A Bayesian Shared Parameter Model for Incomplete Semicontinuous longitudinal Data: An Application To Toenail Dermatophyte Onychomycosis Study

Samaneh Eftekhari Mahabadi
Department of Mathematics, Statistics and Computer Science, Faculty of Science, University of Tehran
Tehran, 14155-6455, Iran
S.Eftekhari@khayam.ut.ac.ir

Received 13 April 2014

Accepted 17 July 2014

Most of statistical analysis for longitudinal data are based on normality assumption for the continuous response of interest which might be violated in some practical areas due to skewed data which possibly contain excess zeros. Some authors have proposed frequentist and Bayesian approaches to model semicontinuous data using a zero-inflated log-normal model which do not consider the problem of incomplete responses which is an almost inevitable complication in drawing inferences for follow up studies. In this article, we will propose a Mixed effect zero inflated log-normal model along with a possibly non-ignorable dropout mechanism by utilizing a practical Bayesian approach for parameter estimation. To account for the possibility of non-ignorable dropout we will use a shared-parameter framework where the outcome and the missingness models are connected by means of common latent variables or random effects. The approach will be illustrated by analyzing a real data set from a longitudinal study for the comparison of two oral treatments for toenail dermatophyte onychomycosis in which the outcome of interest present a typical example of log-normal data with excess zeros. These data have been analyzed by many researchers with the normality assumption for the continuous response of interest which cannot be justified based on the descriptive aspects of the data at hand and the zero-inflated log-normal assumption leads to the better goodness of fit results

Keywords: Longitudinal Studies; Semicontinuous Responses; Non-random Dropout; Bayesian Approach.

2000 Mathematics Subject Classification: 62F15, 62J12, 62P10

1. Introduction

A growing number of researches in public health, medicine, social sciences and economic surveys are performed by means of longitudinal studies that repeatedly measure the outcome of interest over a period of time. Most of statistical analyses for longitudinal data are based on normality assumption for the continuous response of interest which might be violated in some practical areas due to semicontinuous outcomes. A random variable is referred to as semicontinuous when it has a probability mass at a specific point value which is often zero, but the remaining values which are mostly positive follow a continuous distribution. The two main approaches for analyzing cross-sectional semicontinuous data are the Tobit model and the two-part model proposed by [22] and [8], respectively. Later, [5], [16] and [2] extended the two-part model of [8] for the analysis of semicontinuous

Published by Atlantis Press

longitudinal data. Also [14] and [1] developed an estimating equations approach for two-part models with application to clustered data. Recently, [12] has also presented a hierarchical zero-inflated log-normal model for repeated measurements. Compared with the substantial literature on Maximum Likelihood estimation approach for the data with semicontinuous outcomes, few authors have studied the Bayesian approach for modelling these kinds of data. Among them [25] developed a Bayesian two-part model for the analysis of health care data. Also, [19] proposed a hierarchical Bayesian approach to analyze a multivariate two-part model. Later, [9] proposed a flexible class of zero-inflated Bayesian models in a longitudinal setting.

All the above literature on analyzing cross-sectional and longitudinal semicontinuous data has been proposed for the case of complete vector of observed responses. However, an almost inevitable complication in drawing inferences for follow up studies is the subject's attrition from the study prematurely which is known as dropout. In most of the applications, researchers perform a complete case analysis with the ignorability assumption for the dropout mechanism which might be misleading if the missing mechanism has been generated from a non-random process. Actually, in this paper we want to propose a practical Bayesian approach to make inference for incomplete semicontinuous longitudinal data sets.

The dropout or missing data process might be generated from three different mechanisms which should be considered in the data modelling step. The missing process is said to be completely at random (MCAR) if the missingness is wholly unrelated to either the observed or unobserved response variables. If the dropout process is independent of the unobserved data, conditional on the observed ones, the mechanism is known as at random (MAR). If both MCAR and MAR are not valid which means that the dropout depends on the value of the missing responses or on other unobservables even after conditioning on observed data the dropout mechanism is not at random (MNAR). The MCAR and MAR assumptions lead to the ignorable models, which allow valid inferences about parameters to be based on the observed part of likelihood or the posterior function without the need for an explicit dropout model, provided the distinctness of the parameter spaces of the missing mechanism and the response models (and also independence of their prior distributions in the Bayesian approach).

Models for incomplete data are often divided into three different frameworks according to the different factorizations of the joint distribution of the responses and the missingness or dropout process. If the joint distribution is factorized as the conditional distribution of the missingness process given the response variable and the marginal distribution of the response variable, the approach is called selection model (SeM). In the second approach known as pattern mixture model (PMM), the factorization of the joint distribution takes place in the reverse form. Finally, in the third framework it is assumed that the response variable and the missingness process are conditionally independent given a set of shared latent variables, e.g., random effects which are known as shared-parameter model (SPM) ([13] and [7]).

In this article we will propose a class of Bayesian two-part models for incomplete semicontinuous longitudinal data with the possibility of nonrandom dropouts. Our proposed model includes the nonrandom dropout process in a shared parameter framework assuming shared random effects between the two parts of the response model and the dropout mechanism. The proposed model will be applied for analyzing a well known longitudinal data set for the comparison of two oral treatments for toenail dermatophyte onychomycosis. The longitudinal outcome of interest in these data presents a typical example of skewed data with excess zeros. Although these data have been analyzed by many researchers with the normality assumption for the continuous response of interest but this assumption cannot be justified based on the descriptive aspects of the data at hand and it will be shown that the zero-inflated log-normal assumption leads to the better goodness of fit results.

The remainder of the paper is organized as follows. In Section 2, the Bayesian SPM model with possibility of MNAR outcomes and its corresponding posterior function would be presented. The Toenail data and its descriptive aspects will be explained in the first subsection in Section 3. Also Section 3 includes model, computational steps and results of analyzing these data set using the proposed model in Section 2. Finally, Section 4 presents some concluding remarks.

2. Model and Posterior function

Let Y_{it} denote the semicontinuous outcome variable for the i-th subject at time t in a longitudinal study of N subjects where the i-th subject have T_i visits, $1 \leq i \leq N$ and $1 \leq t \leq T_i$. Suppose that the semicontinuous observations are recorded as two variables (W_{it}, Z_{it}), where W_{it} is the indicator variable of zero values in Y_{it}, i.e.,

$$W_{it} = \begin{cases} 1 & if\ Y_{it} = 0 \\ 0 & if\ Y_{it} \neq 0 \end{cases},$$

and $Z_{it} = g(Y_{it})$ for non zero values (positive values) of Y_{it} where $g(.)$ is some monotonically increasing function (e.g. log) chosen to make the nonzero values of Y_{it} approximately normally distributed. To model the vector of repeated measurements for each individual, we assume that conditional on a vector of q dimensional subject-specific random effect parameters $B_i^{'} = (B_{i1}^{'}, B_{i2}^{'})$, the vectors of responses for each individual are independent along the time. Hence, for the joint distribution of the vector of observed outcomes (W_{it}, Z_{it}) given B_i, we assume that the binary indicators W_{it} are Bernoulli variables with the following probability of success:

$$\pi_{it} = P(W_{it} = 1|X_{it}) = h(\alpha_t' X_{it} + B_{i1}' U_{it}),\ \ t = 1, \dots, T_i, \quad i = 1, \dots, N$$

where $h(.)$ is a specified monotonic link function (for example, the logit or probit). Also, X_{it} is the vector of covariates for the i-th subject at time t, and U_{it} is some subset of X_{it} which could for example include time-varying covariates that the researcher believes might have different slopes among sample individuals. Actually when one considers $U_{it} = 1$, there is only a random intercept included in the model varying from one individual to another, while considering $U_{it} = X_{it}$ leads to random slops for all the model covariates. Also it is assumed that Z_{it} has a normal distribution given that $Y_{it} > 0$ as follows:

$$Z_{it} = g(Y_{it})|Y_{it} > 0 \sim N(\mu_{it}, \sigma_z^2),$$
$$\mu_{it} = \beta_t' X_{it} + B_{i2}' U_{it},\ \ t = 1, \dots, T_i, \quad i = 1, \dots, N$$

where,

$$B_i = \begin{bmatrix} B_{1i} \\ B_{2i} \end{bmatrix} \sim MVN_q(0, \Sigma_B), \Sigma_B = \begin{bmatrix} \Sigma_{11}\ \Sigma_{12} \\ \Sigma_{21}\ \Sigma_{22} \end{bmatrix}.$$

Now to allow for the possibility of dropouts in the above semicontinuous model using a shared parameter framework, let R_{it} denote the non-dropout indicator for the i-th individual at time t, defined as follows:

$$R_{it} = \begin{cases} 0 & \quad \textit{if } Y_{it} \textit{ is not observed} \\ 1 & \quad \textit{if } Y_{it} \textit{ is observed.} \end{cases}$$

We restrict our attention to monotone dropout pattern ($R_{it} = 0$ implies that $R_{it'} = 0$ for $t' > t$) in which all subjects are observed at $t = 1$. We consider the Bernoulli distribution for R_{it} with success probability depending on the current values of predictors (X_{it}) and the shared random effect parameters $B_i = (B_{1i}, B_{2i})$ as follows,

$$P(R_{it} = 1|X_{it}, R_{i,t-1} = 1, B_i) = expit(\gamma_{0t}^{'} X_{it} + \gamma_1^{'} B_i)$$

where $expit(a) = exp(a)/(1+exp(a))$ is the CDF for the standard logistic distribution and could be replaced by CDF for normal or extreme value distributions as well. Also, $\gamma_1 = (\gamma_{11}, \ldots, \gamma_{1q})$ is the vector of non-ignorability parameters. In this model, $\gamma_1 = 0_{1\times q}$ lead to MAR dropout mechanism in which the parameter estimates of the response models could be obtained ignoring the missing mechanism (assuming disjoint parameter spaces for response models and the dropout mechanism). Note that using the shared parameter frame work, the vector of random effects B_i in the above missing mechanism includes the same two random effect vectors of the zero inflated model. In this way, the missing mechanism allow the possibility of the correlation between the unobserved responses and the mechanism generating the incompleteness in the data.

It should be mentioned that, to allow time dependency for the model parameters α_t, β_t and γ_{0t} in the three previously mentioned model equations, one can use the interaction effect of time with the other model covariates along with their main effects.

The posterior function of the model parameters $\Theta = \{\alpha_t, \beta_t, \sigma_z^2, \Sigma_B, \gamma_t\}$ would be:

$$\pi(\Theta|Y^{obs}, R; X) = \frac{f(Y^{obs}, R|X, \Theta)\pi(\Theta)}{\int f(Y^{obs}, R|X, \Theta)\pi(\Theta)d\Theta}. \tag{2.1}$$

where,

$$\pi(\Theta) = \pi(\alpha) \times \pi(\beta) \times \pi(\sigma_z^2) \times \pi(\Sigma_B) \times \pi(\gamma),$$

is the product of some low informative independent priors for the vector of model parameters. Actually one can use independent normal priors with some large variances for the elements of the vectors, α, β and γ along with gamma and Wishart priors for $1/\sigma_z^2$ and Σ_B^{-1}, respectively. Also assume that each individual with incomplete visits, has been observed for $M_i < T_i$ times, hence the joint distribution of the observed responses $Y_i^{obs} = (Y_{i1}, \ldots, Y_{i,M_i})$ and the dropout indicators

$R_i = (R_{i2}, \ldots, R_{i,M_i+1})$ for the i-th individual would be as follows:

$$
\begin{aligned}
f(Y_i^{obs}, R_i | X_i, \Theta) &= \int_{B_i} f(Y_i^{obs} | B_i, X_i) f(R_i | B_i, X_i) \phi(B_i) dB_i \\
&= \int_{B_i} \prod_{j=1}^{M_i} f(Y_{ij} | B_i, X_i) \prod_{j=2}^{M_i+1} f(R_{ij} | B_i, X_i) \phi(B_i) dB_i \\
&= \int_{B_i} \prod_{j=1}^{M_i} f(W_{ij} | B_{i1}, X_i) \times f(Z_{ij} | B_{i2}, X_i)^{(1-W_{ij})} \\
&\times \prod_{j=2}^{M_i+1} f(R_{ij} | B_i, X_i) \phi(B_i) dB_i \\
&= \int_{B_i} \prod_{j=1}^{M_i} \pi_{ij}^{W_{ij}} (1 - \pi_{ij})^{(1-W_{ij})} \left[\frac{1}{\sigma_z \sqrt{2\pi}} exp\{ \frac{(z_{ij} - \mu_{ij})^2}{2\sigma_z^2} \} \right]^{(1-W_{ij})} \\
&\times \prod_{j=2}^{M_i+1} f(R_{ij} | B_i, X_i) \phi(B_i) dB_i,
\end{aligned}
$$

where $\phi(B_i)$ is the density function for the $MVN_q(0, \Sigma_B)$ distribution. The first equality in the above equation is a result of shared parameter approach which leads to the conditional independence of responses and their missingness indicators given B_i. Also the second equality is obtained due to conditional independence of the vector of observed responses along time given B_i. Also the joint density function for the completers who have been observed at all T_i visits would be as follows:

$$
\begin{aligned}
f(Y_i, R_i | X_i, \Theta) &= \int_{B_i} f(Y_i | B_i, X_i) f(R_i | B_i, X_i) \phi(B_i) dB_i \\
&= \int_{B_i} \prod_{j=1}^{T_i} \pi_{ij}^{W_{ij}} (1 - \pi_{ij})^{(1-W_{ij})} \left[\frac{1}{\sigma_z \sqrt{2\pi}} exp\{ \frac{(z_{ij} - \mu_{ij})^2}{2\sigma_z^2} \} \right]^{(1-W_{ij})} \\
&\times \prod_{j=2}^{T_i} f(R_{ij} | B_i, X_i) \phi(B_i) dB_i.
\end{aligned}
$$

Consequently the joint distribution of observed responses and the dropout indicators for all study subjects $f(Y^{obs}, R | X, \Theta)$, would be:

$$
f(Y^{obs}, R | X, \Theta) = \prod_{i=1}^{N} f(Y_i^{obs}, R_i | X_i, \Theta),
$$

which should be substituted in to the posterior function of equation 2.1 to draw Bayesian parameter estimations.

3. Application

3.1. *Toenail Data*

The data used in this section are extracted from a randomized, double-blind, parallel group, multicenter study to make a comparison between two oral treatments (in the sequel coded as A and B) for Toenail Dermatophyte Onychomycosis (TDO) [see [6] for more details]. TDO is a common toenail infection, difficult to treat, affecting more than 2 out of 100 persons ([18]). Anti-fungal compounds, classically used for treatment of TDO, need to be taken until the whole nail has grown out healthy. The development of new such compounds, however, has reduced the treatment duration to 3 months. The aim of this study was to compare the efficacy and safety of 12 weeks of continuous therapy with treatment A to that of treatment B.

In this study, 396 patients (198 in each treatment group), distributed over 36 centers, were randomized to be examined. Subjects were followed during 3 months of treatment and followed further, up to a total of 12 months. Measurements were taken at baseline, every month during treatment, and every 3 months afterwards, resulting in a maximum of 7 measurements per subject. At the first occasion, the treating physician indicates one of the affected toenails as the target nail, the nail which will be followed over time. Finally, the resulting sample includes 148 and 150 subjects, in group A and group B, respectively.

The important response of interest in this study is the patient's Unaffected Nail Length (UNL) in millimeters which is measured from the nail bed to the infected part of the nail, which is always at the free end of the nail. Figure 1 shows the histogram of UNL variable during all study visits. According to these plots, in all occasions there are a bulk of zero values along with some other positive values which are rightly skewed. Hence this response variable should be considered as a semi-continuous variable for further analysis of these data set.

Due to a variety of reasons, the vector of outcomes has been completely measured only for 226 (76%) out of the 298 participants and the others have been dropped out from the study at the second visit or after that with a monotone pattern. Figure 2 shows the observed mean profile of the semicontinuous response variable, UNL for different treatment groups. This figure shows that the unaffected nail length was increased in the treatment period (first three months) and continued after it. However, the increase is somehow higher for patients in group B comparing with those in group A. Also Figure 3 displays the observed mean profile of the non-zero response values which have nearly the same pattern as Figure 2 but a little difference is observable in the starting and finishing time visits.

The frequency of zero values of the semicontinuous response variable along the time for different treatment groups are displayed in Figure 4 via bar charts. This figure illustrates the existence of a non-neglible mass of zero values in all visits where the frequencies are reduced through the end of study.

To understand the nature of dropout mechanism based on the available cases, we have considered a binary logistic model for the dropout at the 9th month as a function of previous response of UNL (at month 6) and the treatment group involved. Figure 5 shows that the dropout probability in month 9 slightly increases as the UNL in month 6 increases and this probability is higher for patients in group B compared with those in group A which shows the need for the consideration of the dropout mechanism as a part of model estimation process.

3.2. *Model For The Toenail Data*

The Toenail study is a well-known longitudinal data which have been analyzed by different authors to assess the effect of various therapies on the unaffected nail length (UNL) during the time. All of these analyses involve some linear mixed model with the normality assumption for the UNL variable (For example, [24] and [23]). However, In this section, we attempt to model the evolution of the UNL response variable over the time applying a shared parameter zero-inflated log normal model with possibility of non random dropout via a Bayesian approach. Let the response vector

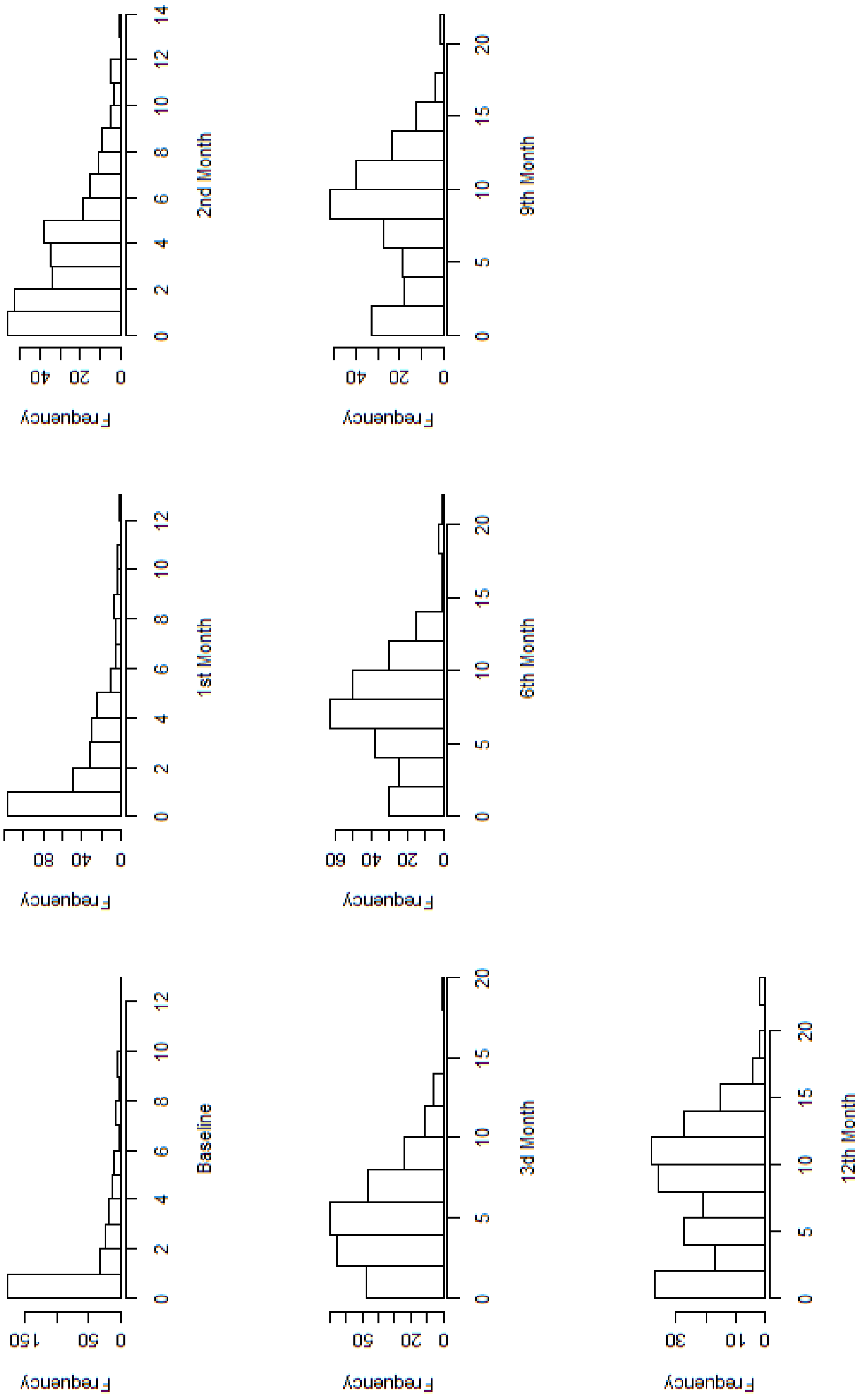

Fig. 1. Histogram of *UNL* response variable during the time

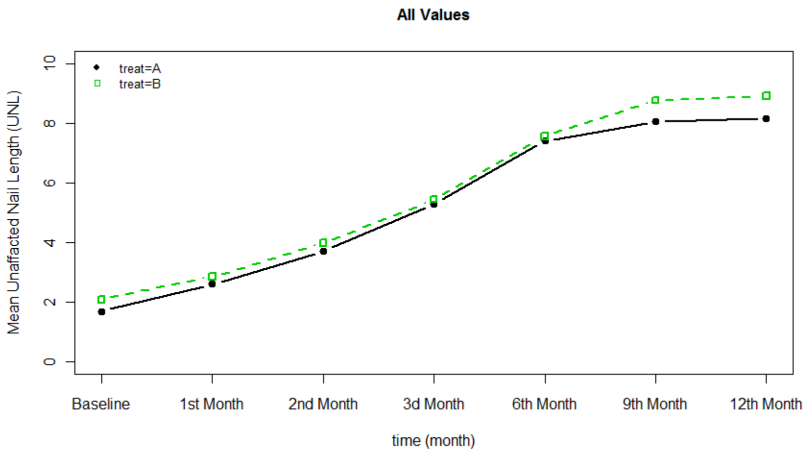

Fig. 2. Observed mean profile of UNL for different treatment groups

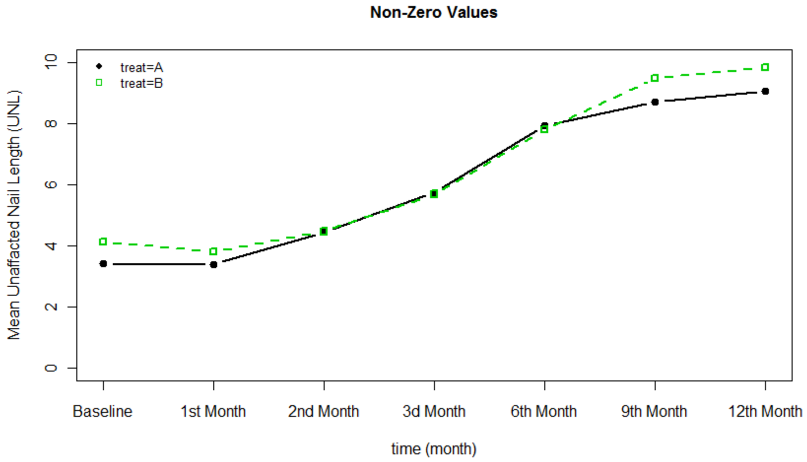

Fig. 3. Observed mean profile of Non-Zero values of UNL for different treatment groups

for the ith individual at time t be denoted by $Y_{it} = (UNL^{*}_{it}, W_{it})$ where UNL^{*}_{it} indicates the nonzero values of UNL_{it} and W_{it} is the zero value indicator for the UNL_{it} variable. Also assume that R_{it} represents the non dropout indicator of Y_{it} as is defined in Section 2. Now we use a Bayesian zero-inflated shared parameter model similar to that proposed in Section 2, with a logistic model for the

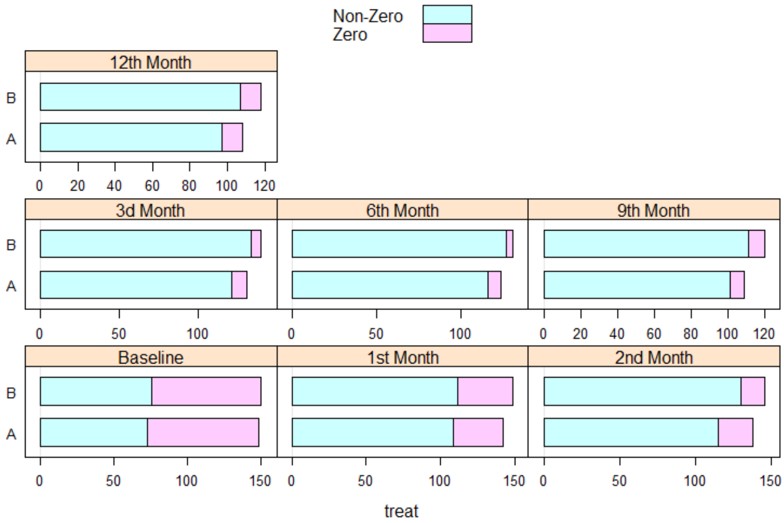

Fig. 4. Zero and Non-Zero Frequencies For UNL in two different treatment groups during the study

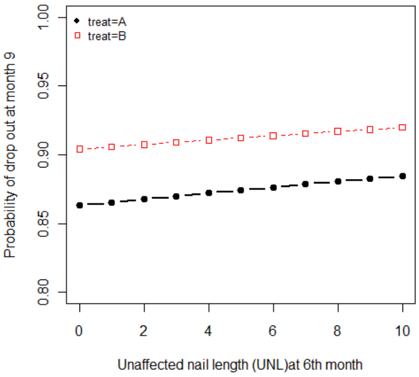

Fig. 5. Probability of Dropout at 9th month versus UNL at month 6 for various treatment groups.

dropout mechanism which can be summarized as follows:

$$\begin{aligned}
&Logit\ P(W_{it}=1|X_{it}) = \alpha_0+\alpha_1 treat_i+\alpha_2 t+\alpha_3 treat_i\times t+b_{1i},\\
&Log\ [UNL^*_{it}]|UNL_{it}>0 \sim N(\mu_{it},\sigma^2),\\
&\mu_{it}=\beta_0+\beta_1 treat_i+\beta_2 t+\beta_3 treat_i\times t+b_{2i},\\
&Logit\ P(R_{it}=1|X_{it},R_{i,t-1}=1,b_{1i},b_{2i}) = \gamma_0+\gamma_1 treat_i+\gamma_2 t\\
&+\gamma_3 treat_i\times t+\gamma_4 b_{1i}+\gamma_5 b_{2i},\\
&\ t=1,\ldots,7,\quad i=1,\ldots,298,
\end{aligned}\tag{3.1}$$

where,

$$b_i = \begin{bmatrix} b_{1i} \\ b_{2i} \end{bmatrix} \sim BVN(0,\Sigma_b),\ \Sigma_b = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

Also the following independent low informative prior distributions are considered for the model parameters:

$$\begin{aligned} \alpha_j,\beta_j &\sim N(0,100),\ j=0,\ldots,3 \\ \gamma_k &\sim N(0,100),\ k=0,\ldots,5 \\ 1/\sigma^2 &\sim \Gamma(0.2,0.001) \\ \Sigma_b^{-1} &\sim Wishart(\begin{bmatrix} 1\,0 \\ 0\,1 \end{bmatrix},2) \end{aligned}$$

Hence, the posterior distribution of the model parameters would be,

$$\pi(\Theta|Y^{obs},R;X) = \frac{f(Y^{obs},R|X,\Theta)\pi(\Theta)}{\int f(Y^{obs},R|X,\Theta)\pi(\Theta)d\Theta}.$$

where,

$$\begin{aligned} f(Y^{obs},R|X,\Theta) = \prod_{i=1}^{298} \int_{b_{1i}} \int_{b_{2i}} \prod_{j=1}^{M_i} \pi_{ij}^{W_{ij}}(1-\pi_{ij})^{(1-W_{ij})} \left[\frac{1}{\sigma\sqrt{2\pi}} exp\{ \frac{(UNL_{ij}^{*}-\mu_{ij})^2}{2\sigma^2} \right]^{(1-W_{ij})} \\ \times \prod_{j=2}^{M_i+1} f(R_{ij}|b_{1i},b_{2i},X_i)\phi(b_{1i},b_{2i})db_{2i}\, db_{1i} \end{aligned}$$

Given the complexity of the model and its posterior function which does not have a closed form, Bayesian inferences need to be based on simulation techniques. For example Gibbs sampling or Markov Chain Monte Carlo (MCMC) methods can be used to make inferences based on values drawn from the joint posterior density which will be described more in the next Section.

3.3. *Posterior Computations based on Gibbs Sampling*

Bayesian parameter estimation for the zero-inflated lognormal model described in the system of equations 3.1 proceeds via drawing samples from the following posterior function:

$$\pi(\alpha,\beta,\gamma,\sigma^2,\Sigma_b|Y^{obs},R,X),$$

which is equivalent to drawing from,

$$\pi(\alpha,\beta,\gamma,\sigma^2,\Sigma_b,b_1,b_2,Y^{miss}|Y^{obs},R,X).$$

These draws are obtained using Gibbs sampling based on the data augmentation algorithm (see [21]) implemented via iteratively sampling the following full conditional distributions of the random and fixed model parameters and the missing variables:

$$
\begin{aligned}
&(i)\ \pi[\alpha,\beta,\gamma,\sigma^2,\Sigma_b,b_1,b_2|Y,R,X]\\
&\quad(i.1)\ \pi[\alpha|\beta,\gamma,\sigma^2,\Sigma_b,Y,R,X,b_1,b_2]\\
&\quad(i.2)\ \pi[\beta|\alpha,\gamma,\sigma^2,\Sigma_b,Y,R,X,b_1,b_2]\\
&\quad(i.3)\ \pi[\gamma|\alpha,\beta,\sigma^2,\Sigma_b,Y,R,X,b_1,b_2]\\
&\quad(i.4)\ \pi[b_1,b_2|\alpha,\beta,\gamma,\sigma^2,\Sigma_b,Y,R,X]\\
&\quad(i.5)\ \pi[\Sigma_b|\alpha,\beta,\gamma,\sigma^2,Y,R,X,b_1,b_2]\\
&(ii)\ \ \pi[Y^{mis}|\alpha,\beta,\gamma,\sigma^2,\Sigma_b,Y^{obs},R,X,b_1,b_2],
\end{aligned}
$$

where the first block represent an outer Gibss Sampling to draw from the posterior function of the model parameters given the full response (augmented) and non dropout vectors and the second block is related to the posterior function of the missing variables given all model parameters and the observed responses. In practical problems, however, not all of the above full conditional posterior distributions are known or have closed form where the rejection sampling ([17], adaptive rejection sampling ([10]), the Metropolis algorithm ([15]), or the Metropolis-Hastings algorithm ([11]) are commonly used for drawing values from these conditional distributions (see also [4]). Here, we will use WinBUGS software ([20]) to implement the above Gibss Sampling procedure to obtain the draws from the posterior function and to derive inferences on parameters of interest. It should be noted that in the use of this software some difficulty arises due to zero-inflated lognormal distribution needed for the vector of response variables which is not pre-designated in this software. We have used the ”zeros trick” procedure in WinBUGS for defining this new sampling distributions.

3.4. *Results for The Toenail Data*

In this Section we will fit four different Bayesian models for the Toenail data to be compared. Model (I) and (II) are shared parameter linear mixed models which are based on the normality assumption for the UNL variable over the time with the MAR and MNAR mechanism, respectively. The following equations are those used in Model (II) which are based on the normality assumption for UNL_{it} variable,

$$
\begin{aligned}
&UNL_{it} \sim N(\mu_{it},\sigma^2),\\
&\mu_{it} = \beta_0 + \beta_1 treat_i + \beta_2 t + \beta_3 treat_i \times t + b_i,\\
&\\
&Logit\ P(R_{it} = 1|X_{it}, R_{i,t-1} = 1, b_{1i}, b_{2i}) = \gamma_0 + \gamma_1 treat_i + \gamma_2 t\\
&+\gamma_3 treat_i \times t + \gamma_5 b_i,\\
&\\
&\quad t = 1,\ldots,7, \quad i = 1,\ldots,298,
\end{aligned}
$$

where,

$$b_i \sim N(0,\sigma_2^2).$$

Also the first model (Model (I)) has the same structure as the above equations where it is assumed that $\gamma_5 = 0$. Model (III) and (IV) are shared parameter zero-inflated lognormal models, with MAR

Table 1. Results of Bayesian analysis of Toenail data corresponding to Model (I)-(IV).

	Model (I)		Model (II)		Model (III)		Model (IV)	
Par.	Est.	S.D	Est.	S.D.	Est.	S.D.	Est.	S.D.
α_0	–	–	–	–	-0.62*	0.25	-0.60*	0.22
α_1	–	–	–	–	-0.19	0.33	-0.18	0.319
α_2	–	–	–	–	-0.65*	0.07	-0.61*	0.07
α_3	–	–	–	–	-0.09	0.11	-0.11	0.10
β_0	1.60*	0.26	1.614*	0.26	0.71*	0.06	0.71*	0.05
β_1	0.22	0.37	0.20	0.36	0.08	0.08	0.10	0.06
β_2	1.20*	0.04	1.19*	0.04	0.24*	0.01	0.24*	0.01
β_3	0.08	0.06	0.08	0.06	0.0002	0.01	-0.003	0.01
γ_0	3.26*	0.38	3.31*	0.40	3.31*	0.39	3.98*	0.60
γ_1	0.87	0.63	0.85	0.65	0.79	0.60	0.91	0.70
γ_2	-0.09	0.09	-0.09	0.10	-0.10	0.09	-0.19	0.12
γ_3	-0.15	0.15	-0.14	0.15	-0.13	0.14	-0.14	0.16
γ_4	–	–	–	–	–	–	-0.99*	0.39
γ_5	–	–	0.06	0.06	–	–	-2.94*	1.22
σ_{UNL}	2.53*	0.05	2.53*	0.05	0.48*	0.01	0.48*	0.01
σ_1	–	–	–	–	4.04*	0.72	3.85*	0.74
σ_2	2.57*	0.12	2.56*	0.12	0.28*	0.03	0.27*	0.03
ρ	–	–	–	–	-0.87*	0.11	-0.81*	0.11

and MNAR assumptions for the dropout mechanism, respectively. The model equation for Model (IV) is presented in equation 3.1. Also Model (III) has the same parameter structure as Model (IV) with the exception of non-ignorability parameters γ_5 and γ_6 which are zero in this ignorable model.

Table 1 presents the results of the Bayesian estimation for Model (I)-(IV). To draw inferences, we have performed the iterative Gibbs sampling procedure in 100,000 iterations, ignoring the first 90,000 iterations as burn-in to get closer to the convergence, so that the inferences about the model parameters are obtained using 10,000 remaining iterations. We use the posterior mean of each parameter as its estimate and the sample standard deviation as the estimated standard deviation of the parameter of interest.

For the comparative study of the above fitted models, the deviance information criterion (*DIC*) measure ([20]) can be used which assess model complexity and is a good measure to compare different models. The default DIC option in WinBUGS is not available for the zero-inflated models presented in this paper. So that we have calculated these amounts outside of WinBUGS by importing the posterior draws of the parameters, random effects and the augmented missing values into the R software (R Development Core Team, 2007), calculating the following two quantities conditional on the random effect values ([3]) :

$$\bar{D}(\Theta) = E(D(\Theta)|y),$$
$$\hat{D}(\Theta) = D(E(\Theta|y)),$$

where $D(\Theta)$ is the deviance function defined as $-2 \times log(likelihood(\Theta))$. Actually, $\bar{D}(\Theta)$ represents the posterior mean of the deviance and $\hat{D}(\Theta)$ indicates a point estimate of the deviance obtained by substituting in the posterior means of Θ. Using the above two quantities yields the P_D (the effective

Table 2. Model comparision statistics for Model (I)-(IV).

Model	$\bar{D}(\Theta)$	P_D	DIC
Model (I)	9629.91	333.43	9963.34
Model (II)	9632.31	334.81	9967.12
Model (III)	3842.09	360.02	4202.11
Model (IV)	3777.07	395.32	4172.39

number of parameters) and *DIC* (Deviance Information Criterion) measures as follows:

$$\begin{aligned} P_D &= \bar{D}(\Theta) - \hat{D}(\Theta), \\ DIC &= P_D + \bar{D}(\Theta). \end{aligned}$$

The model with the smallest DIC is estimated to be the model that would best predict a replicate dataset of the same structure as that currently observed.

The amount of P_D and DIC quantities for Model (I)-(IV) are presented in Table 2. The results show that Model (IV) has the smallest DIC value compared with the three other models which reveals the need for the zero-inflated log normal distribution for the *UNL* response variable along with the MNAR mechanism for the dropout process occurred in these data.

According to the posterior estimates of Model (IV) in Table 1 as the preferred model for these data, the odds of having zero unaffected nail length decreases through the end of study. Also for the patients with non-zero unaffected nail length, the nail length grows more in the later visits. The parameter corresponding to the treatment covariate (β_1) is significant at 0.1 error level for Model (IV) while the results show that the parameters corresponding to the treatment and its interaction with time are not significant for the other three models. The group *B* patients with non-zero *UNL* have 0.1 mm longer unaffected nail length on average when comparing with group *A* patients with non-zero *UNL*. Parameters in the Σ_b matrix for the covariance structure of shared random effects are significantly estimated in this model which means that the occurrence of zero and non-Zero values of UNL variable are correlated and that there is significant correlation among UNL measurements for each patient in different Months. Also, the significant coefficients for the two shared random effect parameters in the dropout mechanism (γ_4 and γ_5) illustrates the non random mechanism for the monotone removal of the patients from the study.

Also for checking the convergence of the MCMC results for Model (IV) which includes Gibss sampling steps along with Metropolis algorithm for unknown conditional posterior functions of the model parameters (see Section 3.3), we should take care about the acceptance rate of the Metropolis algorithm. Figure 6 shows the minimum, maximum and average acceptance rate averaged over 100 iterations as the Metropolis algorithm adapts over the first 4,000 iterations. As it is shown in this plot, the rate lies between the two horizontal lines which shows the accurate use of Metropolis algorithm in the Gibbs sampling steps of the Bayesian Model (IV) .

Also to examine if the posterior simulations of the model parameters have been stabilized, Figure 7 and Figure 8 have been plotted using posterior summaries of the model parameters in the last 10,000 iterations. Actually, Figure 7 plots out the running posterior mean in the last 10,000 iterations, with 95% confidence intervals against iteration number and Figure 8 illustrates the trace plots of the posterior sample values versus iteration for different model parameters. These plots show that for the last 10,000 iterations of the MCMC procedure, the posterior sample values and their means

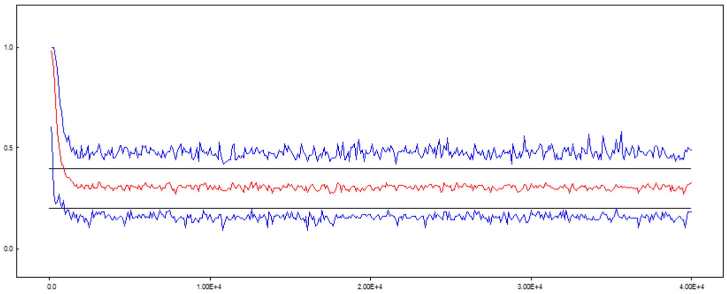

Fig. 6. Minimum, maximum and average Metropolise acceptance rate

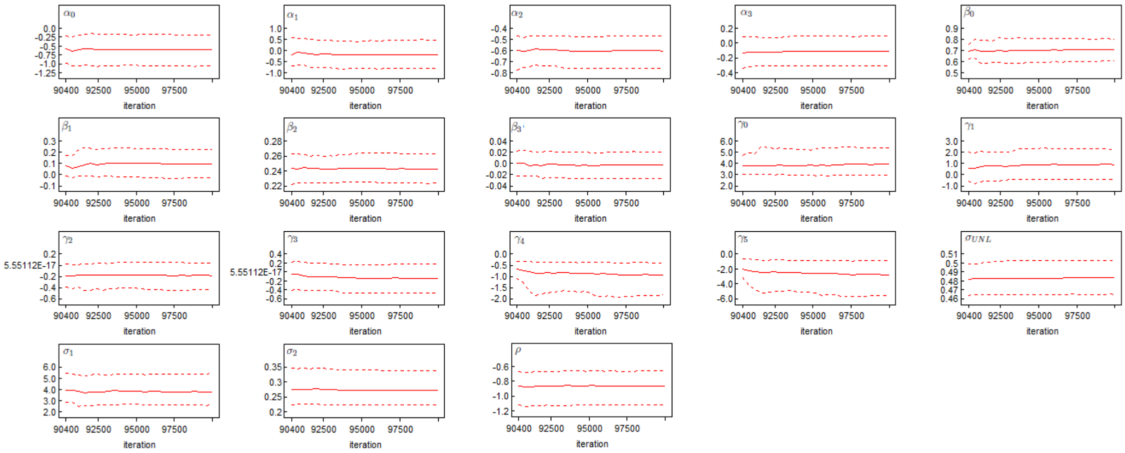

Fig. 7. Runinig posterior mean with 95% confidence intervals against iteration number

for all the model parameters have a stable state with no considerable fluctuations which means that the chain has been converged acceptably.

4. Conclusion

Repeated measurements or follow up studies allow the researcher to assess variations in the interesting variables for each individual as the time increases. The occurrence of missing data is a problem which is commonly encountered in various researches, including both cross-sectional and longitudinal or follow-up studies. Incomplete data in each setting call for some additional challenges in the modeling step where the researcher should care about the possibility of non-random missingness or dropout mechanism. Recently, there has been extensive methodological researches on analyzing semicontinuous responses. However, none of them have considered the possibility of incomplete semicontinuous data in their methods. In this paper, we have presented a flexible Bayesian model

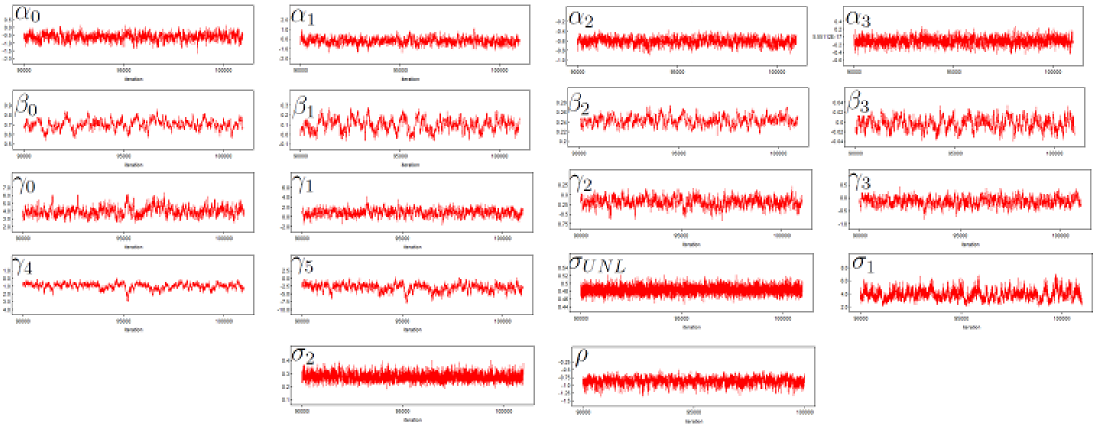

Fig. 8. Trace plots of the posterior sample values against iteration number

for incomplete semicontinuous longitudinal data with the possibility of non-random dropouts. Actually, our model is a two part model along with a dropout mechanism which are correlated due to shared random effect parameters. The proposed approach has been applied for the analysis of a well known longitudinal data set about Toenail infection (TDO) where the response of interest greatly suffer from excess zeros along some positive continuous values at each time visit. We have also compared four different Bayesian models for these data with differing response models and dropout structures where the results show that our proposed model with semicontinuous responses and non-random dropout has the best performance according to some Bayesian goodness of fit indices.

References

[1] P.S. Albert, On the interpretation of marginal inference with a mixture model for clustered semi-continuous data, *Biometrics* **61** (2005) 879–880.

[2] P.S. Albert and J. Shen, Modelling longitudinal semicontinuous emesis volume data with serial correlation in an acupuncture clinical trial, *Journal of the Royal Statistical Society: Series C* **54** (2005) 707–720.

[3] G. Celeux, F. Forbes, C.P. Robert and D.M. Titterington, Deviance information criteria for missing data models, *Bayesian Analysis* **1** (2006) 651–74.

[4] S. Chib and E. Greenberg, Understanding the Metropolis-Hastings Algorithm, *The American Statistician* **49** (1995) 327–335.

[5] M.K. Cowles, B.P. Carlin and J.E. Connett, Bayesian Tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness, *Journal of the American Statistical Association* **91** (1996) 86–98.

[6] M. De Backer, P. De Keyser., C. De Vroey and E. Lesaffre, A 12-week treatment for dermatophyte toe onychomycosis: terbinafine 250mg/day vs. itraconazole 200mg/day a double-blind comparative trial, *British Journal of Dermatology* **134** (1996) 16–17.

[7] P.J. Diggle, P. Heagerty, K.Y. Liang and S.L. Zeger, *Analysis of Longitudinal Data*, (University Press, Oxford, 2002)

[8] N. Duan, W.G.J.r. Manning, C.N. Morris and J.P. Newhouse, A comparison of alternative models for the demand for medical care (Corr: V2 P413). *Journal of Business and Economic Statistics* **1** (1983) 115–126.

[9] P. Ghosh and P.S. Albert, A Bayesian analysis for longitudinal semicontinuous data with an application to an acupuncture clinical trial, *Comput Stat Data Anal.* **53** (2009) 699–706.

[10] W.R. Gilks and P. Wild, Adaptive Rejection Sampling for Gibbs Sampling, *Applied Statistics* **41** (1992) 337-348.

[11] W.K. Hastings, Monte Carlo Sampling Method Using Markov Chains and Their Applications, *Biometrika* **57** (1970) 97–109.

[12] N. Li, D.A. Elashoff, W.A. Robbins and L. Xun, A hierarchical zero-inflated log-normal model for skewed responses, *Statistical Methods in Medical Research* **20** (2011) 175–189.

[13] R.J.A. Little, Modeling the dropout mechanism in repeated-measures studies, *Journal of the American Statistical Association* **90** (1995) 1112–1121.

[14] S.E. Lu, Y. Lin and W.J. Shih, Analyzing Excessive No Changes in Clinical Trials with Clustered Data, *Biometrics* **60** (2005) 257–267.

[15] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller, Equations of State Calculations by Fast Computing Machines, *Journal of Chemical Physics* **21** (1953) 1087–1091.

[16] M.K. Olsen and J.L. Schafer, A two-part random-effects model for semicontinuous longitudinal data, *Journal of the American Statistical Association*, **96** (2001) 730–745.

[17] B. Ripley, *Stochastic Simulation*, (Wiley, New York, 1987)

[18] D.T. Roberts, Prevalence of dermatophyte onychomycosis in the United Kingdom: Results of an omnibus survey, *British Journal of Dermatology 126 Suppl.* **39** (1992) 23–27.

[19] J.W. Robinson, S.L. Zeger and C.B. Forrest, A hierarchical multivariate two-part model for profiling providers' effects on health care charges, *Journal of the American Statistical Association* **101** (2006) 911–923.

[20] D. Spiegelhalter, A. Thomas, N. Best and D. Lunn, MRC Biostatistics Unit Institute of Public Health and Department of Epidemiology Public Health, Imperial College School of Medicine, WinBUGS User Manual, Version 1.4., Available at: http://www.mrc-bsu.cam.ac.uk/bugs (2005)

[21] M. Tanner and W.H. Wong, The Calculation of Posterior Distribution by Data Augmentation (with discussion), *Journal of the American Statistical Association* **82** (1987) 528–550.

[22] J. Tobin, Estimation of relationships for limited dependent variables, *Econometrica* **26** (1958) 24–36.

[23] G. Verbeke, E. Lesaffre, B. Spiessens, The practical use of different strategies to handle dropout in longitudinal studies, *Drug Information Journal* **35** (2001) 419–434.

[24] G. Verbeke and G. Molenberghs, *Linear Mixed Models for Longitudinal Data*, (Springer, New York, 2000).

[25] M. Zhang, R.L. Strawderman, M.E. Cowen and M.t. Wells, Bayesian inference for a two-part hierarchical model: An application to profiling providers in managed health care, *Journal of the American Statistical Association* **101** (2006) 934–945.