Journal of
Medical Statistics and Informatics
ISSN 2053-7662

HOAJ
Herbert Open Access Journals

Methodology Open Access

Joint modeling of longitudinal and event time data: application to HIV study

Hyun J. Lim[1*], Prosanta Mondal[1] and Stuart Skinner[2]
*Correspondence: hyun.lim@usask.ca
[1]Department of Community Health & Epidemiology, College of Medicine, University of Saskatchewan 107 Wiggins Road, Saskatoon, SK S7N 5E5, Canada.
[2]Department of Medicine, College of Medicine, University of Saskatchewan 103 Hospital Drive, Saskatoon, SK, S7N 0W8, Canada.

Abstract

Many clinical studies generate a dataset having longitudinal repeated biomarker measurement data and time to an event data, which often depend on each other. In such studies, characteristics of the pattern of a biomarker change, and the association between the primary survival endpoint and features of the longitudinal profiles are commonly of interest. Often separate analyses using a mixed effects model for the longitudinal outcome and a survival model for the time to event outcome are performed. However, separate models are overly simplified because they do not consider the association between two components of such data and so produce misleading conclusions. An alternative approach is two-stage modeling which allows a separate biomarker pathway for each patient but the parameter estimates are still biased. Joint modeling is the most sophisticated complex approach but enables the repeated biomarker measurements and survival processes to be modelled while accounting for the interrelationship between the two processes. We demonstrate the use of joint modeling in analysis of an HIV dataset with CD4+ count measurements and survival time. In the joint model, we combine a linear Gaussian random effects sub-model for the repeated CD4+ count measurements and Cox or Weibull survival sub-model, linked through their shared dependence on the latent variable. Our study showed that the hazard rate of death depended on the longitudinal progression of CD4+ counts, i.e., a patient's baseline CD4+ count and the rate of change in CD4+ counts significantly impact on his or her survival time.

Keywords: Joint model, random effects model, survival model, longitudinal data, event time data, HIV

Introduction

Patients in many clinical studies are followed over a period of time and biomarkers are collected repeatedly at multiple time points. The data also indicates the time, at which an event of particular interest occurred, eg., relapse of a disease or death. For example, in HIV studies patients are measured for biomarkers such as the CD4+ lymphocyte count or the viral RNA (viral load) until specific outcomes such as AIDS or death. Other applications include cognitive performance and survival in geriatric studies, systolic blood pressure and a coronary event, prostate-specific antigen biomarker and prostate cancer recurrence, and hemoglobin level and survival in type 2 diabetes [1-7]. In such studies, a common primary objective is to investigate the effect of treatment on survival and/or biomarker process. But patterns of change in a biomarker or the association between the primary endpoint and features of the longitudinal profiles are also of interest. The traditional approach uses two models separately, a random effects model that describes the process of the repeated measurements over time and a survival model that examines time to the event. However, separate models are overly simplified because they do not consider the association between the two components of such data, and in failing to take into account all the available information in an integrated manner, provide invalid inferences. One of the alternative approaches is two-stage modeling. In the first stage of two-stage modeling, the repeated biomarker measurements are modeled using a random effect model so that subject-specific values of the covariate may be estimated. The modeled value of individual prediction from the random effects model is substituted into the survival model as time-varying covariate in the second stage. This two-stage approach accounts for some of the measurement error and allows the pathway of the covariates for each patient to be estimated [8]. However, parameter estimates of the two-stage model are still biased and inefficient, even though much of the bias of the uncorrected values is reduced [9]. Since either separate models or the two-stage approach is inappropriate when the longitudinal repeated measurements and survival outcome have an association, others have proposed joint modeling and its extensions, which have obtained great attention in the literature [9-13].

Joint modeling is a sophisticated, complex approach but it enables both longitudinal repeated biomarker measurements and survival processes to be modelled together while taking account of association between them. By including the random effects model for longitudinal data in the survival model, the patterns of a biomarker's performance and the relationship between its progression and survival time can be characterized.

 doi: 10.7243/2053-7662-1-1

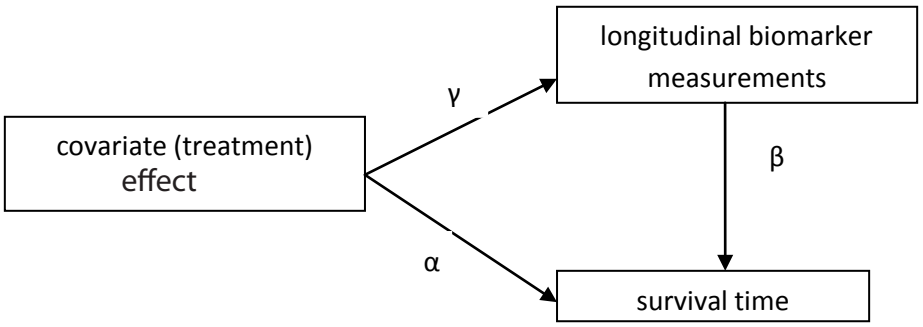


Figure 1. Covariate (treatment) effect on longitudinal biomarker measurements and survival outcome. α is the covariate effect on survival time; γ is the covariate effect on the longitudinal biomarker; β is the effect of the longitudinal biomarker on survival time.

In this way, joint modeling provides less biased estimates and more efficient inferences than separate models or the two-stage approach [9,12,13]. The main advantage of joint modeling is that the effect of a covariate on the longitudinal process can be separated from its effect on survival. Modeling of a covariate effect (eg., treatment) can be described and visualized as in **Figure 1**.

For jointly modeling longitudinal data and event time data, many have proposed a linear or non–linear random effects model for longitudinal measurements and a semi-parametric or parametric survival model for event time data, where a set of random effects is assumed to induce their interdependence [9,14-18]. Wulsohn and Tsiatis [9] proposed a two-step procedure for fitting their model in HIV study (Wulsohn- Tsiatis model). First, they assumed a growth curve random components model with normal errors for true CD4+ counts using the modified EM algorithm for estimation. They then substituted these estimates into the Cox proportional hazards model to obtain estimates of the survival parameters. Other authors have also assumed piecewise constant hazards, parametric Weibull, or accelerated failure time models [17,19-21]. Henderson et al., developed a more general version of the Wulsohn-Tsiatis model and proposed the use of an unobserved or latent bivariate Gaussian process to link longitudinal data and survival data, assuming that the longitudinal and event processes were conditionally independent given the one process covariates [15,22].

Our paper focuses on the application of available methodologies and on interpretation of the results when these are utilized in the analysis of an HIV dataset. In our HIV study, the main objective is to use longitudinal CD4+ measurements to improve prediction of survival prognosis. In the next section, we introduce a description of our HIV study. In Section 3, methods of joint modeling for analysis are reviewed. In Section 4, the conventional separate models and the joint models are applied to the HIV dataset to measure the relative effects of covariates. We examine whether the baseline CD4+ value and the slope of CD4+ counts reveal association with mortality risk. It should be noted, however, our intent is to illustrate the application of joint modeling rather than give universally valid estimates for HIV study outcomes. Section 5 contains a discussion about the results and conclusions.

Study description

Clinical background

CD4+ counts and viral load are considered important biomarkers of HIV disease progression. CD4+ cells are an important part of the immune system, which begin to deplete as the virus infects the body; the amount of decline indicates the degree of immunosuppression. Thus CD4+ count is considered as the primary indicator for prognostic information and a guide for antiretroviral therapy in HIV infected individuals [23,24]. Successful treatment for HIV patients depends on the therapy's ability to restore immune functions. Studies have shown that disease progression among patients with HIV infection is delayed substantially when CD4+counts increase in response to treatment [24,25]. Identification of factors influencing disease progression is vital to effectively care for patients and to improve their survival and quality of life. Detection of not only clinical features and risk factors but also characteristics associated with a more rapid progression of CD4+ counts can help identify patients who may benefit from closer and more frequent clinical follow-up and earlier treatment intervention [24]. We anticipate that changes in biomarkers (CD4+ counts or viral load) will have an impact on the risk of major HIV complications.

Study population

This study was a retrospective, longitudinal case-based investigation. Data was obtained solely by reviewing medical records of HIV patients diagnosed at the Royal University Hospital or the West Side Community Clinic in Saskatoon, Canada between January 1, 2005 and June 1, 2010. A total of 321 adult (age ≥ 18 years) patients who had at least one CD4+ measurement after the first report of HIV diagnosis regardless of clinical stage were eligible for the study. Of them, 1129

doi: 10.7243/2053-7662-1-1

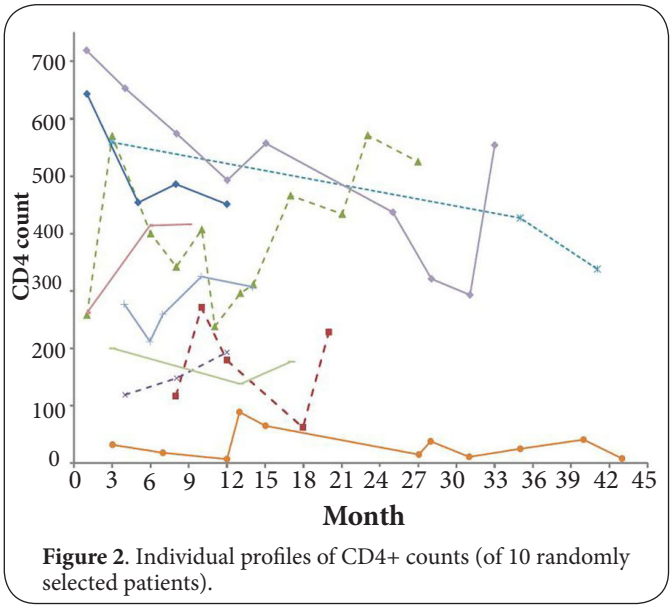


Figure 2. Individual profiles of CD4+ counts (of 10 randomly selected patients).

CD4+ count measurements were observed with median of 3 measurements. The number of measurements per patient varied and the measurement intervals were unequally spaced. Individual CD4+ count profiles of 10 randomly selected patients are shown in **Figure 2**. Plotting of CD4+ counts was grouped according to their date of measurement. The first measurement done within 180 days after study enrollment (first clinic visit) was considered the baseline value. We created a person-period data set for each patient covering 3-month intervals for the first year and then 6-month intervals afterward as CD4+ counts are generally measured every 3 ~ 6 months in clinical practice as the standard of medical care. Our choice of the 3-month or 6-month period for CD4+ counts was supported by clinical practice guidelines and CD4+ data availability in our study. When the CD4+ count was not observed during any particular interval period, we considered the value as missing. We did not anticipate that missing CD4+ counts would depend on the nature of the unobserved CD4+ counts, and thus they are assumed missing at random. For any patients with 2 or more CD4+ counts in a given interval, the average of those CD4+ count measurements was used for that interval. Subjects were followed from the time of entry into the study until death, loss to follow-up, or the end of the study (August 31, 2011).

CD4+ count measurements were limited to 12 per patient in the subsequent analysis as very few subjects had more than 12 measurements. The maximum of 12 measurements is a practical limitation of the data, rather than a methodological one. The other outcome variable considered was time to death, which was defined as time between date of entry into the cohort and date of death, date last seen, or the end of the study. In our study, a 0.05 alpha level was used for statistical significance. Analyses in this study were performed using SAS version 9.2 (The SAS Institute, Cary, NC) or R package.

The study received ethical approval from the University of Saskatchewan institutional review board committee.

Models

Let T be a random variable representing survival time that has the distribution function, $F(t)$. The survival function $S(t)$ at time t is defined to be the probability that the survival time is greater than t, where $S(t) = P(T > t) = 1 - F(t)$. Let T_i be the time when the event occurs for the ith subject and C_i be the corresponding censoring time. When T_i is subject to right censoring, X_i is the observed time which is a minimum of (T_i, C_i), i.e., X_i is equal to T_i if the event was observed and is equal to C_i if it was censored. Let $\delta_i = I(T_i \leq C_i)$, where I(.) is an indicator function and takes the value 1 when $T_i \leq C_i$ and 0 when it is otherwise.

Let Z_{ij} be the observed measurement for the *ith* subject at the *jth* time point, where $i = 1, \ldots, n$ and $j = 1, \ldots, K_i$. Assume that individual i have K_i measurements of the covariate over time and the number of measurements can be different for each individual. The covariate is measured at time $\boldsymbol{t_i} = (t_{ij}: t_{ij} \leq X_i, j = 1, \ldots, K_i)$, where X_i is the observed survival time for subject i. The longitudinal repeated covariate values (biomarker such as CD4+ count or viral load) at times $\boldsymbol{t_i}$ are $\boldsymbol{Z_i} = (Z_{ij}: t_{ij} \leq X_i, j = 1, \ldots, K_i)$. Let $\boldsymbol{W_i}$ be the other covariate vector of p-dimensions for the ith subject. Thus the observed data for subject i is $(X_i, \delta_i, \boldsymbol{Z_i}, \boldsymbol{t_i}, \boldsymbol{W_i})$.

A joint model is comprised of two sub-models, one for the longitudinal process $\boldsymbol{Z_i}(u)$ and the other for the failure time T_i. We assume that the model for the longitudinal process and the hazard for survival time are usually taken to depend jointly on shared, underlying random effects. In the first step of modeling, evolution of the longitudinal biomarker measurements are estimated separately using a random effects model of the linear/polynomial function of time. With this approach, we can estimate changes in biomarkers for each patient at the time of death or censoring. In the second step, a joint modeling approach with Cox or Weibull parametric hazards model is utilized.

The subject-specific random effects model has become a popular tool to analyse various types of longitudinal data as it adequately describes the pattern of repeated measurements over time. In this model, the average progression is described using some function of time, and subject-specific deviations from this average evolution are introduced by using random effects [26,27]. We assume that the longitudinal measurement of the covariate $\boldsymbol{Z_i}$ follows a linear mixed effects model (growth curve model) with random intercept θ_0 and random slope θ_1. So, at times t_{ij}, the linear mixed effect has the structure,

$$Z_{ij} = \boldsymbol{\alpha W}_i + \theta U_i + \boldsymbol{\varepsilon}_i = (\boldsymbol{\alpha}_{0i} + \alpha_{1i} w_{1i} + \ldots + \alpha_{pi} w_{pi}) + (\theta_{0i} + \theta_{1i} t_{ij}) + \varepsilon_{ij},$$

where $\boldsymbol{\alpha}$ is the regression coefficient of covariate vector $\boldsymbol{W}$. θU_i incorporate subject-specific random effects and is sometimes called the trajectory function of the model. ε_{ij}

 doi: 10.7243/2053-7662-1-1

is a random error term that accounts for the unexplained variation in the data. ε_i is assumed to be normally distributed with $N(0, \sigma_\varepsilon^2)$ and cov $(\varepsilon_{ij}, \varepsilon_{il}) = 0$, where $j \neq l$. We also assume that the error ε_{ij} is independent of the random intercept θ_{0i} and random slope θ_{1i}. The intercept and slope, $\theta_i = (\theta_{0i}, \theta_{1i})$, are typically assumed to be random and have a multivariate normal distribution as iid $N(\theta, \boldsymbol{\Sigma})$, i.e.,

$$\begin{pmatrix} \theta_{0i} \\ \theta_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}, \begin{pmatrix} \sigma_0 & \sigma_0 \\ \sigma_0 & \sigma_1 \end{pmatrix} \right).$$

Here the random effect θ_i account for the heterogeneity among the subjects and describes how the profile of the *i*th subject deviates from the average profile. Conditional on bivariate Gaussian latent variable $\theta_i = (\theta_{0i}, \theta_{1i})$ associated with subject *i*, the corresponding longitudinal measurements Z_{ij} at time t_{ij} are mutually independent and

$$Z_{ij} \mid \theta_i \sim N(\boldsymbol{\alpha W}_i + \theta_{0i} + \theta_{1i} t_{ij}, \sigma_\varepsilon^2).$$

The coefficient $\boldsymbol{\alpha}$ assesses the covariate effect (for example, treatment) on the longitudinal biomarker. Possible extensions from a linear mixed effects model allow for more complex relationships such as polynomial specifications of $\boldsymbol{Z}_i$ and $\boldsymbol{W}_i$. The longitudinal and survival components of the joint model are linked through the random effects θ. Assuming the true longitudinal measurement value is given by the growth curve model with random intercept (θ_{0i}), random slope (θ_{1i}) and the covariate $\boldsymbol{W}_i$, the Cox proportional hazards model is

$$h(t \mid \theta_i, \boldsymbol{Z}_i, \boldsymbol{t}_i) = h_0(t) \exp\{\beta(\theta_{0i} + \theta_{1i} t_{ij}) + \boldsymbol{\alpha W}_i\},$$

where $h_0(t)$ is an unspecified baseline hazard and α is the direct covariate effect on survival time. The parameter β measures the association between the longitudinal biomarker and survival time. The form of $\beta(\theta_{0i} + \theta_{1i} t_{ij})$ is subject-specific covariate effects which is often called a random effect *frailty* model [28]. The survival and growth curve densities in this model are both conditional on the unobserved random effect. The hazard function of the survival model can also be a parametric model such as an exponential or a Weibull model. The Weibull hazard model (η, μ_i) at time *t* is given as

$$h(t \mid \theta_i, \boldsymbol{Z}_i, \boldsymbol{t}_i, Y_i) = \eta t^{\eta-1} \exp\{\beta(\theta_{0i} + \theta_{1i} t_{ij}) + \boldsymbol{W\alpha}\},$$

where $Y_i(u) = I(X_i \geq u)$, which is an at risk indicator. Note that if $\eta < 1$, the hazard is decreasing and if $\eta > 1$, the hazard is increasing in *t*. When $\eta = 1$, the hazard is a constant which survival time reduces to the exponential distribution.

The likelihood method is a widely used approach for the parameter estimation in the joint model [9,10,16,18,21]. Assuming that censoring and timing of longitudinal measurements are non-informative, the likelihood function *L* with the observed data for each subject $(X_i, \delta_i, \boldsymbol{Z}_i, \boldsymbol{t}_i)$, is given by [9].

$$L = \prod_{i=1}^{n} \left[\int_{-\infty}^{+\infty} \left\{ \prod_{j=1}^{K_i} f(Z_{ij} \mid \theta_i, \mathbf{t}_i, \sigma_\varepsilon^2) \right\} f(\theta_i \mid \theta, \mathbf{V})\ f(X_i, \delta_i \mid \theta_i, h_0, \beta)\, d\theta_i \right]$$

where

$$f(Z_{ij} \mid \theta_i \sigma_\varepsilon^2) = (2\pi\ \sigma_\varepsilon^2)^{-1/2} \exp\{-(Z_{ij} - \theta_{0i} - \theta_{1i} t_{ij})^2 / 2\sigma_\varepsilon^2)\},$$

$$f(\theta_i \mid \theta, V) = (2\pi |\mathbf{V}|)^{-1/2} \exp\{-(\theta_i - \theta)' \mathbf{V}^{-1} (\theta_i - \theta))/2)\},$$

and

$$f(X_i, \delta_i \mid \theta_i, h_0, \beta) = \left[h_0(X_i)\ \exp\{\beta(\theta_{0i} + \theta_{1i} X_i)\} \right]^{\delta_i} \exp\left[-\int_0^{X_i} h_0(u)\ \exp\{\theta_{0i} + \theta_{1i} u)\}\ du \right]$$

The parameters θ, **V**, σ_ε^2, and β are estimated using parametric likelihood and $h_0(u)$ using nonparametric maximum likelihood. To fit the joint models, the EM algorithm is used for likelihood inferences [29]. The EM algorithm which requires two-dimensional numerical integration corresponding to the dimensionality of $\theta = (\theta_0, \theta_1)$ can be used to estimate the parameters of interest by maximizing the likelihood of the observed data. Monte Carlo method in the E-step to approximate the conditional expectation and the Newton-Raphson method in the M-step is used to estimate the parameters [9,10]. The closed-form maximum likelihood estimates (MLE) are

$$\theta = \sum_{i=1}^{n} \frac{\mathrm{E}_i(\theta_i)}{n}$$

$$\hat{\mathbf{V}} = \sum_{i=1}^{n} \frac{\mathrm{E}_i(\theta_i - \theta)(\theta_i - \theta)'}{n}$$

$$\hat{\sigma}_\varepsilon^2 = \sum_{i=1}^{n} \sum_{j=1}^{K_i} \frac{\mathrm{E}_i(Z_{ij} - \theta_{0i} - \theta_{1i}\ t_{ij})^2}{\sum_{i=1}^{n} K_i}$$

$$\hat{h}_0(u) = \sum_{i=1}^{n} \frac{\delta_i\, I(X_i = u)}{\sum_{j=1}^{n} \mathrm{E}_j\left[\exp\{\beta(\theta_{0i} + \theta_{1i} u\}\right] Y_j(u)}$$

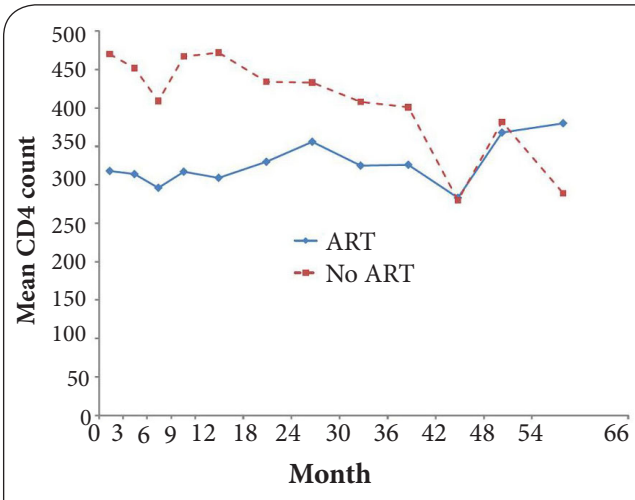
where $Y_j(u) = I(X_j \geq u)$, which is an at risk indicator.

doi: 10.7243/2053-7662-1-1

Table 1: Demographic and clinical characteristics of the study patients. (N=321).

Variable	Number of Patient (%)
Male	165 (51.4%)
Ethnicity	
Aboriginal	16 (71%)
Caucasian	78 (25.7%)
Others	12 (3.3%)
Age	
≤ 25	65 (20.3%)
26 - 30	52 (16.2%)
31 - 35	52 (16.2%)
36 – 40	45 (14%)
41 – 45	43 (13.4%)
> 45	64 (19.9%)
Year of diagnosis	
2005	52 (16.2%)
2006	40 (12.5%)
2007	53 (16.5%)
2008	71 (22.1%)
2009	80 (24.9%)
2010*	25 (7.8%)
Incarceration	97 (30.2%)
Social assistance	105 (32%)
History of IDU	256 (80.8%)
Hepatitis C antibody	252 (79%)
Ever on ART $	163 (58%)
Mean CD4+ at diagnosis (sd)	388 (229)
Mean log(RNA) at diagnosis (sd)	4.3 (1.0)
Death	25 (7.8%)

*up to June 1, 2010
$ART = antiretroviral therapy



# of patients													
	ART:	95	90	84	77	87	74	70	53	37	35	22	17
	No-ART:	72	49	35	28	34	26	17	14	11	5	6	4

Figure 3. Observed mean CD4+ counts profile by antiretroviral therapy (ART) group.

The complete parameter estimation of the EM algorithm was provided in Wulfsohn and Tsiatis's paper [9].

Baysian sampling method using MCMC technique was also used to estimate the parameters [20,30].

Results

A total of 321 HIV infected patients were eligible for the study (**Table 1**). Of them, 165 (51.4%) were males and 216 (71%) were Aboriginal. The mean age was 35 years (SD = 10.5). At baseline, 30% had a history of incarceration, 79% had Hepatitis C antibodies (HCV), 81% had history of injection-drug use (IDU), and 58% had ever been on antiretroviral therapy (ART). The mean CD4+ count at baseline was 388 cells/mm^3 (SD= 229; median=353; interquartile range: 234-536); the mean viral load was 4.3 $\log_{10}$ copies/ml (SD= 1.0; median=4.4; interquartile range: 3.9-4.9). The median follow-up time was 20 months (interquartile range: 9-36 months) and 25 (7.8%) patients died in the study cohort. The overall mean CD4+ counts measurements showed decreasing trend. The overall survival probability was 0.97 at 1 year, 0.95 at 2 years and 0.92 at 3 years, respectively. To normalize CD4+ counts, we used the square root of the CD4+ counts in the models.

Results of random effects model

Figure 3 shows the observed mean CD4+ counts profile as a function of time within separate ART strata. The ART treatment group started with a lower mean CD4+ count than the ART naïve group, but the ART naïve groups had a decreasing mean response while the ART treatment group mean was sustained. The mean response towards the end of the study was unstable because of the small number of patients. At each time point, mean responses were calculated only among those patients who have not yet dropped out of the study. The correct interpretation of the two mean response profiles is subtle because of the potential inter-dependence between a patient's predisposition to drop-out and their associated CD4+ counts. From the univariate model for CD4+ count measurement, HCV and ART treatment were significant (p-values are 0.02 and <0.0001, respectively) while gender, age, ethnicity, clinic site, history of incarceration, and history of IDU were not. Since HCV and ART treatment were significantly associated with longitudinal CD4+ counts measurement, these covariates were included in the random effect model. The random effect model for longitudinal CD4+ count repeated measurements is:

$$\sqrt{CD4+}_{ij} = \boldsymbol{\alpha}_{0i} + \alpha_{1i} HCV_i + \alpha_{2i} ART_i + \boldsymbol{\alpha}_{3i} Time_{ij} + \boldsymbol{\alpha}_{4i} (ART_i * Time_{ij}) + \theta_{0i} + \theta_{1i} Time_{ij} + \varepsilon_{ij}.$$

The results of the random effect model using the unstructured covariance matrix $\boldsymbol{\Sigma}$ are summarized in **Table 2**. The estimated average regression coefficient of *Time* for the ART treatment naïve group is −0.088 (95% CI: -0.14, -0.04; p=0.0008), suggesting a significant decrease in CD4+ count over the

 doi: 10.7243/2053-7662-1-1

Table 2. Estimation of coefficient, 95% confidence interval (C.I), and p-value from separate models: longitudinal random effects model, Cox survival model, Weibull survival model, respectively.

Model	Covariate	Estimate	95%	CI	p-value
Longitudinal random effect	Intercept	22.55	20.96	24.14	< 0.0001
	ART treatment [$]	-4.216	-5.70	-2.74	< 0.0001
	HCV[@]	-2.203	-3.60	-0.81	0.002
	Time	-0.088	-0.14	-0.04	0.0008
	Time*ART treatment	0.089	0.03	0.15	0.005
Cox survival	ART treatment	-1.085	-1.93	-0.24	0.01
	HCV	1.89	-0.14	3.92	0.07
Webull survival	Intercept	-9.279	-12.05	-6.51	<0.0001
	ART treatment	-1.109	-1.94	-0.28	0.013
	HCV	1.79	-0.22	3.80	0.094
	Scale	0.578	0.42	0.-79	--------

[$] ART naïve is the reference group.
[@]no-HCV is the reference group.

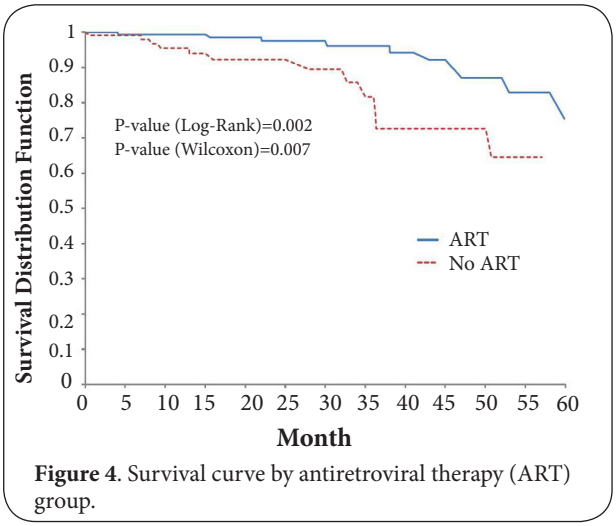


Figure 4. Survival curve by antiretroviral therapy (ART) group.

study period. Average regression coefficient for the ART treatment group is 0.001 and indicates that CD4+ did not decrease over the study period. However, the difference in slope between two groups is statistically significant (95% CI: 0.03 – 0.15; p=0.005). Patients who were co-infected with HCV at study entry had significantly lower CD4+ counts than those without HCV co-infection (p=0.002).

Results of survival model

Figure 4 shows the Kaplan-Meier survival curve estimates by ART groups. The two groups are similar during the first nine months but subsequent survival in the ART treatment group is significantly higher than in the ART naïve group. From univariate analysis for survival time, history of IDU, HCV, and ART treatment were significant (p-values are 0.04, 0.02 and 0.002, respectively) while gender, age, ethnicity, clinic site, and history of incarceration were not. ART treatment and HCV were included in the model because both were significantly associated with survival time as well as CD4+ repeated measurements. No interaction was observed between ART treatment and HCV. The Cox regression model in the absence of random effects is

$$h_i(t) = h_0(t)\ \exp\{\gamma_1 \mathrm{ART}_i + \gamma_2 \mathrm{HCV}_i\}.$$

Under this Cox model, as expected, the ART treatment group had significantly lower hazard rates than the ART naïve group after controlling for HCV (HR=0.34; 95% CI: 0.145 – 0.786; p=0.01). HCV-positive status also resulted in a worse survival than HCV-negative status in the adjusted model, but the difference is marginally significant (p=0.07). The results from the Cox model are similar to the results obtained from the Weibull model. In the Weibull model, the relative hazard rate for patients with ART treatment to ART naïve is exp(-1.109) = 0.33 as compared to 0.34 under the Cox model (**Table 2**).

Results of joint model

Assuming the CD4+ longitudinal value is given by the growth curve model with random intercept and slope, the Cox proportional hazards model for subject *i* is

$$h\,(t \mid \theta_i, \boldsymbol{Z}_i, \boldsymbol{t}_i) = h_0(t)\ \exp\{\alpha_1 \mathrm{ART}_i + \alpha_2\, \mathrm{HCV}_i + \beta\,(\theta_{0i} + \theta_{1i}\, t_{ij})\},$$

where $h_0(t)$ is either unspecified or Weibull baseline hazard. The estimated association parameter in the joint model is -0.102 and is statistically significant in the Cox sub-model (p=0.001). This indicates that there is strong evidence of association between two sub-models. Further investigation of this association showed that both initial level and slope of CD4+ count is negatively associated with the hazard of death (the intercept and slope estimates are -0.06 and -4.45, respectively). Under this joint model, a patient's survival is associated with his/her longitudinal data pattern of the rate at which CD4+ decrease (**Table 3**). The data show that both HCV and ART are associated with CD4+ count over the study period. The rates of CD4+ change between ART groups are significantly different, suggesting a significant decrease in CD4+ count for the ART naïve group compared to the ART group over the study period after adjusting for HCV (p=0.0001). This finding is clinically predictable since patients with more rapid decline in CD4+ would have poorer survival. As expected, the ART treatment was significantly associated with survival after taking into account estimated CD4+ count changes (p= 0.001). With the Weibull sub-model, most results are similar to those determined from the Cox sub-model (**Table 3**). The estimated association parameter is -0.125, which remains statistically significant (p=0.003), as compared to -0.102 under the Cox sub-model. However, the survival rates among those with HCV comparing to those without HCV were lower in using the Weibull sub-model in spite of the lack of statistical significance (p=0.19). This finding was different from what the Cox sub-

doi: 10.7243/2053-7662-1-1

Table 3. Estimation of coefficient, 95% confidence interval (C.I.), and p-value from the joint model with (i) random effects sub-model and Cox survival sub-model and (ii) random effects sub-model and Weibull survival sub-model.

Model	Covariate	Estimate	95%	CI	p-value
(i) Longitudinal random effect sub-model	Intercept	23.79	22.79	24.81	< 0.0001
	ART treatment $	-4.606	-5.83	-3.38	< 0.0001
	HCV@	-3.03	-3.83	-2.23	<0.0001
	Time	-0.093	-0.13	-0.05	<0.0001
	Time*ART treatment	0.097	0.05	0.14	0.0001
Cox survival sub-model	ART treatment	-1.369	-2.19	-0.55	0.001
	HCV	1.462	0.39	2.54	0.008
	Association parameter	-0.102	-0.16	-0.04	0.001
(ii) Longitudinal random effect sub-model	Intercept	21.42	20.31	22.54	< 0.0001
	ART treatment	-3.348	-4.58	-2.11	< 0.0001
	HCV	-1.333	-2.09	-0.58	0.0005
	Time	-0.115	-0.16	-0.07	<0.0001
	Time*ART treatment	0.108	0.05	0.16	0.0001
Weibull survival sub-model	Intercept	-6.288	-9.80	-2.77	0.0005
	ART treatment	-1.389	-2.25	-0.53	0.002
	HCV	1.477	-0.74	3.70	0.19
	Association parameter	-0.125	-0.21	-0.04	0.003

*ART naïve is the reference group.
@no-HCV is the reference group.

model was employed (p=0.0001).

For all estimations in this study, we used SAS PROC MIXED, PROC PHREG, PROC NLMIXED and R package (http://rwiki.sciviews.org/doku.php?id5packages:cran:jm) [31,32].

Discussion

In our study we addressed the relationship between the trajectory of CD4+ counts over time and the risk of death among HIV patients using joint modeling with a longitudinal random effects sub-model and a Cox or Weibull survival sub-model. The joint model revealed that the hazard of death depended on a longitudinal process of an intercept and a slope, i.e., a patient's baseline CD4+ count and the rate of change in CD4+ counts significantly impact on his or her survival time. This observation suggested that the rate of CD4+ count change confers risks of mortality, even after adjustment for HCV co-infection as a risk factor. The parameter estimates of the joint model were different from those of the separate models; this difference might be due to the model correction accounting for the correlation between the longitudinal CD4 + counts and survival time.

Simple methods like separate modeling or two-stage modeling are useful when we explore optional models and select potential covariates. With simple methods, if survival for patients is far longer than the last follow-up time, interpretation should be cautious because extrapolation of the predicted value beyond the observed date may be risky. However, the joint model is able to account for the healthy survivor effect by exploiting these random effects, which account for the dispersion among individual times to death [10,33]. As long as there is sufficient information on the longitudinal repeated measurements, the estimates of the joint model are robust and efficient [Hsieh 2006]. Thus, even if the association with the longitudinal measurements is not of interest, under heavy censoring a joint modelling approach exploits the association to give more efficient inferences for parameters of the survival distribution [34].

As Henderson et al., [22] mentioned, joint modeling is flexible methodology for handling combined longitudinal and event history data. But when the goals of a study concern population-level inferences, and especially when the time-to-event process is not of direct interest, random effect joint modelling may be an over-elaboration [34]. When the focus of interest is the time-to-event process, joint modelling enables longitudinal measurements on each subject to be used as time-varying explanatory variables while recognizing that they may be measured with a non-negligible error. When the association parameter between the longitudinal and survival data is not significant, the joint model analysis should have the same results as would be obtained from separate analyses for each component. However, joint modelling is a valuable technique not only in its efficient use of all available data and its ability to obtain accurate inference, but it is highly recommended when survival time and longitudinal measurements have the same clinical meaning. It is especially applicable to problems involving biomarkers where the focus is on using longitudinal measurements to improve prediction of survival prognosis. If survival time and longitudinal measurements have a different clinical meaning or are not comparable, joint modeling is not appropriate because the result may lead to the conclusion that a covariate effect with worse survival is superior [35].

One of the concerns many authors have raised in making inference on the longitudinal process is that occurrence of the event may induce an informative censoring [36-39]. Subjects with advanced disease may have fewer CD4+ count measurements because they experience death earlier, potentially leading to biased estimation. Valid inference requires a framework in which potential underlying relationships between the death event and the longitudinal process are explicitly acknowledged [10]. The joint model assumes that censoring is independent of the random effects. But if the underlying relationship and censoring process that leads to drop-out is not correctly modeled, the estimate of parameters may be biased [9,20]. Much more work remains to be done for joint models with informative censoring and missing data.

A limitation of this study is the short duration of follow-up time, which might affect the estimates of the covariates. In the data, the median follow-up was 20 months and only 8% of the study patients died. When the follow-up duration is not long enough, it has an impact on the number of CD4+ measurements, possibly leading to less reliable estimation

of the random effect model [40]. The second limitation is that our study was a retrospective cohort, limiting the availability and quality of data collected. We only included HIV positive individuals who attended clinics. Thus the results are likely biased towards a more stable study population and could be potentially biased against rapid disease progressors. For subjects with 2 or more CD4+ values in a given interval, the average of those CD4+ values for that interval was used for simplicity. We did not anticipate that missing CD4+ counts would depend on the unobserved CD4+ counts and thus we assumed missing at random in order to analyze the data. We used generalized linear mixed models which are valid under the assumption of missingness at random [27,41] and believe this assumption to be appropriate for our study population. In our study, a linear growth function was also assumed for CD4+ count trajectories so that the characteristics of intercept and slope could be related to the survival outcome; however, a linear trend may not be adequate to describe the time course of the CD4+ counts. If the CD4+ count trajectories were not linear, model misspecification may lead to biased results. With no general, user-friendly commercial software available as of yet for this technique, the use of joint modeling is limited in practice.

In summary, the analysis of the data using a joint model with latent variables that link the longitudinal models and the survival models together will result in unbiased and more efficient estimates. Joint model is intuitive and enable prognostic information to be collected from longitudinal measurements. With such prognostic information, clinicians make better informed decisions for specific patients based on their longitudinal biomarkers of disease. Thus we recommend that the joint modeling of CD4+ count progression and survival time should be performed to obtain a clear picture of the effect of specific covariate. This model provides useful predictive information on future CD4+ counts when the current value and the rate change of CD4+ count are known. As a flexible modeling approach, joint modeling can be used for patient-specific treatment strategies and future clinical interventions.

Competing interests

The authors declare that they have no competing interests. The data collected for this study was obtained solely from patients records.

Authors' contributions

Authors' contributions	HJL	PM	SS
Research concept and design	√	--	--
Collection and/or assembly of data	--	--	√
Data analysis and interpretation	√	√	--
Writing the article	√	--	--
Critical revision of the article	√	--	--
Final approval of article	√	√	√
Statistical analysis	--	√	--

Acknowledgement

We thank Ms. Stephanie Konrad who assisted data collection.

Publication history

Editors: Jimmy Efird(EIC) East Carolina University, USA.
Zhongxue Chen(Editor), Indiana University Bloomington, USA.
Received: 12-Sep-2013 Revised: 19-Sep-2013
Accepted: 21-Sep-2013 Published: 28-Sep-2013

References

1. Zhang JP, Kahana B, Kahana E, Hu B and Pozuelo L. **Joint modeling of longitudinal changes in depressive symptoms and mortality in a sample of community-dwelling elderly people**. *Psychosom Med.* 2009; **71**:704-14. | Article | PubMed Abstract | PubMed Full Text
2. Murphy TE, Han L, Allore HG, Peduzzi PN, Gill TM and Lin H. **Treatment of death in the analysis of longitudinal studies of gerontological outcomes**. *J Gerontol A Biol Sci Med Sci.* 2011; **66**:109-14. | Article | PubMed Abstract | PubMed Full Text
3. Bowman FD and Manatunga AK. **A joint model for longitudinal data profiles and associated event risks with application to a depression study**. Journal of the Royal Statistical Society: Series C 2005; 54:301–316. | Article
4. Ghisletta P. **Application of a joint multivariate longitudinal-survival analysis to examine the terminal decline hypothesis in the Swiss Interdisciplinary Longitudinal Study on the Oldest Old**. *J Gerontol B Psychol Sci Soc Sci.* 2008; **63**:P185-92. | Article | PubMed
5. Gebregziabher M, Egede LE, Lynch CP, Echols C and Zhao Y. **Effect of trajectories of glycemic control on mortality in type 2 diabetes: a semiparametric joint modeling approach**. *Am J Epidemiol.* 2010; **171**:1090-8. | Article | PubMed Abstract | PubMed Full Text
6. Taylor JM, Yu M and Sandler HM. **Individualized predictions of disease progression following radiation therapy for prostate cancer**. *J Clin Oncol.* 2005; **23**:816-25. | Article | PubMed
7. Proust-Lima C and Taylor JM. **Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach**. *Biostatistics.* 2009; **10**:535-49. | Article | PubMed Abstract | PubMed Full Text
8. Morrell CH and Brant LJ. **A two-stage linear mixed-effects/Coxmodel for longitudinal data with measurement error and survival.** *Proceedings of the Biometrics Section of the American Statistical Association* 2000; 198-203. | Pdf
9. Wulfsohn MS and Tsiatis AA. **A joint model for survival and longitudinal data measured with error**. *Biometrics.* 1997; **53**:330-9. | Article | PubMed
10. Tsiatis AA and Davidian M. **Joint modeling of longitudinal and time-to-event data: an overview.** *Statistica Sinica.* 2004; **14**:809–834. | Article
11. Ye W, Lin X and Taylor JM. **Semiparametric modeling of longitudinal measurements and time-to-event data--a two-stage regression calibration approach**. *Biometrics.* 2008; **64**:1238-46. | Article | PubMed
12. Albert PS and Shih JH. **On estimating the relationship between longitudinal measurements and time-to-event data using a simple two-stage procedure**. *Biometrics.* 2010; **66**:983-7; discussion 987-91. | Article | PubMed
13. Tsiatis AA, Degruttola V and Wulfsohn MS. **Modeling the relationship of survival to longitudinal data measured with error: applications to survival and CD4 counts in patients with AIDS**. *Journal of the American Statistical Association* 1995; **90**:27–37. | Article
14. De Gruttola V and Tu XM. **Modelling progression of CD4-lymphocyte count and its relationship to survival time**. *Biometrics.* 1994; **50**:1003-14. | Article | PubMed
15. Henderson R, Diggle P and Dobson A. **Identification and efficacy of longitudinal markers for survival**. *Biostatistics.* 2002; **3**:33-50. | Article | PubMed
16. Song X, Davidian M and Tsiatis AA. **A semiparametric likelihood**

approach to joint modeling of longitudinal and time-to-event data. *Biometrics.* 2002; **58**:742-53. | Article | PubMed

17. Vonesh EF, Greene T and Schluchter MD. **Shared parameter models for the joint analysis of longitudinal data and event times**. *Stat Med.* 2006; **25**:143-63. | Article | PubMed
18. Wu, L. **A Joint Model for Nonlinear Mixed-Effects Models with Censoring and Covariates Measured with Error, with Application to AIDS Studies.** *Journal of the American Statistical Association*, 2002; **97**:955-964. | Article
19. Pawitan Y and Self S. **Modelling disease marker processes in AIDS.** *Journal of the American Statistical Association* 1993; **88**:719–726. | Article
20. Faucett CL and Thomas DC. **Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach**. *Stat Med.* 1996; **15**:1663-85. | Article | PubMed
21. Tseng YK, Hsieh F and Wang JL. **Joint modeling of accelerated failure time and longitudinal data**. *Biometrika* 2005; **92**:587-603. | Article
22. Henderson R, Diggle P and Dobson A. **Joint modelling of longitudinal measurements and event time data**. *Biostatistics.* 2000; **1**:465-80. | Article | PubMed
23. Kartikeyan S, Bharmal RN, Tiwari RP abd Bisen PS. **HIV and AIDS: Basic elements and Priorities [monograph online]**. The Netherlands: Springer; 2007 [cited 2010 June 3]. Available from: MyLibrary.
24. Nelson K and Williams CM. **Infectious Disease Epidemiology**. Theory and Practice. **2nd ed**. 2007.
25. Langford SE, Ananworanich J and Cooper DA. **Predictors of disease progression in HIV infection: a review**. *AIDS Res Ther.* 2007; **4**:11. | Article | PubMed Abstract | PubMed Full Text
26. Laird NM and Ware JH. **Random-effects models for longitudinal data**. *Biometrics.* 1982; **38**:963-74. | Article | PubMed
27. Diggle PJ, Heagerty P, Liang K-Y. **Analysis of Longitudinal Data**. **2nd ed**. Oxford, United Kingdom: Oxford University Press; 2002.
28. Kalbfleisch JD and Prentice RL. The Statistical Analysis of Failure Time Data New York: Wiley, **2nd Edition**. 2002.
29. Dempster AP, Laird NM and Rubin DB. **Maximum Likelihood from Incomplete Data via the EM Algorithm.** *Journal of the Royal Statistical Society. Series B (Methodological)* 1977; **39**:1-38. | Website
30. Wang Y and Taylor JMG. **Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome**. *Journal of the American Statistical Association* 2001; **96:**895–905. | Article
31. Guo X and Carlin BP. **Separate and joint modeling of longitudinal and event time data using standard computer packages**. *American Statistician*. 2004; **58**:16–24.
32. Rizopoulos D. **JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data**. *Journal of Statistical Software*. 2010; **35**:1-33.
33. Lin H, McCulloch CE and Mayne ST. **Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables**. *Stat Med.* 2002; **21**:2369-82. | Article | PubMed
34. Diggle PJ, Sousa I and Chetwynd AG. **Joint modelling of repeated measurements and time-to-event outcomes: the fourth Armitage lecture**. *Stat Med.* 2008; **27**:2981-98. | Article | PubMed
35. Finkelstein DM and Schoenfeld DA. **Combining mortality and longitudinal measures in clinical trials**. *Stat Med.* 1999; **18**:1341-54. | Article | PubMed
36. Schluchter MD. **Methods for the analysis of informatively censored longitudinal data**. *Stat Med.* 1992; **11**:1861-70. | Article | PubMed
37. Schluchter MD, Greene T and Beck GJ. **Analysis of change in the presence of informative censoring: application to a longitudinal clinical trial of progressive renal disease**. *Stat Med.* 2001; **20**:989-1007. | Article | PubMed
38. Touloumi G, Babiker AG, Pocock SJ and Darbyshire JH. **Impact of missing data due to drop-outs on estimators for rates of change in longitudinal studies: a simulation study**. *Stat Med.* 2001; **20**:3715-28. | Article | PubMed
39. Scharfstein DO, Rotnitzky A and Robins JM. **Adjusting for nonignorable drop-out using semiparametric nonresponse models**. *Journal of the American Statistical Association* 1999; **94**:1096–1146. | Article
40. Kenward MG and Rosenkranz GK. **Joint modeling of outcome, observation time, and missingness**. *J Biopharm Stat.* 2011; **21**:252-62. | Article | PubMed
41. Ibrahim JG and Molenberghs G. **Missing data methods in longitudinal studies: a review**. *Test (Madr).* 2009; **18**:1-43. | Article | PubMed Abstract | PubMed Full Text

Citation:
Lim HJ, Mondal P and Skinner S. **Joint modeling of longitudinal and event time data: application to HIV study**. *J Med Stat Inform*. 2013; **1**:1.
http://dx.doi.org/10.7243/2053-7662-1-1