Published in final edited form as:

Biometrics. 2017 June ; 73(2): 666–677. doi:10.1111/biom.12588.

A Joint Modeling Approach for Multivariate Survival Data with Random Length

Shuling Liu¹, Amita K. Manatunga¹, Limin Peng¹, and Michele Marcus²

¹Department of Biotatistics and Bioinformatics, Emory University, Atlanta, Georgia, U.S.A.

²Department of Epidemiology, Emory University, Atlanta, Georgia, U.S.A.

Summary

In many biomedical studies that involve correlated data, an outcome is often repeatedly measured for each individual subject along with the number of these measurements, which is also treated as an observed outcome. This type of data has been referred as multivariate random length data by Barnhart and Sampson (1995). A common approach to handling such type of data is to jointly model the multiple measurements and the random length. In previous literature, a key assumption is the multivariate normality for the multiple measurements. Motivated by a reproductive study, we propose a new copula-based joint model which relaxes the normality assumption. Specifically, we adopt the Clayton-Oakes model for multiple measurements with flexible marginal distributions specified as semi-parametric transformation models. The random length is modeled via a generalized linear model. We develop an approximate EM algorithm to derive parameter estimators and standard errors of the estimators are obtained through bootstrapping procedures and the finite-sample performance of the proposed method is investigated using simulation studies. We apply our method to the Mount Sinai Study of Women Office Workers (MSSWOW), where women were prospectively followed for one year for studying fertility.

Keywords

Clayton-Oakes model; Approximate EM algorithm; Joint models; Menstrual cycle length; Random length data; Semi-parametric transformation model; Time-to-pregnancy

1. Introduction

There has been extensive literature for the analysis of correlated data with different types of outcomes including continuous, ordinal and survival outcomes etc. (Laird and Ware, 1982; Cai and Prentice, 1995; Liang and Zeger, 1995). In such literature, multiple measurements from each subject are correlated and the number of these measurements is treated as fixed. When the number of measurements is a random variable and may be an outcome of interest, joint models with random effects have been considered to model both outcomes. These joint models aim to determine the influence of covariates on both outcomes, the within-subject correlation as well as the association between the two outcomes (Hogan and Laird, 1997; Dunson et al., 2003). Alternatively, Barnhart and Sampson (1995) described this type of data as random length data where a vector of observations is collected for each subject and the length of the vector is also a random variable. They considered a multivariate normal

distribution for the correlated outcomes, a generalized linear model for the length of the vector and both distributions are parameterized in terms of treatment effects.

In this manuscript, we study the modeling approach as in Barnhart and Sampson (1995), but we consider the copula model for repeated measures to relax the multivariate normality assumption and the distribution of the length of the vector is specified via a generalized linear model. Our work is motivated by a reproductive study called the Mount Sinai Study of Women Office Workers (MSSWOW). In this study, women were followed prospectively for one year in order to study fertility until a clinical pregnancy or the end of the study. Multiple measures of menstrual cycle lengths (MCLs) for each subject were observed. Time-to-pregnancy (TTP) was defined as the number of menstrual cycles taken to conceive excluding the conception cycle. When a woman gets pregnant and hence is no longer at risk of menstrual bleeding, MCL at the conception cycle cannot be observed. Therefore, it is natural to view MCL collected from the first attempt to actual conception as a vector containing multiple measurements of MCL with random vector length equal to TTP, which is similar to the terminology of random length data in Barnhart and Sampson (1995). Most existing statistical approaches are concentrated on modeling either TTP or MCL separately for investigating the covariate effects (Harlow and Zeger, 1991; Scheike and Jensen, 1997; Guo et al., 2006). Methods are limited in studying both MCL and TTP simultaneously and the relationship between MCL and TTP has not been well investigated. We take an approach that considers MCL as a mediator variable which describes how the effect of covariate may influence the TTP. That is, the covariates affect MCL, and consequently, the changes in the distribution of MCL will affect the outcome TTP, which facilitates the relationship between MCL and TTP.

As in many other epidemiologic studies, several complicated issues have been raised by MSSWOW study. First, there is evidence showing that the distribution of MCL has a long right tail and therefore a normal distribution is not adequate (See Figure 1; Harlow and Zeger, 1991; Murphy et al., 1995; Harlow et al., 2000; Guo et al., 2006). Second, when a subject does not conceive at the end of the study, the subject's TTP is censored and the remaining MCLs until pregnancy are missing. Third, since some subjects were trying to get pregnant before entering the study, TTP is left truncated and MCLs before entry are also missing. These aforementioned challenging and complex features of TTP and MCLs data indicate the need for flexible statistical methods. McLain et al. (2012) considered a Bayesian framework for modeling MCL and TTP simultaneously, with a mixture of normal and Weibull (or extreme-value) distribution to handle the skewness of the distribution of MCLs. However, with skewed continuous outcomes, parametric methods with mixture distributions impose computational difficulties and raise additional questions about the robustness of inference to its assumptions; therefore, a more flexible semi-parametric modeling is desired. This motivates us to specify the marginal distribution of MCLs via a general class of semi-parametric transformation model (Zeng and Lin, 2007; Chen and Yu, 2012), which includes proportional hazards model and proportional odds model as two special cases.

We consider a joint model where the repeated measurements are assumed to follow a special type of copula model known as the Clayton-Oakes model (Clayton, 1978; Oakes, 1982) and the marginal distribution of the Clayton-Oakes model follows the semi-parametric

transformation model. Furthermore, the length of the vector, which corresponds to TTP, is modeled via a discrete time hazard model (Scheike and Jensen, 1997). To understand the relationship between multiple measurements and the random length (e.g., MCL and TTP), shared parameters are imposed on both the Clayton-Oakes model and the discrete survival time model. With regard to the Clayton-Oakes model, Glidden and Self (1999) have proposed a semi-parametric estimation approach when the length of vector is fixed and the marginal distribution follows a proportional hazards model. We adopt a similar estimation method, but in our case we have a shared-parameter joint model, with a more general specification for the marginal distributions (transformation models) and a generalized linear model with a complementary log-log link function for the random length.

In Section 2, we describe our joint modeling framework and marginal model specification as well as properties of the model. Section 3 proposes an approximate EM algorithm to derive generalized maximum likelihood estimators (Gill, 1989) for the parameters in the joint model. We provide semiparametric likelihood that allows for censoring and left truncation in both outcomes (repeated measures and random length). In Section 4, we conduct general simulation studies to evaluate the performance of the proposed method. We apply our method to the MSSWOW study in Section 5.

2. The Models

Suppose that we have m subjects. Let i index the subject and j index the measurement for each subject. Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{iN_i})'$ denote a vector of multiple measurements on a quantitative variable with a random length N_i for the i -th subject, $i = 1, \dots, m$. In addition, a p -dimensional covariate vector $\mathbf{X}_i$ is collected for each subject which may affect both the repeatedly measured outcome and the random length. The joint model is based on the factorization of the joint distribution of $(\mathbf{Y}_i, N_i)$. First, we define the model for the vector of multiple measurements $\mathbf{Y}_i$ given N_i . Conditional on $N_i = n_i$, we assume that $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})'$ has a multivariate survival distribution that follows the Clayton-Oakes model (Clayton, 1978; Oakes, 1982) as

$$S_j(\mathbf{y}_i | N_i = n_i, \mathbf{X}_i; \boldsymbol{\beta}) = \Pr\{Y_{i1} > y_{i1}, \dots, Y_{in_i} > y_{in_i} | N_i = n_i, \mathbf{X}_i; \boldsymbol{\beta}\} = \left[\sum_{j=1}^{n_i} S(y_{ij} | \mathbf{X}_i; \boldsymbol{\beta})^{-\theta} - n_i + 1 \right]^{-\frac{1}{\theta}} \quad (1)$$

where $S_j(\cdot)$ indicates a joint distribution of the multiple measurements, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression parameters associated with $\mathbf{X}_i$, $S(y_{ij} | \mathbf{X}_i; \boldsymbol{\beta})$ represents the marginal survivor distribution of the j -th observation for the i -th subject given $\mathbf{X}_i$ and θ depicts the within-subject dependence. As θ approaches 0, the observations within the same subject become independent, and the joint survival function is simply the product of the marginal survival functions. When θ goes to $+\infty$, the joint survival function converges to its upper Frechet bound and $S_j(\mathbf{y}_i | N_i = n_i, \mathbf{X}_i; \boldsymbol{\beta}) = \min\{S(y_{ij} | \mathbf{X}_i; \boldsymbol{\beta}), j = 1, \dots, n_i\}$. In addition, θ is related to Kendall's (1962) coefficient of concordance known as the Kendall's τ which can

be expressed as $\tau = \theta/(\theta + 2)$. The model implies that the association between any two measurements from the same subject is constant.

Previous literature has shown that parametric distributions such as Weibull distribution or mix of normal and Weibull distribution can only partially capture the long tail of menstrual lengths (Harlow and Zeger, 1991; Guo et al, 2006; McLain et al., 2012).

Therefore, for the marginal distribution of each Y_{ij} , we consider a class of semi-parametric linear transformation models (Cheng et al., 1995; Chen et al., 2002; Lu, 2005; Martinussen and Scheike, 2006). We specify $S(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta})$ in model (1) where Y_{ij} depends on the covariates via an unknown function $q(\cdot)$ as

$$q(Y_{ij}) = -\mathbf{X}_i^T \boldsymbol{\beta} + \varepsilon_{ij}, \quad (2)$$

and $q(\cdot)$ is a completely unspecified and strictly increasing function and ε_{ij} is a random error with a known distribution function denoted by F_ε . Let $S_\varepsilon = 1 - F_\varepsilon$ be the survivor function for ε , then the marginal survival function of Y_{ij} given $\mathbf{X}_i$ can be written as

$S(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta}) = S_\varepsilon(q(y_{ij}) + \mathbf{X}_i^T \boldsymbol{\beta})$ and the hazard function has a form of

$h(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta}) = \partial q(y_{ij}) / \partial y_{ij} \cdot h_\varepsilon(q(y_{ij}) + \mathbf{X}_i^T \boldsymbol{\beta})$. If we reparameterize the transformation model as $\Phi(y_{ij}) = \exp(q(y_{ij}))$ where $\Phi(0) = 0$ and $\lim_{y \rightarrow \infty} \Phi(y) = \infty$ (Martinussen and Scheike, 2006), we can write the hazard function for Y_{ij} given $\mathbf{X}_i$ as

$$h(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta}) = \phi(y_{ij}) \exp(\mathbf{X}_i^T \boldsymbol{\beta}) h_0(\Phi(y_{ij}) \exp(\mathbf{X}_i^T \boldsymbol{\beta})) \quad (3)$$

where $\phi(y_{ij}) = \Phi'(y_{ij})$ and $h_0(\cdot)$ is the hazard function associated with $\exp(\varepsilon)$.

Various choices for S_ε will generate different marginal models. For example, if S_ε follows the extreme value distribution as $S_\varepsilon(s) = \exp(-\exp(s))$, model (2) becomes the familiar proportional hazards model where the hazard function $h(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta}) = \phi(y_{ij}) \exp(\mathbf{X}_i^T \boldsymbol{\beta})$. If ε_{ij} has a standard logistic distribution with $S_\varepsilon(s) = 1/(1 + \exp(s))$, model (2) yields the proportional odds model that has a form of $\text{logit}(1 - S(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta})) = \text{log}(\Phi(y_{ij})) + \mathbf{X}_i^T \boldsymbol{\beta}$. A more general form for S_ε can be expressed as the class of logarithmic transformations

$S_\varepsilon(s) = [1 + r \exp(s)]^{-\frac{1}{r}}$, $r > 0$ (Chen and Yu, 2012), where $r = 0$ corresponds to the proportional hazards model and $r = 1$ yields the proportional odds model.

For the random length N_i which may capture a discrete survival outcome such as TTP, it is reasonable to assume that N_i follows a discrete distribution with a general form of

$$\Pr\{N_i = n_i | \mathbf{X}_i^T \boldsymbol{\beta}, \mathbf{Z}_i; \gamma, \boldsymbol{\alpha}, \boldsymbol{\xi}\} = \nu(\alpha(n_i) + \gamma \mathbf{X}_i^T \boldsymbol{\beta} + \mathbf{Z}_i^T \boldsymbol{\xi}), n_i = 0, 1, \dots, M, \quad (4)$$

where $\nu(\cdot)$ is a probability density function, $M > 0$ is a known positive integer, and $\nu(\alpha(n_i))$ denotes the baseline probability density. The parameter $\boldsymbol{\beta}$ is shared by the Clayton-Oakes

model (1) and the discrete model (4), and γ is a scaling parameter that evaluates the impact of shifts or changes in the distribution of $\mathbf{Y}_i$ with respect to covariates $\mathbf{X}_i$ on the distribution of N_i . The association between $\mathbf{Y}_i$ and N_i is induced by the covariates $\mathbf{X}_i$. We exclude the cases where vector $\boldsymbol{\beta} = \mathbf{0}$ with probability one, which can occur when all the components of $\boldsymbol{\beta}$ is zero. In this situation, γ cannot be identified and it is not meaningful to examine the mediation effect of $\mathbf{Y}_i$ on N_i . The parameterization also allows the two models contain different covariates. For example, the predictor vector $\mathbf{Z}_i$ is included in the discrete model but not in the Clayton-Oakes model.

Combining the Clayton-Oakes model (1) with a marginal transformation model (2) and the discrete model (4), a general form of the joint model for the multiple measurements and the random length $(\mathbf{Y}_i, N_i)$ is derived as

$$\Pr\{Y_{i1} > y_{i1}, \dots, Y_{in_i} > y_{in_i}, N_i = n_i | \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\pi}\} = \left[\sum_{j=1}^{n_i} S_{\varepsilon}(q(y_{ij}) + \mathbf{X}_i^T \boldsymbol{\beta})^{-\theta} - n_i + 1 \right]^{-\frac{1}{\theta}} \cdot \nu(\alpha(n_i) + \gamma \mathbf{X}_i^T \boldsymbol{\beta} + \mathbf{Z}_i^T \boldsymbol{\xi})$$

(5)

where $\boldsymbol{\pi} = (\theta, \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}, \boldsymbol{\xi})$ is the vector of parameters to estimate.

Our joint model (5) for MCLs and TTP is constructed in the spirit of modeling the “mediation effect” of MCL on TTP. That is, covariates affect the MCL and part of their effects on TTP is manifested through MCL. The common term in both models, $\mathbf{X}_i^T \boldsymbol{\beta}$, entails the influence of MCL on the pregnancy outcome, while the parameter γ moderates the strength of the connection between the change in distribution of MCL and TTP and is easy to interpret. For example, the situation with $\gamma = 0$ implies that there is no association between MCL and TTP. In addition, model (5) has some appealing properties (Barnhart and Sampson, 1995). First, zero random length is allowed for some subjects. In MSSWOW data, some women may get pregnant at the first menstrual cycle. In this case, no measurements of MCLs are observed, which results in a random length of zero. Second, a stochastic ordering property has been imposed by specifying the shared parameters and the scaling parameter, which can describe the association between the multiple measurements and the random length. Specifically, if the probability density function stochastically increases with the increasing of n_i , it implies that for a covariate X_{ib} , $b = 1, \dots, p$, when the scaling parameter $\gamma > 0$, the larger the value of $\beta_b X_{ib}$ is, the more likely a larger random length as well as a larger value of each component in the vector of multiple measurements would be observed, and vice versa for $\gamma < 0$. In addition, adopting the Clayton-Oakes model for the multivariate measurements of MCL allows us to obtain a neat closed form for the likelihood contribution from the observed MCL measurements. To explain, we can first formulate our MCL data in a multivariate/longitudinal scenario, where the availability of a MCL measurement is captured by $I(L_i < N_i - C_i)$. Under model (2) and the assumption that (L_i, C_i) is independent of N_i given $\mathbf{Z}_i$, the missing probability of a MCL measurement only depends on the observed data $(L_i, C_i, \mathbf{Z}_i)$; therefore the missing mechanism in our example is missing at

random (MAR). Consequently, the likelihood for the observed MCL measurements can be expressed based on the distribution of a sub-vector of $\mathbf{Y}_i$, which, according to the probabilistic property of the Clayton-Oakes model, still takes the Clayton-Oakes form. This special feature facilitates the likelihood derivation in our problem setting and can be useful in other multivariate settings with missing data.

In particular, we consider a generalized linear model for the random length. In the context of MSSWOW data, the random length N_i is TTP, which is a discrete survival time. The random length $N_i = n_i$ indicates that a subject conceives at the $(n_i + 1)$ -th cycle after having observed n_i MCLs. We denote the probability of a woman getting pregnant at the $(n_i + 1)$ -th cycle given that she had not been pregnant during the first n_i cycles as $\lambda(n_i + 1 | \mathbf{X}_i^T \boldsymbol{\beta}, \mathbf{Z}_i; \gamma, \boldsymbol{\alpha}, \boldsymbol{\xi})$ (i.e., the hazard rate of getting pregnant). Similar to previous modeling approach for TTP (Scheike and Jensen, 1997; Kalbeish and Prentice, 2002), we parameterize the hazard rate $\lambda(n_i + 1 | \mathbf{X}_i^T \boldsymbol{\beta}, \mathbf{Z}_i; \gamma, \boldsymbol{\alpha}, \boldsymbol{\xi})$ using the complementary log-log (CLL) function as

$$\lambda(n_i + 1 | \mathbf{X}_i^T \boldsymbol{\beta}, \mathbf{Z}_i; \gamma, \boldsymbol{\alpha}, \boldsymbol{\xi}) = 1 - \exp \left(-\exp(\alpha(n_i) + \gamma \mathbf{X}_i^T \boldsymbol{\beta} + \mathbf{Z}_i^T \boldsymbol{\xi}) \right), n_i = 0, 1, \dots, M, \quad (6)$$

where all the parameters are as defined as before. Then, we write the probability of a woman getting pregnant at the $(n_i + 1)$ -th cycle (i.e., $N_i = n_i$) as

$$\Pr\{N_i = n_i | \mathbf{X}_i^T \boldsymbol{\beta}, \mathbf{Z}_i; \gamma, \boldsymbol{\alpha}, \boldsymbol{\xi}\} = \lambda(n_i + 1 | \mathbf{X}_i^T \boldsymbol{\beta}, \mathbf{Z}_i; \gamma, \boldsymbol{\alpha}, \boldsymbol{\xi}) \prod_{j=1}^{n_i} \left(1 - \lambda(j | \mathbf{X}_i^T \boldsymbol{\beta}, \mathbf{Z}_i; \gamma, \boldsymbol{\alpha}, \boldsymbol{\xi}) \right).$$

3. Parameter Estimation

A key challenge of the semi-parametric estimation for the joint model (5) is the estimation of unknown baseline function $\Phi(\cdot)$ in the marginal distribution of $S_\varepsilon(q(y_{ij}) + \mathbf{X}_i^T \boldsymbol{\beta})$ in the Clayton-Oakes model. We adopt an approximate EM algorithm for parameter estimation in the spirit of the estimation methods for the Clayton-Oakes model with a marginal proportional hazards model (Glidden and Self, 1999). We start with constructing the joint likelihood by exploiting the equivalence of Clayton-Oakes model and gamma frailty model (Nielsen et al., 1992; Klein, 1992). Next we propose to use a two-level approximate EM algorithm. At the first level, the association parameter θ is fixed and an approximate EM algorithm is used to derive generalized maximum likelihood estimators (GMLEs) of other parameters in the joint model. In the E-step, the expected values of the latent frailties are calculated conditional on the observed data and the M-step involves the maximization of the full joint likelihood function and the unknown baseline function is estimated by a Breslow-type estimator. At the second level of the iteration, the profile likelihood of θ is maximized to obtain GMLE of θ . The steps are iterated until convergence is achieved. We describe the key steps below.

3.1 Likelihood Construction

The form of joint model (5) implies that the likelihood from the joint density contains two components, denoted by LF_1 and LF_2 , respectively. The first component LF_1 is the likelihood contribution from the vector of multiple measurements $\mathbf{Y}_i$ conditioning on N_i and LF_2 denotes the likelihood function for the random length N_i . The full joint likelihood function for $\boldsymbol{\pi} = (\boldsymbol{\beta}, \theta, \gamma, \boldsymbol{\alpha}, \boldsymbol{\xi})$ is given by $LR(\boldsymbol{\pi}|\mathbf{Y}, N) = LF_1(\boldsymbol{\beta}, \theta|\mathbf{Y}, N) \cdot LF_2(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}, \boldsymbol{\xi}|N)$.

First, we consider the likelihood contribution from the random length N_i . As a time-to-event variable, the random length N_i involves truncation and censoring issues. Let L_i denote the left truncation variable and C_i be the right censoring time, which is assumed to be independent of N_i . Conditioning on that $N_i > L_i$, the observed data for the i -th subject on the discrete time to event N_i consists of $(\tilde{N}_i, \delta_i)$, where $\tilde{N}_i = \min(N_i, C_i)$ and $\delta_i = I(N_i \leq C_i)$ is the censoring indicator. Assuming that a subject enters the study after time L_i , i.e., the left truncation $L_i = I_i$, the likelihood contribution from N_i taking into account the left truncation and right censoring is given by

$$LF_2(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}, \boldsymbol{\xi}|N) = \prod_{i=1}^m \frac{\prod_{j=1}^{n_i+1} \lambda(j|\mathbf{X}_i^T \boldsymbol{\beta}, \mathbf{Z}_i; \gamma, \boldsymbol{\alpha}, \boldsymbol{\xi})^{\eta_{ij}} (1 - \lambda(j|\mathbf{X}_i^T \boldsymbol{\beta}, \mathbf{Z}_i; \gamma, \boldsymbol{\alpha}, \boldsymbol{\xi}))^{1-\eta_{ij}}}{\prod_{k=1}^{l_i} (1 - \lambda(k|\mathbf{X}_i^T \boldsymbol{\beta}, \mathbf{Z}_i; \gamma, \boldsymbol{\alpha}, \boldsymbol{\xi}))} \quad (7)$$

where $\eta_{ij} = 1$ if an event occurs at the j -th time for the i -th individual, and $\eta_{ij} = 0$ otherwise.

In order to evaluate the likelihood $LF_1(\boldsymbol{\beta}, \theta|\mathbf{Y}, N)$ from the Clayton-Oakes model part, we utilize a key feature that the Clayton-Oakes model can be obtained as a gamma frailty model (Clayton, 1978; Oakes, 1982; Glidden and Self, 1999). Assuming that the observations from the i -th subject are independent conditional on a latent frailty denoted by μ_i , we write the hazard rate for Y_{ij} as

$$\lim_{s \rightarrow 0} \frac{\Pr(y_{ij} \leq Y_{ij} < y_{ij} + s | Y_{ij} \geq y_{ij}, \mu_i, \mathbf{X}_i; \boldsymbol{\beta}, \theta)}{s} = \mu_i h^*(y_{ij} | \mathbf{X}_i; \boldsymbol{\beta}) \quad (8)$$

where $h^*(y_{ij} | \mathbf{X}_i; \boldsymbol{\beta})$ is called the basic hazard functions, and μ_i has a gamma distribution with mean one and variance θ . The representation of (8) provides a gamma frailty model (Vaupel et al., 1979; Glidden and Self, 1999). It has been shown that a gamma frailty model has a joint survival function in the form of the Clayton-Oakes model if the basic hazard function is written as

$$h^*(y_{ij} | \mathbf{X}_i; \boldsymbol{\beta}) = h(y_{ij} | \mathbf{X}_i; \boldsymbol{\beta}) \exp \{ \theta H(y_{ij} - | \mathbf{X}_i; \boldsymbol{\beta}) \}$$

where $h(y_{ij} | \mathbf{X}_i; \boldsymbol{\beta})$ is the hazard function of Y_{ij} associated with the marginal distribution $S(y_{ij} | \mathbf{X}_i; \boldsymbol{\beta})$ in the Clayton-Oakes model and $H(y_{ij} | \mathbf{X}_i; \boldsymbol{\beta}) = \int_0^{y_{ij}} h(s | \mathbf{X}_i; \boldsymbol{\beta}) ds$ (Clayton, 1978; Oakes, 1982). Based on the hazard function $h(y_{ij} | \mathbf{X}_i; \boldsymbol{\beta})$ defined in (3), the basic hazard function in the gamma frailty model can be written as

$$h^*(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta}) = \phi(y_{ij}) \exp(\mathbf{X}_i^T \boldsymbol{\beta}) h_0(\Phi(y_{ij}) \exp(\mathbf{X}_i^T \boldsymbol{\beta})) \exp\{\theta H_0(\Phi(y_{ij}) \exp(\mathbf{X}_i^T \boldsymbol{\beta}))\} \quad (9)$$

where $H_0(y) = \int_0^y h_0(s) ds$.

We take equation (9) as our basic model for the intensity of the associated counting processes for Y_{ij} where $\phi(y_{ij})$ is treated as the unknown baseline function. Define $K_{ij}(y) = I(Y_{ij} \leq y)$ and $N_{ij}(y) = K_{ij}(y) - K_{ij}(y^-)$, $y \in [0, \tau]$ where τ is the upper bound for Y_{ij} . Given $L_i = l_i$ and $N_i = n_i$, the “complete data” for $\mathbf{Y}_i$ is defined as a filtration of $\mathcal{G} = \{\mu_i, N_{ij}(s), K_{ij}(s+), Y_{ij}, \mathbf{X}_i, 0 \leq s \leq y, j = l_i + 1, \dots, n_i, i = 1, \dots, m\}$ which contains the unobservable frailty μ_i for each subject. The “incomplete data” is observations of the filtration of $\mathcal{F} = \{N_{ij}(s), K_{ij}(s+), Y_{ij}, \mathbf{X}_i, 0 \leq s \leq y, j = l_i + 1, \dots, n_i, i = 1, \dots, m\}$. The likelihood of “complete data” for the Clayton-Oakes model part is

$$LF_1^{\mathcal{G}}(\boldsymbol{\beta}, \theta | \mathbf{Y}, N) = \prod_{i=1}^m \prod_{j=l_i+1}^{n_i} g(\mu_i; \theta) \exp \left(-\mu_i \int_0^\tau K_{ij}(s) dH_{ij}^*(s | \mathbf{X}_i; \boldsymbol{\beta}) \right) \int_0^\tau \mu_i K_{ij}(y) h_{ij}^*(y | \mathbf{X}_i; \boldsymbol{\beta}) dN_{ij}(y)$$

where $g(\cdot; \theta)$ is the density for a gamma random variable with mean one and variance θ and $H^*(y) = \int_0^y h^*(s) ds$. Under regularity conditions, integrating over the frailties gives the likelihood for the “incomplete data” of $\mathcal{F}$

$$LF_1^{\mathcal{F}}(\boldsymbol{\beta}, \theta | \mathbf{Y}, N) = \prod_{i=1}^m \frac{\theta^{-\theta-1} \Gamma(\theta^{-1} + N_{i\cdot}(\tau))}{\Gamma(\theta^{-1}) \left(\theta^{-1} + \sum_{j=1}^{n_i} \int_0^\tau K_{ij}(y) dH_{ij}^*(y | \mathbf{X}_i; \boldsymbol{\beta}) \right)^{\theta^{-1} + N_{i\cdot}(\tau)}} \prod_{j=l_i+1}^{n_i} \int_0^\tau K_{ij}(y) h_{ij}^*(y | \mathbf{X}_i; \boldsymbol{\beta}) dN_{ij}(y)$$

where “ $\cdot$ ” in the subscript indicates a sum over the corresponding index and Γ is the gamma function. Then the (partial) likelihood of the joint model for the “complete data” can be written as

$$\begin{aligned} LF^{\mathcal{G}}(\boldsymbol{\pi}) & \propto \prod_{i=1}^m g(\mu_i; \theta) \prod_{j=l_i+1}^{n_i} \exp \left(-\mu_i \int_0^\tau K_{ij}(y) dH_{ij}^*(y | \mathbf{X}_i; \boldsymbol{\beta}) \right) \int_0^\tau \mu_i K_{ij}(y) h_{ij}^*(y | \mathbf{X}_i; \boldsymbol{\beta}) dN_{ij}(y) \\ & \cdot \prod_{i=1}^m \frac{\prod_{j=1}^{n_i+1} (1 - \exp(-\exp(\alpha(j) + \gamma \mathbf{X}_i^T \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\xi})))^{\eta_{ij}} \exp(-\exp(\alpha(j) + \gamma \mathbf{X}_i^T \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\xi}))^{1-\eta_{ij}}}{\prod_{k=1}^{l_i} \exp(-\exp(\alpha(j) + \gamma \mathbf{X}_i^T \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\xi}))} \end{aligned} \quad (10)$$

The (partial) likelihood for the “incomplete data” is given by

$$\begin{aligned}
LF^{\mathcal{F}}(\boldsymbol{\pi}) & \propto \prod_{i=1}^m \frac{\theta^{-\theta^{-1}} \Gamma(\theta^{-1} + N_{i\cdot}(\tau))}{\Gamma(\theta^{-1}) \left(\theta^{-1} + \sum_{j=1}^{n_i} \int_0^\tau K_{ij}(y) dH_{ij}^*(y | \mathbf{X}_i; \boldsymbol{\beta}) \right)^{\theta^{-1} + N_{i\cdot}(\tau)}} \prod_{j=l_i+1}^{n_i} \int_0^\tau K_{ij}(y) h_{ij}^*(y | \mathbf{X}_i; \boldsymbol{\beta}) dN_{ij}(y) \\
& \cdot \frac{\prod_{j=1}^{n_i+1} (1 - \exp(-\exp(\alpha(j) + \gamma \mathbf{X}_i^T \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\xi})))^{n_{ij}} \exp(-\exp(\alpha(j) + \gamma \mathbf{X}_i^T \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\xi}))^{1-\eta_{ij}}}{\prod_{k=1}^{l_i} \exp(-\exp(\alpha(j) + \gamma \mathbf{X}_i^T \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\xi}))}
\end{aligned}
\tag{11}$$

Note that if the hazard function $h(y_{ij} | \mathbf{X}_i; \boldsymbol{\beta})$ is fully parametric, the likelihood of (11) can be directly maximized. However, for semi-parametric basic hazard models, maximization of (11) can be computationally prohibitive, while the maximization of (10) based on the “complete data” are feasible. Since the structure of the “complete data” likelihood (10) contains the “missing data” of frailties μ_i , an approximate EM algorithm approach is proposed.

3.2 An Approximate EM algorithm

The approximate EM algorithm involves two levels of iteration. At the first level, θ is treated as a fixed value at $\tilde{\theta}$, and an approximate EM algorithm is iterated to convergence to maximize (10) with respect to $(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}, \boldsymbol{\xi})$ which leads to the GMLEs $(\hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\xi}}) | \tilde{\theta}$. The estimated unknown baseline function $\hat{\Phi}(\cdot) | \tilde{\theta}$ is obtained by a Breslow-type estimator. Substituting these estimators in the likelihood (10) gives us the profile likelihood for θ . Repeated evaluations of this profile likelihood provide us the GMLE for θ . Therefore, the parameter estimation procedure involves repeated assessment of the profile likelihood for θ and each assessment requires iteration of an approximate EM algorithm to obtain GMLEs for $(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}, \boldsymbol{\xi}, \Phi(\cdot))$.

For E-step of the approximate EM algorithm, we take the logarithm of likelihood function (10) which results in a function of the latent frailty μ_i that still has a form of gamma distribution conditional on the observed data (Nielsen et al., 1992). Then the expectation of the latent frailties given the observed data are obtained as

$$E(\mu_i | \mathcal{F}) = \frac{1 + \theta N_{i\cdot}(y)}{1 + \theta \sum_{j=l_i+1}^{n_i} \int_0^y K_{ij}(s) dH_{ij}^*(s | \mathbf{X}_i; \boldsymbol{\beta})} \tag{12}$$

This is the key basis for the E-step where the frailties in the complete data log-likelihood are replaced by this conditional expectation. In particular, the basic cumulative hazard function $H_{ij}^*(s | \mathbf{X}_i; \boldsymbol{\beta})$ in this expectation involves with the unknown baseline function $\Phi(\cdot)$. Therefore, $\Phi(\cdot)$ is treated as nuisance parameters and needs to be estimated in the M-step as if the frailties μ_i were observed. However, the M-step for deriving the basic hazard functions is complex due to that there is no closed-form expression for $\Phi(\cdot)$. Using the similar technique

as Glidden and Self (1999), we propose to use an approximate GMLE for the unknown baseline function. Specifically, we only utilize part of the estimating equation that gives us a closed-form estimator for $\Phi(\cdot)$ and it can be shown that the estimator is still unbiased.

Assuming that θ is fixed at the value $\tilde{\theta}$ and given the initial values $(\hat{\beta}^{(0)}, \hat{\gamma}^{(0)}, \hat{\alpha}^{(0)}, \hat{\Phi}(\cdot)^{(0)})$, the $(I+1)$ -th iteration of the approximate EM algorithm has a structure as follows.

Step0: Calculation of Basic Cumulative Hazard—To obtain the expectation of the frailties, the basic cumulative hazard function needs to be calculated. Based on equation (9) and given the values of $(\hat{\beta}^{(l)}, \hat{\gamma}^{(l)}, \hat{\Phi}(\cdot)^{(l)})$, the estimator of $H_{ij}^*(s|\mathbf{X}_i; \beta)$ in the $I+1$ -th iteration can be expressed as

$$\hat{H}_{ij}^{*(l)}(y|\mathbf{X}_i; \hat{\beta}^{(l)}) = \int_0^y \exp \left\{ \tilde{\theta} \hat{H}_{0,ij}^{(l)}(\hat{\Phi}_{ij}^{(l)}(s-) \exp(\mathbf{X}_i \hat{\beta}^{(l)})) \right\} d\hat{H}_{0,ij}^{(l)}(\hat{\Phi}_{ij}^{(l)}(s) \exp(\mathbf{X}_i \hat{\beta}^{(l)}))$$

where $\hat{\Phi}(\cdot)$ is given in the M-step.

E-step: Posterior Expectation of the latent frailty μ_i —In the E-step, equation (12) is evaluated under current parameter estimates as

$$\hat{\mu}_i^{(l+1)} = \frac{1 + \tilde{\theta} N_{i \cdot}(\tau)}{1 + \tilde{\theta} \sum_{j=l_i+1}^{n_i} \int_0^\tau K_{ij}(s) d\hat{H}_{ij}^{*(l)}(s|\mathbf{X}_i; \beta)}$$

The unknown frailties in the complete data log-likelihood are replaced by those estimated $\hat{\mu}_i^{(l+1)}$'s.

M-step: Estimation of (β, γ, α) and $\Phi(\cdot)$ —The M-step is involved with maximization of the likelihood function (10) with respect to (β, γ, α) and at the same time the unknown baseline function $\Phi(\cdot)$ is estimated by a Breslow-type estimator by keeping β fixed.

M1: M-step for (β, γ, α)

$$(\hat{\beta}, \hat{\gamma}, \hat{\alpha})^{(l+1)} = \arg \max_{(\beta, \gamma, \alpha)} \cdot \prod_{i=1}^m \frac{\prod_{j=1}^{n_i+1} (1 - \exp(-\exp(\alpha(j) + \gamma \mathbf{X}_i^T \beta + \mathbf{Z}_i \xi)))^{n_{ij}} \exp(-\exp(\alpha(j) + \gamma \mathbf{X}_i^T \beta + \mathbf{Z}_i \xi))^{1-n_{ij}}}{\prod_{k=1}^{l_i} \exp(-\exp(\alpha(j) + \gamma \mathbf{X}_i^T \beta + \mathbf{Z}_i \xi))} \\ \cdot \int_0^\tau \frac{\hat{\mu}_i^{(l+1)} \exp\{\tilde{\theta} \hat{H}_{0,ij}(\hat{\Phi}_{ij}(y-) \exp(\mathbf{X}_i^T \beta))\} \hat{h}_{0,ij}(\hat{\Phi}_{ij}(y) \exp(\mathbf{X}_i^T \beta)) \exp(\mathbf{X}_i^T \beta)}{\sum_{k=1}^m \sum_{l=1}^{t_k} \hat{\mu}_k^{(l+1)} \exp\{\tilde{\theta} \hat{H}_{0,kl}(\hat{\Phi}_{kl}(y-) \exp(\mathbf{X}_k \beta))\} \hat{h}_{0,kl}(\hat{\Phi}_{kl}(y) \exp(\mathbf{X}_k \beta)) \exp(\mathbf{X}_k \beta) K_{kl}(y)} dN_{ij}(y)$$

M2: Approximate M-step for $\hat{\Phi}(\cdot)$ given $(\hat{\beta}, \hat{\gamma}, \hat{\alpha}, \hat{\mu})^{(l+1)}$.

The unknown baseline function $\Phi(\cdot)$ can be obtained by solving the following step function inductively with starting value of $\Phi(0) = 0$.

$$\hat{\Phi}^{(l+1)}(y) = \int_0^y \left[\sum_{i=1}^m \sum_{j=1}^{n_i} \hat{\mu}_i^{(l+1)} R \left(\hat{\Phi}^{(l+1)}(s-), \hat{\beta}^{(l+1)} | \tilde{\theta}, \mathbf{X}_i \right) K_{ij}(s) \right]^{-1} dN_{..}(s)$$

where

$$\begin{aligned} & R \left(\hat{\Phi}^{(l+1)}(s-), \hat{\beta}^{(l+1)} | \tilde{\theta}, \mathbf{X}_i \right) \\ &= \exp \{ \tilde{\theta} \hat{H}_{0,ij}(\hat{\Phi}^{(l+1)}(s-)) \exp(\mathbf{X}_i \hat{\beta}^{(l+1)}) \} \hat{h}_{0,ij}(\hat{\Phi}^{(l+1)}(s-)) \exp(\mathbf{X}_i \hat{\beta}^{(l+1)}) \\ &= \exp \{ \tilde{\theta} \hat{H}_{0,ij}(\hat{\Phi}^{(l+1)}(s-)) \exp(\mathbf{X}_i \hat{\beta}^{(l+1)}) \} \exp(\mathbf{X}_i \hat{\beta}^{(l+1)}). \end{aligned}$$

Note that the solution for $\hat{\Phi}(\cdot)$ in this M-step is an approximation. In the E-step, the expectation is estimated by replacing $dH_{ij}^*(\cdot)$ with its estimator $d\hat{H}_{ij}^*(\cdot)$ and by fixing θ at $\tilde{\theta}$. For the M-step, we first obtain the GMLEs for the parameters (β, γ, α) by maximizing the likelihood of (10) where the frailties μ_i 's are substituted by their estimates from the E-step and $H_{ij}^*(\cdot)$ and $dH_{ij}^*(\cdot)$ are replaced with $\hat{H}_{ij}^*(\cdot)$ and $d\hat{H}_{ij}^*(\cdot)$, respectively. Using the estimators for (β, γ, α) , the estimator for the baseline function $\Phi(\cdot)$ is also updated. Therefore, the M-step is involved with maximization of likelihood function of (10) with respect to (β, γ, α) as well as the baseline function $\Phi(\cdot)$.

Step0, E-step, M-step1 and M-step2 are repeated until convergence is obtained. The values of β and $\Phi(\cdot)$ at convergence, denoted by $\hat{\beta}|\tilde{\theta}$ and $\hat{\Phi}(\cdot)|\tilde{\theta}$, are used to maximize the profile likelihood of θ using one dimensional optimization method. Due to the complexity of the estimators $(\hat{\beta}, \hat{\gamma}, \hat{\alpha}, \hat{\Phi}(\cdot))$, we propose to use a bootstrap procedure to derive the estimates of variance of the estimators as well as the confidence intervals. Specifically, bootstrapped samples are drawn randomly with replacement from the observed data $(\mathbf{Y}_i, \tilde{N}_i, \mathbf{X}_i)$, $i = 1, \dots, n$ and the bootstrap sampling is based on subjects. Then we apply the estimation approach for each bootstrapped sample and obtain the parameter estimates. And the variance estimate is calculated as the empirical variance of the 500 bootstrap parameter estimates.

With regard to joint models, a computationally simpler two-stage estimation procedure is commonly studied in literature (Spiekerman and Lin, 1998; Glidden, 2000; Chen et al., 2002; Lu, 2005; Chen and Yu, 2012). For the joint model (5), in the first stage, estimators of the marginal parameters $(\beta, \gamma, \alpha, \xi)$ are derived under the independence working assumption where the repeated measurements are independent (i.e., $\theta = 0$). In the second stage, a pseudo likelihood for the dependence parameter θ is constructed by replacing the marginal parameters in the full likelihood by the estimators obtained in the first stage. The asymptotic theory of the two-stage estimator for our model follows from the work of Spiekerman and Lin (1998) and Glidden (2000). Although two-stage estimation methods are computationally simpler in copula models, they are known to be less efficient particularly when the association is very strong (Joe, 2005).

4. Simulation Studies

To evaluate the performance of the joint modeling procedure, we conducted simulation studies in different settings. One thousand replicates were performed for each set. Each simulation sample consisted of m subjects ($m = 200$ and 400). First, random lengths were generated as a time-to-event process from the complementary log-log model with a constant baseline hazard $\lambda(j|\mathbf{X}_i^T\boldsymbol{\beta};\gamma) = 1 - \exp(-\exp(\alpha_j + \gamma(\beta_1 X_{1i} + \beta_2 X_{2i})))$ where α_j represents the baseline hazard rate associated with j and X_1 and X_2 represent dummy variables of a categorical variable with three levels. In addition, N_j 's were subject to independent left truncation and right censoring. The left truncation time was generated from a discrete uniform distribution $U[0, 4]$ and censoring time was simulated from a complementary log-log model with intercept only. Note that if an event occurred before entering the study (i.e., N_j is less than truncation time), then this subject would be excluded.

The vector of multiple measures $\mathbf{Y}_j$ for each subject was generated from the Clayton-Oakes model with a marginal transformation model, and the length of vector $\mathbf{Y}_j$ was equal to the time-to-event N_j . For the marginal model, we considered the class of logarithmic

transformations of the form of $S(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta}) = [1 + r \exp(q(y_{ij}) + \beta_1 X_{1i} + \beta_2 X_{2i})]^{-\frac{1}{r}}$ with $r = 0$ corresponding to the marginal proportional hazards model and $r = 1$ corresponding to the marginal proportional odds model (Dabrowska and Doksum, 1988; Chen and Yu, 2012).

The parameters were chosen to be similar to those estimated in the MSSWOW study presented in Section 5. Specifically, the following values were set for the vector of parameters $(\beta_1, \beta_2, \gamma, \alpha) = (0.3, 0.5, 4, -4)$. Approximately, these values provided the mean of the observed random vector length equal to 5 with 10% of left truncation and 65% of censoring, and the mean of multiple measurements equal to 24. The dependence parameter θ was set to be 0.5 and 3.0, corresponding to Kendall's tau of 0.2 and 0.6, respectively. Each data set was analyzed using the joint modeling procedure and the approximate EM algorithm described in Section 2 and 3. Biases for β_1, β_2, γ and θ were evaluated from the simulations. Bootstrapping standard errors of the parameters, average simulation standard errors, and 95% coverage probabilities were also calculated. The simulation results are summarized in Table 1.

The results show that the biases for estimating β_1, β_2 and γ are small, particularly when sample size is large. The largest biases (between 2% and 4%) are seen in the estimators of β_1, β_2 and γ in the scenario where the sample size is small with $m = 200$ and at the same time the dependence parameter is large for the multiple measurements with $\theta = 3.0$. There exists a slight negative bias for estimation of θ , which becomes smaller with decreasing of association and increasing of sample size. This negative bias has been observed previously with maximizing the profile likelihood (Nielsen et al., 1992; Glidden and Self, 1999). The bootstrap standard errors agree well with the Monte Carlo simulation standard errors. When dependence parameter decreases and sample size increases, the difference between the two standard errors becomes negligible. The 95% coverage probability for the regression parameters (β_1, β_2) maintains near the nominal level in all cases. Most of 95% confidence intervals for γ have reasonable coverage probabilities except for a few cases when $r = 0$ and

$\theta = 3.0$ (coverage probability for γ is 93%). The coverage probability for θ is slightly lower than 95% in the case where $\theta = 3.0$ (coverage probability for θ is 93%). A similar phenomenon has been reported in previous literature for transformation model (Chen and Yu, 2012). In addition, as sample size increases, the bias and standard errors of all the parameters decrease. Overall, the simulation study illustrates that the generalized maximum likelihood estimators based on the approximate EM algorithm perform reasonably well even when within-subject correlation is relatively high and sample size is moderate.

5. Application to MSSWOW Study

MSSWOW was a prospective cohort study conducted from 1991 to 1994. Women who were between the ages of 19 and 40 and at risk for pregnancy (sexually active and not consistently using birth control) were eligible for the study and a total of 470 women were finally enrolled and the participants were followed with menstrual diaries and urine samples for up to one year until a clinical pregnancy or the end of the study. Women kept a daily record of menstrual bleeding and unprotected intercourse during follow-up and collected urine samples which were tested for pregnancy. Time-to-pregnancy (TTP) was recorded as the number of cycles for a woman taken to conceive (not including the conception cycle) from her starting point. Some subjects had already been trying to get pregnant when the study began, and the time that a subject had been trying until the study entry is also recorded (i.e., TTP were left truncated). Therefore, subjects were only included in the study if their TTP was greater than time from the first attempt to the start of the study. Of all participants, 61.3% of them had left truncations with an average of 5.9 cycles. By the end of the study, 179 (38.1%) of the participants got pregnant. Menstrual cycle lengths (MCLs) were calculated from the first day of menstrual bleeding until the day before the next onset of menses. A total of 3689 MCLs were recorded, with a median of 28 days (interquartile range, 25–31 days; range, 1–189 days). Of those MCLs, 6.2% were missing and 8.9% were censored.

This study has been originally designed to investigate the effect of Video Display Terminal on spontaneous abortion (Marcus, 1990), but it provided the opportunity to explore the role of age on reproductive health outcomes of MCLs and TTP. In previous work, menstrual cycle characteristics, including cycle length and bleeding length, were found to be associated with a woman's likelihood of becoming pregnant as well as pregnancy outcome (Small et al., 2006). Guo et al. (2006) demonstrated the inadequacy of normal distribution for MCLs and showed that MCLs were distributed differently among various sub-populations of different age groups, but the impact of age on TTP was not investigated. Similar to previous work (Guo et al., 2006), we categorize age into four groups (19–25, 26–30, 31–35 and 36–41). Our goal is to assess the influence of aging effect on MCLs and then evaluate the effect of the potential changes in MCLs due to age on TTP.

Table 2 presents the results from the MSSWOW data analysis including the parameter estimates and their bootstrapped standard errors based on the approximate EM algorithm (column 3 and 4). For a comparison, we also present estimators based on the two-stage estimation procedure (column 6 and 7). The parameter estimates based on those two different methods are very similar to each other. By comparing the standard errors, it can be

seen that the estimates based on the approximate EM algorithm have smaller standard errors than those from the two-stage approach, which implies that there is an efficiency loss of the two-stage estimators although the computational burden is reduced. Based on the results from the EM algorithm, with different marginal distributions for menstrual lengths (proportional hazards model and proportional odds model etc.), the results indicate that there exists a significant age effect on the distribution of MCLs. Specifically, the first three groups are not significantly different in terms of the distributions of MCLs and they are significantly different from the older age group 36–41. The size of the effects gets larger as women get younger in reference to the oldest age group (36–41). The estimated association parameter $\hat{\theta}$ (Kendall's tau $\hat{\tau} = 0.136, 0.152, \text{ and } 0.232$ in different models, respectively) was found to be significantly greater than zero, indicating a modest correlation was observed among MCLs from the same woman. The estimated scaling parameter $\hat{\gamma}$ is negative and is significantly different from zero in all three marginal models ($r=0, 0.5, 1$), implying the effect of MCLs is significantly associated with TTP through underlying aging effects. In other words, the risk of getting pregnant stochastically decreases with the increasing of the MCLs. This means that with the increasing of a woman's age, we would observe longer MCLs and lower probability of getting pregnant. Adjusting for the impact of MCLs, women between 19 and 25 have the highest chance of getting pregnant, followed by those between 26 and 30 and those between 31 and 35. The women in age group 36 to 41 have the lowest pregnancy rate. The status of unprotected sex is found to have significant influence on the higher risk of getting pregnant. (Although all eligible women reported having unprotected sex in the previous three months at entry to the study, not all women had unprotected sex during each menstrual cycle.)

To evaluate the fitted models, we compare the Akaike Information Criterion (AIC) across the models, which shows that the joint model based on the Clayton-Oakes model with a marginal proportional odds model (AIC=23477.50) is preferable to the other two models (AIC=24484.71 for proportional hazards model and AIC=23850.31 for transformation model with $r = 0.5$).

In particular, one can expect that the impact of age on MCLs reduces over time when women reach a certain age. For the proportional odds model, the covariate effects are specified as a multiplicative factor on the baseline odds function (Bennett, 1983), which indicates that the difference in hazards by covariates diminishes over time. That is, the covariate effects diminish with time. This property of the proportional odds model seems intuitively reasonable compared to proportional hazards model for interpreting the age effects on menstrual lengths. Furthermore, we plotted the estimated survival function based on joint models and compared with Kaplan-Meier (KM) estimator (see Figure 2). It is shown that the estimated survival functions based on the proportional odds model assumption are closer to KM curves compared to those based on proportional hazards models. This also demonstrates the appropriateness of the proportional odds model for MCLs, while other plots (not presented here) also agree with this conclusion.

6. Discussion

In this paper, we propose a joint modeling framework for the analysis of multivariate random length data (Barnhart and Sampson, 1995), where multiple measurements are modeled via a Clayton-Oakes model and the random length is modeled via a generalized model. Particularly, we specify a flexible class of semi-parametric transformation models for the marginal models in the Clayton-Oakes part of the joint model. Consistent with MSSWOW data, we model the random length as a discrete survival outcome, via a complementary log-log link, although the method is generally applicable to any discrete distribution for the random length. Under the joint modeling assumptions, an approximate EM algorithm based on a gamma frailty model is developed to derive the semi-parametric inferences for the parameters.

Our proposed method provide a flexible modeling framework to analyze two related outcomes jointly. First, our joint model can appropriately handle missing outcomes and censoring and truncation issues of the random vector length. Second, we use semi-parametric transformation models for MCLs which include the commonly used proportional hazards model and the proportional odds model as special cases. Third, the estimation procedure has been developed to obtain generalized maximum likelihood estimators of all the model parameters including the regression coefficients in the joint model, dependence parameter and the unknown baseline function in the transformation model. Existing estimation methods primarily focus on the Clayton-Oakes model with a Cox proportional hazard model as the marginal model (Glidden and Self, 1999). To our knowledge, no work has been done to perform statistical inference using the approximate EM algorithm where marginal distributions in the Clayton-Oakes model are specified as transformation models. AIC is used to assess the global goodness-of-fit for the joint model, however, more work is needed in determining the goodness-of-fit of these models.

MSSWOW study was conducted in a prospective manner and hence has several advantages with regards to the analysis of TTP. Retrospectively collected data may be biased by the outcome, which is known by the woman when reporting TTP (Scheike and Keiding, 2006). In addition, it is more difficult to recall accurately whether unprotected intercourse occurred during each of the menstrual cycles. Because approximately 50% of pregnancies in the United States are unintended, the MSSWOW population is more likely to represent the population at risk of pregnancy than a retrospective study where women who may have terminated a pregnancy are excluded. For prospectively collected TTP data, TTP is observed as a waiting time that can be any positive number. Therefore, conditional on initiation time and covariates, TTP can be modeled via the conditional distribution as $\Pr\{N = n | N \leq n\}$. That is, the hazard of TTP given covariates can be directly observed. For retrospective study, we only observe the sample from a certain period $[0, T]$ instead of the entire timeline. In this case, TTP is both right and left-truncated conditional on initiation time. TTP from retrospective sample has a conditional distribution of $\Pr\{N = n | 0 \leq N \leq T\}$. Extensive details about the analysis of TTP using prospective and retrospective methods can be found in Scheike and Keiding's paper (2006). Our joint model proposed here is designed for the analysis of prospectively collected TTP data.

Acknowledgments

This research is partly supported by NIH grant R01HL113548.

References

- Barnhart HX, Kosinski AS, Sampson AR. A Regression Model for Multivariate Random Length Data. *Statistics in Medicine*. 1999; 18:199–211. [PubMed: 10028140]
- Barnhart HX, Sampson AR. Multiple Population Models for Multivariate Random Length Data-With Applications in Clinical Trials. *Biometrics*. 1995; 51:195–204. [PubMed: 7766774]
- Bennett S. Analysis of Survival Data by the Proportional Odds Model. *Statistics in Medicine*. 1983; 2:273–277. [PubMed: 6648142]
- Louis, GBM., Platt, RW., editors. Reproductive and perinatal epidemiology. New York: Oxford University Press; 2011. 2011
- Cai J, Prentice RL. Estimating Equations for Hazard Ratio Parameters Based on Correlated Failure Time Data. *Biometrika*. 1995; 82(1):151–164.
- Chen K, Jin Z, Ying Z. Semiparametric Analysis of Transformation Models with Censored Data. *Biometrika*. 2002; 89:659668.
- Chen CM, Yu CY. A Two-Stage Estimation in the Clayton-Oakes Model with Marginal Linear Transformation Models for Multivariate Failure Time Data. *Lifetime Data Analysis*. 2012; 18:94–115. [PubMed: 21983914]
- Cheng SC, Wei LJ, Ying Z. Analysis of transformation models with censored data. *Biometrika*. 1995; 82(4):835–845.
- Clayton DG. A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. *Biometrika*. 1978; 65(1):141–151.
- Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society*. 1972; 34(2): 187–220.
- Dunson DB, Chen Z, Harry J. A Bayesian Approach for Joint Modeling of Cluster Size and Subunit-Specific Outcomes. *Biometrics*. 2003; 59:521–530. [PubMed: 14601753]
- Efron, B., Tibshirani, RJ. An Introduction to the Bootstrap. New York. Z: Chapman and Hall; 1993.
- Gill RD. Non- and Semi-Parametric Maximum Likelihood and the von Mises Method (Part 1). *Scandinavian Journal of Statistics*. 1989; 16:97–128.
- Ghosh D, Lin DY. Semi-parametric Analysis of Recurrent Events Data in the Presence of Dependent Censoring. *Biometrics*. 2003; 59:877–885. [PubMed: 14969466]
- Glidden DV, Self SG. Semi-parametric Likelihood Estimation in the Clayton- Oakes Failure Time Model. *Scandinavian Journal of Statistics*. 1999; 26(3):363–372.
- Guo Y, Manatunga AK, Chen S, Marcus M. Modeling Menstrual Cycle Length Using a Mixture Distribution. *Biostatistics*. 2006; 7(1):100–114. [PubMed: 16020617]
- Harlow SD, Lin X, Ho MJ. Analysis of menstrual diary data across the reproductive life span Applicability of the bipartite model approach and the importance of within-woman variance. *Journal of Clinical Epidemiology*. 2000; 53(7):722–733. [PubMed: 10941950]
- Harlow SD, Matanoski GM. The Association between Weight, Physical Activity, and Stress and Variation in the Length of the Menstrual Cycle. *American Journal of Epidemiology*. 1991; 133(1): 38–49. [PubMed: 1983897]
- Harlow SD, Zeger SL. An Application of Longitudinal Methods to the Analysis of Menstrual Diary Data. *Journal of Clinical Epidemiology*. 1991; 44(10):1015–1025. [PubMed: 1940994]
- Hogan JW, Laird NM. Mixture Models For The Joint Distribution Of Repeated Measures And Event Times. *Statistics in Medicine*. 1997; 16:239–257. [PubMed: 9004395]
- Huang X, Elliott MR, Harlow SD. Modelling Menstrual Cycle Length and Variability at the Approach of Menopause by Using Hierarchical Change Point Models. *Journal of Royal Statistical Society: Series C (Applied Statistics)*. 2014; 63(3):445–466.
- Joe H. Asymptotic Efficiency of the Two-Stage Estimation Method for Copula-Based Models. *Journal of Multivariate Analysis*. 2005; 94:401–419.

- Kalbeisch, JD., Prentice, RL. The Statistical Analysis of Failure Time Data. 2nd. Wiley; 2002.
- Keiding N, Kvist K, Hartvig H, Tvede M. Estimating Time to Pregnancy from Current Durations in a Cross-Sectional Sample. *Biometrics*. 2002; 3(4):565–578.
- Kendall, MG. Rank correlation methods. 3rd. London: Griffin & Company; 1962.
- Klein JP. Semi-Parametric Estimation of Random Effects Using the Cox Model Based on the EM Algorithm. *Biometrics*. 1992; 48:795–806. [PubMed: 1420842]
- Laird NM, Ware JH. Random-Effects Models for Longitudinal Data. *Biometrics*. 1982; 38:963–974. [PubMed: 7168798]
- Liang KY, Zeger SL. Inference Based on Estimating Functions in the Presence of Nuisance Parameters. *Statistical Science*. 1995; 10:158–199.
- Lin H, McCulloch CE, Mayne ST. Maximum Likelihood Estimation in the Joint Analysis of Time-to-Event and Multiple Longitudinal Variables. *Statistics in Medicine*. 2002; 21:2369–2382. [PubMed: 12210621]
- Lin X, Raz J, Harlow SD. Linear Mixed Models with Heterogeneous within-Cluster Variances. *Biometrics*. 1997; 53(3):910–923. [PubMed: 9290222]
- Lu W. Marginal Regression of Multivariate Event Times Based on Linear Transformation Models. *Lifetime Data Analysis*. 2005; 11:389404.
- Lum KJ, Sundaram R, Louis TA. Accounting for Length-Bias and Selection Effects in Estimating the Distribution of Menstrual Cycle Length. *Biostatistics*. 2015; 16(1):113–128. [PubMed: 25027273]
- Marcus M. Epidemiologic Studies of VDT Use And Pregnancy Outcome. *Reproductive Toxicology*. 1990; 4:51–56. [PubMed: 2136021]
- Martinussen, T., Scheike, TH. Dynamic regression models for survival data. 1st. Springer; 2006.
- McLain AC, Lum KJ, Sundaram R. A Joint Mixed Effects Dispersion Model for Menstrual Cycle Length and Time-to-Pregnancy. *Biometrics*. 2012; 68:648–656. [PubMed: 22321128]
- Murphy SA, Bentley GR, O'Hanesian MA. An Analysis for Menstrual Data with Time-Varying Covariates. *Statistics in Medicine*. 1995; 14:1843–1857. [PubMed: 8532979]
- Nielsen GG, Gill RD, Andersen PK, Sorensen TIA. A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models. *Scandinavian Journal of Statistics*. 1992; 19:25–43.
- Oakes D. A Model for Association in Bivariate Survival Data. *Journal of the Royal Statistical Society*. 1982; 44:414–422.
- Scheike TH, Jensen TK. A Discrete Survival Model with Random Effects: An Application to Time to Pregnancy. *Biometrics*. 1997; 53(421):318–329. [PubMed: 9147597]
- Scheike TH, Keiding N. Design and Analysis of Time-to-Pregnancy. *Statistical Methods in Medical Research*. 2006; 15:127–140. [PubMed: 16615653]
- Small CM, Manatunga AK, Klein M, Feigelson HS, Dominguez CE, McChesney R, et al. Menstrual Cycle Characteristics: Associations with Fertility and Spontaneous Abortion. *Epidemiology*. 2006; 17:52–60. [PubMed: 16357595]
- Vaupel JW, Manton KG, Stallard E. The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality. *Demography*. 1979; 16:439–454. [PubMed: 510638]
- Yang S, Prentice RL. Semi-Parametric Inference in the Proportional Odds Regression Model. *Journal of the American Statistical Association*. 1999; 94:125–136.
- Zeng D, Cai J. Asymptotic Results for Maximum Likelihood Estimators in Joint Analysis of Repeated Measurements and Survival Time. *The Annals of Statistics*. 2005; 33(5):2132–2163.
- Zeng D, Lin DY. Efficient Estimation for the Accelerated Failure Time Model. *Journal of the American Statistical Association*. 2007b; 102:13871396.

Figure 1.
Distribution of Menstrual Cycle Lengths

Figure 2.
Plot of Estimated Survival Function vs. Cycle Lengths

Table 1
Simulation Results for Estimating the Parameters Using the Approximate EM Algorithm with 1000 Replicates.⁷

		$\theta = 0.5$					$\theta = 3.0$				
Scenarios	Params(True)	Bias	SD ²	SE ³	95% Cp ⁴	Bias	SD	SE	95% CP		
$m = 200$											
r=0	$\beta_1(0.3)$	0.0061	0.1395	0.1405	0.96	0.0081	0.1953	0.1948	0.93		
	$\beta_2(0.5)$	-0.0041	0.1342	0.1432	0.95	0.0082	0.1986	0.1985	0.94		
	$\gamma(4.0)$	0.0502	0.8538	0.8921	0.94	0.0633	0.6374	0.6387	0.94		
	$\theta(0.5/3.0)$	-0.0162	0.0650	0.0687	0.94	-0.0855	0.2272	0.2489	0.93		
r=0.5	$\beta_1(0.3)$	0.0103	0.0781	0.0784	0.95	0.0115	0.1036	0.1022	0.95		
	$\beta_2(0.5)$	0.0114	0.0871	0.0885	0.96	0.0126	0.1102	0.1125	0.94		
	$\gamma(4.0)$	-0.0607	0.6850	0.6901	0.94	-0.0779	0.6834	0.6938	0.95		
	$\theta(0.5/3.0)$	-0.0063	0.0383	0.0377	0.95	-0.0813	0.1957	0.1987	0.93		
r=1	$\beta_1(0.3)$	-0.0075	0.0747	0.0759	0.96	0.0112	0.1016	0.1039	0.93		
	$\beta_2(0.5)$	-0.0104	0.0736	0.0724	0.96	0.0135	0.1026	0.1034	0.94		
	$\gamma(4.0)$	0.0808	0.8348	0.8535	0.93	0.0738	0.8039	0.8066	0.93		
	$\theta(0.5/3.0)$	-0.0030	0.0473	0.0439	0.95	-0.0935	0.2556	0.2531	0.95		
$m = 400$											
r=0	$\beta_1(0.3)$	0.0019	0.0954	0.0949	0.94	0.0014	0.1015	0.1056	0.95		
	$\beta_2(0.5)$	0.0021	0.0984	0.0958	0.95	0.0028	0.1032	0.1035	0.95		
	$\gamma(4.0)$	0.0287	0.5581	0.5360	0.93	0.0399	0.5307	0.5320	0.94		
	$\theta(0.5/3.0)$	-0.0125	0.0553	0.0557	0.95	-0.0886	0.1926	0.1922	0.93		
r=0.5	$\beta_1(0.3)$	0.0027	0.0717	0.0711	0.96	0.0058	0.0583	0.0616	0.94		
	$\beta_2(0.5)$	0.0053	0.0826	0.0839	0.96	0.0063	0.0764	0.0741	0.93		
	$\gamma(4.0)$	-0.0446	0.6119	0.6204	0.95	-0.0579	0.6431	0.6449	0.94		
	$\theta(0.5/3.0)$	-0.0061	0.0281	0.0252	0.95	-0.1003	0.1596	0.1553	0.93		
r=1	$\beta_1(0.3)$	-0.0018	0.0641	0.0659	0.96	0.0024	0.0936	0.0912	0.93		
	$\beta_2(0.5)$	0.0025	0.0715	0.0727	0.95	0.0031	0.1004	0.0997	0.95		
	$\gamma(4.0)$	-0.0606	0.4231	0.4506	0.96	-0.0746	0.5119	0.5145	0.94		

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

$\theta = 3.0$									
$\theta = 0.5$									
Scenarios	Params(True)	Bias	SD ²	SE ³	95% CP ⁴	Bias	SD	SE	95% CP
$\theta(0.5/3.0)$									
		-0.0049	0.0404	0.0403	0.96	-0.0939	0.1907	0.2005	0.95

¹ Truncation percentage is 10% and censoring percentage is about 65%.

² The standard error is based on the bootstrap sampling.

³ This is the Monte Carlo standard error based on the simulations.

⁴ CP stands for the coverage probability.

Table 2

Model Results for MSSWOW Data (m=470)

Model	Effects ¹	Estimate ²	SE ²	P-value	Estimate ³	SE ³
<i>r</i> = 0						
Joint Models	Age group 19–25	β_{11}	-0.386	0.040	<.001	-0.365 0.052
	Age group 26–30	β_{12}	-0.349	0.034	<.001	-0.347 0.042
	Age group 31–35	β_{13}	-0.244	0.041	0.011	-0.243 0.051
	Age group 36–41	-	-	-	-	-
	Scaling parameter	γ	-1.414	0.282	<.001	-1.435 0.361
Association among cycle length	Unsafe Sex	β_2	0.079	0.011	<.001	0.079 0.011
	Association	θ	0.315	0.015	<.001	0.338 0.034
	Kendall's tau	τ	0.136	0.006	<.001	0.145 0.012
<i>r</i> = 0.5						
Joint Models	Age group 19–25	β_{11}	-0.483	0.067	<.001	-0.479 0.084
	Age group 26–30	β_{12}	-0.302	0.051	<.001	-0.307 0.066
	Age group 31–35	β_{13}	-0.194	0.066	0.003	-0.203 0.077
	Age group 36–41	-	-	-	-	-
	Scaling parameter	γ	-1.033	0.249	0.036	-1.024 0.336
Association among MCL	Unsafe Sex	β_2	0.079	0.010	<.001	0.079 0.011
	Association	θ	0.359	0.013	<.001	0.351 0.032
	Kendall's tau	τ	0.152	0.005	<.001	0.149 0.012
<i>r</i> = 1						
Joint Models	Age group 19–25	β_{11}	-0.979	0.112	<.001	-0.990 0.135
	Age group 26–30	β_{12}	-0.674	0.088	<.001	-0.678 0.109
	Age group 31–35	β_{13}	-0.377	0.086	<.001	-0.430 0.094
	Age group 36–41	-	-	-	-	-
	Scaling parameter	γ	-0.496	0.220	0.024	-0.483 0.297
Association among	Unsafe Sex	β_2	0.079	0.010	<.001	0.079 0.011
	Association	θ	0.623	0.055	<.001	0.618 0.119

Model	Effects ¹	Estimate ²	SE ²	P-value	Estimate ³	SE ³
MCL	Kendall's tau	0.238	0.016	<.001	0.239	0.035

¹The estimates for parameters associated with baseline hazard of TTP are given as 1) for model $r=0$, 2: $\alpha_1 = -3.440(0.107)$, $3 \sim 8$: $\alpha_2 = -3.568(0.091)$, $=9$: $\alpha_3 = -4.798(0.425)$, $10 \sim 12$: $\alpha_4 = -3.195(0.102)$, and 13: $\alpha_5 = -3.948(0.094)$; 2) for model $r=0.5$, 2: $\alpha_1 = -3.610(0.161)$, $3 \sim 8$: $\alpha_2 = -3.733(0.158)$, $=9$: $\alpha_3 = -4.955(0.268)$, $10 \sim 12$: $\alpha_4 = -3.370(0.155)$, and 13: $\alpha_5 = -4.104(0.130)$; and 3) for model $r=1$, 2: $\alpha_1 = -3.305(0.548)$, $3 \sim 8$: $\alpha_2 = -3.435(0.539)$, $=9$: $\alpha_3 = -4.674(0.730)$, $10 \sim 12$: $\alpha_4 = -3.065(0.538)$, and 13: $\alpha_5 = -3.835(0.494)$.

²The parameter estimates and standard errors are based on the approximate EM algorithm.

³The parameter estimates and standard errors are obtained using two-stage estimation procedure.