Joint modeling of longitudinal functional feature and time to event: an application to fecundity studies

Ling Ma
Department of Mathematical Sciences, Clemson University, Clemson, South Carolina, U.S.A.

Abhisek Saha
Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, DHHS, Maryland, U.S.A.

Animik Biswas
Department of Mathematics and Statistics, University of Maryland Baltimore County, Maryland, U.S.A.

Rajeshwari Sundaram †
Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, DHHS, Maryland, U.S.A.

E-mail: sundaramr2@mail.nih.gov

Summary. In many longitudinal studies, it is often of interest to investigate how the geometric functional features (such as the curvature, location and height of a peak), of a marker’s measurement process is associated with the time to event being studied. We propose a joint model for certain geometric functional features of a longitudinal process and a time to event, making use of B-splines to smoothly approximate the infinite dimensional functional data. The proposed approach allows for prediction of survival probabilities for future subjects based on their available longitudinal measurements. We illustrate the performance of our proposed model on a prospective pregnancy study, namely Stress and Time to

†Address for correspondence: Rajeshwari Sundaram, Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, DHHS, 6710B Rockledge Blvd Room 3232, Bethesda, Maryland, U.S.A. Email: sundaramr2@mail.nih.gov
Pregnancy, a component of Oxford Conception Study, where hormonal measurements of luteinizing hormone (LH) and estrogen indicate timing of ovulation, and whether ovulation is going to occur, in a menstrual cycle. A joint modeling approach was used to assess whether the functional features of the hormonal measurements, such as the peak of the hormonal profile and its curvature, are associated with time to pregnancy. Our simulation studies indicate reasonable performance of the proposed approach.

Keywords: Joint modeling; TTP; Functional Data; Hormonal profile; Longitudinal data; Curvature; Peak

1. Introduction

In biomedical studies, information is routinely collected longitudinally on various biomarkers up to a time-to-event (usually censored), along with additional covariates. An often cited example of such data is from HIV clinical trial with the key longitudinal process of interest being CD4 counts. Many other examples from cancer studies, reproductive health etc. have been the motivation for the development of various methods in this area of research. In joint modeling, one is typically interested in (i) how to model the pattern of change of the longitudinal process, and (ii) to characterize the relationship between the survival event, the longitudinal process, and the covariates.

Most of the literature on joint modeling of longitudinal process and time to event have focused on modeling the mean of the longitudinal process, where the dependence between the longitudinal and time to event process is through the mean process with subject-specific random effect(s) (Wu and Carroll, 1988; Tsiatis and Davidian, 2004). Recently, some authors have also considered modeling the dispersions of the longitudinal process and time to event (McLain et al., 2012). In many situations (e.g. the motivating example discussed below), it may be more appropriate to view the longitudinal process as a functional data (Li and Luo, 2017). However, most of these existing literature focused on the situations where the longitudinal process and the time-to-event are on the same time scale, while there are examples where this assumption may not be true. For instance, in our motivating example, the longitudinal process is measured in a daily scale while the time-to-pregnancy is measured in menstrual cycles. In this paper, our
Joint modeling of longitudinal functional feature

The main objective is to jointly model certain geometric features of the observed longitudinal process and a time-to-event, so that the geometric features and the time-to-event are on the same time scale, with a view towards studying the effect of the said geometric features on the time-to-event of interest.

There is a large literature on functional data analysis, particularly in the context of dynamic modeling and smoothing (e.g. Cao and Ramsay, 2009; Ramsay et al., 2007; Cao et al., 2011; see also Ramsay and Silverman, 2010; and Ferraty and Vieu, 2006; and the references there in). There is also some literature focusing on functional principal component analysis and functional linear regression for longitudinal data (e.g. Yao et al., 2005a, b; Hall et al., 2006; and the references there in). However, joint modeling of geometric features of functional data and time-to-event, considered in this paper, appears to be novel.

We present our motivating example from reproductive epidemiology. Reproductive hormones, like the luteinizing hormone (LH), estrogen and its metabolites (e.g., estrone-3-glucuronide etc) patterns play an important role in the study of conception, infertility and other chronic disease (Mumford et al., 2012; Parazzini et al., 1993; Terry et al., 2005; Whelan et al., 1994; Solomon et al., 2001, 2002). However, data on these endogenous hormones is difficult to quantify due to complex cyclical patterns of hormones, the need for timed collection, and the cost required for multiple sample collections. Consequently, menstrual cycle characteristics such as cycle length are often used as proxies for cumulative hormonal exposure and/or hormonal patterns as they can be easily assessed in population studies. Short and long or irregular cycles have been associated with increased risks for breast cancers, osteoporosis, type 2 diabetes mellitus, cardiovascular diseases etc. Consequently, it is important to be able to study the hormonal profiles directly. However, LH has important role in the luteinizing of the follicle and the functional maturation of the nucleus of the oocyte (Verpoest et al., 2000). Abnormalities in the LH surge may impair the development of the oocyte and consequently its fertilization ability. Additionally, LH surge abnormalities, such as reduced peak values, have been associated with infertility (Cohlen et al., 1993; Cahill et al., 1997), indicating that shape of the hormonal curve plays an important biological role as well. Moreover, the
pattern of the LH profile is highly variable in normal menstruating women. Motivated by these issues, we are interested in modeling the LH surge and its relationship with time-to-pregnancy. A source of such data arises in the prospective pregnancy studies, where couples are followed from the time they go off contraception (with the intention of becoming pregnant) until they get pregnant and collect information on these hormones daily. This provides an opportunity to study the patterns of these hormones and their relationship with time-to-pregnancy. We focus on three functional features, namely, the value of LH peak, the curvature at the LH peak and the average curvature of the LH profile within fertile window from the longitudinally measured LH values.

Motivated by the example described above, we consider joint modeling of the longitudinal process (“hormonal profile”) and a discrete survival time (“time-to-pregnancy”), where one is interested in modeling various geometric functional features of the longitudinal process (e.g., value at the peak, curvature at the peak, average curvature within fertile window of a cycle) and the relationship between these characteristics and the time to event. In the next section, we introduce our data and the modeling framework. In Section 3, we provide the estimation approach for the parameters of interest and also give out the prediction approach for the time-to-event distribution of a new subject given its longitudinal measurement history. We assess the performance of the proposed estimates through simulation studies in Section 4. In Section 5, we present analysis of the Stress and Time to pregnancy, a sub-component of the Oxford Conception Study (McLain et al., 2015).

2. Model and Notation

For subject $i$, $i = 1, \ldots, n$, denote $T_i$ as the (discrete) time to pregnancy, i.e., the number of menstrual cycles it took the $i$th woman to get pregnant. As in many time to event studies, $T_i$ is assumed to be subject to right censoring and one observes $X_i = \min(T_i, \tau_i)$ and $\delta_i = I(T_i \leq \tau_i)$, where $\tau_i$ denotes the censoring time and $I(\cdot)$ denotes the indicator function. Throughout this article, we assume that the censoring time $\tau_i$ is independent of $T_i$.

Denote the hormonal profile in a cycle as $h(t)$, $t \in [1, \tau]$ where $\tau$ is the cycle length in
days and \( \hat{t} \) is the time point where the hormonal profile reaches the highest level (peak). Also define the fertile window in a cycle as \([\hat{t} - 5, \hat{t} + 1]\). The three functional features of hormonal profile that are of interest here are: (1) curvature at the hormone peak, 
\[
k(\hat{t}) = |h''(\hat{t})|\{1 + h'(\hat{t})\}^{-3/2},
\] which measures the sharpness of the hormonal profile at the peak time; (2) hormone peak value, \( h(\hat{t}) \); (3) the average curvature of the hormone profile within fertile window, 
\[
\sum_{t=\hat{t}-5}^{\hat{t}+1} k(t)/6; \text{ within each cycle.}
\]

Let \( \tilde{Y}_{ij} \) be the true functional feature of hormonal profile of interest for subject \( i \) in menstrual cycle \( j \) and let \( Z_{ij} \) be the vector of covariates for \( \tilde{Y}_{ij} \). Hormone levels are measured on daily basis in each cycle for each subject. The observed hormonal profiles and functional features for each cycle are then calculated using some smoothing technique, e.g. B-splines. Denote by \( Y_{ij} \) as the observed functional feature of hormonal profile for subject \( i \) in menstrual cycle \( j \).

We relate the true functional feature \( \tilde{Y}_{ij} \) to the covariates \( Z_{ij} \) through a linear model with subject-specific random intercept \( b_{Y,i} \),
\[
\tilde{Y}_{ij} = Z_{ij}' \beta + b_{Y,i}.
\]

The observed hormonal functional features \( Y_{ij} \) are then modeled by
\[
Y_{ij} = \tilde{Y}_{ij} + \epsilon_{ij},
\]
where \( \epsilon_{ij} \) are all independent and identically distributed and follow \( N(0, \sigma^2) \) distribution.

We use the discrete survival model by Sundaram et al. (2012) for TTP where the hazard of discrete survival time is related linearly to the covariates when transformed by a complementary log-log function. It also accounts for the fact that, the hazard for conception in a cycle is zero if the couple does not have any intercourse in the fertile window of that cycle. The model is given by:
\[
\lambda_{i}(j|b_{T,i}, U_{ij}) = 1 - \exp \left[ -A_{ij} \exp\{b_{T,i} + \rho_j + U_{ij}' \gamma\} \right] \quad (2)
\]
where for subject \( i \) in cycle \( j \), \( U_{ij} \) is the vector of covariates for TTP (which could have overlap with \( Z_{ij} \), the covariates for the functional features), \( A_{ij} \) is the indicator of intercourse within fertile window, \( \rho_j \) is the baseline effect for cycle \( j \) and \( b_{T,i} \) is a subject-specific random effect. Note that the fertile window refers to the days in a menstrual.
cycle around the day of ovulation when a couple having intercourse can potentially conceive; $A_{ij} = 0$ means that couple $i$ did not have intercourse during the fertile window of cycle $j$, which implies that there is no risk of pregnancy in that cycle.

To study the association between a woman’s hormonal profile and TTP, we take into account the cycle level functional feature of a woman’s hormonal profile, $\tilde{Y}_{ij}$, in (2). Recalling that $\tilde{Y}_{ij} = Z_{ij}'\beta + b_{Y,i}$, we propose the following model

$$
\lambda_i(j|b_{T,i}, b_{Y,i}, U_{ij}) = 1 - \exp\left\{ - A_{ij} \exp\{b_{T,i} + \rho_j + U_{ij}^i\gamma + \psi_j\tilde{Y}_{ij}\} \right\}, (3)
$$

where $\psi_j$ is the regression coefficient of $\tilde{Y}_{ij}$, and assume that $i \equiv (b_{Y,i}, b_{T,i}) \sim \text{MVN}(0, D)$ with

$$
D = \begin{pmatrix}
\sigma_1^2 & \zeta \sigma_1 \sigma_2 \\
\zeta \sigma_1 \sigma_2 & \sigma_2^2
\end{pmatrix},
$$

and $i$’s are independent and identically distributed (iid) and independent of $\epsilon_{ij}$. The association between the hormonal profile and TTP is taken into account not only through the fact that the cycle level functional feature is included in the model for TTP as a predictor, but also through the possible correlation between subject specific random effects $b_{Y,i}$ and $b_{T,i}$.

### 2.1. Estimation and Prediction

Denote the observed data for subject $i$ as $O_i = (X_i, \delta_i, Y_{ij}, Z_{ij}, U_{ij})$. The observed data log likelihood can be written as

$$
l(\theta) = \sum_{i=1}^{n} \log \left\{ \int f_i(X_i|\delta) S_i(X_i|\delta)^{1-\delta_i} \prod_{j=1}^{X_i} f_{Y_{ij}}(Y_{ij}|f_b) d\right\}, (4)
$$

where $\theta = (\beta, \gamma, \rho, \psi, \sigma_1^2, \sigma_2^2, \zeta)' = (b_Y, b_T)$,

$$
S_i(j) = \exp\left\{ - \sum_{k=1}^{j} A_{ik} \exp(b_T + \rho_k + U_{ik}'\gamma + \psi_k(Z_{ik}'\beta + b_Y)) \right\},
$$

$$
f_i(j) = S_i(j-1) - S_i(j),$$

and $f_{Y_{ij}}(Y_{ij}|)$ is the density function of normal distribution with mean $(Z_{ij}'\beta + b_Y)$ and variance $\sigma^2$. 
One natural way of finding estimates for $\theta$ is to maximize the log likelihood function with respect to $\theta$. However, the two-dimensional integration with respect to the random effects does not have a closed form. We propose to use Gaussian quadrature for approximation. Specifically, let $\tilde{Z} = ZR$ where $R$ is the Cholesky square root of the covariance matrix $D$ (e.g. $D = R' R$) and $\tilde{Z}$ is a two-dimensional row vector of independent standard normal variables. Let $\{ (\tilde{Z}_k, w_k), k = 1, \ldots, K \}$ be the $K$ Gaussian nodes and weights for a standard normal variable, then the $K^2$ nodes of may be constructed by

$$k, s = (R_{11} \tilde{Z}_k + R_{21} \tilde{Z}_s, R_{12} \tilde{Z}_k + R_{22} \tilde{Z}_s), k, s = 1, \ldots, K,$$

where $R_{ks}$ is the $(k, s)$th element of $R$, and the associated weight is calculated by $w_k w_s$.

Then the likelihood contribution of the $i$th subject could be approximated by

$$\int f_i(X_i) \delta_s(X_i) [1 - \delta_i] \prod_{j=1}^{X_i} f_{Y_{ij}}(Y_{ij}) f_b() d \approx \sum_{k=1}^{K} \sum_{s=1}^{K} f_i(X_{i|k,s}) S_i(X_{i|k,s}) [1 - \delta_i] \prod_{j=1}^{X_i} f_{Y_{ij}}(Y_{ij|k,s}) w_k w_s.$$

We then maximize the approximated log likelihood function to get estimate of $\theta$, denoted by $\hat{\theta}$. The covariance matrix $\Sigma$ of $\theta$ is estimated by using the observed information matrix.

One nice feature of using the joint modeling approach is that we could use hormonal profile characteristics to predict time to pregnancy distribution. Based on a joint model fitted on a sample of size $n$, we are interested in predicting the time-to-event distribution for a new subject $i$ that has provided a set of longitudinal measurements up to cycle $j_0$. Denote $\hat{\Sigma} = \text{var}(\hat{\theta})$ as the estimated variance covariance matrix for $\theta$. The partial information for the new subject $i$ is denoted by $D_i(j_0) = \{Y_i(j_0), U_i(j_0), Z_i(j_0)\}$, where $Y_i(j_0) = \{Y_{ij}, j \leq j_0\}$, $U_i(j_0) = \{U_{ij}, j \leq j_0\}$ and $Z_i(j_0) = \{Z_{ij}, j \leq j_0\}$. Prediction of the conditional probability of surviving cycle $j$ is of interest only if the couple have not achieved pregnancy at cycle $j_0$. Hence we focus on the conditional probability of surviving cycle $j$ given survival up to cycle $j_0$

$$\pi_i(j | j_0) \equiv \Pr(T_i \geq j | T_i > j_0, D_i(j_0), D_n)$$

$$= \int \Pr(T_i \geq j | T_i > j_0, D_i(j_0), D_n, \theta) p(\theta | D_n) d\theta,$$

(5)
where \( D_n \) denotes the sample on which the joint model was fitted and on which the predictions will be based. The first part of the integrand can be written as

\[
\Pr(T_i \geq j \mid T_i > j_0, D_i(j_0), D_n, \theta)
= \int \Pr(T_i \geq j \mid T_i > j_0, D_i(j_0), \cdot, \theta)p(i \mid T_i > j_0, D_i(j_0), \theta)d_i
= \int \Pr(T_i \geq j \mid T_i > j_0, \cdot, \theta)p(i \mid T_i > j_0, D_i(j_0), \theta)d_i
= \int S_i(j \mid i, \theta)p(i \mid T_i > j_0, D_i(j_0), \theta)d_i.
\]

(6)

The second part is the posterior distribution of the parameters given the observed data. By using arguments of standard asymptotic Bayesian theory and assuming that the sample size \( n \) is sufficiently large, we approximate the distribution of \( \{\theta \mid D_n\} \) by \( N(\hat{\theta}, \hat{\Sigma}) \).

Given \( D_i(j_0) \) and \( \theta \), the posterior distribution of \( i \) is

\[
f_i(i \mid T_i > j_0, D_i(j_0), \theta) \propto p(T_i > j_0, i, D_i(j_0) \mid \theta)
= S_i(j_0 \mid i, \theta) \frac{\prod_{t=1}^{j_0} f_Y(Y_{it} \mid i, \cdot, \theta)}{S_i(j_0 \mid i, \theta)}f_i(i) d_i.
\]

(7)

This posterior distribution of the random effects is of nonstandard form.

Following Rizopoulos (2011), we make use of a Metropolis-Hastings algorithm with independent proposals from a multivariate \( t \) distribution centered at the empirical Bayes estimates \( \hat{i} = \arg \max \{\log p(T_i > j_0, i, D_i(j_0) \mid \hat{\theta})\} \), with scale matrix

\[
\text{vár}(\hat{i}) = \{-\partial^2 \log p(T_i > j_0, i, D_i(j_0) \mid \hat{\theta}) / \partial \theta \mid \hat{i}\},
\]

and four degrees of freedom.

A Monte Carlo sample of \( \pi_i(j \mid j_0) \) can be obtained using the following simulation scheme:

• Step 1. Draw \( \theta^{(l)} \sim N(\hat{\theta}, \hat{\Sigma}) \).

• Step 2. Draw \( i^{(l)} \sim t_4 \{\hat{i}, \text{vár}(\hat{i})\} \).

• Step 3. Compute \( \pi_i^{(l)}(j \mid j_0) = S_i\{j \mid \hat{i}^{(l)}, \cdot, \theta^{(l)}\} / S_i\{j_0 \mid \hat{i}^{(l)}, \theta^{(l)}\} \).

Repeat Steps 1–3 for \( l = 1, \ldots, L \) times, where \( L \) denotes the Monte Carlo sample size. In our prediction analysis, we used \( L = 500 \) samples to estimate the mean and 95\% quantile based confidence intervals.
3. Simulation Studies

In this section, we conducted simulation studies to investigate the performance of the proposed estimates using likelihood-based approach.

For simplicity, we assumed covariates $U_{ij} = U_i$ and generated $U_i$ from normal distribution with mean 2 and variance 1. Let $Z_{ij} = Z_i = (1, U_i)$. $i = (b_{Y,i}, b_{T,i})$ were generated from multivariate normal distribution with mean 0 and variance covariance matrix

$$D = \begin{pmatrix} \sigma_1^2 & \zeta \sigma_1 \sigma_2 \\ \zeta \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}.$$  

The random error $\epsilon_{ij}$ were generated from $N(0, 0.9^2)$ and $Y_{ij}$ were generated using equation (1).

The cycle-level intercourse indicator $A_{ij}$ were generated from binomial distribution with success probability 0.92. That is, there about 8% of the cycles without intercourse during the fertile window. Subjects who had not experienced an event at $j = 6$ were censored.

Table 3 presents the estimation results using the proposed likelihood based approach. The results were based on samples of size $n = 250$ and $n = 500$ with 1000 replicates. Estimation used Gaussian quadrature with 50 nodes and the simulation was conducted in software R. We present in Table 3 the true value (True) of the parameter, the estimation bias (Bias) which equals to the average of the difference between the estimated parameter value and the true value, the standard deviation of the estimates (SD), and the 95% coverage probability.

The simulation results show that the proposed estimation approach works well in the situation considered here with reasonably small bias and coverage probabilities close to the nominal level.

4. Analysis of the Oxford Conception Study

We illustrate our proposed method by analyzing the Oxford Conception Study, Stress and Time-to-Pregnancy component [Pyper et al. 2006], which is one of a few prospective cohort studies with preconception enrollment of women aged 18-40 years who were
Table 1. Estimation results of the simulation studies including estimation bias (Bias) of the proposed estimates, the standard deviation of the estimates (SD) and the 95% coverage probability (CP) for sample sizes of 250 and 500.

| Parameter | True | Bias | SD | CP | Bias | SD | CP |
|-----------|------|------|----|----|------|----|----|
| $\beta_1$ | 4    | 0.011| 0.119 | 0.947 | 0.008 | 0.089 | 0.947 |
| $\beta_2$ | $-0.5$ | $-0.005$ | 0.048 | 0.942 | $-0.004$ | 0.035 | 0.945 |
| $\gamma_1$ | $-27$ | $-0.152$ | 1.900 | 0.963 | $-0.082$ | 1.406 | 0.961 |
| $\rho_1$ | $-6.5$ | 0.375 | 2.343 | 0.964 | 0.416 | 1.863 | 0.945 |
| $\rho_2$ | 10 | 0.396 | 1.956 | 0.969 | 0.455 | 1.470 | 0.947 |
| $\rho_3$ | 17 | 0.453 | 1.964 | 0.947 | 0.444 | 1.526 | 0.901 |
| $\rho_4$ | 21 | 0.489 | 2.058 | 0.948 | 0.486 | 1.452 | 0.934 |
| $\rho_5$ | 24 | 0.566 | 2.055 | 0.960 | 0.542 | 1.420 | 0.937 |
| $\rho_6$ | 25 | 0.449 | 2.081 | 0.969 | 0.456 | 1.452 | 0.942 |
| $\psi_\mu$ | 18 | $-0.002$ | 1.771 | 0.962 | $-0.087$ | 1.354 | 0.955 |
| $\sigma_1$ | 0.3 | 0.008 | 0.067 | 0.907 | 0.006 | 0.050 | 0.927 |
| $\sigma_2$ | 3 | 0.698 | 2.214 | 0.943 | 0.383 | 1.412 | 0.951 |
| $\sigma$ | 0.9 | $-0.003$ | 0.028 | 0.949 | 0.000 | 0.020 | 0.954 |
| $\zeta$ | $-0.2$ | $-0.004$ | 0.432 | 0.919 | 0.003 | 0.323 | 0.925 |
attempting to become pregnant. The women in the Oxford Conception Study (hereafter OCS) provided daily level information on reproductive hormones, in particular, luteinizing hormone, intercourse acts, and host of other lifestyle variables, along with couple level baseline covariates. The luteinizing hormone was observed through the fertility monitors used by the woman to track her daily fertility level, the values provided by the monitor were monotonically transformed values of the luteinizing hormones. These women were prospectively followed for the number of menstrual cycles it took them to become pregnant (i.e., human-chorionic gonadotropin confirmed pregnancy on the day of expected menses), known as their TTP, or a maximum of 6 menstrual cycles. Other examples of prospective pregnancy cohort studies’ include the Longitudinal Investigation of Fertility and Environment (Louis et al., 2011), Fertili (Colombo and Masarotto, 2000), and Billings (Colombo et al., 2006).

The data consisted of 337 women with Luteinizing hormone measurements, resulting in a total of 1023 menstrual cycles. We randomly selected 225 (about two thirds) of the women with 686 cycles in the training set and the rest of the women were taken to be the prediction set. The hormonal measurements were taken daily, with possible missing values on some days. For each cycle of each woman, we used B-spline functions to smooth the hormone data. We considered three functional features of the hormone profiles: curvature at the Luteinizing hormone (LH) peak which measures the sharpness of the LH profile at peak, LH peak value and the average LH profile curvature within fertile window. Figure 1 gives an example of the LH measurements for 4 randomly selected subject in the data set. Due to the fact that the three functional features considered here are highly correlated, instead of putting all of them simultaneously in one joint model, we model TTP with each one of the functional features using the training dataset and evaluate their prediction abilities, respectively. In our analysis, we focus on the following covariates: female age, couple average age, female minus male age difference, female’s body mass index (BMI) (categorized as underweight or normal weight if BMI<25; overweight if 25 ≤BMI<30; obese if 30 ≤BMI)), female’s smoking category (smoke.n if not smoking; smoke.m if smoking but average cigarette smoked per day ≤10; smoke.h if average cigarette smoked per day >10), stress level measured by
the salivary biomarker alpha amylase and parity (nulliparous or multiparous; nulliparous women are those who haven’t had live birth before). The three age related covariates as well as alpha amylase level were scaled by suitable constants in the analysis and were denoted as Age*, avg.age*, dif.age* and Alp*, respectively. The functional features of hormonal profiles were also scaled by adequate constants in the analysis. Subjects with BMI smaller than 25 who are non-smokers and nulliparous were considered as the reference group.

Table 4 presents the estimates and 95% confidence intervals for the covariate effects on the curvature at LH peak, covariate effects on TTP, cycle-specific baseline effect on TTP, effect of curvature at LH peak on TTP as well as the variance covariance parameters for the random effects. Notice that all covariates seem non-significant for the curvature at LH peak. A couple’s average age and female-male age difference seem to have significant negative effect on TTP. Women who smoke but smoke no more than 10 cigarettes per day have significantly longer TTP compared to non-smoking women while the difference between the heavy smoking women (> 10 cigarettes per day) and non-smoking women was not significant. Multiparous women seem to have significantly shorter TTP than nulliparous women. We also found that the curvature at LH peak has significant positive effect on TTP. This implies that the sharper a woman’s LH peak is, the shorter her TTP tends to be.

We also present the estimation results for jointly modeling LH peak value and TTP in Table 4. The conclusions of covariate effects were generally consistent with that in Table 4 except that obesity seems to have protective effect on TTP and heavy smoking women also seem to have significantly longer TTP than non-smoking women.

The estimation results for joint modeling of average curvature of LH profile within fertile window and TTP are shown in Table 4. The overall trend of covariate effects is consistent with that in Table 4 for curvature at LH peak, except that overweight women seem to have significant longer TTP than underweight/normal weight women, women with higher stress level (alpha amylase) seem to have longer TTP, while the effect of smoking on TTP does not seem to be significant.

Now that we have fitted the joint models using the training dataset, we are interested
Fig. 1. LH measurements profile curves by days for 4 randomly selected subjects in the data set.
Table 2. Joint model estimation results on curvature at LH peak.

| Parameter | Estimate (95%CI) | Parameter | Estimate (95%CI) |
|-----------|------------------|-----------|------------------|
| \( \beta \): Age* | -0.51 (-2.73, 1.72) | \( \gamma \): avg.age* | -27.89 (-42.76, -13.02) |
| \( \beta \): Age*² | 0.09 (-0.62, 0.80) | \( \gamma \): dif.age* | -5.17 (-8.30, -2.03) |
| \( \beta \): Overweight | 0.04 (-0.17, 0.25) | \( \gamma \): Overweight | -2.99 (-7.49, 1.51) |
| \( \beta \): Obese | -0.19 (-0.45, 0.07) | \( \gamma \): Obese | 5.66 (-0.20, 11.51) |
| \( \beta \): Smoke.m | 0.02 (-0.21, 0.25) | \( \gamma \): Smoke.m | -7.59 (-13.45, -1.72) |
| \( \beta \): Smoke.h | 0.05 (-0.46, 0.55) | \( \gamma \): Smoke.h | -5.23 (-15.15, 4.69) |
| \( \beta \): Alp* | -0.15 (-0.47, 0.18) | \( \gamma \): Alp* | -3.50 (-9.82, 2.81) |
| \( \beta \): Alp*² | 0.02 (-0.07, 0.12) | \( \gamma \): Alp*² | 1.80 (-0.16, 3.76) |
| \( \beta \): Parity | -0.08 (-0.27, 0.10) | \( \gamma \): Parity | 9.11 (3.10, 15.13) |
| \( \rho_1 \) | -6.49 (-16.49, 3.50) | \( \rho_2 \) | 10.01 (1.77, 12.25) |
| \( \rho_3 \) | 17.71 (14.96, 20.47) | \( \rho_4 \) | 21.03 (16.41, 25.65) |
| \( \rho_5 \) | 24.36 (18.25, 30.47) | \( \rho_6 \) | 24.96 (18.06, 31.86) |
| \( \rho_7 \) | -2.94 (-2.94, -2.94) | \( \psi_\mu \) | 18.63 (3.93, 33.33) |
| \( \sigma_1 \) | 0.35 (0.25, 0.50) | \( \sigma_2 \) | 30.44 (18.23, 50.83) |
| \( \sigma \) | 0.96 (0.90, 1.02) | \( \zeta \) | -0.20 (-0.31, -0.08) |
Table 3. Joint model estimation results on LH peak value.

| Parameter | Estimate (95%CI) | Parameter | Estimate (95%CI) |
|-----------|------------------|-----------|------------------|
| $\beta$: Age* | $-0.03 \ (-1.45, 1.38)$ | $\gamma$: avg.age* | $-25.41 \ (-39.73, -11.09)$ |
| $\beta$: Age$^{*2}$ | $-0.01 \ (-0.47, 0.46)$ | $\gamma$: dif.age* | $-5.31 \ (-8.68, -1.93)$ |
| $\beta$: Overweight | $-0.05 \ (-0.17, 0.06)$ | $\gamma$: Overweight | $0.12 \ (-2.81, 3.05)$ |
| $\beta$: Obese | $-0.25 \ (-0.39, -0.12)$ | $\gamma$: Obese | $7.89 \ (2.47, 13.31)$ |
| $\beta$: Smoke.m | $0.06 \ (-0.06, 0.18)$ | $\gamma$: Smoke.m | $-7.82 \ (-12.94, -2.69)$ |
| $\beta$: Smoke.h | $0.21 \ (-0.06, 0.48)$ | $\gamma$: Smoke.h | $-8.86 \ (-17.29, -0.44)$ |
| $\beta$: Alp* | $-0.18 \ (-0.35, -0.01)$ | $\gamma$: Alp* | $2.50 \ (-2.26, 7.27)$ |
| $\beta$: Alp$^{*2}$ | $0.04 \ (-0.01, 0.09)$ | $\gamma$: Alp$^{*2}$ | $-0.41 \ (-1.75, 0.92)$ |
| $\beta$: Parity | $0.02 \ (-0.07, 0.12)$ | $\gamma$: Parity | $6.07 \ (1.98, 10.16)$ |
| $\rho_1$ | $-11.50 \ (-18.68, -4.33)$ | $\rho_2$ | $1.34 \ (-0.42, 3.10)$ |
| $\rho_3$ | $7.97 \ (4.62, 11.32)$ | $\rho_4$ | $10.35 \ (5.73, 14.97)$ |
| $\rho_5$ | $13.48 \ (7.06, 19.89)$ | $\rho_6$ | $13.81 \ (6.97, 20.65)$ |
| $\rho_7$ | $-5.05 \ (-5.05, -5.05)$ | $\psi_{\mu}$ | $24.63 \ (9.67, 39.59)$ |
| $\sigma_1$ | $0.24 \ (0.20, 0.29)$ | $\sigma_2$ | $25.30 \ (15.01, 42.65)$ |
| $\sigma$ | $0.43 \ (0.40, 0.45)$ | $\zeta$ | $-0.22 \ (-0.29, -0.14)$ |
Table 4. Joint model estimation results on average curvature of LH profile within fertile window.

| Parameter | Estimate (95%CI) | Parameter | Estimate (95%CI) |
|-----------|-----------------|-----------|-----------------|
| $\beta$: Age* | $-0.27 (-2.32, 1.78)$ | $\gamma$: avg.age* | $-23.27 (-38.97, -7.58)$ |
| $\beta$: Age*<sup>2</sup> | $0.10 (-0.57, 0.77)$ | $\gamma$: dif.age* | $-8.61 (-14.08, -3.13)$ |
| $\beta$: Overweight | $-0.07 (-0.24, 0.09)$ | $\gamma$: Overweight | $-8.04 (-13.80, -2.28)$ |
| $\beta$: Obese | $-0.08 (-0.28, 0.12)$ | $\gamma$: Obese | $2.76 (-2.05, 7.57)$ |
| $\beta$: Smoke.m | $0.06 (-0.12, 0.24)$ | $\gamma$: Smoke.m | $-3.98 (-8.49, 0.53)$ |
| $\beta$: Smoke.h | $0.07 (-0.34, 0.49)$ | $\gamma$: Smoke.h | $2.93 (-7.44, 13.30)$ |
| $\beta$: Alp* | $-0.18 (-0.44, 0.08)$ | $\gamma$: Alp* | $-14.97 (-25.34, -4.59)$ |
| $\beta$: Alp*<sup>2</sup> | $0.04 (-0.04, 0.12)$ | $\gamma$: Alp*<sup>2</sup> | $5.45 (1.89, 9.01)$ |
| $\beta$: Parity | $0.02 (-0.12, 0.17)$ | $\gamma$: Parity | $20.65 (9.02, 32.29)$ |
| $\rho_1$ | $-27.04 (-44.87, -9.21)$ | $\rho_2$ | $-5.70 (-12.00, 0.60)$ |
| $\rho_3$ | $8.28 (5.64, 10.92)$ | $\rho_4$ | $13.38 (8.07, 18.69)$ |
| $\rho_5$ | $22.83 (12.21, 33.45)$ | $\rho_6$ | $23.56 (11.93, 35.19)$ |
| $\rho_7$ | $-0.02 (-0.02, -0.02)$ | $\psi_\mu$ | $18.53 (4.67, 32.38)$ |
| $\sigma_1$ | $0.36 (0.30, 0.44)$ | $\sigma_2$ | $43.56 (24.77, 76.59)$ |
| $\sigma$ | $0.60 (0.56, 0.64)$ | $\zeta$ | $-0.17 (-0.24, -0.10)$ |
Joint modeling of longitudinal functional feature

in predicting the probability of subfertility (i.e., TTP > 6 cycles) given survival past one cycle. That is, for all women in the prediction set with \( T_i > 1 \), we wish to classify \( I(T_i > 6) \) by the conditional survival probability \( \pi_i(6|1) \). To measure the classification rate, we empirically estimated the sensitivity: \( P(\hat{\pi}_i(6|1) > c|T_i > 6) \), and specificity \( P(\hat{\pi}_i(6|1) \leq c|T_i \leq 6) \), where \( \hat{\pi}_i(6|1) \) is the mean of the Monte Carlo sample \( \{\pi_i^{(l)}, l = 1, \ldots, L\} \) obtained following the steps in Section 2.1. Of the 79 women available for prediction, 20 were censored between 1 and 6 cycles and removed from the prediction analysis. The classification measures using ROC for the models with curvature at LH peak, LH peak value and average curvature of LH profile within fertile window, are displayed in Figure 2(a),(b),(c), respectively, for all \( c \in [0, 1] \). The AUC is 0.650 for the model which includes the curvature at LH peak, 0.646 for the model which includes the LH peak value and 0.614 for the model which includes the average curvature of LH profile within fertile window. This indicates that the prediction ability of the three models are moderately good.

As mentioned in Section 2.1 one could use the fitted joint model and hormonal measurement history up to cycle 1, to predict the conditional survival probability \( \pi_i(j|1) \), for any \( j > 1 \) if that is of interest.

The data used in the above analysis are available from the corresponding author upon reasonable request.

5. Discussion

We have proposed a joint modeling approach to assess the association between functional features of a woman’s hormonal profile and her fecundity measured by TTP. A likelihood based approach with the use of Gauss quadrature approximation was proposed for estimation of the unknown parameters. Simulation studies have demonstrated that the proposed estimation approach works reasonable well in the situation similar to the Oxford data, with reasonably small bias and coverage probabilities close to the nominal level. With the estimates from the joint models, we also derived the approach to predict individual characteristics of TTP given a set of longitudinal measurements up to a certain cycle. The prediction of the probability that a woman was subinfertile
Fig. 2. ROC curves for classifying $I(T_i > 6)$ by $\hat{\pi}_i(6|1)$, for models with: (a) curvature at LH peak; (b) LH peak value; (c) average LH curvature within fertile window.
given her past one menstrual cycle behavior without getting pregnant, was moderately accurate for model with any one of the three functional features of LH profile, especially with curvature at hormonal profile peak (AUC around 0.65).

The analysis of Oxford data found that couple average age and female minus male age difference are significantly associated with TTP in the sense that older couples and couples with larger female minus male age difference have significantly longer TTP. The association between BMI and TTP seem to be marginal with moderate evidence that overweight women have a slower rate of pregnancy while obese women have a faster rate of pregnancy. Multiparous women were found to have significantly shorter TTP than nulliparous women. Furthermore, we found that women with sharper LH peaks, higher LH peaks and overall more curved LH profiles within fertile window tend to have significantly shorter TTP.

We are now jointly modeling TTP and one functional feature at a time. Further investigation is needed to address the issue with multiple correlated functional features and identify their correlation with TTP jointly. Finally, our approach though motivated by reproductive epidemiology may be appliable to examples arising in other disciplines. For example, one may be interested in understanding the association between the high blood pressure and risk for heart attack and not just on their average blood pressure measurements. In conclusion, this paper studies a novel model for joint models of longitudinal and survival data where one is interested in the longitudinally varying functional features with survival data.

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health, Eunice Kennedy Shriver National Institute of Child Health and Human Development. The authors would like to acknowledge that this study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD [http://biowulf.nih.gov]. The research of Animikh Biswas was supported in part by the NSF grant DMS 1517027. Conflict of Interest: None declared.
Cahill, D., Wardle, P., Maile, L., Harlow, C. and Hull, M. (1997) Ovarian dysfunction in endometriosis-associated and unexplained infertility. *Journal of assisted reproduction and genetics, 14*, 554–557.

Cao, J. and Ramsay, J. O. (2009) Generalized profiling estimation for global and adaptive penalized spline smoothing. *Computational Statistics & Data Analysis, 53*, 2550–2562.

Cao, J., Wang, L. and Xu, J. (2011) Robust estimation for ordinary differential equation models. *Biometrics, 67*, 1305–1313.

Cohlen, B. J., te Velde, E. R., Scheffer, G., van Kooij, R. J., de Brouwer, C. P. M. and van Zonneveld, P. (1993) The pattern of the luteinizing hormone surge in spontaneous cycles is related to the probability of conception. *Fertility and sterility, 60*, 413–417.

Colombo, B. and Masarotto, G. (2000) Daily fecundability: first results from a new database. *Demographic research, 3*.

Colombo, B., Mion, A., Passarin, K. and Scarpa, B. (2006) Cervical mucus symptom and daily fecundability: first results from a new database. *Statistical Methods in Medical Research, 15*, 161–180.

Ferraty, F. and Vieu, P. (2006) *Nonparametric functional data analysis: theory and practice*. Springer Series in Statistics.

Hall, P., Müller, H.-G. and Wang, J.-L. (2006) Properties of principal component methods for functional and longitudinal data analysis. *The annals of statistics*, 1493–1517.

Li, K. and Luo, S. (2017) Functional joint model for longitudinal and time-to-event data: an application to alzheimer’s disease. *Statistics in medicine, 36*, 3560–3572.

Louis, G. M. B., Rios, L. I., McLain, A., Cooney, M. A., Kostyniak, P. J. and Sundaram, R. (2011) Persistent organochlorine pollutants and menstrual cycle characteristics. *Chemosphere, 85*, 1742–1748.

McLain, A. C., Lum, K. J. and Sundaram, R. (2012) A joint mixed effects dispersion model for menstrual cycle length and time-to-pregnancy. *Biometrics, 68*, 648–656.
McLain, A. C., Sundaram, R., Louis, B. and Germaine, M. (2015) Joint analysis of longitudinal and survival data measured on nested timescales by using shared parameter models: an application to fecundity data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64, 339–357.

Mumford, S. L., Steiner, A. Z., Pollack, A. Z., Perkins, N. J., Filiberto, A. C., Albert, P. S., Mattison, D. R., Wactawski-Wende, J. and Schisterman, E. F. (2012) The utility of menstrual cycle length as an indicator of cumulative hormonal exposure. *The Journal of Clinical Endocrinology & Metabolism*, 97, E1871–E1879.

Parazzini, F., La Vecchia, C., Negri, E., Franceschi, S. and Tozzi, L. (1993) Lifelong menstrual pattern and risk of breast cancer. *Oncology*, 50, 222–225.

Pyper, C., Bromhall, L., Dummett, S., Altman, D. G., Brownbill, P. and Murphy, M. (2006) The oxford conception study design and recruitment experience. *Paediatric and perinatal epidemiology*, 20, 51–59.

Ramsay, J. O., Hooker, G., Campbell, D. and Cao, J. (2007) Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 741–796.

Ramsay, J. O. and Silverman, B. W. (2010) *Functional data analysis*. Springer Series in Statistics.

Rizopoulos, D. (2011) Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67, 819–829.

Solomon, C. G., Hu, F. B., Dunaif, A., Rich-Edwards, J., Willett, W. C., Hunter, D. J., Colditz, G. A., Speizer, F. E. and Manson, J. E. (2001) Long or highly irregular menstrual cycles as a marker for risk of type 2 diabetes mellitus. *Jama*, 286, 2421–2426.

Solomon, C. G., Hu, F. B., Dunaif, A., Rich-Edwards, J. E., Stampfer, M. J., Willett, W. C., Speizer, F. E. and Manson, J. E. (2002) Menstrual cycle irregularity and risk for future cardiovascular disease. *The Journal of Clinical Endocrinology & Metabolism*, 87, 2013–2017.
Sundaram, R., McLain, A. C. and Louis, G. M. B. (2012) A survival analysis approach to modeling human fecundity. *Biostatistics*, 13, 4–17.

Terry, K. L., Willett, W. C., Rich-Edwards, J. W., Hunter, D. J. and Michels, K. B. (2005) Menstrual cycle characteristics and incidence of premenopausal breast cancer. *Cancer Epidemiology and Prevention Biomarkers*, 14, 1509–1513.

Tsiatis, A. A. and Davidian, M. (2004) Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 809–834.

Verpoest, W. M., Cahill, D. J., Harlow, C. R. and Hull, M. G. (2000) Relationship between midcycle luteinizing hormone surge quality and oocyte fertilization. *Fertility and Sterility*, 73, 75–77.

Whelan, E. A., Sandler, D. P., Root, J. L., Smith, K. R. and Weinberg, C. R. (1994) Menstrual cycle patterns and risk of breast cancer. *American Journal of Epidemiology*, 140, 1081–1090.

Wu, M. C. and Carroll, R. J. (1988) Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 175–188.