REVISITING METHODS FOR MODELING LONGITUDINAL AND SURVIVAL DATA: THE FRAMINGHAM HEART STUDY

Julius S Ngwa[1 2*], Howard J. Cabral[1], Debbie M. Cheng[1], David R. Gagnon[1], Michael P. LaValley[1], L. Adrienne Cupples[1, 3*]

*Correspondence to Julius S Ngwa; Email: jngwa1@jhu.edu; Phone: (330)-519-2920 or L. Adrienne Cupples; Email: adrienne@bu.edu; Phone: (617)-638-5176

Author Affiliations

[1]Department of Biostatistics, Boston University, School of Public Health, 801 Massachusetts Ave, CT 3rd Floor, Boston, MA 02118, U.S.A

[2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe St, Baltimore, MD 21205, U.S.A

[3]National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA 01702, U.S.A

Corresponding Author:
Julius S. Ngwa, Ph.D
Department of Biostatistics,
Johns Hopkins Bloomberg School of Public Health,
615 N Wolfe St E3009,
Baltimore, MD 21205,
Phone: 410-955-5808
Email: jngwa1@jhu.edu

ABSTRACT

Background: Statistical methods for modeling longitudinal and time-to-event data has received much attention in medical research and is becoming increasingly useful. In clinical studies, such as cancer and AIDS, longitudinal biomarkers are used to monitor disease progression and to predict survival. These longitudinal measures are often missing at failure times and may be prone to measurement errors. More importantly, time-dependent survival models that include the raw longitudinal measurements may lead to biased results. In previous studies these two types of data are frequently analyzed separately where a mixed effects model is used for the longitudinal data and a survival model is applied to the event outcome.

Methods: In this paper we compare joint maximum likelihood methods, a two-step approach and a time dependent covariate method that link longitudinal data to survival data with emphasis on using longitudinal measures to predict survival. We apply a Bayesian semi-parametric joint method and maximum likelihood joint method that maximizes the joint likelihood of the time-to-event and longitudinal measures. We also implement the Two-Step approach, which estimates random effects separately, and a classic Time Dependent Covariate Model. We use simulation studies to assess bias, accuracy and coverage probabilities for the estimates of the link parameter that connects the longitudinal measures to survival times.

Results: Simulation results demonstrate that Two-Step approach performed best at estimating the link parameter when variability in the longitudinal measure is low but is somewhat biased downwards when the variability is high. Bayesian semi-parametric and maximum likelihood joint methods yield higher link parameter estimates with low and high variability in the longitudinal measure. Time Dependent Covariate method resulted in consistent underestimation of the link parameter. We illustrate these methods using data from the Framingham Heart Study in which lipid measurements and Myocardial Infarction data were collected over a period of 26 years.

Conclusions: Traditional methods for modeling longitudinal and survival data, such as time dependent covariate method, that use the observed longitudinal data, tend to provide downward bias estimates. Two-step approach and joint models provide better estimates, although a comparison of these methods may depend on the underlying residual variance.

KEY WORDS: Joint longitudinal and Survival Model, Cox Model, Two-step Approach, Mixed Effect Modeling, Time Dependent Covariate Models, Weibull Distribution, Residual Variance.

1. BACKGROUND

Statistical methods for modeling longitudinal and time-to-event data has received much attention recently in medical research and is becoming increasingly useful. A common objective in this research is to characterize the relationship between the longitudinal response and the time-to-event[1, 2]. Typical settings where these may occur include HIV studies in which baseline characteristics are recorded and immunological measures such as CD4+ lymphocyte counts or viral load are measured repeatedly to assess patients' health until HIV conversion[3]. We consider data from the Framingham Heart Study in which lipid measurements and myocardial infarction (MI) data were collected over a period of 26 years. The Framingham Heart Study (http://www.framinghamheartstudy.org/) is a well-known longitudinal study that identifies potential risk factors for the development of cardiovascular disease. In this study high-density lipoprotein (HDL), low density lipoprotein (LDL) and triglycerides (TG) were measured at generally comparable time intervals over the 26-years. Time to MI was also recorded for each participant, although some subjects were censored by the end of the study period or due to death from other causes. We assess methods that characterize associations between the longitudinal lipid measures and time to MI with an emphasis on precise estimation of the parameters linking longitudinal risk factors with time to MI.

There is extensive literature and a wide range of statistical packages for jointly modeling longitudinal and survival data. Some recent work includes those of Brown and Ibrahim[4]; Zeng and Cai[5]; Tseng, Hsieh, and Wang[6]; Ye, Lin, and Taylor[7]; Ibrahim, Chu, and Chen[8]; Rizopoulos[9], among others. Tsiatis and Davidian[2] provide a comprehensive overview of earlier articles addressing a number of methods for jointly modeling longitudinal and survival data. These articles include work on joint models by Robins and Tsiatis[10]; DeGruttola and Tu[11]; Tsiatis, DeGruttola, and Wulfsohn[12]; LaValley and Degruttola[13]; Faucett and Thomas[14]; Wulfsohn and Tsiatis[15]; Wang and Taylor[16]; Xu and Zeger[17]. Sweeting and Thompson[18] provide a comparison of a shared random effects model with a naïve time-dependent covariate model and a two-stage joint model for modelling the

association between the longitudinal process and the time-to-event outcome. They conducted a simulation study to contrast these three approaches in their ability to estimate the true association between a longitudinal process and the event hazard. In their simulations they assumed a constant baseline hazard model to simulate a relatively rare event with a modest correlation between the longitudinal and survival process[18]. Our paper builds upon the Sweeting and Thompson paper and implements a new data generation scheme, assuming a Weibull distribution for the survival process and considers a wide range of scenarios for the event rates and the residual variances.

Our main objectives in this paper are to (i) evaluate a Bayesian semi-parametric joint model (BSJM) and a maximum likelihood joint model approach (MLA) that link longitudinal trajectories to survival and (ii) compare these joint maximum likelihood methods with the Two-Step Approach (TSA), and the Cox Time Dependent Covariate Model (TDCM) [4, 12, 15, 19]. In these methods the BSJM and MLA maximize the joint likelihood of the longitudinal and survival data. In the TSA, random effects are estimated separately in the first stage and the predicted longitudinal measures are then substituted into the second stage survival analysis. The argument in favor of the joint model has been the efficient use of the data as the survival information goes into modeling the longitudinal process and vice versa. The TDCM implements time varying covariate approaches in which the observed longitudinal measures are applied in the survival model.

2. METHODS

In this section, we describe the methods for jointly modeling longitudinal data and survival data using a joint likelihood. In such modeling, the main focus may be on the longitudinal component, the survival component, or both, depending on the objectives of the studies. When the focus is on one aspect, the other component is then secondary; so its parameters may be viewed as nuisance parameters[20]. Our goal is to characterize the relation between the time-to-event (primary) outcome and the longitudinal measures (secondary), adjusting for covariates in the model.

2.1 Joint Likelihood Model

We consider a joint likelihood model similar to that of Brown and Ibrahim[4], which links the longitudinal trajectories of each subject to survival data. The longitudinal responses, $Y_{i,}$ are linked to the time-to-event model using a Cox proportional hazard

model[21]. For each individual i, we let S_i be the survival time and C_i censoring time, with T_i the observed survival time. Due to censoring we observe $T_i = min(S_i, C_i)$. Let $\delta_i = I(S_i \leq C_i)$, denote the event indicator:

$$\delta_i = \begin{cases} 1 \;\; if \;\; S_i \leq C_i \\ 0 \;\; if \;\; S_i > C_i \end{cases}$$

The joint likelihood for subject i can be constructed as a product of the longitudinal model and survival model conditional on the longitudinal measures.

$$f(Y_i, t_i, \delta_i) = f(Y_i) * f(t_i, \delta_i | Y_i)$$

We begin with the likelihood model for the longitudinal measures and implement a random effects approach that models the longitudinal measures with possible measurement errors. Each subject has m_i measures denoted by Y_{ij} , $j = 1, 2, \ldots, m_i$, where Y_{ij} represents the observed outcome for the i^{th} subject at the j^{th} time point. We denote $Y_{ij}{}^*$ as the true unobserved measure value such that:

$$Y_{ij} = Y_{ij}{}^* + \epsilon_{ij} , \qquad \epsilon_{ij} \sim N(0, \sigma^2) \tag{1}$$

Y_{ij}^* is also known as the trajectory function. The trajectory can be modeled in a linear form[4] or quadratic form[12], or spline, and other time series forms can be implemented to capture the trajectory of the longitudinal measures but the trade-off is the complexity and interpretation of the model. These longitudinal measures are often missing at failure times and may be prone to measurement error. Including the raw longitudinal measurements in analysis may lead to bias if the measures are related to the censoring process[1].

In this paper we use a linear mixed effects model (LME), which allows individual or subject-specific inference following the approach of Laird and Ware[22], to fit data from longitudinal response processes.

$$E\big(Y_{ij}\big) = Y_i^*(t_{ij}) = U_{i1} + U_{i2} * t_{ij} \tag{2}$$

Here U_{i1} and U_{i2} are random effects, representing the subject-specific intercepts and slopes, and are usually assumed to be multivariate normally distributed. The variable t_{ij} represents the times for the i^{th} subject at the j^{th} time at which the longitudinal measures are recorded during the follow-up period.

For the survival model we consider the Cox model that links the time-to-event data and the longitudinal trajectories through the hazard function. For each individual $i, i =$

$1, 2, \ldots, n$, we let S_i denote the survival or event time and C_i denote the censoring time respectively. We assume that the censoring process is random or non-informative.

For individual i, t_i denotes the failure time, which may be right censored. The Cox model that links the time-to-event outcome and the longitudinal trajectories through the hazard function can be written as:

$$h_i(t) = h_0(t)\exp\{Y_i^*(t)\boldsymbol{\gamma} + X_i{}^T\alpha\} \quad (3)$$

The parameter γ is a scalar (link) parameter that links the predicted longitudinal trajectories to the hazard function; α is a vector of unknown parameters for the time independent covariate measures $X_i{}^T$; $h_0(t)$ is the baseline hazard function. From (3) one can generate a Cox partial likelihood[21] from which statistical inferences can be derived if Y_i^* were observed:

$$L_p^* = \prod_{i=1}^{n}\left\{\frac{exp\left(Y_i^*(t_i)\boldsymbol{\gamma} + X_i{}^T\alpha\right)}{\sum_{k=1}^{n} exp\left(Y_k^*(t_i)\boldsymbol{\gamma} + X_i{}^T\alpha\right)I(t_k \geq t_i)}\right\}^{\delta_i} \quad (4)$$

In (4) two assumptions are made: (i) survival times are independent and without ties; (ii) the longitudinal measures are available at each event time for all individuals. Assumption (ii) is often not true as covariate measurement times may not coincide with event times, leading to missing data in time-dependent covariates in the survival model. Such missingness may be assumed to be ignorable or MAR where the missingness is not related to the missing values that would be observed[23]. The last-value-carried-forward (LVCF) method for missing longitudinal measures has been widely used to impute the missing covariates, but may lead to bias in the estimates[1].

The hazard function for the survival component (3) given the longitudinal trajectory function yields:

$$f(t_i, \delta_i) \propto f(t_i)^{\delta_i}S(t_i)^{1-\delta_i} = h(t_i)^{\delta_i}S(t_i), \qquad where\ S(t) = exp\left(-\int_0^t h(u)du\right)$$

$$f(t_i, \delta_i|Y_i^*, X_i) = h_0(t_i)^{\delta_i}\exp\{\delta_i\left(Y_i^*(t_i)\boldsymbol{\gamma} + X_i{}^T\alpha\right)\} \quad (5)$$

$$* exp\left\{-\int_0^{t_i} h_0(u)\exp\left[Y_i^*(u)\boldsymbol{\gamma} + X_i{}^T\alpha\right]du\right\}$$

Statistical inference based on the above likelihood function is potentially computationally intensive. A non-adaptive Gauss-Kronrod integration can be employed to numerically calculate the integral[24]. The maximum likelihood parameter estimates can be obtained using the Expectation–Maximization (EM) algorithm and Newton-Raphson

approximation[25]. Or a Monte Carlo Markov Chain (MCMC) approach can be implemented in a Bayesian framework[14]. Tsiatis and Davidian[2] applied a different approach to the issue by proposing a conditional score that is also efficient and yields consistent and asymptotically normal estimators.

2.1.1 Bayesian Semi-Parametric Joint Model (BSJM)

Faucett and Thomas[14] applied a Bayesian approach to estimate the parameters of the longitudinal process (σ^2, mean and covariance of U_{i1} and U_{i2}) and the proportional hazard model (λ_0, γ and α) in the joint likelihood framework via MCMC. They specified non-informative priors for the parameters to obtain results similar to the maximum likelihood estimates. Wang and Taylor[16] applied a similar approach to model the survival component and used a more flexible approach to the longitudinal part by incorporating a stochastic process into the longitudinal trajectory. Brown and Ibrahim[4] considered a semi-parametric Bayesian joint model in which they relax the distributional assumptions for the longitudinal model using Dirichlet process priors on the random effect parameters.

We apply a Bayesian approach for jointly modeling longitudinal and survival data similar to that of Brown and Ibrahim[4]. For the survival model, we specify normal priors for the parameters γ and α. We use MCMC methods to obtain posterior distributions of the parameters given the data. For the longitudinal component, we consider the LME model in which we assume prior distributions for the mean parameters, with variance components. From (1) and (2) we have that:

$$Y_{ij} = Y_{ij}^{*} + \epsilon_{ij}; \quad Y_{ij}^{*} = U_{i1} + U_{i2} * t_{ij}; \quad i = 1, 2, \dots, n; \quad j = 1, 2, \dots, m_i$$

$$\epsilon_{ij} \sim N(0, R_i)$$

$$Y_{ij}|U_i \sim N\left(Y_{ij}^{*}, V\right), \qquad V = Z_i G Z_i' + R_i, \qquad R_i = \sigma^2 I_{m_i}$$

We model the longitudinal trajectories using a linear growth curve model with U_{i1} and U_{i2} representing the random intercepts and slopes. G denotes the covariance matrix of the random effects and Z_i is a diagonal matrix of the longitudinal time points. A subject $i's$ contribution to joint likelihood function can be written as:

$$f(Y_i, t_i, \delta_i) = h_0(t_i)^{\delta_i} exp\left\{\delta_i\left(Y_i^{*}(t_i)\gamma + X_i^{T}\alpha\right)\right\} * exp\left\{-\int_0^{t_i} h_0(u)\, exp\left[Y_i^{*}(u)\gamma + X_i^{T}\alpha\right] du\right\} * \frac{1}{(2\pi\sigma^2)^{\frac{m_i}{2}}} exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{m_i}\left(Y_{ij} - Y_{ij}^{*}\right)^2\right\} \quad (6)$$

A typical approach assumes a constant hazard ratio for the baseline hazard and normal prior distributions for the random effects and the unknown regression parameter (α). The variance-covariance matrix (V) is assigned a Wishart distribution through a precision matrix P:

$$P = V^{-1} \sim Wishart(Q^{-1}, v) \quad \text{and}$$
$$\sigma^2 \sim IG(a, b)$$

In the Wishart distribution, Q denotes the scale matrix and v denotes the degrees of freedom. *IG* represents the inverse gamma distribution with shape a and scale parameter b.

We employed R and WinBUGS to obtain parameter estimates and credible intervals from the posterior distribution of the Bayesian modeling using standard Gibbs sampling MCMC methods[26, 27]. Once the model, data and initial values are specified in WinBUGS, the parameters can be monitored until convergence is attained. At the end of the MCMC process we obtain plots and diagnostic statistics to assess convergence in the parameters. One shortcoming of the BSJM is the computing time involved in estimating the parameters.

2.1.2 Maximum Likelihood Approach (MLA)

The maximum likelihood approach for jointly modeling longitudinal and survival data was described by Rizopoulos[9]. This method, employing the same general approach as Wulfsohn and Tsiatis[15], implements a shared parameter model for the joint modeling of longitudinal responses and time-to-event data.

The joint distribution for the longitudinal and survival model can be written as:

$$f(Y_i, T_i, \delta_i) = \int f(Y_i|U_i) * \{h_i(T_i|U_i)^{\delta_i} S(T_i|U_i)\} * f(U_i)\, dU_i \quad (7)$$

where $S(T_i|U_i)$ and $h_i(T_i|U_i)$ represent the survival function and the hazard function respectively conditional on the random effects, and $f(Y_i, T_i, \delta_i)$ is the density function. The distributional form in (7) assumes that given the random effects, the longitudinal measures are independent of the time-to-event outcome and are independent of each other. The true unobserved values of the longitudinal measures $\left(Y_{ij}^*(t)\right)$ are associated with the event outcome (T_i) through the hazard function:

$$h_i\big(t|Y_i(t)\big) = h_0(t) exp\{Y_{ij}^*(t)\boldsymbol{\gamma} + X_i{}^T \alpha\}$$

Parameter estimates, from (7), can be obtained using the JM package in R. The package fits a shared parameter model for the joint modeling of normally distributed longitudinal responses and event time under the likelihood approach. The maximum likelihood estimation for joint models is based on the maximization of the log-likelihood corresponding to the joint distribution of the time-to-event and longitudinal outcomes[9]. The maximization is challenging as the integral of the survival function has no analytic solution. Following Rizopoulos, we implemented a Weibull model using Gauss-Hermite integration to approximate the integral. In the estimation process a hybrid optimization approach is employed that starts with EM and then continues with direct maximization. The procedure for the EM algorithm uses a fixed number of iterations and if convergence is not achieved it switches to a quasi-Newton algorithm until convergence is attained.

2.2 Two Step Approach (TSA)

In the TSA the parameters of the longitudinal process are estimated separately, and the estimated random effects are substituted directly into the survival model. This approach was first implemented by Tsiatis[12] in which a linear mixed effect model is fit to the longitudinal measures and the fitted values are inserted to the Cox Proportional Hazard model in the second stage as time dependent covariate measures. Ye, Lin, and Taylor[7] proposed two approaches for modeling the TSA called risk set regression calibration (RRC) and ordinary regression calibration (ORC). In the first approach the LME model is fit using the observed longitudinal data only among subjects with the event. This approach can be implemented if the longitudinal trajectories of subjects who experienced the event may be different from those who did not.

In the second approach the LME model is fit using observed longitudinal data from all subjects. In the first step the longitudinal process is estimated using the LME model in (1) and (2) and estimates of the random effects are used to obtain predicted values of the longitudinal measures at event times. In the second step the predicted longitudinal measures are used in the Cox Model to estimate the hazard for the event. The variance estimates for the parameters are obtained from the observed information of the partial likelihood function. See equation (3). X_i represents the time independent covariate measures in the model and the $Y_i^*(t)$ represents the predicted values of the longitudinal measures at event time t. The link parameter (γ) in this approach can be interpreted as the association between the longitudinal measures at event time t and the survival time. The

estimation and inference for the hazard model (3) can be performed by using the partial likelihood theory proposed by Cox[21].

The main advantage of this approach is that it is simple and can be implemented using existing statistical packages. Tsiatis et al[12] argue that a repeated measures random effects model for the covariate process is superior to naive methods where one maximizes the partial likelihood of the Cox using the observed covariates values. Ye, Lin, and Taylor[7] argue that there are two main disadvantages of a simple TSA; 1) it may provide biased estimates especially when the longitudinal process and the survival process are strongly associated; and 2) it does not incorporate the uncertainty of estimation in the first stage into the second stage, possibly leading to under-estimation of the standard errors. We evaluate a number of scenarios to assess the validity of these assumptions using simulation studies.

2.3 Time Dependent Covariate Modeling (TDCM)

A time dependent explanatory variable is one whose value for any given subject may change over the period of time that the subject is observed[28]**.** The TDCM employs the observed longitudinal measures to predict an event. Therneau and Grambsch[19] considered a well-known example of TDCM using the Stanford Heart Transplant Program. Data for a subject is presented as multiple observations, each of which applies to an interval of observation. A proportional hazards model is often used to analyze covariate information that change over time. The hazard may be thought of as being proportional to the instantaneous probability of an event at a particular time[21].

We consider sample of size n, consisting of $[T_i, \delta_i, [Y_i(t), 0 \leq t \leq T_i], i = 1, 2, \dots, n]$, where T_i is the time-to-event for the i^{th} subject, δ_i is the event indicator. The vector $Y_i(t) = \left[Y_{i1}(t), Y_{i2}(t), \dots, Y_{ig}(t)\right]^t$ is a set of observed longitudinal measures, and $m_i \leq g$ is the number of times intervals for the i^{th} subjects and g is the maximum.

$$\delta_i = \begin{Bmatrix} 1 \;\; if \;\; T_i \leq C_i \\ 0 \;\; if \;\; T_i > C_i \end{Bmatrix}$$

The hazard for this model at time t can be written as:

$$h(t) = h_0(t) * exp\left(\beta Y_i(t) + X_i{}^T \alpha\right) = h_0(t) * exp\left(\sum_{k=1}^{m_i} \beta_k Y_{ik}(t) + X_i{}^T \alpha\right) \quad (8)$$

The vector $Y_i(t) = \left[Y_{i1}(t), \ldots, Y_{im_i}(t)\right]$ is a set of covariates and m_i is the number of longitudinal measures for the i^{th} subject. We define $t_1 < t_2 < t_3 < \cdots < t_D$ as a set of ordered event times and $Y_i(t_i)$ as the time-dependent covariate associated with the individual whose failure time is t_i. The risk set $R(t_i)$ at time t_i is the set of all individuals who are still under study at a time just prior to t_i. The partial likelihood based on the hazard function specified (9) can be written as:

$$L(\alpha, \beta) = \prod_{i=1}^{D} \left\{ \frac{exp\left(\sum_{k=1}^{m_i} \beta_k Y_{ik}(t) + X_i{}^T \alpha\right)}{\sum_{l \epsilon R(t_i)} exp\left[\sum_{k=1}^{m_i} \beta_k Y_{lk}(t_i) + X_l{}^T \alpha\right]} \right\} \quad (9)$$

In most applications, $\beta_k = 0$ for intervals which are not under consideration. The estimates can be obtained by maximizing the likelihood specified in (9).

3. SIMULATIONS

In this section, we carried out a series of simulations to compare the performance of these four methods for modeling longitudinal and survival data described above. The performance of these methods was assessed with Type I error, bias, accuracy, and coverage probabilities for the link parameter. We simulated longitudinal and survival data to resemble data from the Framingham Heart Study. In Table 1, we highlight the simulation model and the required parameters (residual error, random effect means for the longitudinal process, covariance of random effects and coefficients for Age and Sex) for simulating the longitudinal and survival data. The longitudinal trajectories were generated from a linear model adjusting for the age at baseline of the participants. The survival time was generated to depend on the longitudinal measures and a set of covariates (Age at baseline and Sex).

In our previous paper on Generating Survival Times with Time-Varying Covariates[29] we provide a detailed algorithm for generating survival times for time-varying Cox Exponential and Weibull models using Lambert's W function. The simulation requires the specification of the longitudinal measures and the distribution of the survival data. The longitudinal measures can be obtained from a mixed effects model by introducing the random effects as shown in (1) and (2). We focus on a simple linear mixed effect model (LME), which allows individual or subject-specific inference[19] to fit data from longitudinal response process. For the survival data, two independent Weibull distributions were simulated; (i) the survival times that would be observed if the follow-up had been

sufficiently long to reach the event and (ii) the censoring mechanism. The survival distribution was generated to depend on the longitudinal measures and a set of covariates in accordance with the model in (5). If the survival time is less than or equal to the censored time, then the event is considered to be observed and the time-to-event equals the survival time; otherwise the event is considered censored and the time-to-event equals the censored time[30]. We assume random non-informative right censoring and employ a uniform distribution for censoring that allows a maximum follow-up time of 30 years. We use the Weibull distribution to generate the survival data. We simulated 1000 independent multivariate datasets consisting of longitudinal measures, time-to-event outcomes, and additional covariates. We present a general formula (see Table 1) which links the survival time of the Cox model and the random effects of the longitudinal model. We use the Weibull distribution to generate survival times from the longitudinal data in the simulation studies by using the Lambert function[29].

We applied the following methods to each of 1000 replicates (10,000 for Type I error): (1) Bayesian semi-parametric joint model (BSJM); (2) Maximum likelihood approach (MLA); (3) Two-step approach (TSA); (4) Time dependent covariate model (TDCM). In the BSJM method several criteria were considered in the MCMC runs including: 1) the number of chains for each run, 2) correlation between successive draws and 3) the length of burn-in time. A total of 101,000 iterations were run with a thinning of 50 and a burn-in of 1000 for 4 chains each, thus providing a sample of 500 iterations per chain. Empirical means and standard deviations for each variable were estimated. The quantiles for each variable are estimated in WinBUGS and used to compute credible intervals. Two shortcomings of the BSJM are the computing time involved in estimating the parameters and the lack of convergence in some of the MCMC runs.

The variable Sex is considered a fixed covariate at each exam in all the methods. The baseline Age is also included in the model; the data structure is a single row per subject where longitudinal measures, covariates and the overall survival/censoring time are specified for each subject. The statistical analyses were performed using SAS Software (version 9.3; SAS Institute, Cary, NC) and the computing environment R (R Development Core Team, 2012). The Bayesian analysis was conducted in R using WinBUGS version 1.4.3, MRC Biostatistics Unit, Cambridge, UK.

4. RESULTS

We compute Type I error for the link parameter to assess all four methods (Table 2) for a sample size of 100 with 10,000 replicates. The methods appear to provide Type I error rates close to the nominal level (0.050) for both the Exponential and Weibull models with two exceptions. The BSJM shows deflated type I error for both Exponential and Weibull with high censoring rates and elevated Type I error for Weibull with low censoring. The MLA shows elevated Type I errors for Exponential with high censoring.

Table 3 presents the estimates, SEs, coverage probability (CP), bias and mean square error (MSE) for the comparison of the longitudinal effect on survival using the Weibull distribution for n = 100. The MLA and BSJM provide lower standard errors and shorter confidence intervals for the link estimate (Table 3) compared to the other methods. The simulation scenarios with higher censoring rates show higher standard errors and larger confidence intervals. The results suggest the TSA performs best at estimating the link parameter, as it has lower bias and higher coverage probability (CP) compared to the other methods. For example, with 10% censoring and $\gamma = 0.00$, the TSA has a bias of 0.000 with a CP value of 95.2%. The TSA provides larger standard errors for the link estimate with larger confidence intervals and CP values close to the nominal level of 95%. The TDCM yields a negative bias with low CP values when there is a strong effect of the longitudinal measure on the outcome.

The performance of the methods in estimating the simulated effects of Age ($\alpha_1 = 0.050$) and Sex ($\alpha_2 = -0.500$) was also assessed. As the effect of the longitudinal measure on survival becomes stronger, the effect of Age becomes weaker for all the methods except the TSA. The SE's, CP and confidence intervals tend to be smaller when n = 1000, as expected. The results for the sex effect show that all methods provide precise estimates for all the scenarios considered (See Supplemental Figure S4). From the simulation results we see that the standard errors and confidence intervals become larger with higher censoring rates. The sex effect is fairly consistent among the different methods for all simulation scenarios considered.

We also implemented a data generation scheme by Rizopoulos to confirm our results across different settings. We compared the methods using FHS parameters specified in Table 1. The BSJM was not included in this setting due to the computing time involved in estimating the parameters. We used a sample size of 1000 with censoring set at 90% and a link parameter of 0.500. As shown in Table 4, with a residual variance of the longitudinal

trajectories ($\sigma^2 = 0.1161$), the TSA showed low bias of 0.006 with higher coverage probability 0.952 compared to the other methods. The TSA also provided larger standard errors for the link estimate with larger confidence intervals. The TDCM provided a negative bias of -0.137 with low CP values of 0.870. The MLA provided the highest estimate in the link parameter with a positive bias of 0.082 and a coverage probability of 0.882. When we applied the above residual variance ($\sigma^2 = 0.1161$), the pattern in the results provided similar findings as our data generation scheme. When the residual error was increased ($\sigma^2 = 0.396$), to reflect the errors of the joint model by Wulfsohn and Tsiatis[15], we observed a similar trend in the results with the TSA providing the least biased estimate, and nearly 95% coverage, but larger SE's (Table 4). The link parameter estimates (0.191) for the TDCM were highly attenuated in this scenario. With a larger residual variance, the TSA showed modest bias (0.452) in the link parameter compared to the scenario with lower residual variance. As seen earlier, the MLA provided larger estimates in the link parameter (0.580) compared to the other methods.

5. APPLICATION TO FRAMINGHAM HEART STUDY

To illustrate the performance of these methods, we examined data from the Framingham Heart Study (FHS) in which lipid measurements and Myocardial Infarction (MI) data were collected over a period of 26 years. The FHS is a widely known longitudinal study that seeks to identify common factors contributing to cardiovascular disease (CVD). Since 1948 three generations of participants have been recruited and followed over the years: Original cohort (recruited in 1948), Offspring (recruited in 1971) and third generation (recruited in 2002). Among the offspring participants, high-density lipoprotein (HDL), low density lipoprotein (LDL) and triglycerides (TG) were measured at fairly similar time intervals over a period of 26 years. The time to myocardial infarction was recorded for each participant, although some subjects were censored by the end of the study period or due to death from other causes. We log transformed the TG measures in our analysis to reduce skewness in TG measures. A total of 2262 subjects with complete data, until event or death, were followed from 1979 to 2005 and data was collected at the start of each exam (Table 5). The mean age at baseline was 43.3 years. Six exams were considered in the analysis from Exam 2 to Exam 7. The mean triglyceride measures at each exam were calculated with values ranging from 100.49 to 158.70, showing an increased trend from Exam 2 (1979 – 1983) to Exam 7 (1998 – 2002). The mean (SD) follow-up time was 22.80

(5.13). The cumulative event rate for the 26-year period was 3.71%. The proportion of female participants was 51%.

We fit the FHS data to the models described in Section 2 and characterize the association between the longitudinal measures and time-to-event response. We used log TG at each exam for the longitudinal part of the model assuming a linear trend and survival time measured from exam 2 to MI or loss to follow up (up to 2005). We adjusted for Sex and baseline Age in all the models. In Figure 1 we show the distribution of time-to-event among the 2262 subjects with complete data. The survival distribution among subjects with events was fairly uniform and the distribution of survival times for censored subjects was skewed to the left with most censoring times occurring at the tail end of the distribution (20 – 26 years).

In Table 6 we present the estimates for Age, Sex and the link parameter (γ) for each method. The link parameter provides a measure of the association between the longitudinal TG measures and the risk of MI; γ is the log hazard ratio for a one unit increase in the log TG at time t in the survival model. The link parameter refers to the longitudinal levels at time t and does not relate to changes in the longitudinal model over time. The results suggest a higher estimate in the link parameter for the MLA and BSJM (γ = 0.9764 and 1.0263). These results are similar to the findings by Wulfsohn and Tsiatis[15] with higher estimates in the joint maximization compared to the two-step approach. The TDCM method shows a lower link estimate compared to the TSA and the joint likelihood methods. The TSA and the MLA provided higher standard errors for the link parameter compared to the other methods. Tsiatis[12] argued that the standard error for the link parameter is greater when using the joint estimation procedure compared to the TSA because the random effects are assumed to be influenced by the uncertainty in the estimated growth curve parameters; thus, more variability is incorporated. We do not see this, however, in these results. The Age effects and standard errors were similar among the methods with estimates ranging from 0.050 to 0.065. The Sex effect was fairly consistent among the different methods ranging from -1.025 to -0.999. Using a 0.05 level of significance the Age, Sex and Log of the triglyceride measures were significantly associated with risk of myocardial infarction. The results of all methods suggest that repeated measures of triglyceride levels are significantly associated with the risk of myocardial infarction in the Framingham Heart Study Cohort.

6. DISCUSSION

In this paper, we compared longitudinal and survival data methods that link longitudinal trajectories to survival. These methods quantify the link parameter as the association between the current level of a longitudinal process and the survival outcome. We analyzed data from FHS in which triglyceride measurements and Myocardial Infarction (MI) data were collected over a period of 26 years. We used a simulation study to assess the performance of these methods with respect to bias, accuracy and coverage probabilities. We compared the TSA to the MLA, BSJM and TDCM methods using a derivation of the survival time function for modeling time dependent covariate data.

Based on the simulation studies the TSA, which uses the predicted longitudinal measures, performed best at estimating the link parameter for moderate residual variance. The joint likelihood methods provided upwardly biased estimates in the link parameter, similar to the findings by Wulfsohn and Tsiatis[15]. The TDCM that uses the observed longitudinal measures as time-dependent covariate measures in the survival analysis resulted in underestimation of the true parameters. These results are similar to the findings by Sweeting and Thompson[18]. They recommended the use of a shared random effects model. In most of our models the age effect was attenuated depending on the association of the longitudinal measures on survival. This result was expected as age was associated with the longitudinal measures. The time independent covariates (baseline Age and Sex) were unbiased when there was no association between the longitudinal measures and survival ($\gamma = 0.00$). Comparison of the methods in Framingham Heart Study revealed similar patterns.

We implemented a data generation scheme by Rizopoulos in order to confirm our results across the different settings. The results showed similar findings with the two-step approach performing best at estimating the link parameter connecting the longitudinal measures to the event time. For larger residual errors, we found that the TDCM methods had attenuated estimates of the link parameter and the MLA provided upwardly biased estimates. The TSA also yielded biased estimates in the link parameter when there are larger residual errors. These results were similar to the findings by Dafni and Tsiatis[31]. They used different values for the residual error ($\sigma^2 = 0.32, 0.62, 1.24$). In their simulations with larger measurement error, the bias of the estimates based on the observed longitudinal measures increased dramatically in the positive direction. They indicated that the two-step model yielded parameter estimates that were somewhat biased towards the

null (for larger residual errors). We explored several other scenarios not presented in this paper where we varied the residual error in the longitudinal measures. Our results show that with low residual errors in the longitudinal measures, the TSA provides results similar to time dependent covariate methods that use the observed longitudinal measures. We interpreted this to mean that a small residual error results in low measurement error and the observed values are comparable to the predicted values.

One limitation of our study is the linearity assumption in the longitudinal measures. The trajectory can be modeled in a linear form[4] or quadratic form[12]. Splines and other time series forms can also be implemented to better capture the trajectory of the longitudinal measures, but the trade-off is the complexity of the model and interpretability. The simulation data generation scheme was also based on the two-step approach; this may provide more precise estimates when analyzing the simulated data using the two-step model. Another limitation is the number of time points used for the longitudinal measures. With more exam visit time-points the LME model becomes computationally intensive. In addition, the use of distributions other than the Exponential and Weibull is indispensable in investigating the characteristics of the Cox proportional hazard model. There is the need for the use of empirical distributions to handle flexibly parameterized proportional hazard models[32]. Despite these limitations this paper strengthens the current knowledge on methods for jointly modeling longitudinal and survival data.

7. CONCLUSION

Traditional methods, such as TDCM that use observed data, tend to provide downward bias towards the null. The TSA and joint models provide better estimates, although a comparison of these methods may depend on the underlying residual variance. Hence, an avenue for future exploration is to evaluate the degree of attenuation in the link parameter by the magnitude of the residual variance. Joint modeling for longitudinal and survival data has also recently received attention in statistical genetics research. In genome wide association studies and gene expression studies, longitudinal and survival measures are often collected over time. The development of new methods to handle these high dimensional data is essential. Binary longitudinal measures, multiple survival endpoints and missing data analyses are also important areas for further investigation using the methods for jointly modeling longitudinal and survival data.

8. DECLARATIONS

Ethics, Approval and Consent: The research protocols of the FHS are reviewed annually by the Institutional Review Board of the Boston University Medical Center and by the Observational Studies Monitoring Board of the National Heart, Lung and Blood Institute. Since 1971, written consent has been obtained from participants before each examination. Information about the content of the Framingham Heart Study research examinations is presented to the participants at each examination cycle in the text of the corresponding consent form and in a discussion with a trained admitting coordinator at the beginning of the scheduled appointment. Information from every completed consent form is coded and recorded in a database. Questions regarding the ethical conduct of research are presented by FHS investigators to the FHS Ethics Advisory Board. The Advisory Board reviews, discusses, and make recommendations regarding these questions.

Consent to Publish: Manuscript does not contain any individual person's data.

Availability of Data: The data used in this study are available at NIH BioLINCC (https://biolincc.nhlbi.nih.gov/home/). They can also be provided to interested researchers on written request to FHS. Request for FHS data may be done by submitting a proposal through the FHS web-based research application. A catalogue of the FHS data repository may be accessed through the FHS website: www.framinghamheartstudy.org/researchers/description-data/

Competing Interests: There were no competing interests in the conduct of this study.

Acknowledgements: We thank Dr. Timothy Heeren and Dr. Michael Pencina for their input in the simulation analysis.

Authors' Contributions: JSN and LAC conceived the study. The data were cleaned by JSN and LAC. JSN, HJC, DMC, DRG, MPL and LAC authors contributed towards the design of

the simulation scenarios implemented in the study. The analysis was carried out by JSN. The manuscript was written by JSN. JSN, HJC, DMC, DRG, MPL and LAC contributed toward reviewing and revising the final manuscript.

Funding: This work was supported by the National Heart, Lung and Blood Institute's Framingham Heart Study (Contracts No. N01-HC-25195 and HHSN268201500001I) of the National Institutes of Health and Boston University School of Medicine. A portion of this research was conducted using Boston University Linux Cluster for Genetic Analysis (LinGA) funded by the National Center for Research Resources (NIH NCRR).

LIST OF ABBREVIATIONS: BSJM: Bayesian Semi-Parametric Joint Modeling; CP: Coverage Probability; LME: Linear Mixed Effects; MLA: Maximum Likelihood Approach; TDCM: Time Dependent Covariate Modeling; TSA: Two Step Approach; MSE: Mean Square Error

REFERENCES

1. Prentice R. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*. 1982; 69: 331-42.
2. Anastasios A. Tsiatis MD. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*. 2004.
3. Ngwa JS, Cabral HJ, Cheng DM, et al. A comparison of time dependent Cox regression, pooled logistic regression and cross sectional pooling with simulations and an application to the Framingham Heart Study. *BMC Medical Research Methodology*. 2016; 16: 148.
4. R Brown E and G Ibrahim J. A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*. 2003; 59: 221-8.
5. Zeng D and Cai J. Simultaneous modelling of survival and longitudinal data with an application to repeated quality of life measures. *Lifetime data analysis*. 2005; 11: 151-74.
6. Tseng Y-K, Hsieh F and Wang J-L. Joint modelling of accelerated failure time and longitudinal data. *Biometrika*. 2005; 92: 587-603.
7. 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.
8. Ibrahim JG, Chu H and Chen LM. Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*. 2010; 28: 2796-801.
9. Rizopoulos D. JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software (Online)*. 2010; 35: 1-33.
10. Robins J and Tsiatis AA. Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika*. 1992; 79: 311-9.
11. De Gruttola V and Tu XM. Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*. 1994: 1003-14.
12. Tsiatis A, Degruttola V and Wulfsohn M. 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.
13. LAVALLEY MP and DEGRUTTOLA V. Models for empirical Bayes estimators of longitudinal CD4 counts. *Statistics in Medicine*. 1996; 15: 2289-305.

14. Faucett CL and Thomas DC. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in medicine*. 1996; 15: 1663-85.
15. Wulfsohn MS and Tsiatis AA. A joint model for survival and longitudinal data measured with error. *Biometrics*. 1997: 330-9.
16. 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.
17. Xu J and Zeger SL. Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2001; 50: 375-87.
18. Sweeting MJ and Thompson SG. Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*. 2011; 53: 750-63.
19. Therneau TM and Grambsch PM. *Modeling survival data: extending the Cox model*. Springer Science & Business Media, 2013.
20. Wu L. *Mixed effects models for complex data*. CRC Press, 2009.
21. Cox DR. Models and life-tables regression. *JR Stat Soc Ser B*. 1972; 34: 187-220.
22. Laird NM and Ware JH. Random-effects models for longitudinal data. *Biometrics*. 1982: 963-74.
23. Little RJ and Rubin DB. *Statistical analysis with missing data*. John Wiley & Sons, 2014.
24. Piessens R, de Doncker-Kapenga E, Überhuber CW and Kahaner DK. *QUADPACK: a subroutine package for automatic integration*. Springer Science & Business Media, 2012.
25. Herzet C, Wautelet X, Ramon V and Vandendorpe L. Iterative synchronization: EM algorithm versus Newton-Raphson method. *Acoustics, Speech and Signal Processing, 2006 ICASSP 2006 Proceedings 2006 IEEE International Conference on*. IEEE, 2006, p. IV-IV.
26. Venables WN and Smith DM. The R development core team. *An Introduction to R R Foundation for Statistical Computing, Vienna, Austria*. 2006.
27. Lunn DJ, Thomas A, Best N and Spiegelhalter D. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*. 2000; 10: 325-37.
28. Cox DR and Oakes D. *Analysis of survival data*. CRC Press, 1984.

29. Ngwa JS, Cabral HJ, Cheng DM, Gagnon DR, LaValley MP and Cupples LA. Generating survival times with time-varying covariates using the Lambert W Function. *Communications in Statistics-Simulation and Computation*. 2019: 1-19.
30. Burton A, Altman DG, Royston P and Holder RL. The design of simulation studies in medical statistics. *Statistics in medicine*. 2006; 25: 4279-92.
31. Dafni UG and Tsiatis AA. Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*. 1998: 1445-62.
32. Bender R, Augustin T and Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine*. 2005; 24: 1713-23.