Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Jaffa et al., J Biomet Biostat 2014, 5:4
DOI: 10.472/2155-6180.1000203

Research Article Open Access

A Joint Modeling Approach for Right Censored High Dimensional Multivariate Longitudinal Data

Miran A Jaffa[1]*, Mulugeta Gebregziabher[2] and Ayad A Jaffa[3,4]

[1]*Epidemiology and Population Health Department, Faculty of Health Sciences, American University of Beirut, Beirut, Lebanon, P.O.Box 11-0236 Riad El-Solh/Beirut, Lebanon 1107 2020, Lebanon*
[2]*Department of Public Health Sciences, Medical University of South Carolina, Charleston, South Carolina, 135 Cannon Street, Suite 303, Charleston, SC 2942., USA*
[3]*Department of Biochemistry and Molecular Genetics, Faculty of Medicine, American University of Beirut, Beirut, Lebanon, P.O.Box 11-0236 Riad El-Solh/Beirut, Lebanon 1107 2020, Lebanon*
[4]*Department of Medicine, Medical University of South Carolina, Charleston, SC 29425, USA*

Abstract

Analysis of multivariate longitudinal data becomes complicated when the outcomes are of high dimension and informative right censoring is prevailing. Here, we propose a likelihood based approach for high dimensional outcomes wherein we jointly model the censoring process along with the slopes of the multivariate outcomes in the same likelihood function. We utilized pseudo likelihood function to generate parameter estimates for the population slopes and Empirical Bayes estimates for the individual slopes. The proposed approach was applied to jointly model longitudinal measures of blood urea nitrogen, plasma creatinine, and estimated glomerular filtration rate which are key markers of kidney function in a cohort of renal transplant patients followed from kidney transplant to kidney failure. Feasibility of the proposed joint model for high dimensional multivariate outcomes was successfully demonstrated and its performance was compared to that of a pairwise bivariate model. Our simulation study results suggested that there was a significant reduction in bias and mean squared errors associated with the joint model compared to the pairwise bivariate model.

Keywords: Informative right censoring; Joint modelling; Likelihood based approach; Multivariate longitudinal outcomes; Random effect; Slope estimation

Introduction

Kidney function is assessed using the markers serum creatinine, blood urea nitrogen (BUN), and estimated glomerular filtration rate (eGFR). All three markers are needed to assess kidney function since each marker has its own limitation. For instance, serum creatinine varies inversely with glomerular filtration rate (GFR) and creatinine levels are influenced by age and gender [1]. Whereas, BUN levels could fluctuate with protein intake, catabolism and tubular reabsorption of urea [2] and eGFR could be less accurate in obese individuals and those with normal or near normal GFR [3-5]. Also in most clinical settings the exact values of BUN, creatinine and eGFR provide little information on disease severity. What is more important is monitoring the rate of change in serum creatinine, BUN and eGFR over time to determine disease progression or to ascertain if state of disease is stable or changing [6]. This is done by taking repeated measures of these markers on the same patient and by calculating the rate of change or slopes for each of these markers to provide an evaluation of disease progression over time. This is critically significant for patients who undergo kidney transplant where a routine follow-up evaluation for their kidney function is mandated to determine how well their kidneys are functioning post-transplantation and to verify if graft failure is likely to transpire. For the cohort of renal transplantation we considered in this study, longitudinal measures of creatinine and BUN are recorded and eGFR levels are computed post transplant repeatedly over time till patient experiences renal graft failure. Patients who experience graft failure will therefore have an incomplete set of repeated measures on their creatinine, BUN and eGFR, a situation referred to as informative right censoring.

Slope estimation for these outcomes is complicated in the presence of informative right censoring and special method of analysis that accounts for this problem should be conducted so that valid inferences are reached. If this type of censoring is ignored or treated as non-informative, it could result in biased estimates and lead to inaccurate inferences [7]. Since informative right censoring is widespread in longitudinal studies several statistical methods were developed for slope estimation that account for right censoring [8-16]. However, all these methodological approaches have been developed for a single longitudinal outcome with informative right censoring. Although multiple outcomes are commonly encountered in medical research setting methodological approaches for slope estimation for multivariate longitudinal outcomes with informative right censoring are still not well developed [17]. This paucity is due to the level of complexity that accompanies such approaches where informative right censoring and the different correlations should be adjusted for [18]. This problem becomes much more compounded when the outcomes are of high dimension which results in an increase in the number of parameters in the variance-covariance matrix and in the number of estimates, and hence could lead to convergence problems in many situations. Few studies have developed methods for slope estimation for bivariate longitudinal outcomes adjusting for informative right censoring [19-21]. For example, a joint model for a time to clinical event and for repeated measures over time on surrogate outcomes was presented by Xu and Zeger [22,23]. In this study a multivariate mixed model was used for the joint analysis of multivariate repeated measures data and times to an event with an underlying assumption of conditional independence between the censoring time and the biomarkers given

***Corresponding author:** Miran A Jaffa, Epidemiology and Population Health Department, Faculty of Health Sciences, American University of Beirut, Beirut, Lebanon, P.O.Box 11-0236 Riad El-Solh/Beirut, Lebanon 1107 2020. Lebanon, Tel: +961-1-350000 Ext: 4603; Fax: +961-1-744470; E-mail: ms148@aub.edu.lb

Received July 02, 2014; **Accepted** July 27, 2014; **Published** July 30, 2014

Citation: Jaffa MA, Gebregziabher M, Jaffa AA (2014) A Joint Modeling Approach for Right Censored High Dimensional Multivariate Longitudinal Data. J Biomet Biostat 5: 203. doi:10.4172/2155-6180.1000203

the latent process for each outcome. A pairwise fitting model was proposed to analyze multivariate outcomes [24,25]; but this approach could lead to efficiency loss. He and Luo developed a joint model of the multilevel item response theory (MLIRT) and Cox's proportional hazard model for the time to the dependent terminal event with shared random effects to link the two models [26].

Joint modeling and maximization of the full multivariate likelihood function if feasible is a favorable approach [24]. However, this approach becomes difficult to implement with high dimensional outcomes given the computational complexity and convergence problem that could be encountered in such a setting. In the current article we aim to extend the bivariate model developed by Jaffa et al. [21] to high dimensional multivariate outcomes and to demonstrate its implementation feasibility through its successful convergence. Specifically, we propose a likelihood based approach where we jointly model the censoring process along with the slopes of the multivariate longitudinal outcomes and their variance-covariance matrix in the same multivariate likelihood function. This likelihood function is then maximized to generate slope estimates for the population as well as individual subjects. Estimating the individual slopes provides an assessment for the rate of change for every subject and therefore a case by case prognosis of the disease. This approach accounts for the informative right censoring and the correlation between the longitudinal outcomes. Specifically, the number of visits (before censoring happens due to kidney failure) for every patient is modeled using a discrete probability model with the individual slopes of the outcomes as covariates in the model. The innovation in the proposed model resides in its feasibility to handle high dimensional outcomes by jointly modeling all the slopes, their correlations and the censoring process in one likelihood function and casting the problem in such a way that standard software could be used to generate estimates for multivariate longitudinal outcomes of high dimension. We first used simulated data to assess the performance of the proposed method in terms of bias and efficiency and made comparison with the bivariate approach proposed in Jaffa et al. [21] applied in a pairwise fashion. This comparison enables us to determine if incorporating all the correlations concomitantly in the same likelihood function leads to a better precision than that of the bivariate modeling with pairwise joint modeling of the correlations. Moreover, a small simulation study was conducted on outcomes with different dimensions (7 outcomes and 10 outcomes) to confirm the feasibility of the model for high dimensional data. We then used data from a cohort of renal transplant patients at the Medical University of South Carolina where the markers of interest: creatinine, BUN and eGFR, are measured for each patient in a longitudinal fashion [27]. The objective of this study is to assess kidney function over time by estimating the population and individual slopes corresponding to these kidney markers. The baseline measures for BUN, creatinine, and eGFR recorded prior to the transplantation do not have any impact on kidney function after the transplantation since a new organ is transplanted and post operative measures of these markers determine the progression of disease. This is demonstrated by Kaplan Meier survival analysis which confirmed that there is no significant difference in survival between the group of patients whose baseline pre-transplant eGFR levels is less than 15 ml/min/1.73 m^2 and those who are above this cutoff point (P-value = 0.5). Those with less than 15 ml/min are patients with kidney failure while those with more than 15 correspond to those who have severe to mild kidney damage prior to transplantation. This classification of patients is based on that of Perazella and Reilly [28]. Failing to capture a significant difference in survival between the two groups of patients indicates that the pre-transplant baseline eGFR levels do not affect kidney function post-transplantation and intercept could therefore be discarded in such a clinical setting and the corresponding statistical model could focus on the slopes only.

Multivariate Model

The multivariate model proposed in this manuscript is an extension of the bivariate approach [21]. Specifically the multivariate model could be specified as follows:

Consider a set of i=1,…,n independent subjects, with k=1,…,q multivariate correlated outcomes ($y_{ij1,}$ $y_{ij2,}$…, y_{ijq}) recorded at j=1,..,p predetermined times denoted as $(t_{1k}, t_{2k}, ..., t_{pk})$. These p times are not necessarily equally spaced. Because of right censoring, the ith individual has m_i number of visits or observations, where $m_i \le p$, corresponding to times of measurement $(t_{i1k}, t_{i2k}, ..., t_{im_ik})$. We assume that ($y_{ij1,}$ $y_{ij2,}$…,y_{ijq}) are all observed at all time points before dropout, i.e., the q outcomes are either all measured or all are missing. For example in the renal transplantation example, a patient has measurements on BUN, creatinine, and eGFR till renal failure. In this case measurements are no longer taken on any of the outcomes after failure.

The underlying model for (y_{ij1}, $y_{ij2,}$…, y_{ijq}) at time t_{ik} is assumed to be

$$y_{ijk} = \alpha_{ik} + \beta_{ik} t_{ik} + e_{ijk}, \quad (1)$$

for the ith individual, jth observation and kth outcome, with random effects α_{ik} and β_{ik}, and random error e_{ijk} which is assumed to be normal with mean zero and variance $\sigma^2_{\varepsilon_k}$. We note here that, in a model with informative right censoring, the probability of being censored depends on unobserved data. In our informative censoring model, the probability of being censored depends on the random slopes $(\beta_{i1}, \beta_{i2}, ..., \beta_{iq})$, that will be incorporated in the likelihood function described later in this section. Specifically, our interest is in estimating multivariate population slopes $(\beta_1, \beta_2, ... \beta_q)$ and predicting the multivariate individual random effects $(\beta_{i1}, \beta_{i2}, ..., \beta_{iq})$ that are assumed to be correlated with covariance $\sigma^2_{\beta_{kl}} = \rho_{\beta_k \beta_l} * \sigma_{\beta_k} * \sigma_{\beta_l}$ and follow a multivariate normal distribution:

$$\sim \underline{\beta}_i' = (\beta_{i1}, \beta_{i2}, ..., \beta_{iq}) \sim MVN\left\{ \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{bmatrix}, \begin{bmatrix} \sigma^2_{\beta_1} & \sigma^2_{\beta_{12}} & \cdots & \sigma^2_{\beta_{1q}} \\ & \sigma^2_{\beta_2} & \cdots & \sigma^2_{\beta_{2q}} \\ & & \ddots & \vdots \\ & & & \sigma^2_{\beta_q} \end{bmatrix} \right\}. \quad (2)$$

Hence by having the slopes of the outcomes correlated and incorporating these correlations in the likelihood function (described below), we are therefore accounting for the correlations between the outcomes. Moreover, since interest is in the slopes, the individual's observations y_{ijk} pertaining for very outcome are reduced to the generalized least square estimates (GLS) denoted as b_{ik} that are obtained using SAS proc mixed. The likelihood is a function of the following:

Assumption 1:

$m_i \mid \underline{\beta}_i'$ having Truncated Discrete Distribution with

$$\Pr(M = m_i) = G(\gamma_0 + \gamma_{11}\beta_{i1} + \gamma_{12}\beta_{i2} + \ldots + \gamma_{1q}\beta_{iq}); \quad (3)$$

This assumption states that the number of observations for each individual follows a discrete probability distribution with probability dependent on the individual slopes $(\beta_{i1}, \beta_{i2}, ..., \beta_{iq})$. Also, γ_{10},γ_{11},γ_{12},… ,γ_{1q} are the parameters of the censoring distribution. This discrete censoring distribution can be right-truncated since the study may terminate before observing the withdrawal of all the subjects and

left truncated since at least two observations need to be recorded for every individual so that individual slopes can be estimated. A special case of the discrete censoring distribution is the truncated geometric distribution.

Assumption 1 can then be defined as follows:

$m_i|\underline{\beta}_i'$ ~ Truncated geometric with

$$P_i = F(\gamma_0 + \gamma_{11}\beta_{i1} + \gamma_{12}\beta_{i2} + \ldots + \gamma_{1q}\beta_{iq}) \quad (4)$$

with probability model

$$\Pr(M = m_i) = \left\{1 - F(\gamma_0 + \gamma_{11}\beta_{i1} + \gamma_{12}\beta_{i2} + \ldots + \gamma_{1q}\beta_{iq})\right\}^{m_i - 2}\left\{F(\gamma_0 + \gamma_{11}\beta_{i1} + \gamma_{12}\beta_{i2} + \ldots + \gamma_{1q}\beta_{iq})\right\}^{R_i} \quad (5)$$

With m_i = 2,3,…,p; p is the number of prespecified measurement time points and $F(\gamma_0 + \gamma_{11}\beta_{i1} + \gamma_{12}\beta_{i2} + \ldots + \gamma_{1q}\beta_{iq})$ is a function of individual slopes $\underline{\beta}_i'$. To account for right truncation the geometric distribution was modified by introducing an indicator variable denoted as R_i to the geometric probability function. The indicator variable R_i=1 if censoring occurred and R_i=0 otherwise. In a previous study [16] it was suggested that logistic model may be employed for the sake of simplicity and the function $F(\gamma_0 + \gamma_{11}\beta_{i1} + \gamma_{12}\beta_{i2} + \ldots + \gamma_{1q}\beta_{iq})$ is therefore defined as follows:

$$F(\gamma_0 + \gamma_{11}\beta_{i1} + \gamma_{12}\beta_{i2} + \ldots + \gamma_{1q}\beta_{iq}) = \frac{1}{1 + \exp\left[-\left(\gamma_0 + \gamma_{11}\beta_{i1} + \gamma_{12}\beta_{i2} + \ldots + \gamma_{1q}\beta_{iq}\right)\right]} \quad (6)$$

Note that if censoring parameters $\gamma_{11},\gamma_{12},\ldots,\gamma_{1q}$ are all equal to zero then the drop-out process does not depend on any of the outcomes and is therefore non-informative. In case of dropout, it is assumed that no more measurements are recorded on all the outcomes. Thus each subject will have the same number of observations for all q outcomes.

Assumption 2:

Given m_i, the observed individuals' GLS estimates b_{ik} for every outcome are assumed to be normally distributed with mean β_{ik} and variance $\sigma^2_{b_{ik}}$

The censoring process is viewed as informative or non-ignorable in the sense that the censoring mechanism is dependent on the unobserved random vector $\underline{\beta}_i'$. In this context, $\underline{\beta}_i'$ is both a parameter in the distribution of the vector of GLS estimates $\underline{b}_i'$ and is unobserved itself. The marginal likelihood, integrated over the unobserved random effects $\underline{\beta}_i'$ is maximized to obtain the maximum likelihood estimates for the population slopes and censoring parameters, and empirical Bayes estimates for the individual slopes $\underline{\beta}_i'$. The joint distribution of m_i, and $\underline{b}_i'$ is used in the likelihood function as follows:

$$L = L\left(\underline{\beta}, \sigma^2_{\beta_1}, \sigma^2_{\beta_2}, \ldots, \sigma^2_{\beta_q}, \sigma^2_{\beta_{12}}, \sigma^2_{\beta_{13}}, \ldots, \sigma^2_{\beta_{uq}}, \gamma_0, \gamma_{11}, \gamma_{12}, \ldots \gamma_{1q}\right)$$

$$= \prod_{i=1}^{n} \int f\left(m_i, \underline{b}_i'\right) f\left(\underline{\beta}_i'\right) d\beta_i$$

$$= \prod_{i=1}^{n} \int f\left(\underline{\beta}_i'\right) f\left(m_i \mid \underline{\beta}_i'\right) f\left(\underline{b}_i' \mid m_i\right) d\beta_i,$$

$$= \prod_{i=1}^{n} \int \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\left[\left(\beta_{ij} - \beta_j\right)' \Sigma^{-1}\left(\beta_{ij} - \beta_j\right)\right]\right)$$

$$* \left\{1 - F(\gamma_0 + \gamma_{11}\beta_{i1} + \gamma_{12}\beta_{i2} + \ldots + \gamma_{1q}\beta_{iq})\right\}^{m_i - 2}\left\{F(\gamma_0 + \gamma_{11}\beta_{i1} + \gamma_{12}\beta_{i2} + \ldots + \gamma_{1q}\beta_{iq})\right\}^{R_i} \quad (7)$$

$$* \frac{1}{2\pi\sigma_{b_{i1}}\sigma_{b_{i2}}\ldots\sigma_{b_{iq}}} \exp\left\{\frac{-1}{2}\left[\left(\frac{b_{i1} - \beta_{i1}}{\sigma_{b_{i1}}}\right)^2 + \left(\frac{b_{i2} - \beta_{i2}}{\sigma_{b_{i2}}}\right)^2 + \cdots + \left(\frac{b_{iq} - \beta_{iq}}{\sigma_{b_{iq}}}\right)^2\right]\right\} d\beta_i$$

The log of this likelihood is maximized to obtain estimates of the population slopes $(\beta_1,\beta_2,\ldots,\beta_q)$ in addition to the parameters of the dropout model $(\gamma_0,\gamma_{11},\gamma_{12},\gamma_{1q})$. The Empirical Bayes estimates of the individual random slopes $(\beta_{i1},\beta_{i2},\ldots,\beta_{iq})$ are obtained via the approach described [29]. The slopes $\underline{\beta}_i'$ of the q multivariate longitudinal outcomes are assumed to follow a multivariate normal distribution and the correlations between these slopes are included in the variance-covariance matrix denoted as Σ^{-1} and hence are incorporated in the likelihood function.

SAS procedure NLMIXED (SAS 9.3 Institute Inc.) is used to implement the proposed approach [30].

Simulation Study

A simulation study was conducted to assess the performance of the multivariate model with three outcomes in comparison to the bivariate model described by Jaffa [21] assuming geometric distribution for the patients' number of visits. Performance was determined by bias, and means squared errors for the population and individual slopes denoted as MSEa and MSEb respectively and evaluated as follows:

$$MSEa = \frac{1}{r}\sum_{s=1}^{r}\left(\hat{\beta}_s - \beta\right)^2 \text{ and } MSEb = \frac{1}{nr}\sum_{s=1}^{r}\sum_{i=1}^{n}\left(\hat{\beta}_{is} - \beta_{is}\right)^2 \quad (8)$$

with n being the number of subjects in each data set and r being the total number of replications. The bivariate model was used in a pairwise fashion for all possible pairwise combination of outcomes. The average of bias and that of MSEa and MSEb were computed respectively for every outcome. In the proposed multivariate model the slopes of the three outcomes and their correlations were all concurrently incorporated in the same likelihood function which is maximized to obtain the slopes estimates. This simulation study enables us to determine whether maximizing the multivariate likelihood function which incorporates all the slopes for the outcomes on which the censoring process depends has any effect on the accuracy of the estimates compared to the bivariate model that accounts for only two of these outcomes and ignores the rest. Specifically this comparison enables us to determine whether incorporating all slopes simultaneously in the likelihood function has any effect on the precision and accuracy of the slope estimates assessed by bias and mean squared errors compared to the pairwise bivariate model. The number of visits per individual was randomly generated from the truncated geometric distribution that depended on three censoring parameters γ_{11},γ_{12}, and γ_{13}. The number of observations ranged from 2 to 7 recorded at prespecified time points (t_{ij}) 0, 1, 3, 6, 12, 24, and 36 months and a linear relationship was assumed between each outcome and $\log(t_{ijk}+1)$. Thus, the model for the outcomes is $y_{ijk} = \alpha_{ik} + \beta_{ik}\log\left(t_{ijk}+1\right) + e_{ijk}$ for k=1,2,3. Errors e_{ijk} were assumed to be normally distributed with mean zero and variance $\sigma^2_{\varepsilon k}$. The parameters used in the simulation study were the ones acquired from the renal transplant with estimated slopes $\beta_1 = \beta_{\text{creatinine}}$=-0.089, $\beta_2 = \beta_{\text{BUN}}$=-0.142, and $\beta_3 = \beta_{\text{eGFR}}$=0.302, slope variances of $\sigma^2_{\beta_1} = 0.255, \sigma^2_{\beta_2} = 0.269$, and $\sigma^2_{\beta_3} = 1.417$, and error variances $\sigma^2_{\varepsilon_1} = 0.078, \sigma^2_{\varepsilon_2} = 0.08$ and $\sigma^2_{\varepsilon_3} = 0.14$. Different values of correlation coefficient between the three outcomes were considered ranging from low to high correlations and 2000 replications each with size of 200 were generated. In specific, low correlations with $\rho_{\beta_1\beta_2} = 0.1$, $\rho_{\beta_1\beta_3} = 0.2$, $\rho_{\beta_2\beta_3} = 0.3$, mid correlation with $\rho_{\beta_1\beta_2} = 0.4$, $\rho_{\beta_1\beta_3} = 0.4$, $\rho_{\beta_2\beta_3} = 0.5$ and high correlation with $\rho_{\beta_1\beta_2} = 0.9$, $\rho_{\beta_1\beta_3} = 0.85$, $\rho_{\beta_2\beta_3} = 0.8$ were examined along with different censoring levels. Incorporating different levels of correlation between the outcomes and allowing the censoring process to vary in magnitude help us understand the impact of the correlation and censoring on the accuracy of the estimates.

Tables 1-3 present the bias and mean square errors for the slopes estimates for the three outcomes under the trivariate and bivariate models applied in a pairwise fashion. The reported means for bias for the first outcome in the bivariate model correspond to the mean of bias associated with the bivariate outcomes one and two (first pair), and one and three (second pair). The same computation was followed for the bias for outcomes two and three. Mean MSEa and mean MSEb for the bivariate model were computed similarly to the mean bias. We will start first by discussing the results reported in Table 1 wherein low correlations between the outcomes were assumed. In this context, when the censoring level was low (γ_1=3.5, γ_2=2.7 and γ_3=3.0) bias for the three outcomes was decreased by 52% on average under the multivariate model compared to the pairwise bivariate model, MSEa by 23% and MSEb by 54%. When the censoring level increased to a midlevel with γ_1=4.2, γ_2=4.8 and γ_3=5.1 bias decreased on average for the three outcomes by 54%, MSEa by 10% and MSEb by 26% for the multivariate model compared to the pairwise bivariate model. When the censoring level increased to high censoring withγ_1=7.8, γ_2=7.5 and γ_3=7.2 bias decreased by 26%, MSEa by 41% and MSEb by 20%. These results indicate that regardless of the censoring levels the estimates generated under the multivariate model had better precision and accuracy compared to the bivariate model. When the correlation levels between the outcomes increased to mid levels with ρ_{12}=0.4, ρ_{13}=0.4, and ρ_{23}=0.5 (Table 2), bias across all levels of censoring decreased by 40%, MSEa by about 15% and MSEb by 28% for multivariate compared to bivariate model and this decrease was demonstrated across all levels of censoring. Similar results were observed with high correlation levels of ρ_{12}=0.9, ρ_{13}=0.85, and ρ_{23}=0.8 (Table 3). In this context, under the multivariate model we observed a decrease in bias, MSEa and MSEb by an average of 40%, 15% and 20% respectively. These results suggest that regardless of the correlation and censoring levels modeling all outcomes simultaneously in the likelihood function and accounting

	Parameter	Trivariate model	Bivariate model†
γ_1*=3.5, γ_2*=2.7, γ_3*=3.0	Bias_1*100	-0.007	-0.021
	Bias_2*100	-0.028	-0.076
	Bias_3*100	-0.119	0.163
	MSEa_1*10³	0.144	0.205
	MSEa_2*10³	0.853	0.909
	MSEa_3*100	0.127	0.187
	MSEb_1*100	0.102	0.113
	MSEb_2*10	0.091	0.094
	MSEb_3*100	0.215	0.567
γ_1*=4.2, γ_2*=4.8, γ_3*=5.1	Bias_1*100	-0.012	0.022
	Bias_2*100	-0.151	0.343
	Bias_3*100	-0.132	0.346
	MSEa_1*10³	0.140	0.150
	MSEa_2*10³	0.805	0.856
	MSEa_3*100	0.154	0.173
	MSEb_1*100	0.104	0.116
	MSEb_2*10	0.097	0.110
	MSEb_3*100	0.266	0.611
γ_1*=7.8, γ_2*=7.5, γ_3*=7.2	Bias_1*100	-0.073	0.101
	Bias_2*100	-0.303	-0.310
	Bias_3*100	-0.126	-0.237
	MSEa_1*10³	0.132	0.161
	MSEa_2*10³	0.083	0.921
	MSEa_3*100	0.182	0.214
	MSEb_1*100	0.107	0.115
	MSEb_2*10	0.099	0.109
	MSEb_3*100	0.335	0.588

$\gamma_0 = -3.0, \sigma^2_{\beta_1} = 0.255, \sigma^2_{\beta_2} = 0.269, \sigma^2_{\beta_3} = 1.417, \sigma^2_{\varepsilon_1} = 0.078, \sigma^2_{\varepsilon_2} = 0.08, \sigma^2_{\varepsilon_3} = 0.14$

1*, first outcome; 2*, second outcome; 3*, third outcome; Bivariate model† mean bias, mean MSEa, mean MSEb of the pairwise bivariate model estimates.

Table 1: Comparisons of the performance of the trivariate and bivariate models, $\rho_{\beta_1\beta_2} = 0.1,\ \rho_{\beta_1\beta_3} = 0.2,\ \rho_{\beta_2\beta_3} = 0.3$.

	Parameter	Trivariate model	Bivariate model†
γ_1*=3.5, γ_2*=2.7, γ_3*=3.0	Bias_1*100	-0.011	0.015
	Bias_2*100	0.015	0.141
	Bias_3*100	0.041	0.144
	MSEa_1*10³	0.133	0.146
	MSEa_2*10³	0.797	0.846
	MSEa_3*100	0.125	0.169
	MSEb_1*100	0.103	0.112
	MSEb_2*10	0.098	0.111
	MSEb_3*100	0.219	0.593
γ_1*=4.2, γ_2*=4.8, γ_3*=5.1	Bias_1*100	-0.016	0.054
	Bias_2*100	-0.204	-0.265
	Bias_3*100	-0.147	-0.177
	MSEa_1*10³	0.142	0.158
	MSEa_2*10³	0.841	0.922
	MSEa_3*100	0.135	0.168
	MSEb_1*100	0.105	0.115
	MSEb_2*10	0.099	0.111
	MSEb_3*100	0.234	0.611
γ_1*=7.8, γ_2*=7.5, γ_3*=7.2	Bias_1*100	-0.114	-0.123
	Bias_2*100	-0.413	-0.513
	Bias_3*100	-0.158	-0.234
	MSEa_1*10³	0.133	0.140
	MSEa_2*10³	0.899	0.921
	MSEa_3*100	0.129	0.199
	MSEb_1*100	0.105	0.114
	MSEb_2*10	0.102	0.113
	MSEb_3*100	0.228	0.636

$\gamma_0 = -3.0, \sigma^2_{\beta_1} = 0.255, \sigma^2_{\beta_2} = 0.269, \sigma^2_{\beta_3} = 1.417, \sigma^2_{\varepsilon_1} = 0.078, \sigma^2_{\varepsilon_2} = 0.08, \sigma^2_{\varepsilon_3} = 0.14$

1*, first outcome; 2*, second outcome; 3*, third outcome; Bivariate model† mean bias, mean MSEa, mean MSEb of the pairwise bivariate model estimates.

Table 2: Comparisons of the performance of the trivariate and bivariate models, $\rho_{\beta_1\beta_2} = 0.4,\ \rho_{\beta_1\beta_3} = 0.4,\ \rho_{\beta_2\beta_3} = 0.5$.

	Parameter	Trivariate model	Bivariate model†
γ_1*=3.5, γ_2*=2.7, γ_3*=3.0	Bias_1*100	-0.064	-0.107
	Bias_2*100	-0.078	-0.128
	Bias_3*100	-0.071	0.091
	MSEa_1*10³	0.143	0.146
	MSEa_2*10³	0.777	0. 810
	MSEa_3*100	0.132	0.198
	MSEb_1*100	0.116	0.121
	MSEb_2*10	0.121	0.122
	MSEb_3*100	0.270	0.617
γ_1*=4.2, γ_2*=4.8, γ_3*=5.1	Bias_1*100	-0.055	-0.111
	Bias_2*100	-0.178	-0.443
	Bias_3*100	-0.048	-0.141
	MSEa_1*10³	0.139	0.152
	MSEa_2*10³	0.839	0.892
	MSEa_3*100	0.148	0.205
	MSEb_1*100	0.119	0.122
	MSEb_2*10	0.124	0.124
	MSEb_3*100	0.264	0.615
γ_1*=7.8, γ_2*=7.5, γ_3*=7.2	Bias_1*100	-0.135	-0.183
	Bias_2*100	-0.479	-0.648
	Bias_3*100	-0.216	-0.241
	MSEa_1*10³	0.140	0.154
	MSEa_2*10³	0.801	0.958
	MSEa_3*100	0.165	0.195
	MSEb_1*100	0.120	0.121
	MSEb_2*10	0.119	0.125
	MSEb_3*100	0.317	0.666

$\gamma_0 = -3.0, \sigma^2_{\beta_1} = 0.255, \sigma^2_{\beta_2} = 0.269, \sigma^2_{\beta_3} = 1.417, \sigma^2_{\varepsilon_1} = 0.078, \sigma^2_{\varepsilon_2} = 0.08, \sigma^2_{\varepsilon_3} = 0.14$

1*, first outcome; 2*, second outcome; 3*, third outcome; Bivariate model† mean bias, mean MSEa, mean MSEb of the pairwise bivariate model estimates.

Table 3: Comparisons of the performance of the trivariate and bivariate models, $\rho_{\beta_1\beta_2} = 0.9,\ \rho_{\beta_1\beta_3} = 0.85,\ \rho_{\beta_2\beta_3} = 0.8$.

for the correlations among them reduced the bias and MSEs associated with slope estimates and thus increased accuracy of estimation in comparison to modeling only two of the outcomes and ignoring the rest. Feasibility of the model to fit high dimensional outcomes was also verified via a small simulation study conducted at a high level of censoring. In this context we showed successful convergence of our model and slope estimates for 7 outcomes were attained with average bias of 0.006, MSEa of 0.0017 and MSEb of 0.012, as well as those for 10 outcomes with associated average bias of 0.008, MSEa of 0.0097, and MSEb of 0.059. Thus, joint modeling of multivariate slope outcomes, their correlations, and the censoring process was implemented and convergence on different dimensions of the outcomes was successfully achieved.

Cohort of Renal Transplant

The multivariate model was illustrated using a cohort of renal transplantation at the Medical University of South Carolina. A total of 110 patients who underwent kidney transplant in the calendar year 2000 were followed and demographical information along with a three year repeated measures of their kidney function using the markers serum creatinine, BUN and eGFR were recorded. The normal values for creatinine are in the range of 0.7 to 1.3 mg per 100 ml of blood, for BUN are between 8 and 25 mg per 100 ml of blood and for eGFR is approximately 90 and 120 mL/min/1.73 m^2 for men and women respectively. A decrease in eGFR and an increase in creatinine and BUN could be an indication of disease progression. Repeated measures on these markers were recorded between the years 2000 and 2003 yielding seven data points registered at baseline pre-transplantation (month 0), and post-transplantation at months 1, 3, 6,12, 24 and 36. Patients were followed at predetermined time points following a pre-specified schedule for all outcomes and measurements were taken repeatedly until graft failure is encountered. In this situation, patients revert back to dialysis treatment and acquisition of their renal markers is no longer possible. This leads to patient dropout from the follow-up study resulting in 19% informative right censoring due to graft failure.

For every individual we use the baseline adjusted model by incorporating the baseline measure as a covariate in the model as such: $y_{ijk} = \beta_{1ik} t_{ijk} + \beta_{2k} y_{0ik} + \beta_{3ik} t_{ijk}^2 + e_{ijk}$ where i=1,…,n, j=1,…, m_i and k=1,…,q. In Figure 1a and 1b) we present the mean levels of BUN and eGFR, and those of creatinine respectively against follow-up months. A curvilinear relationship between follow-up months and each marker was shown, so in order to linearize it we log transformed and we introduced a quadratic term in the model's equation. The logarithmic transformation linearized this relationship to a certain extent but not fully so we still needed to include the quadratic term.

The mean levels of the outcomes (BUN, eGFR, and creatinine) were plotted against the number of visits that range from 2 to 7 Figure 2a for BUN and eGFR and Figure 2b for creatinine. This figure shows that patients with high number of visits (5 and above) appear to have lower BUN and creatinine levels and higher eGFR levels on average compared to those with lower number of visits (below 5 visits). Since, the lower the creatinine and BUN, and the higher the eGFR the better the prognosis of kidney disease, this figure therefore indicates that the length of stay in the study measured by number of visits is determined by the levels of these markers.

Our proposed multivariate model was applied to the renal transplant dataset, to estimate population and individual slopes pertaining to the renal markers creatinine, BUN and eGFR. We assumed that the number of visits follow a truncated geometric distribution. The

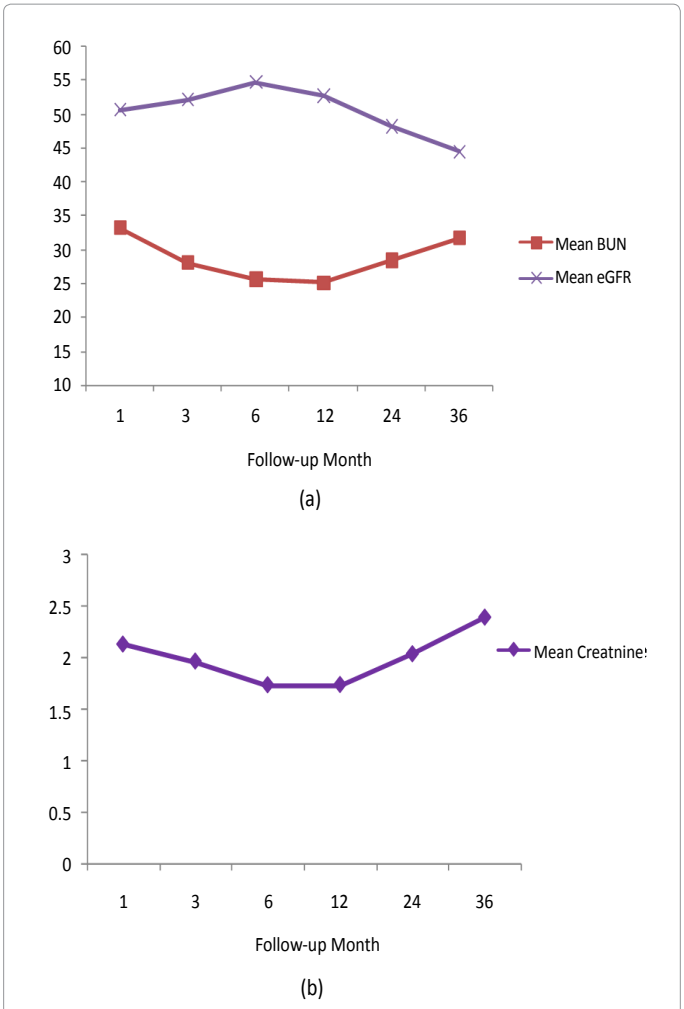


Figure 1: BUN, eGFR (a), and Serum creatinine mean levels (b) respectively plotted against follow-up months.

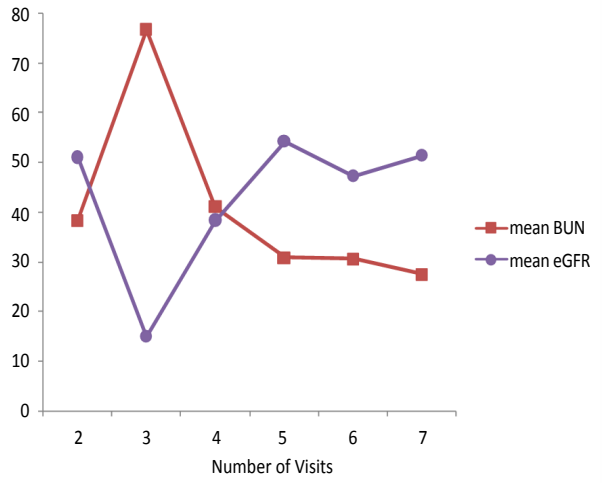
associated goodness-of-fit test confirmed that the number of visits is adequately modeled using the specified geometric distribution with test statistics P-value=0.6, X^2=0.3 DF=1. The estimated population slope for creatinine, BUN and eGFR were respectively $\beta_{creatinine}$=-0.089 (SE=0.0117), β_{BUN}=-0.142 (SE=0.011), and β_{eGFR}=0.302 (SE=0.0155) with P-value<0.0001. The estimated variance-covariance matrix of $\underline{\beta}_i'$ was

$$\sigma_{cr}^2 = 0.255,\ \sigma_{BUN}^2 = 0.269,\ \sigma_{eGFR}^2 = 1.417,\ \sigma_{cr,BUN}^2 = 0.148,\ \sigma_{cr,eGFR}^2 = -0.4,\ \sigma_{BUN,eGFR}^2 = -0.3.$$

The mean values of the estimated individual slopes were $mean_{\beta_i Creatinine}$ of -0.086 ± 0.494, $mean_{\beta_i BUN}$ of -0.20 ± 0.515, and $mean_{\beta_i eGFR}$ of 0.348 ± 1.178. The censoring parameters for all three markers were $\gamma_{creatinine}$=-0.093, γ_{BUN}=-0.0584, γ_{eGFR}=0.091 (P-value<0.0001). The significance in the censoring parameters for all three markers indicates that the censoring process was informative and significantly dependent on all three markers.

Discussion

In this article we propose a joint likelihood based approach for a multivariate mixed model that is an extension of the bivariate model by Jaffa et al. [21]. The proposed multivariate model generates maximum likelihood estimates for the population slopes and empirical Bayes estimates for the individual slopes that are adjusted for



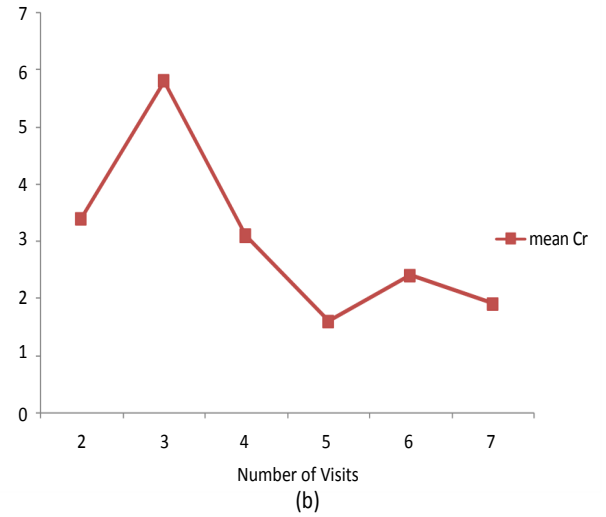


Figure 2: BUN, eGFR (a), and Serum creatinine mean levels (b) respectively plotted against number of visits.

informative right censoring. To account for this type of censoring a discrete distribution for the number of visits was assumed. Random slopes for the outcomes and the correlation between them were also concomitantly incorporated in the maximum likelihood function. The proposed multivariate model was applied to the cohort of renal transplant patients and the three biomarkers serum creatinine, BUN and eGFR that are typically measured to assess renal function over time following transplantation were modeled and their corresponding slopes were estimated. Moreover, individual slope for every patient was obtained thus making it viable to monitor disease progression on an individual and population basis. In this study we aimed at demonstrating the feasibility of the joint multivariate approach to fit the likelihood function for high dimensional outcomes (3, 7 and 10 outcomes) using standard software and successfully obtaining the population and individual slopes with minimal bias and MSEs for the different dimension of the multivariate outcomes. The model is not necessarily just limited to up to 10 outcomes but could be also extended to higher dimensions. Joint modeling approach proposed in this research work has always been considered advantageous yet challenging to implement compared to other methods such as pairwise model fitting proposed to handle situations of high dimensional outcomes. Specifically, our proposed joint modeling approach is favored over the pairwise model fitting since it increases efficiency and accuracy of the estimates [24]. In this context, our simulation study showed that fitting the full multivariate model had negligible bias and mean squared errors associated with slope estimates and that accuracy of estimates increased with this approach compared to the pairwise bivariate model. Joint modeling of the multivariate outcomes has its established advantages compared to the univariate separate analysis. The latter approach ignores the intrinsic correlations between the outcomes while the multivariate joint analysis exploits this correlation to generate more accurate estimates, and controls for type 1 error that might emanate from the univariate analysis when conducted without accounting for multiple comparisons [31]. Moreover, we have recently shown that joint bivariate analysis results in more precise estimates compared to the univariate separate analysis and the same conclusion should follow in the multivariate setting [21]. In a recent study, Jaffa conducted a sensitivity analysis on this model and it was shown that the proposed model is robust for assumptions about the underlying distributions for the number of visits m_i [32]. In this regard, an average increase of 18% was detected for bias and MSEs when the underlying distribution of the number of visits was mispecified and wrong assumptions were considered. In addition, we were also able to verify that violation of the normality assumption for the outcomes had a minor effect on the accuracy of the estimates and less than 10% increase in bias and MSEs was captured for the non-normal compared to normal outcomes. This robustness to assumptions misspecification makes it plausible for the proposed model to be applied to a wider range of datasets with multivariate high dimensional longitudinal outcomes.

Acknowledgment

This work was supported by the National Institutes of Health Grants HL077192 and HL087986 (AAJ).

References

1. James GD, Sealey JE, Alderman M, Ljungman S, Mueller FB, et al. (1988) A longitudinal study of urinary creatinine and creatinine clearance in normal subjects. Race, sex, and age differences. Am J Hypertens 1: 124-131.
2. Schrier RW (2008) Blood urea nitrogen and serum creatinine: not married in heart failure. Circ Heart Fail 1: 2-5.
3. Cirillo M, Anastasio P, De Santo NG (2005) Relationship of gender, age, and body mass index to errors in predicted kidney function. Nephrol Dial Transplant 20: 1791-1798.
4. Froissart M, Rossert J, Jacquot C, Paillard M, Houillier P (2005) Predictive performance of the modification of diet in renal disease and Cockcroft-Gault equations for estimating renal function. J Am Soc Nephrol 16: 763-773.
5. Verhave JC, Fesler P, Ribstein J, du Cailar G, Mimran A (2005) Estimation of renal function in subjects with normal serum creatinine levels: influence of age and body mass index. Am J Kidney Dis 46: 233-241.
6. Traynor J, Mactier R, Geddes CC, Fox JG (2006) How to measure renal function in clinical practice. BMJ 333: 733-737.
7. Schluchter MD (1992) Methods for the analysis of informatively censored longitudinal data. Stat Med 11: 1861-1870.
8. Wu MC, Bailey KR (1989) Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. Biometrics 45: 939-955.
9. Wu MC, Carroll RJ (1988) Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. Biometrics 44: 175-188.
10. Johnson BA, Tsiatis AA (2004) Estimating mean response as a function of treatment duration in an observational study, where duration may be informatively censored. Biometrics 60: 315-323.

11. Robins JM, Hernán MA, Brumback B (2000) Marginal structural models and causal inference in epidemiology. Epidemiology 11: 550-560.

12. Robins JM, Rotnizky A, Zhao LP (1994) Estimation of regression coefficients when some regressors are not always observed. Journal of American Statistical Association 89: 846-866.

13. Hernan MA, Brumback B, Robins JM (2001) Marginal structural models to estimate the joint causal effect on nonrandomized treatments. Journal of American Statistical Association 96: 440-448.

14. Hernán MA, Brumback B, Robins JM (2000) Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology 11: 561-570.

15. Satten GA, Datta S, Robins JM (2001) Estimating the marginal survival function in the presence of time dependent covariates. Statistics and Probability Letters 54: 397-403.

16. Mori M, Woolson RF, Woodworth GG (1994) Slope estimation in the presence of informative right censoring: modeling the number of observations as a geometric random variable. Biometrics 50: 39-50.

17. Sammel M, Lin X, Ryan L (1999) Multivariate linear mixed models for multiple outcomes. Stat Med 18: 2479-2492.

18. Gueorguieva RV (2005) Comments about Joint Modeling of Cluster Size and Binary and Continuous Subunit-Specific Outcomes. Biometrics 61: 862-866.

19. Ibrahim JG, Chen M, Sinha D (2004) Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. Statistica Sinica 14: 863-883.

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

21. Jaffa MA, Woolson RF, Lipsitz SR (2011) Slope estimation for bivariate longitudinal outcomes adjusting for informative right censoring by using a discrete survival model: application to the renal transplant cohort. Journal of the Royal Statistical Society A 174: 387-402.

22. Xu J, Zeger SL (2001) The evaluation of multiple surrogate endpoints. Biometrics 57: 81-87.

23. Xu J, Zeger SL (2001) Joint analysis of longitudinal data comprising repeated measures and times to events. Applied Statistics 50: 375-387.

24. Fieuws S, Verbeke G (2006) Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. Biometrics 62: 424-431.

25. Fieuws S, Verbeke G, Maes B, Vanrenterghem Y (2008) Predicting renal graft failure using multivariate longitudinal profiles. Biostatistics 9: 419-431.

26. He B, Luo S (2013) Joint modeling of multivariate longitudinal measurements and survival data with applications to Parkinson's disease. Stat Methods Med Res.

27. Jaffa MA, Woolson RF, Lipsitz SR, Baliga PK, Lopes-Virella M, et al. (2010) Analyses of renal outcome following transplantation adjusting for informative right censoring and demographic factors: A longitudinal study. Ren Fail 32: 691-698.

28. Perazella MA, Reilly RF (2003) Chronic kidney disease: A new classification and staging system. Hospital Physician 39: 18-45.

29. Pinheiro JC, Bates DM (1995) Approximations to the log-likelihood function in the nonlinear mixed -effects model. Journal of Computational and Graphical Statistics 4: 12-35.

30. SAS 9.3, Institue Inc (2010) Cary, N.C., USA.

31. Fitzmaurice GM, Laird NM (1997) Regression models for mixed discrete and continuous responses with potentially missing values. Biometrics 53: 110-122.

32. Jaffa MA, Lipsitz S, Woolson RF (2011) Slope Estimation for Informatively Right Censored Longitudinal Data Modeling the Number of Observations Using Geometric and Poisson Distributions: Application to Renal Transplant Cohort. Statistical Methods in Medical Research (In press, Published on line first on Dec 4, 2011).

Citation: Jaffa MA, Gebregziabher M, Jaffa AA (2014) A Joint Modeling Approach for Right Censored High Dimensional Multivariate Longitudinal Data. J Biomet Biostat 5: 203. doi:10.4172/2155-6180.1000203