o})\right]du\right\}\right\}, \quad (3.7) \end{aligned}$$

and

$$\begin{aligned} B(\beta) \quad &= E\left\{\frac{\partial}{\partial\beta}\left\{\sum_{i=1}^{N} -log(2\pi) - \frac{1}{2}log|\hat{\eta}_i^o| - \frac{1}{2}(Y_i^o - X\beta)^T\hat{\eta}_i^{o^{-1}}(Y_i - X\beta)\right.\right. \\ &\left. +log\int_{-\infty}^{c_{i1}}\int_{-\infty}^{c_{i2}}\cdots\int_{-\infty}^{c_{in_i^c}}\frac{1}{2\pi|\hat{\eta}_i^{c|o}|^{1/2}}exp\left[\frac{-1}{2}(u-\mu_i^{c|o})^T\hat{\eta}_i^{c|o^{-1}}(u-\mu_i^{c|o})\right]du\right\} \\ &\frac{\partial}{\partial\beta'}\left\{\sum_{i=1}^{N} -log(2\pi) - \frac{1}{2}log|\hat{\eta}_i^o| - \frac{1}{2}(Y_i^o - X\beta)^T\hat{\eta}_i^{o^{-1}}(Y_i - X\beta)\right. \\ &\left.\left. +log\int_{-\infty}^{c_{i1}}\int_{-\infty}^{c_{i2}}\cdots\int_{-\infty}^{c_{in_i^c}}\frac{1}{2\pi|\hat{\eta}_i^{c|o}|^{1/2}}exp\left[\frac{-1}{2}(u-\mu_i^{c|o})^T\hat{\eta}_i^{c|o^{-1}}(u-\mu_i^{c|o})\right]du\right\}\right\}. \quad (3.8) \end{aligned}$$

We developed a SAS macro to estimate the model parameters and to obtain the components of the robust variance-covariance estimator. The corrected variance estimate $A(\beta)^{-1}B(\beta)A(\beta)^{-1}$ is computed by dropping the expectation and replacing the unknown parameter β by the PMLE $\hat{\beta}$.

The methods proposed by Jacqmin-Gadda et.al(2000) and Thiebaut et.al (2002) are based on maximizing the full likelihood in (3.3) and parameter estimation involves high dimensional multiple integrations. In their methods the likelihood in (3.3) requires n_i^c number of multiple integrals of the multivariate normal density to be numerically calculated. The PMLE method avoids such complex high dimensional integration by reducing the integration to a single dimension where censored observations require only computation of the univariate normal distribution for which an efficient numerical algorithm is used.

3.2.3 Computational Details

We developed a SAS macro to obtain the pseudo-likelihood estimates of model parameters by maximizing the log pseudo likelihood in (3.4) using the dual Quasi-Newton optimization method in SAS Proc NLMIXED. The NLMIXED procedure provides improved ML estimates and unlike other procedures, it allows for the explicit modeling of random effects by allowing the user to write his/her own function. But NLMIXED does not have an option for adjusting standard errors, so we developed a SAS macro that runs NLMIXED and calculates the corrected standard error using the robust(sandwich) estimator given in (3.6) (SAS macro and other codes developed for parameter estimation and the sandwich estimator are given in appendix). Another limitation of the NLMIXED procedure is that it lacks a REPEATED statement and so has limited capacities for modeling the covariance structure of correlated data, however in modeling longitudinal data in which there is not a high degree of serial correlation this limitation may not be serious.

3.3 SIMULATION STUDY

To explore the performance of the PMLE, we conducted a simulation study. We used the simulation study to determine if the obtained estimators are indeed unbiased, if the standard errors are correct and if there is any loss in efficiency compared to the full likelihood approach.

Data were generated as follows, for $i = 1, \ldots, N$: we first generated all of the components of the design matrix X, that is, a binary variable is randomly generated using Bernoulli(0.5) and is assigned as a covariate variable for each subject. For times of measurements random numbers (similar to the real data measurement time) uniformly distributed between 1 and 5 were selected. Censoring values were chosen independently of time and subject for each marker and parameter values were selected to be close to those obtained from the real data analysis with the following parameter values being used:

Table 6: Computational time comparisons according method used. Methods based on the two full Maximum likelihood based approaches(ML-CENSAD and ML-NLMIXED) and the Pseudo likelihood approach (PMLE) proposed in this study. Time given is in cpu seconds*.

Method	*Model with 2 Random Effects*	*Model with 4 Random Effects*
ML-CENSAD	720	1856
ML-NLMIXED	780	Didn't converge
PMLE	510	820

* time given is for a single run

$$\beta^1 = \begin{pmatrix} 0.70 \\ -0.15 \\ -0.2 \\ 0.03 \end{pmatrix}, \beta^2 = \begin{pmatrix} 0.50 \\ -0.15 \\ 0.15 \\ 0.05 \end{pmatrix}.$$

Using these specifications 1000 measurements of 200 subjects were simulated according two different models. The first model is a model with two random effects, one random slope for each response (marker)

$$Y_{ij}^k = \beta_0^k + \beta_1^k t_{ij} + \beta_2^k X_i + \beta_3^k t_{ij} * X_i + \gamma_{1i}^k t_{ij} + \epsilon_{ij}, \tag{3.9}$$

where β and γ are vectors, ϵ is a matrix as defined in section 2, and X_i is a binary covariate variable.

The second model includes a random slope and a random intercept for each response for a total of four random effects in the model,

$$Y_{ij}^k = \beta_0^k + \beta_1^k t_{ij} + \beta_2^k X_i + \beta_3^k t_{ij} * X_i + \gamma_{0i}^k + \gamma_{1i}^k t_{ij} + \epsilon_{ij}. \tag{3.10}$$

In both models a binary variable X_i is used as a covariate variable and its effect over time is also studied by including an interaction term in the model. For all of the simulation studies 25% of the measures were censored.

Table 7: Selected simulation results comparing the performance of the PMLE approach with the ML approach. Bias and S.E. for fixed effect parameters of the linear mixed effect model with a random slope (model (9)) estimated by ML-NLMIXED and by PMLE. Values reported are for the mean of 500 replications.

		Method					
		ML-NLMIXED		*PMLE*[1]		*PMLE*[2]	
Parameter	*True Value*	*Bias*	*SE*	*Bias*	*SE*	*Bias*	*SE*
Time[1] (β_1^1)	-0.15	-0.0059	0.0340	-0.005	0.036	0.0040	0.0336
Covariate[1] (β_2^1)	-0.20	0.0009	0.0432	0.002	0.035	0.0009	0.0433
Interaction[1] (β_3^1)	0.03	0.0087	0.0468	-0.002	0.049	-0.0091	0.0469
Time[2] (β_1^2)	-0.15	-0.0242	0.0320	-0.027	0.037	0.0281	0.0354
Covariate[2] (β_2^2)	0.15	-0.0179	0.0426	0.023	0.045	0.0181	0.0426
Interaction[2] (β_3^2)	0.05	-0.0061	0.0491	-0.008	0.050	-0.0077	0.0490

The proposed method is compared to two different existing full likelihood methods for efficiency and accuracy. The first method used for the comparison is a method proposed by Jacqumin-Gadda et.al (2000), in which parameter estimation is carried out by maximizing the full likelihood using a Marquardt algorithm and other multiple iterative process. They used a FORTRAN program called CENSAD (and hence we labeled this method as ML-CENSAD in the results presented in the tables). The second method used is that of Thiebaut and Jacqmin-Gadda (2004) which is also a full likelihood based method,the authors used SAS Proc Nlmixed procedure for maximization (results from this method are lebeled as ML-NLMIXED in tables). Both of these methods are compared with the proposed method for computational time, efficiency and bias.

For the first step in the estimation of the PMLE both the method of moment estimators and the second method described in section 3.2.2 were used. Estimates obtained by these methods are labeled as PMLE[1] and PMLE[2], respectively, in tables of results.

In Table 6 we present a summary of computation time by each method for both models

Table 8: Selected simulation results comparing the performance of the PMLE approach with the ML approach. Bias and S.E. for fixed effect parameters of the linear mixed effect model with a random intercept and a random slope (model (10)) estimated by ML-CENSAD and by PMLE. Values reported are for the mean of 500 replications.

		Method					
		ML-CENSAD		*PMLE*[1]		*PMLE*[2]	
Parameter	*True Value*	*Bias*	*SE*	*Bias*	*SE*	*Bias*	*SE*
Time[1] (β_1^1)	-0.15	0.0311	0.0018	0.031	0.001	0.0284	0.0023
Covariate[1] (β_2^1)	-0.20	0.0512	0.0788	0.033	0.079	0.0323	0.0812
Interaction[1] (β_3^1)	0.03	0.0225	0.0095	0.023	0.014	0.0234	0.0136
Time[2] (β_1^2)	-0.15	-0.0401	0.0010	-0.050	0.001	-0.0483	0.0010
Covariate[2] (β_2^2)	0.15	-0.0410	0.0273	-0.047	0.030	-0.0472	0.0291
Interaction[2] (β_3^2)	0.05	0.0110	0.0019	0.014	0.002	0.0123	0.0023

(3.9) and (3.10). The proposed PMLE method converges in substantially less time than the other two full likelihood based methods. Comparing the two methods used for the first step of the estimation of the PMLE, the method of moment estimators takes less time than the second method. The ML-CENSAD converges a little bit faster than the ML-NLMixed, but is significantly slower than the PMLE-method. For the model with four random effects, model (3.10), the ML-NLMIXED did not converge and it was stopped after an hour (3600 cpu seconds). Comparing ML-CENSAD and PMLE for model (3.10), again the PMLE converges significantly faster than the ML-CENSAD. Generally speaking, the computation time significantly increased based on the data structure and model used for both full likelihood based methods, but the PMLE is not significantly affected by these changes. For instance, the ML-CENSAD is very slow to converge when the number of measures n_i for each subject increases (the results reported in table 1 are for $n_i = 5\ \forall_i$) and the ML-Nlmixed has convergence difficulties when the number of random effects in the model increases (for the model with 4 random effects the method does not converge for 3600 cpu seconds and was stopped before convergence).

Tables 7 and 8 display the bias and the standard error obtained from each method for models (3.9) and (3.10) respectively. As can be seen from these tables the proposed PMLE method produces estimates comparable to both of the full likelihood methods with less bias and significantly shorter computation time. Both the method of moments and the second method used in the first step of the estimation of the PMLE gives similar results with no significant difference in the bias and standard error.

3.4 APPLICATION

We applied the proposed method and the other two existing methods to analyze the GenIMS data descriped earlier in chapter 2. One aim in the GenIMS study is to examine the relationship between a set of inflammatory markers and to determine if changes in these markers over time were related to mortality and/or development of sepsis. A total of 2320 patients were recruited to the study with inflammatory markers measured on a subset of 1797

subjects (see section 2.6.1 and table 3 for in depth description of the data). The biomarkers of greatest interest are IL-6 (interleukin-6) and tumor necrosis factor (TNF) as markers of pro-inflammatory and IL-10 (interleukin-10), as a marker of the anti-inflammatory cytokine response.

To determine if there are specific patterns of the circulating levels of these makers associated with severe sepsis and death, we fit a linear mixed model with random intercept and random slope for each biomarker. Before applying the methods a normalizing transformation is considered to assure normality, and measurements are transformed using the log transformation function.

$$Y_{ij}^1 = \beta_0^1 + \beta_1^1 t_{ij} + \beta_2^1 Mortality_i + \beta_3^1 (t_{ij} * Mortality_i) + \gamma_{0i}^1 + \gamma_{1i}^1 t_{ij} + \epsilon_{ij}.$$
$$Y_{ij}^2 = \beta_0^2 + \beta_1^2 t_{ij} + \beta_2^2 Mortality_i + \beta_3^2 (t_{ij} * Mortality_i) + \gamma_{0i}^2 + \gamma_{1i}^2 t_{ij} + \epsilon_{ij}.$$

where the superscript 1 and 2 represents a cytokine response (IL-6, TNF or IL-10).

We conducted multiple analyses using different combinations of the pro-inflammatory and anti-inflammatory markers using the proposed PMLE and the ML-CENSAD methods to estimate model parameters. The ML-NLMIXED method is not used here since it does not converge for a model with four random effects.

We present results of three analysis using mortality, severe sepsis and both mortality and severe sepsis together in tables 9, 10 and 11 respectively, comparing analysis using separate univariate models and bivariate model considering two markers simultaneously.

When comparing the estimates obtained using ML-CENSAD versus the PMLE method, the standard errors of the PMLE estimates are larger. However, the inferences are the same in both cases. In all cases, the computation time for the PMLE method is substantially less than that of the ML-CENSAD.

The joint modeling of the two biomarkers using a bivariate model allows us to study the correlation between the markers over time which can be of importance when understanding the role of biomarkers in the development of sepsis. Moreover, the bivariate model takes into account the estimation of the correlation matrix between random effects for the estimation of other model parameters. For example the first analysis (using IL-6 and IL-10), estimated

Table 9: Parameter estimates and S.E. of fixed effects for the linear mixed model of IL-6 and IL-10 of the GenIMS data including mortality as a covariate variable according to method (ML-CENSAD and PMLE) and model (Univariate Vs. Multivariate) used. Responses are log(IL-6) and log(IL-10) and time is measured in days

	Two Separate	*Bivariate Model according to Method*	
Parameter	*Univariate Models*	*ML-CENSAD*	*PMLE*
Slope time-IL-6	-0.329(0.160)	-0.2339(0.0093)	-0.2738(0.0122)
Slope time-IL-10	-3.279(0.616)	-3.0955(0.6549)	-3.0939(0.6649)
Mortality-IL-6	1.277(0.228)	1.1937(0.1758)	1.1944(0.1811)
Mortality-IL-10	0.496(0.267)	0.5319(0.1272)	0.5317(0.1329)
*Mortality*time-IL-6*	0.160(0.349)	0.6597(0.2495)	0.6641(0.2578)
*Mortality*time-IL-10*	0.036(0.125)	0.7035(0.1001)	0.6952(0.1068)

Table 10: Parameter estimates and S.E. of fixed effects for the linear mixed model of TNF and IL-10 of the GenIMS data including severe sepsis as a covariate variable according to a model (Univariate Vs. Multivariate) used. Responses are log(TNF) and log(IL-10) and time is measured in days

Selected	*Two Separate*	*Bivariate*
Parameter	*Univariate Models*	*Model*
Slope time-TNF	-0.228(0.0296)	-0.094(0.0110)
Slope time-IL-10	-0.649(0.0306)	-0.308(0.0153)
Severe Sepsis-TNF	0.264(0.0478)	0.259(0.0440)
Severe Sepsis-IL-10	0.209(0.0792)	0.315(0.0728)
*Severe Sepsis*time-TNF*	0.004(0.0307)	-0.005(0.0178)
*Severe Sepsis*time-IL-10*	0.167(0.0386)	0.055(0.0236)

Table 11: Parameter estimates and S.E. of fixed effects for the linear mixed model of TNF and IL-10 of the GenIMS data including mortality and severe sepsis as covariate variables according to a model (Univariate Vs. Multivariate) used. Responses are log(TNF) and log(IL-10) and time is measured in days

Selected Parameter	*Two Separate Univariate Models*	*Bivariate Model*
Slope time-TNF	-0.226(0.0212)	-0.093(0.0111)
Slope time-IL-10	-0.641(0.0303)	-0.313(0.0153)
SS-TNF	0.212(0.0495)	0.2034(0.0458)
SS-IL-10	0.173(0.0833)	0.245(0.0767)
Mortality-TNF	0.337(0.0897)	0.356(0.0826)
Mortality-IL-10	0.226(0.1320)	0.419(0.1250)
*SS*time-TNF*	0.006(0.0315)	-0.0001(0.0184)
*SS*time-IL-10*	0.111(0.0394)	0.031(0.0249)
*Mortality*time-TNF*	-0.0189(0.0604)	-0.027(0.0340)
*Mortality*time-IL-10*	0.312(0.0635)	0.148(0.0389)

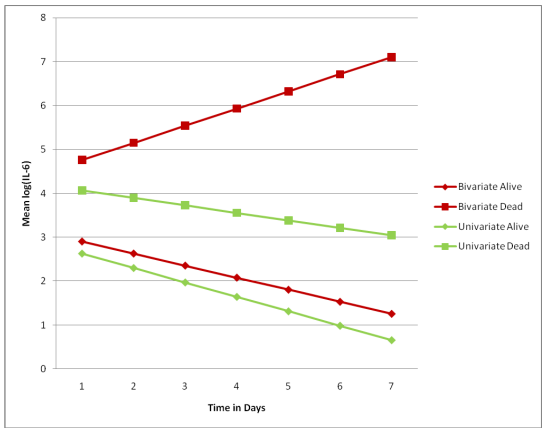
standard error for the slope of time IL-6 is 0.0122 (table 9) using a bivariate model, smaller than the estimated standard error using a univariate model (0.160), underlining the fact that information provided by the second marker (IL-10) data in the bivariate model contributes to the estimation of the first marker. In addition, use of the bivariate model resulted in a different relationship between mortality and IL-6 values when compared to modeling IL-6 as a single outcome, and statistical significance of the interaction term for mortality and time for IL-10 differs depending on the model used, with the term being non-significant for the single outcome model and highly significant for the bivariate model.

Plots of the estimated means from the three analysis, using a univariate and bivariate models are presented in Figures 1-3. The plotted values are the estimated mean levels on each of the seven days using both univariate and bivariate models. The estimated mean level of the markers using a bivariate model are different compared to the levels of the univariate model. Generally the estimated mean concentration is higher in day 1 and reduced in the subsequent days for all the markers.

The estimated mean IL-6 concentrations were higher for subjects who died compared to those who did not, and for the subjects who developed severe sepsis compared with those who did not. The pattern was similar for TNF, the mean estimated level was higher for subjects who developed sever sepsis (both survivors and non survivors) compared with those who did not develop sever sepsis.

The mean estimated IL-10 concentration were lower than those observed for IL-6. Higher levels were associated with dead subjects compared to survivors and with subjects with severe sepsis compared with subjects without severe sepsis. Comparing the bivariate model (red line) and the univariate model(green line) (fig.1 b), estimated mean concentration of IL-10 for non-survivors is significantly higher than survivors in the bivariate model, but not significantly different using a univariate model.

Figure 1: Estimated model means of log(IL-6) and log(IL-10) for the GenIMS data, using the PMLE parameter estimation by mortality status (dead or alive) and model used (univariate or bivariate).



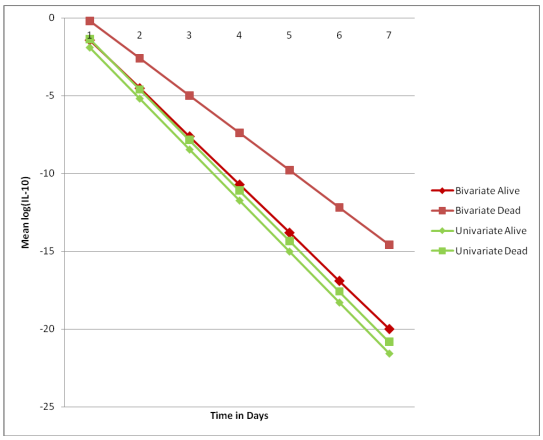
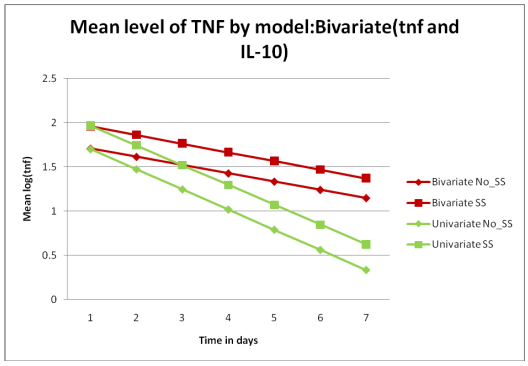


Figure 2: Estimated model means of log(TNF) and log(IL-10) for the GenIMS data, using the PMLE parameter estimation method by model(univariate or bivariate) using severe sepsis as a covariate variable.



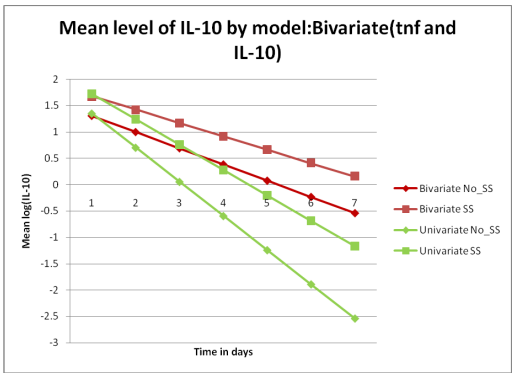
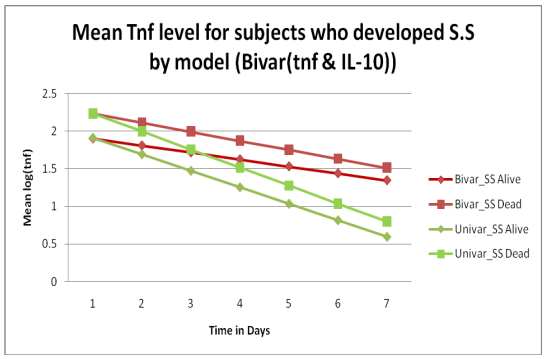
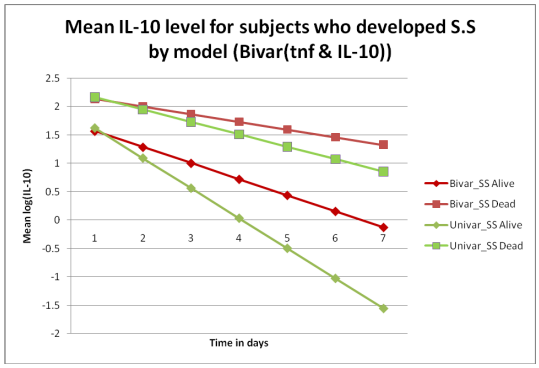
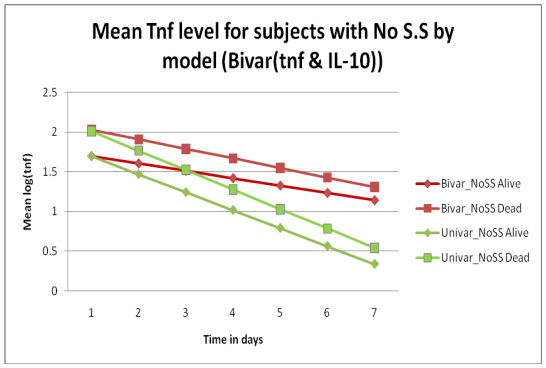
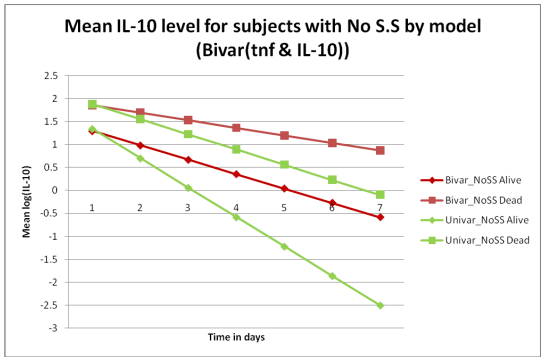


Figure 3: Estimated model means of log(TNF) and log(IL-10) for the GenIMS data, using the PMLE parameter estimation method by model(univariate or bivariate) using mortality and severe sepsis as covariate variables.









3.5 DISCUSSION

We proposed and evaluated a pseudo maximum likelihood estimator for the analysis of multivariate longitudinal left-censored data that simplifies computational complexities and can be applied to different models and different data structures. The major advantage of the pseudo likelihood estimator is its computational simplicity when compared with the full likelihood method currently used for modeling multivariate longitudinal data. The proposed method significantly eases the numerical complexities of the full likelihood approach by reducing high-dimensional integration to integration of a single dimension. Further, it alleviates the need to specify and estimate many nuisance parameters that are required in a full likelihood approach. As is demonstrated by the simulation and the real life data studies, the pseudo likelihood approach yields estimates with small bias and robust standard errors.

For longitudinal data with high rate of censoring, like the GenIMS data analyzed in this study, the pseudo likelihood method is recommended since it dramatically decreases the computation time. The full likelihood approach methods are limited by the rate of censoring as these methods require numerical evaluation of multiple integrals of a multivariate normal density whose dimension is equal to that of the number of censored measures. Whereas, the pseudo likelihood approach avoids numerical evaluation of multivariate integrals, since filling in censored observations requires computation of the univariate normal distribution function for which efficient numerical algorithms are available.

4.0 DISCUSSION AND FUTURE WORK

4.1 DISCUSSION

We presented two approaches to fit linear mixed model accounting for left-censoring of the response. In the maximum likelihood approach to estimate mixed effects linear models with left-censored longitudinal data, the simulation study showed that the bias and the mean squared error of parameter estimates obtained by this method are smaller than those obtained by imputing the censoring limit or half of the limit.

The proposed extension of the maximum likelihood method to handle a multivariate model when multiple outcomes are needed to be studied simultaneously yields better estimations than the two univariate models. Moreover the correlation matrix between random effects could be very informative as illustrated in the application of this study.

Limitations of the full maximum likelihood are its computational complexities, it involves numerical complexities that require high dimensional integrations and the convergence and other problems related when the data to be analyze involve high rate of censoring. The pseudo-maximum likelihood method proposed simplifies these and other problems. As is demonstrated by the simulation and the analysis of the GenIMS data studies, the pseudo maximum likelihood approach yields estimates with small bias and robust standard errors in a significantly less time than the full likelihood. It also significantly eases the numerical complexities by reducing the high dimensional integration to integration of a single dimension.

We recommend the pseudo maximum likelihood method for longitudinal data with high rate of censoring, like the GenIMS data analyzed in this study, not only because it significantly decreases the computation time, but also unlike the full likelihood method, it can be

applied to a data with numerous censored measures.

Unlike the full likelihood, the pseudo-likelihood requires specification of the distribution for the data at times on the same subject. Further, compared to maximum likelihood, which requires the full likelihood to be correctly specified in order to obtain consistent estimates, the pseudo- likelihood estimates are consistent as long as the marginal distributions are correctly specified.

4.2 FUTURE WORK

The future work plan is to apply the review of outcomes from this study to analysis of a single outcome model. Develop a methodology based on the pseudo maximum likelihood to a single outcome analysis to evaluate and compare the performance with existing methods.

The model developed in this study assumes ignorable drop outs, and did not take into account informative drop out. In the future I plan to work on developing a joint model of longitudinal data with informative drop out and censoring.

APPENDIX

A.1 PROOF OF CONSISTENCY

We demonstrate consistency of $\hat{\eta}$ obtained by the second method described in section 2.2 in the following theorem.

Let η_o denote the true value of η, and $\tilde{\beta}$ a consistent estimator of β we will make use of the following regularity conditions.

(C1) The parameter space θ is a compact subset of the Euclidean p-space (R^p) (the value η_o is an interior point of θ).

(C2) $\ell_N(\tilde{\beta}, \eta)$ is a measurable function for all $\eta \in \theta$ and $\frac{\partial}{\partial \eta}\ell_N(\tilde{\beta}, \eta)$ exists and is continuous in an open neighborhood of η_o.

(C3) $\frac{1}{N}\ell_N(\tilde{\beta}, \eta)$ converges in probability uniformly to a function $\ell(\tilde{\beta}, \eta)$ in an open neighborhood of η_o, and $\ell(\tilde{\beta}, \eta)$ attains a local maximum at η_o.

Condition (C1) is one of the assumptions used to develop our method, while (C2) can be verified easily as the first-order derivatives of $\ell_N(\tilde{\beta}, \eta)$ are bounded in a neighborhood of η_o, and that $E|\frac{1}{N}\ell_N(\tilde{\beta}, \eta)| \leq K$, on a neighborhood of η_o. (C3) can be verified using the first-order Taylor's expansion and the law of large numbers.

Theorem: Under the above regularity conditions, let $\hat{\eta}_N$ be a root of the equation

$$\frac{\partial}{\partial \eta}\ell_N(\tilde{\beta}, \eta) = 0$$

for which $|\hat{\eta}_N - \eta_o| < \epsilon$ for $\epsilon > 0$. Then $\hat{\eta}_N \xrightarrow{p} \eta_o$.

Note: if $\hat{\eta}_N$ is not unique, we appropriately choose one such value in such a way that $\hat{\eta}_N$ is a measurable function. This is possible by a theorem of Jennrich (1969, p. 637).

Proof: Let K be an open neighborhood in R^p containing η_o. Then $\bar{K} \cap \theta$, where $\bar{K}$ is the complement of K in R^p, is compact. Therefore $\max_{\eta \in \bar{K} \cap \theta} \ell(\tilde{\beta}, \eta)$ exists. Define

$$\epsilon = \ell(\tilde{\beta}, \eta_o) - \max_{\eta \in \bar{K} \cap \theta} \ell(\tilde{\beta}, \eta). \tag{.1}$$

Let A_N be the event that $|\frac{1}{N}\ell_N(\tilde{\beta}, \eta) - \ell(\tilde{\beta}, \eta)| < \epsilon/2$ for all η. Then

$$A_N \Rightarrow \ell(\tilde{\beta}, \hat{\eta}_N) > \frac{1}{N}\ell_N(\tilde{\beta}, \hat{\eta}_N) - \epsilon/2, \tag{.2}$$

and

$$A_N \Rightarrow \frac{1}{N}\ell_N(\tilde{\beta}, \eta_o) > \ell(\tilde{\beta}, \eta_o) - \epsilon/2. \tag{.3}$$

But then since $\ell_N(\tilde{\beta}, \hat{\eta}_N) \geq \ell_N(\tilde{\beta}, \eta_o)$, from (A.2) we have

$$A_N \Rightarrow \ell(\tilde{\beta}, \hat{\eta}_N) > \frac{1}{N}\ell_N(\tilde{\beta}, \eta_o) - \epsilon/2. \tag{.4}$$

Therefore, adding both sides of the inequalities in (A.3) and (A.4), we obtain

$$\begin{aligned} A_N &\Rightarrow \frac{1}{N}\ell_N(\tilde{\beta}, \eta_o) + \ell(\tilde{\beta}, \hat{\eta}_N) > \ell(\tilde{\beta}, \eta_o) - \epsilon/2 + \frac{1}{N}\ell_N(\tilde{\beta}, \eta_o) - \epsilon/2 \\ &\Rightarrow \ell(\tilde{\beta}, \hat{\eta}_N) > \ell(\tilde{\beta}, \eta_o) - \epsilon. \end{aligned} \tag{.5}$$

Hence, from (A.1) and (A.5) we conclude that $A_N \Rightarrow \hat{\eta}_N \in K$, which implies

$$P(A_N) \leq P(\hat{\eta}_N \in K).$$

But, since by condition (C3) $\lim_{N \to \infty} P(A_N) = 1$,

$$\Rightarrow \hat{\eta}_N \xrightarrow{p} \eta_o.$$

Bibliography

[1] Amemiya, T. (1985). Advanced Econometrics. *Harvard University Press, Cambridge, MA.*

[2] Berntsen, J., Espelid, T.O. and Genz, A. (1991). Algorithm 698: DCUHRE-an adaptive multidimensional integration routine for a vector of integrals. *ACM Transactions on Mathematical Software*, **17,** 452-456.

[3] Billingsley, P. (1995). Probability and Measure, 3^{rd} ed. *Wiley, NY.*

[4] Casella, G. and Berger, R. (2002). Statistical Inference 2^{nd} ed. *Pacific Grove, CA : Thomson Learning.*

[5] Carriquiry, A.L., Gianola, D. and Fernando,R.L. (1987) Mixed model analysis of a censored normal distribution with reference to animal breeding. *Biometrics,* **43,** 929-939.

[6] Genz, A. (1992) Numerical computation of multivariate normal probabilities, *J. Comput. Graph. Statist.* **1** 141-149.

[7] Genz, A. (1993) Comparison of methods for the computation of multivariate normal probabilities, *Comput. Sci. Stat.* **25** 400-413.

[8] Gong, G. and Samaniego, F. (1981). Pseudo Maximum Likelihood Estimation: Theory and Application. *The Annals of Statistics* **9,** 861–869.

[9] Gourieroux, C., Monfort, A. and Trognon, A. (1984). Pseudo Maximum Likelihood Methods: Theory. *Econometrica* **52,** 681–700.

[10] Hornung,R.W., Reed,L.D.,(1990). Estimation of average concentration in the presence of nondetectable values. *Appl. Occup. Environ Hyg.* **5**, 4651.

[11] Hughes,J.P. (1999). Mixed effects models with censored data with application to HIV RNA levels, *Biometrics* **55,** 625–629.

[12] 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.

[13] Keet,I.P., Janssen,M., Veugelers,P.J., Miedema, F., Klein,M.R, Goudsmit,J., Coutinho, R.A., de Wolf,F.(1997). Longitudinal analysis of CD4 T cell counts, T cell reactivity, and human immunodeficiency virus type 1 RNA levels in persons remaining AIDS-free despite CD4 cell counts less than 200 for more than 5 years. *J. Infect. Dis.* **176,** 665-671.

[14] Laird N.M and Ware JH.(1982). Random-effects models for longitudinal data. *Biometrics* **38,** 963–974.

[15] Lubin,J.H, Colt J.S, Camann D, Davis S, Cerhan J, Severson RK, et al.(2004). Epidemiologic evaluation of measurement data in the presence of detection limits. *Environ Health Perspect.* **112,**16911696.

[16] Lyles,R.H., Lyles,C.M., Taylor, D.J.(2000) Random regression models for human immunodeficiency virus ribonucleic acid data subject to left censoring and informative drop-outs, *J. R.Stat. Soc. C* **49,** 485–497.

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

[18] O'Brien,T.R., Rosenberg,P.S., Yellin, F. and Goedert, J.J.(1998) Longitudinal HIV-1 RNA levels in a cohort of homosexual men, *J. AIDS.* **18**, 155-161.

[19] Paxton,W., Coombs, R., McElrath, J., Keefers, M., Sinangil, F., Williams, B., Chernoff, D., Hughes, J., Corey, L. (1997). Longitudinal analysis of quantitative virologic measures in HIV-1 infected individuals with greater than 400 CD4+ cells/microliter. *J. Infect. Dis.,* **175**, 274-254.

[20] Sy, J.P., Taylor, J.M., Cumberland, W.G. (1997). A stochastic model for the analysis of the bivariate longitudinal AIDS data. *Biometrics,* **53,** 542-555.

[21] Thiebaut,R., Jacqmin-Gadda,H., Chene, G., Leport,C.(2002) Commenges, D. Bivariate linear mixed models using SAS Proc MIXED, *Comput. Methods Programs Biomed.* **69** 249–256.

[22] Thibaut, R., Jacqmin-Gadda, H., Leport, C. et al.(2003) Bivariate longitudinal model for the analysis of the evolution of HIV RNA and CD4 cell count in HIV infection taking into account left censoring of HIV RNA measures, *J. Biopharm. Stat.* **13,** 271–282.

[23] Thibaut, R., Jacqmin-Gadda, H.(2004) Mixed models for longitudinal left-censored repeated measures, Comput. Methods Programs Biomed. **74,** 255–260.

[24] White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometica* **50,** 1–25

[25] Zeger, S. L. and Liang, K. -Y. (1986). Longitudinal analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-130.