Published in final edited form as:

J R Stat Soc Ser C Appl Stat. 2015 February ; 64(2): 339–357. doi:10.1111/rssc.12075.

Joint analysis of longitudinal and survival data measured on nested timescales by using shared parameter models: an application to fecundity data

Alexander C. McLain, Rajeshwari Sundaram, and Germaine M. Buck Louis

Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, USA

Summary

We consider the joint modelling, analysis and prediction of a longitudinal binary process and a discrete time-to-event outcome. We consider data from a prospective pregnancy study, which provides day level information regarding the behaviour of couples attempting to conceive. Reproductive epidemiologists are particularly interested in developing a model for individualized predictions of time to pregnancy (TTP). A couple's intercourse behaviour should be an integral part of such a model and is one of the main focuses of the paper. In our motivating data, the intercourse observations are a long series of binary data with a periodic probability of success and the amount of available intercourse data is a function of both the menstrual cycle length and TTP. Moreover, these variables are dependent and observed on different, and nested, timescales (TTP is measured in menstrual cycles whereas intercourse is measured on days within a menstrual cycle) further complicating its analysis. Here, we propose a semiparametric shared parameter model for the joint modelling of the binary longitudinal data (intercourse behaviour) and the discrete survival outcome (TTP). Further, we develop couple-based dynamic predictions for the intercourse profiles, which in turn are used to assess the risk for subfertility (i.e. TTP longer than six menstrual cycles).

Keywords

Binary data; Discrete survival; Empirical Bayes predictions; Semiparametric models; Shared parameter models

1. Introduction

Fecundity is defined as the biologic potential of men and women for reproduction and is often measured by estimating the probability of pregnancy in each menstrual cycle among

---

Address for correspondence: Rajeshwari Sundaram, Biostatistics and Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, 6100 Executive Boulevard, Rockville, MD 20852, USA. sundaramr2@mail.nih.gov.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary materials for: "Joint analysis of longitudinal and survival data measured on nested time-scales using shared parameter models: an application to fecundity data"'.
'

couples having regular unprotected intercourse (Gini, 1928). Estimating fecundity is challenging, in part, given the effect that varying patterns of sexual intercourse may have on the length of an attempt at pregnancy (Stanford and Dunson, 2007). Clinical guidance is sometimes sought to aid couples in timing acts of intercourse around ovulation to minimize the time that is needed to achieve pregnancy. There is little empirical evidence delineating the timing of intercourse relative to ovulation, resulting in a generalized clinical recommendation to have intercourse every other day (Practice Committee of the American Society for Reproductive Medicine, 2013). Understanding the relationship between fecundity, intercourse behaviour and other relevant covariates is increasingly relevant given population level changes in the sociodemographic characteristics of reproductive-aged couples such as an increase in age at first pregnancy (Mathews and Hamilton, 2002). This may be associated with reduced intercourse activity, longer time to pregnancy (TTP), an increased prevalence of infertility or a combination of all these factors (de La Rochebrochard and Thonneau, 2003; Dunson *et al.*, 2004).

Our main objective in this paper is to model intercourse behaviour, a binary longitudinal process, and TTP, a survival outcome, jointly, with a view towards prediction of both processes. Our modelling procedure is motivated by a scientific interest in understanding the change in the probability of intercourse over time, the distribution of TTP and in predicting a couple's risk for subfertility (TTP more than six cycles).

Thus far, TTP prediction models have not incorporated day level intercourse information despite its being a necessary criterion for pregnancy. Bortot *et al.* (2010) used a model for the length of the follicular phase of the menstrual cycle to predict the probability of conception. A limitation of their prediction model was that they treated the intercourse behaviour as fixed, as no model for intercourse behaviour over the whole menstrual cycle was available. McLain *et al.* (2012) predicted TTP by using the estimated average number of acts of intercourse in the fertile window, which is defined as the days of the menstrual cycle where intercourse can lead to pregnancy. Consequently, one needs to develop methods that will assess the intercourse behaviour over the entire menstrual cycle and to study its effect on TTP, so that couples requiring clinical investigation or intervention can be helped.

To address the objectives of this paper, we utilize data from a prospective pregnancy cohort study that we introduce in Section 1.1. We also discuss the analytical challenges in considering intercourse behaviours over the entire menstrual cycle.

1.1. Prospective pregnancy studies and analytical issues with intercourse data

We utilized data from the Oxford conception study (OCS) (Pyper *et al.*, 2006), which is one of a few prospective cohort studies with preconception enrolment of women who were attempting to become pregnant. The women in the OCS provided daily level information on acts of inter-course, menstrual bleeding, smoking and alcohol consumption, along with couple level baseline covariates. These women were prospectively followed for the number of menstrual cycles that it took them to become pregnant (i.e. human chorionic gonadotropin confirmed pregnancy on the day of expected menses), which is known as their TTP, or a maximum of six menstrual cycles. Other examples of prospective pregnancy cohort studies include the 'Longitudinal investigation of fertility and environment' (Buck Louis *et al.*,

2011b), the fertili study (Colombo and Masarotto, 2000) and the Billings study (Colombo *et al.*, 2006).

Next, we discuss key methodological issues that arise in the analysis of intercourse data in prospective pregnancy studies. Each woman is observed for a random number of menstrual cycles (i.e. her TTP, which is subject to right censoring due to loss to follow-up), with a random number of observations per menstrual cycle. Here, TTP and the probability of intercourse depend on various unobserved variables that include libido and daily stressors, as well as a host of external factors like endocrine disrupting pollutants that potentially impact both male intercourse behaviour (Burnett, 2008) and female fecundity (Buck *et al.*, 2002; Buck Louis *et al.*, 2011a). Moreover, TTP and menstrual cycle length are dependent possibly because of common unmeasured covariates (Jensen *et al.*, 1999; Small *et al.*, 2010). Thus, the total number of observations of intercourse is a function of two random processes (i.e. menstrual cycle lengths and TTP) which are dependent.

Previous analyses of day level intercourse data have demonstrated that couples are more likely to have intercourse around ovulation regardless of their intentions about pregnancy (Bullivant *et al.*, 2004; Wilcox *et al.*, 2004), possibly because of an increase in the level of libido or testosterone before ovulation. This results in a periodic pattern of the probability of intercourse, where each menstrual cycle may have a maximum probability of intercourse close to the day of ovulation. As can be seen in Fig. 1, the relationship between day relative to ovulation and the probability of intercourse is non-linear. Since the structure of the relationship is unknown, we use semiparametric cubic *B*-splines to relate day relative to ovulation and probability of intercourse. To account for the heterogeneity of the probability of intercourse profile between women (possibly caused by heterogeneity in their hormone profiles), we allow for subject-specific random effects for the height and peak (i.e. amplitude) of the cubic *B*-spline curves.

The day relative to ovulation is equal to 0 for the day of ovulation, -1 for the day before ovulation, 1 for the day after ovulation, etc. For each cycle, a fertility monitor uses urinary measurements of luteinizing hormone and oestrone-3-glucuronide to identify impending ovulation within 24 h (see Fig. 1). In approximately 10% of cycles, however, the fertility monitor did not detect a day of ovulation. For cycles with unknown day of ovulation, the day relative to ovulation (which is used in our model) cannot be calculated. A missing day of ovulation is an indication that a cycle may have low hormone variability, which in turn could result in a less variable pattern of the probability of intercourse across the menstrual cycle. Ignoring such cycles (i.e. using listwise deletion) may bias results towards cycles with more variable probabilities of intercourse. Our modelling procedure accounts for cycles with unknown day of ovulation to avoid biased results.

The outline of the paper is as follows. In Section 2, we present our data structure and notation. In Section 3, we give our proposed models for intercourse and TTP and the shared parameter model framework. In Section 4, we discuss how to implement joint predictions with the shared parameter model framework using a Monte Carlo empirical Bayes technique. In Section 5, we present our analysis and prediction of the OCS data by splitting it into training and test sets. In Appendix A, we present some technical derivations for our

proposed shared parameter model, and in Appendix B we present the details of our estimation procedure.

2. Data structure and notation

This section introduces the data structure and the associated notation that will be used through-out the paper. We present below the fully observed data from the OCS prospective pregnancy study that will be used in the methods proposed.

- a. Typically, women who have enrolled in prospective pregnancy studies (e.g. the OCS) are followed from baseline up to a maximum time point τ (six menstrual cycles in the OCS) at which observation is administratively censored.
- b. During the follow-up, daily intercourse behaviour is measured with X_d denoting the indicator of an act of intercourse on day d of the study.
- c. During the follow-up, the TTP is also evaluated. The quantity is a discrete survival out-come T , indicating the number of menstrual cycles that it took to become pregnant.
- d. The menstrual cycle timescale can be ‘linked’ to the daily timescale through the observed lengths of each menstrual cycle measured in days. The length (in days) of the j th menstrual cycle is denoted by K_j (which is also the maximum number of observations of intercourse in menstrual cycle j). We denote by Y_j the follicular phase length that is associated with the j th menstrual cycle.
- e. Let T^* denote the TTP and C denote the time to censoring, both measured in menstrual cycles. Note that the latter corresponds to the number of menstrual cycles for which a woman is observed without pregnancy until time τ : Let $T = \min(T^*, C)$ and $I = I(T^* > C)$ where $I(\cdot)$ is the indicator function.
- f. Let $N = \sum_{j=1}^T K_j$ denote the total number of observable days of intercourse behaviours. Note that no intercourse data are observed after the day of a positive pregnancy test. Furthermore, let $\mathbf{X} = \{X_1, \dots, X_N\}$ denote the collection of observable intercourse indicators. However, as shown in Fig. 1 we want to account for the nesting of daily timescales within a menstrual cycle in modelling intercourse. Accordingly, we relabel intercourse information with a nested time index, i.e. $\mathbf{X} = \{X_{jk} : 1 \leq k \leq K_j; 1 \leq j \leq T\}$ where X_{jk} denotes the intercourse indicator on the k th day of the j th menstrual cycle. We use similar time indexing for all daily level covariates of interest.

3. Shared parameter model for longitudinal binary and discrete survival outcomes

We present a shared parameter framework for joint modelling of longitudinal binary data and discrete survival which incorporates the underlying characteristics that are exhibited by these processes. Shared parameter models were introduced in Wu and Carroll (1988) in the general framework of longitudinal data with missingness (see also Ten Have *et al.* (1998)

and Pulkstenis *et al.* (1998) for related work). One advantage of the shared parameter model framework is that the joint models for intercourse and TTP can be used to predict jointly the probability of intercourse on a given day and the probability distribution of the TTPs. In this section, we give the details of the specific submodels for the intercourse $f_X(\mathbf{X}|\mathbf{b}; \theta)$ and the TTP $f_T(T|\mathbf{b}; \kappa)$, where θ and κ denote the unknown parameters and $\mathbf{b}$ denotes a shared random effect.

3.1. Model for the pattern of intercourse

Here, we seek to develop a model for intercourse that allows the probability to be periodic, varying according to an unknown function within each menstrual cycle. An important aspect of our intercourse model is the cubic B -splines component to account for the periodic non-linear relationship between the probability of intercourse and time, and a mixed model to account for between-subject heterogeneity in the height and peak of the B -splines.

Let

$$f_X(\mathbf{X}_i|\mathbf{Z}_i, \mathbf{W}_i, \mathbf{Y}_i, \mathbf{b}_i; \theta) = \prod_{j=1}^{T_i} \prod_{k=1}^{K_{ij}} p_{ijk}^{X_{ijk}} (1-p_{ijk})^{1-X_{ijk}},$$

where X_{ijk} is the intercourse indicator for the i th subject on the k th day of the j th menstrual cycle and $p_{ijk} = \Pr(X_{ijk} = 1 | \mathbf{Z}_{ijk}, \mathbf{W}_{ij}, Y_{ij}, \mathbf{b}_i; \theta)$ has the functional form

$$\text{logit}(p_{ijk}) = \mu_{ijk} + \varphi_{ij} g(k - Y_{ij}), \quad (3.1)$$

with

$$\begin{aligned} \mu_{ijk} &= \boldsymbol{\beta}' \mathbf{Z}_{ijk} + \sigma_{b_1} b_{i1}, \\ \varphi_{ij} &= \exp(\boldsymbol{\phi}' \mathbf{W}_{ij} + \sigma_{b_2} b_{i2}). \end{aligned}$$

Here $\mathbf{Z}_{ijk}$ and $\mathbf{W}_{ij}$ are covariate vectors, b_{i1} and b_{i2} are random effects, $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ are vectors of regression parameters, $g(\cdot)$ is a smooth function, Y_{ij} is the follicular phase length and

$\text{logit}(p) = \log\{p/(1-p)\}$. We estimate g by using cubic B -splines with $g(y) = \sum_{l=1}^L \alpha_l B_l(y)$, Where $\{B_1(\cdot), \dots, B_L(\cdot)\}$ are B -spline basis functions of degree 3.

As discussed in Section 1, variation in hormones across the menstrual cycle contributes to a heterogeneous within-cycle probability of intercourse (i.e. variability in p_{ijk} by k ; see Fig. 1). This is represented in equation (3.1) by $g(k - Y_{ij})$ and ϕ_{ij} where $k - Y_{ij}$ is the day relative to ovulation and Y_{ij} is the follicular phase length. Here, ϕ_{ij} represents the peakedness parameter, and b_{i2} the peakedness random effect. Note that ϕ_{ij} is constrained to be positive; the minimum value of 0 indicates that the probability of intercourse has no within-cycle variability. Similar shape invariant mixed models of human hormonal profiles have been

considered by Wang *et al.* (2003) and Albert and Hunsberger (2005) (see also Ding and Wang (2008)).

3.2. Model for time to pregnancy

Next we give a model for the TTP which is a discrete survival model that modifies the hazard function as in Sundaram *et al.* (2012) accounting for whether a couple put themselves at risk for pregnancy. We assume that the survival outcome, namely TTP, has a discrete hazard

$$\lambda_i(j|\mathbf{b};\kappa)=\Pr(T_i^*=j|T_i^*\geq j, V_{ij}, U_{ij}, b_{i1}, b_{i2};\kappa)$$

with a complementary log–log–link

$$\lambda_i(j|\mathbf{b};\kappa)=1-\exp[-V_{ij}\exp\{\boldsymbol{\gamma}'\mathbf{U}_{ij}+\boldsymbol{\xi}'(b_{i1}, b_{i2})\}], \quad (3.2)$$

where $\mathbf{U}_{ij}$ is a covariate vector, $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ are vectors of regression parameters and V_{ij} is an indicator that subject i is at risk for an event in menstrual cycle j . As proposed in Sundaram

et al. (2012), we modify the hazard by using $V_{ij}=I(\sum_{k=-6}^1 X_{ij}(Y_{ij}+k)>0)$ to identify whether a couple were at risk for pregnancy if intercourse occurred in the fertile window, which is defined here as from day –6 to day 1 relative to ovulation. It is well studied through day-specific conception probabilities that acts on differing days relative to ovulation have a different influence on the probability of conception in a cycle; see for example Dunson and Stanford (2005). So, additionally to the V_{ij} -term we assess the effect of intercourse behaviour through the term $\boldsymbol{\xi}$. The parameters $\boldsymbol{\xi}=(\xi_1, \xi_2)$ capture the effect of the intercourse profile on hazard for conception in that cycle. Here, ξ_1 is associated with b_{i1} and captures the effect of the overall chance of having intercourse on a day whereas ξ_2 is associated with b_{i2} and captures the effect of a surge (or lack thereof) in intercourse activity around ovulation. We also include a time-dependent intercept in $\mathbf{U}_{ij}$ giving model (3.2) a flexible semiparametric structure. This model is equivalent to the grouped version of the continuous time proportional hazards model (Fahrmeir and Tutz, 2001). The corresponding probability mass and survivor functions are given respectively by

$$f_T(t|\mathbf{b};\kappa)=\lambda(t|\mathbf{b};\kappa)\prod_{j=1}^{t-1}\{1-\lambda(j|\mathbf{b};\kappa)\}$$

and

$$\overline{F}_T(j|\mathbf{b};\kappa)\equiv\prod_{l=1}^j\{1-\lambda(l|\mathbf{b};\kappa)\}.$$

In the absence of the at-risk indicator, the complementary log–log-link has also been used in TTP modelling by Scheike and Jensen (1997) and Scheike *et al.* (1999).

3.3. Joint model for intercourse and time to pregnancy

We next present our proposed joint model for intercourse and TTP. In our context, the number of longitudinal observations N and the intercourse behaviour $\mathbf{X}$ are dependent. More precisely, a woman's value of N is related to her intercourse behaviour $\mathbf{X}$ through her TTP; see point (f) in Section 2, i.e. N is a random variable that may potentially be informative for $\mathbf{X}$: Using arguments similar to those in section 5.16 of Cox (2006), by conditioning on N and the random effects, and assuming that the parameters for the distribution of N and those of the TTP and the intercourse behaviour are variation independent, we have

$$f(X_1, \dots, X_N, N, T | \mathbf{b}; \boldsymbol{\eta}) \propto f_X(X_1, \dots, X_N | \mathbf{b}, T; \boldsymbol{\theta}) f_T(T | \mathbf{b}; \boldsymbol{\kappa}), \quad (3.3)$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\kappa}$ are the unknown parameters in the models for intercourse and TTP respectively. Observe that expression (3.3) essentially says that the contribution of N to the log-likelihood of $\mathbf{X}$ and T is a term that is independent of the parameters of interest ($\boldsymbol{\theta}$, $\boldsymbol{\kappa}$), and in deriving expression (3.3) we have also used the framework of the shared parameter models. We provide the details of the derivation of the above expression specifically for our setting in Appendix A.

3.4. Missing data

In practice, observations are subject to missingness. We focus on two types of missingness that are encountered in our data. These two types of missingness are

- a. missingness of ovulation, i.e. the length of follicular phase in some menstrual cycles is unknown, and
- b. intermittent missingness of the intercourse indicators $\mathbf{X}$.

We first focus on the follicular phase length missingness. For cycles with known follicular phase length, we assume that Y_{ij} is observed without error. Let $\mathbf{Y}_i = \{\mathbf{Y}_{i\text{obs}}, \mathbf{Y}_{i\text{mis}}\}$ where $\mathbf{Y}_{i\text{obs}}$ and $\mathbf{Y}_{i\text{mis}}$ denote the observed and missing Y_{ij} respectively, and $R_{ij} = 1$ if the i th woman's j th follicular phase length is missing and $R_{ij} = 0$ otherwise. We include a model on the Y_{ij} s due to their informative missingness (as discussed in Section 1.1), and their presence in p_{ijk} . We assume that $Y_{ij} \sim \text{log-normal}(\nu_i, \sigma^2)$ where $\nu_i = \nu + \sigma_{b_3} b_{i3}$ and b_{i3} is a random effect. For a cycle with $R_{ij} = 1$, we integrate over the follicular phase lengths with respect to the log-normal distribution. Note that, biologically, follicular phase length cannot be longer than the menstrual cycle length; however, we found it reasonable to ignore this restriction. We assume that $\mathbf{b}_i = (b_{i1}, b_{i2}, b_{i3}) \sim \text{MVN}(\mathbf{0}, \mathbf{D})$, where $\mathbf{D}$ is a correlation matrix with off-diagonal elements denoted by ρ_{jk} . The Gaussian random-effects assumption has been shown to be robust to misspecification for estimation (see Song *et al.* (2002), Hsieh *et al.* (2006) and Rizopoulos *et al.* (2008)) and prediction (see McCulloch and Neuhaus (2011) and Albert (2012)). In Appendix A, we present details of an adaptive Gaussian quadrature inference method that was used in our data analysis, which we describe briefly below.

To estimate the likelihood of the observed intercourse values, TTP and follicular phase lengths we integrate over unobserved variables. If subject i has no missing follicular phase lengths, this amounts to integrating over $\mathbf{b}_i$ with respect to $f_b(\cdot; \mathbf{D})$. If subject i has missing follicular phase lengths, we integrate with respect to $\{\mathbf{b}_i, \mathbf{Y}_{i\text{mis}}\}$ with respect to $f_b(\cdot; \mathbf{D})$ and the distribution of $\mathbf{Y}_{i\text{mis}}|\mathbf{b}_i$. As a result, the likely values for $\mathbf{Y}_{i\text{mis}}$ are determined by the intercourse behaviour (through the p_{ijk}^s) and $\mathbf{Y}_{i\text{obs}}$ (through b_{i3}). Here, $\{\mathbf{b}_i, \mathbf{Y}_{i\text{mis}}\}$ can be integrated out by using adaptive Gaussian quadrature. An analytic file with all the programs and data used to implement our analyses can be obtained from

<http://wileyonlinelibrary.com/journal/rss-datasets>

Another type of missingness that occurs in practice is the intermittent missingness of the longitudinal process $\mathbf{X}$: We assume that, given the TTP and the menstrual cycle lengths, the intermittent missingness is independent of $\mathbf{b}$ as well as the other relevant processes. This is quite reasonable in our motivating example because intermittently missed observations were sporadic and predominantly due to missed diary entries on random days (e.g. due to forgetting to record information). Moreover, the fraction of days where such missingness occurred was small (less than 1%). So our results will probably be robust to a violation of the above assumption. Note that, under such missingness, expression (3.3) still holds.

In light of the various modelling assumptions that are made in our procedure, sections B and C of the on-line supplementary materials contain discussion and exploratory analysis of the model assumptions.

4. Prediction

In this section, we develop methods for predictions of longitudinal and survival outcomes with shared parameter models when time varying covariates are present via a Monte Carlo empirical Bayes procedure. Our method of prediction is applicable when the model proposed has been fitted by using training data, and partial information on the subjects is available. The partial information is denoted by $\mathcal{D}_{ij_0}$ for the history up to menstrual cycle j_0 for subject i . Let $\hat{\psi}$ denote an estimate of the parameter vector $\psi = \{\theta, \kappa, \mathbf{D}, \nu, \sigma\}$, with variance estimate denoted by $\hat{\Sigma}$. Prediction for menstrual cycle $j_0 + 1$ is of interest only if the couple have not achieved pregnancy, so we use $T_i = j_0$ and $i = 0$. We assume that $R_{ij} = 0$ for $j = 1, \dots, j_0$; the extension to include missing Y_{ij} is straightforward.

Given $\mathcal{D}_{ij_0}$ and $\hat{\psi}$, the posterior distribution of $\mathbf{b}_i$ is

$$f_b(\mathbf{b}_i | \mathcal{D}_{ij_0}; \hat{\psi}) \propto \bar{F}_T(T_i | \mathbf{b}_i; \hat{\kappa}) \prod_{j=1}^{j_0} f_{Y_i}(Y_{ij}; \hat{\nu}_i, \hat{\sigma}) \prod_{k=1}^{K_{ij}} \hat{p}_{ijk}^{x_{ijk}} (1 - \hat{p}_{ijk})^{1-x_{ijk}} f_b(\mathbf{b}_i; \hat{\mathbf{D}}). \quad (4.1)$$

Estimates of $\mathbf{b}_i$ can be made via the posterior mode

$$\tilde{\mathbf{b}}_{ij_0} = \arg \max_{\mathbf{b} \in \mathbb{R}^3} \log \{f_b(\mathbf{b}_i | \mathcal{D}_{ij_0}; \hat{\psi})\}.$$

The covariance matrix of $\mathbf{b}_{ijk|sub>0</sub>}$, which is denoted by Ψ_{ijQ} can be estimated from the Hessian matrix of the log-posterior distribution evaluated at $\mathbf{b}_{ijk|sub>0</sub>}$. We use the posterior mode, instead of the posterior mean, because it does not require the estimation of the normalizing constant of the posterior density.

Given ψ , $\hat{\mathbf{b}}$ and Y_{ij} , prediction of p_{ijk} , which is denoted as $p_{ijk}^{\sim}$, can be made by inputting the estimates into equation (3.1). In our data analysis we include the act of intercourse on the previous day in the fixed effect design matrix. Prediction will be done for an entire cycle of data, so lagged intercourse observations will be unknown for all except the first day. As a result, we use $p_{ijk+1|l} = P(X_{ijk+1}=1|X_{ijk}=l, \mathbf{b}; \hat{\theta})$; then given $p_{ijk}^{\sim}$ predictions can be made sequentially via $p_{ijk+1}^{\sim} = p_{ijk}^{\sim} p_{ijk+1|1}^{\sim} + (1 - p_{ijk}^{\sim}) p_{ijk+1|0}^{\sim}$. For the survival outcome, we focus on

$$\pi_i(j|j_0) = P(T_i > j | T_i > j_0, \mathcal{D}_{ij_0}, \mathbf{b}_i; \kappa) = \frac{\bar{F}_T(j | \mathcal{D}_{ij_0}, \mathbf{b}_i; \kappa)}{\bar{F}_T(j_0 | \mathcal{D}_{ij_0}, \mathbf{b}_i; \kappa)} \quad \text{for } j > j_0, \quad (4.2)$$

the conditional survivor function, which can be estimated by using $\lambda_i(j|\mathbf{b}; \kappa)$ in expression (3.2). Here, $\lambda_i(j|\mathbf{b}; \kappa) = 0$ if $V_{ij} = 0$, where V_{ij} is a function of the X_{ijk} . As a result, we estimate

$$\tilde{V}_{ij} = \Pr(V_{ij} = 1 | \tilde{\mathbf{b}}; \hat{\theta}) = 1 - (1 - \tilde{p}_{ij(Y_{ij}-6)}) \prod_{k=-5}^1 (1 - \tilde{p}_{ij(Y_{ij}+k)|0}).$$

The predicted hazard for the i th subject's j th cycle is

$$\tilde{\lambda}_i(j | \tilde{\mathbf{b}}; \hat{\psi}) = \tilde{V}_{ij} (1 - \exp[-\exp\{\hat{\gamma}' \mathbf{U}_{ij} + \hat{\phi}' (\tilde{b}_{i1j_0} \tilde{b}_{i2j_0})\}]).$$

We then estimate the conditional survivor by using the first-order approximation

$$\hat{\pi}_i(j|j_0) = \prod_{l=j_0+1}^j \{1 - \tilde{\lambda}_i(l | \tilde{\mathbf{b}}; \hat{\psi})\}.$$

Owing to the unknown follicular phase lengths, we use a Monte Carlo scheme to estimate the marginal expectations of p_{ijk} and $\pi_i(j|j_0)$. This method is also useful for estimating confidence intervals that are associated with the prediction estimates. To do so, we must account for the uncertainty that is associated with ψ , the empirical Bayes estimate $\hat{\mathbf{b}}_{ij_0}$ and the future follicular phase lengths Y_{ij} . Following arguments given in Rizopoulos (2001), we estimate the distribution of p_{ijk} and $\pi_i(j|j_0)$ by assuming that $\psi \sim \text{MVN}(\psi, \hat{\Sigma})$, and $\mathbf{b}_i \sim t_4(\mathbf{b}_{ij_0}, \hat{\psi}_{ij_0})$ where t_4 denotes a multivariate t -distribution with 4 degrees of freedom. Using this assumption, Monte Carlo samples of $p_{ijk}^{\sim}$ and $\pi_i(j|j_0)$ for $j=j_0+1, \dots, j_{\max}$ can be made as follows.

- a. Draw $\psi_{(l)} \sim \text{MVN}(\psi, \hat{\Sigma})$ and $\mathbf{b}_{(l)i} \sim t_4(\mathbf{b}_{ij0}, \psi_{ij0})$.
- b. Draw $Y_{(l)ij} \sim \text{log-normal}(\nu_{(l)} + b_{(l)i3}, \sigma_{(l)}^2)$ for $j=j_0+1, \dots, j_{\max}$.
- c. Compute $p_{(l)ijk}$ and $\pi_{(l)}(\hat{t}_{j0})$ by using $\psi_{(l)}$, $\mathbf{b}_{(l)i}$ and $(Y_{(l)ij_0+1}, \dots, Y_{(l)ij_{\max}})$.

We repeat (a)–(c) for $l=1, \dots, L$. The prediction analysis in Section 5.2 used $L=2000$. We present the mean values of the Monte Carlo samples, which we denoted by p_{ijk} and $\pi(\hat{t}_{j0})$, along with 95%-quantile-based confidence intervals.

Once values have been predicted for the test data set, the prediction performance is quantified with receiver operating characteristic (ROC) curves of empirically estimated sensitivity, $P(p_{ijk} > c | X_{ijk}=1)$, versus specificity, $P(p_{ijk} < c | X_{ijk}=0)$, for all $c \in [0, 1]$, and the area under the ROC curve, AUC. Further, we look at the calibration of the prediction by plotting

$$\text{PP}_l = \frac{\sum_{i,j,k} \hat{p}_{ijk} I(c_{l-1} \leq \hat{p}_{ijk} < c_l)}{\sum_{i,j,k} I(c_{l-1} \leq \hat{p}_{ijk} < c_l)} \quad \text{versus} \quad \text{OP}_l = \frac{\sum_{i,j,k} X_{ijk} I(c_{l-1} \leq \hat{p}_{ijk} < c_l)}{\sum_{i,j,k} I(c_{l-1} \leq \hat{p}_{ijk} < c_l)}, \quad (4.3)$$

for $c_l = 20$ with $l=1, \dots, 20$. Here, PP_l and OP_l are the empirical estimated $E(p_{ijk} | c_{l-1} \leq p_{ijk} < c_l)$ and $E(X_{ijk} | c_{l-1} \leq p_{ijk} < c_l)$ respectively. The prediction estimates are properly calibrated if $\text{PP}_l \approx \text{OP}_l$ for all l .

5. Analysis of the Oxford stress and time-to-pregnancy study

In this section, we present an analysis of a subset of the OCS data that were discussed in Section 1.1. The data that are presented here include $n = 338$ women, 1064 cycles and 24214 days of information. This analysis does not include 36 women who were excluded from the analysis because of missing values for maternal age, parity or pregnancy status. In this analysis, we are interested in assessing the association between exposures, probability of intercourse and TTP, along with predicting women's intercourse profiles and TTP survival estimates. To estimate the predictive error in our model, we randomly split the data into training (two-thirds) and test (one-third) sets. First, we analyse the training set comprising 225 women contributing 703 cycles and 16289 days (an average of 72.7 observations per woman). Secondly, we use the model fitted to present dynamic predictive intercourse profiles, and risk for subfertility for the remaining 113 woman and 7925 days in the test set.

The model that is presented for the analysis of the test data was chosen with a combination of Akaike information criterion AIC, ROC curves and AUC. AIC was used to determine whether covariates improved the model fit. We found that including some covariates that are unknown prospectively (e.g. lag variables) improved AIC but did not result in a gain in AUC for the test data, since they are unknown and must be averaged over while predicting prospectively in time. Sections C and D of the on-line supplementary materials contain details on how the final model was selected (including knot selection) and a full list of parameter estimates with standard errors respectively.

5.1. Analysis of the training set

Using the notation of Section 3, the probability that intercourse occurs on cycle j , day k , for woman i was modelled by using the form in equation (3.1) with $\mathbf{Z}_{ijk} = (1, X_{ijk-1}, \text{SMK}_{ijk}, \text{ALC}_{ijk}, \text{BLD}_{ijk}, \text{WKND}_{ijk}, \text{PAR}_i, \text{AGE}_i, \text{PAR}_i^* \text{AGE}_i)'$ (see Laumann *et al.* (2000) and Wilcox *et al.* (2004) for discussion on factors that are associated with intercourse). The definition of all variables is contained in Table 1. We modelled the distribution of TTP with model (3.2), where $\mathbf{U}_{ij} = \{I(j=1), \dots, I(j=6), \text{PAR}_i, \text{AGE}_i\}$, and

$V_{ij} = I(\sum_{k=-6}^1 X_{ij(Y_{ij}+k)} > 0)$ For follicular phase lengths, we assume $Y_i \sim \text{IID log-normal}(\nu_i, \sigma^2)$ where $\nu_i = \nu + \sigma b_3 b_{\beta}$. We found that setting $\rho_{13} = \rho_{23} = 0$ in $\mathbf{D}$ had a better fit than a model that included ρ_{13} and ρ_{23} . In the model fitting portion we tested for the presence of various trends in intercourse behaviour by cycle number. The final model includes covariates $j^* = j - 1$ and j^{*2} in the peakedness portion of the intercourse model (j^* is used to help the identifiability of the peakedness parameters). The final model had 10 interior knots at $\{-13, -10, -7, -4, -2, 1, 4, 7, 10, 13\}$ days relative to ovulation, corresponding to the percentiles of the observed $k - Y_{ij}$ (see section C of the on-line supplementary materials for further details). We assumed that $g(\cdot)$ was constant outside -20 and 50 days relative to ovulation. Standard errors of estimates were calculated from a numerical approximation to the observed Fisher information matrix. We compare our results with those of a generalized linear model (GLM) and a generalized non-linear mixed model (GNLMM) that included all covariates and the cubic B -splines function. For these models for cycles with missing follicular phase length we use a fixed $Y_{ij} = 14$ days corresponding to the follicular phase length in an ideal menstrual cycle. This is motivated by common practice of assuming that an ideal menstrual cycle is of length 28 days and the ovulation is assumed to occur mid-cycle (approximately around day 14) (Dominik *et al.*, 2001). The GNLMM also included a random intercept.

In Table 2, we present the intercourse submodel estimates, standard errors and 95% Wald-type confidence intervals for the method proposed, a GLM and a GNLMM. The biggest factor in the probability of intercourse was BLD_{ijk} , the menstrual bleeding indicator. The other covariates that had a significant effect on the probability of intercourse were X_{ijk-1} , ALC_{ijk} and $\text{PAR}_i^* \text{AGE}_i$. There was no significant cycle effect on the peakedness, but AIC did prefer a model with these components to a model without. There are noticeable differences between the model proposed *versus* the GLM and GNLMM in the value of the intercept estimate $\hat{\beta}_0$, which is lower in the models that do not account for joint modelling with TTP. This was expected since people having less intercourse contribute more cycles of data (longer TTPs). Another marked difference is in the direction of the lag 1 value in the model proposed and the GNLMM, *versus* the GLM. The positive direction for the lag parameter in the GLM is not unexpected since, without a baseline random effect, it is the only way to increase the marginal probability of intercourse $E(p_{ijk})$ for highly active couples. We also believe that the performance of the GLM and GNLMM would also be impacted by not accounting for variability in follicular phase length.

In Table 3, we display the results for the standard deviations and correlation of the random effects, and the TTP parameters. All the standard deviations were significantly greater than 0, and there was a substantial negative correlation between b_1 and b_2 . The negative

correlation is a result of couples who target their act of intercourse acts to fall within the fertile window and have little intercourse otherwise, *versus* couples who have consistent intercourse throughout the menstrual cycle. The estimates of ξ_1 and ξ_2 show that there is a significant dependence between the baseline b_1 and peakedness b_2 random effects and TTP. The estimates suggest that those with a positive baseline or peakedness random effect become pregnant faster. However, those women random effects that are large in absolute value but opposite in sign (e.g. $b_1 = c$ and $b_2 = -c$) have TTP characteristics that are not markedly different from a woman with $b_1 = b_2 = 0$. The results indicate that, for a fixed $b_1 = 0$, a 35-year-old nulliparous (i.e. no previous live birth) woman with better-than-average intercourse behaviour ($b_2 = 1$) has a median TTP of four cycles. However, if she has lower-than-average intercourse behaviour ($b_2 = -1$) the median TTP is 10 cycles (assuming that $\gamma_j = \gamma_6$ for $j = 7, 8, \dots$). Further graphical examples are given in section D of the on-line supplementary materials.

5.2. Prediction for test data set

Using estimates for the models that were built in the previous section, we predicted intercourse probability profiles and TTP characteristics for the test data set. We use the first cycle of data to implement the Monte Carlo empirical Bayes estimation procedure that was detailed in Section 4. For the GLM and GNLMM we used a fixed 14-day follicular phase length. The GNLMM used empirical Bayes estimates of the baseline random effect. For time varying covariates SMK_{ijk} and ALC_{ijk} , we imputed the empirical probabilities of these indicators by using $\mathcal{D}_{i1}$. For BLD_{ijk} we used the bleeding pattern that was observed in the first cycle.

The performance of the prediction is quantified by using calibration curves, ROC curves and AUC. The AUC and ROC curve will quantify the quality of the rankings that are generated by each specific model, whereas the calibration curves assess any bias in the predictive estimates. In Fig. 2(a), we present a calibration curve, which plots PP_i against OP_i given in expression (4.3), and in Fig. 2(b) an ROC curve, for the GLM, GNLMM and the model proposed. The calibration of the model proposed and the GNLMM are similar, and both are superior to the GLM. The rankings of the probabilities improved substantially with the proposed method *versus* the standard approaches. The AUC of the models was 0.796, 0.725 and 0.679 for the model proposed, GNLMM and GLM respectively. Using the model proposed, which incorporates a model for follicular phase length to account for missingness, leads to a 17.3% and 9.8% improvement in AUC over respectively the GLM and GNLMM with a fixed 14 days to account for missing follicular phase lengths.

Using the methods in Section 4, we estimated predictive intercourse profiles with 95% confidence intervals, for the second and third cycles of two women with identification numbers 1135 and 1468 (Fig. 3). Woman 1135 has a particularly low value of $\tilde{b}_{12}$, indicating a flatter, more consistent probability profile, whereas woman 1468 has a particularly high value of $\tilde{b}_{12}$, indicating a highly peaked probability profile. The empirical Bayes estimates were based on $\mathcal{D}_{i1}$ and $\mathcal{D}_{i2}$ for the second and third cycles respectively. Menstrual bleeding's effect on the probability of intercourse can be seen in the sharp decrease in p for the beginning days for woman 1135. These figures demonstrate the flexibility of the proposed

semiparametric mixed model of binary data. Allowing for heterogeneity in the peakedness is critical for our outcome.

We were further interested in predicting the probability of subfertility (i.e. TTP longer than 6 cycles). Let $\tilde{\mathbf{U}}_{ij}$ denote the predicted covariates for cycle j . The form of the predicted conditional survivor function is given by

$$\hat{\pi}_i(j|j_0) = \prod_{l=j_0+1}^j \{1 - \tilde{V}_{il}(1 - \exp[-\exp\{\hat{\gamma}'\mathbf{U}_{il} + \hat{\phi}'(\tilde{b}_{i1j_0}\tilde{b}_{i2j_0})\}])\}. \quad (5.1)$$

We predict the probability that a couple are subfertile conditionally on one and two cycles of information, i.e. $\pi(6|1)$ and $\pi(6|2)$ respectively. We did not include women with $T_i = 1$ or $T_i < 6$ with $j = 0$, which left 65 women available for the prediction analysis. We compare the predictions that are given in equation (5.1) with a semiparametric discrete survival model with complementary log-log-form, where the sum of the intercourse values over the fertile window $\sum_{k=-6}^1 X_{ij}(Y_{ij}+k)$, parity and age were included as covariates, along with the V_{ij} at-risk indicator. Cycles with missing Y_{ij} used a fixed 14 days. We used the average number of acts of intercourse over the fertile window of the first j_0 menstrual cycles to predict $\pi(j|j_0)$ with the discrete survival model.

The sensitivity $P\{\hat{\pi}_i(6|j) > c | T_i > 6\}$ and specificity $P\{\hat{\pi}_i(6|j) \leq c | T_i \leq 6\}$ for $j=1, 2$ $c \in [0, 1]$, are displayed in Fig. 4 for the method proposed and the standard discrete survival model. For the method proposed AUC for $\hat{\pi}_i(6|1)$ and $\hat{\pi}_i(6|2)$ is 0.707 and 0.714 respectively; the corresponding values for the standard method are 0.565 and 0.588. This indicates that AUC for the method proposed is 25.1% and 21.4% higher than that for the standard approach, when conditioning on one and two cycles of data. For the method proposed $\hat{\pi}_i(6|1)$ ranks women with subfertility higher than women without 70.7% of the time. This shows the efficacy of the proposed model in accurately predicting longitudinal and survival processes.

6. Summary and discussion

In this paper, we have proposed a semiparametric shared parameter model for the analysis and prediction of intercourse behaviour (a longitudinal process) and TTP (a discrete time-to-event process). The total number of observations of intercourse can be viewed as having nested time-scales, where we observe a series of intervals with varying numbers of observations per interval. The shared parameter model framework was beneficial in our scenario, since it allowed for the joint analysis and prediction of the longitudinal and survival processes.

The training set portion of our data analysis found that there was a significant dependence between the shared baseline and peakedness intercourse random effects and TTP. This dependence was in the hypothesized direction, where people with higher rates of intercourse behaviour became pregnant faster than those with lower rates of intercourse behaviour. Our proposed semiparametric peakedness model successfully accounted for the non-linear trend between the probability of intercourse and day relative to ovulation (see Fig. 3). By

including a model on the follicular phase lengths we could integrate out their effect when unknown, and in the prediction analysis. The test portion showed how a Monte Carlo empirical Bayesian procedure could be used to obtain point predictions and to estimate predictive error. Such an approach, where the estimation of the parameters is done via a likelihood-based method and prediction via a Bayesian approach, is quite common in practice and has been shown to have desirable properties (Proust-Lima and Taylor, 2009; Rizopoulos, 2011) in the context of joint modelling of longitudinal and time-to-event data analysis. However, a full Bayesian approach and a comparison *vis à vis* our approach including computational cost is worth investigating in the future. The predictive model of intercourse behaviour can be used in future attempts of TTP prediction (see the on-line supplementary materials for a full list of parameter estimates and standard errors). One could alternatively predict TTP by using the intercourse observations in the fertile window and the survival model that was proposed by Sundaram *et al.* (2012). However, owing to the unknown follicular phase lengths, this model would still require a model for intercourse observations outside the fertile window, adding to its complexity.

Our work is relevant for the development of couple-specific intercourse and TTP prediction models and offers promise to help couples to minimize the time required to achieve pregnancy by addressing sexual intercourse behaviours. Further, preconception counsellors can use current methods of targeting acts of intercourse (see Scarpa and Dunson (2007) and Bortot *et al.* (2010)) for those with predicted intercourse or fecundity problems to maximize their success in pregnancy. AUC for our TTP prediction was slightly lower than the AUC of 0.77 that was found in the TTP prediction in McLain *et al.* (2012) that was based on menstrual cycle characteristics. Their results predicted the TTP for a different data set and focused on the relationship between menstrual cycle length and TTP. Morework is needed in evaluating the predictive benefits of multiple longitudinal processes with a relationship to a survival outcome (Rizopoulos and Ghosh, 2011).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institute of Health, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. This work was completed while the first author was a Research Fellow with Dr Sundaram. The authors also acknowledge that this study utilized the high performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Maryland (<http://biowulf.nih.gov>). The authors thank the Joint Editor, the Associate Editor and two reviewers for comments and suggestions that considerably improved this paper.

References

- Albert PS. A linear mixed model for predicting a binary event from longitudinal data under random effects misspecification. *Statist Med.* 2012; 31:143–154.
- Albert PS, Hunsberger S. On analyzing circadian rhythms data using non-linear mixed models with harmonic terms. *Biometrics.* 2005; 61:1115–1120. [PubMed: 16401286]
- Bortot P, Masarotto G, Scarpa B. Sequential predictions of menstrual cycle lengths. *Biostatistics.* 2010; 11:741–755. [PubMed: 20400622]

- Buck GM, Vena JE, Greizerstein HB, Weiner JM, McGuinness B, Mendola P, Kostyniak PJ, Swanson M, Bloom MS, Olson JR. Pcb congeners and pesticides and female fecundity, New York state angler prospective pregnancy study. *Environ Toxicol Pharmacol*. 2002; 12:83–92.
- Buck Louis GM, Rios LI, McLain AC, Cooney MA, Kostyniak PJ, Sundaram R. Persistent organochlorine pollutants and menstrual cycle characteristics. *Chemosphere*. 2011a; 85:1742–1748. [PubMed: 22018858]
- Buck Louis GM, Schisterman EF, Sweeney AM, Wilcosky TC, Gore-Langton RE, Lynch CD, Boyd Barr D, Schrader SM, Kim S, Chen Z, Sundaram R. on behalf of the LIFE Study. Designing prospective cohort studies for assessing reproductive and developmental toxicity during sensitive windows of human reproduction and development—the LIFE Study. *Paed Perntl Epidem*. 2011b; 25:413–424.
- Bullivant SB, Sellergren SA, Stern K, Spencer NA, Jacob S, Mennella JA, McClintock MK. Women's sexual experience during the menstrual cycle: identification of the sexual phase by noninvasive measurement of luteinizing hormone. *J Sex Res*. 2004; 41:82–93. [PubMed: 15216427]
- Burnett AL. Environmental erectile dysfunction: Can the environment really be hazardous to your erectile health? *J Androl*. 2008; 29:229–236. [PubMed: 18187396]
- Colombo B, Masarotto G. Daily fecundability: first results from a new data base. *Demog Res*. 2000; 3(5)
- Colombo B, Mion A, Passarin K, Scarpa B. Cervical mucus symptom and daily fecundability: first results from a new database. *Statist Meth Med Res*. 2006; 15:161–180.
- Cox, DR. *Principles of Statistical Inference*. Cambridge: Cambridge University Press; 2006.
- Ding J, Wang JL. Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics*. 2008; 64:546–556. [PubMed: 17888040]
- Dominik R, Zhou H, Cai J. A statistical model for the evaluation of barrier contraceptive efficacy. *Statist Med*. 2001; 20:3279–3294.
- Dunson DB, Baird DD, Colombo B. Increased infertility with age in men and women. *Obstet Gyn*. 2004; 103:51–56.
- Dunson DB, Stanford JB. Bayesian inferences on predictors of conception probabilities. *Biometrics*. 2005; 61:126–133. [PubMed: 15737085]
- Fahrmeir, L.; Tutz, G. *Multivariate Statistical Modelling based on Generalized Linear Models*. 2. New York: Springer; 2001.
- Gini, C. In: Fields, JC., editor. *Premières recherches sur la fécondabilité de la femme*; Proc. Int. Math. Congr; Toronto. Aug. 11th–16th, 1924; Toronto: University of Toronto Press; 1928. p. 889–892.
- Hsieh F, Tseng YK, Wang JL. Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics*. 2006; 62:1037–1043. [PubMed: 17156277]
- Jensen TK, Scheike T, Keiding N, Schaumburg I, Grandjean P. Fecundability in relation to body mass and menstrual cycle patterns. *Epidemiology*. 1999; 10:422–428. [PubMed: 10401878]
- de La Rochebrochard E, Thonneau P. Paternal age > 40 years: an important risk factor for infertility. *Am J Obstet Gyn*. 2003; 189:901–905.
- Laumann, E.; Gagnon, J.; Michael, R.; Michaels, S. *The Social Organization of Sexuality: Sexual Practices in the United States*. Chicago: University of Chicago Press; 2000.
- Mathews T, Hamilton BE. Mean age of mother, 1970–2000. *Natn Vitl Statist Rep*. 2002; 51:1–13.
- McCulloch CE, Neuhaus JM. Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics*. 2011; 67:270–279. [PubMed: 20528860]
- 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]
- Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J Computnl Graph Statist*. 1995; 4:12–35.
- Practice Committee of the American Society for Reproductive Medicine. Optimizing natural fertility: a committee opinion. *Fertil Steril*. 2013; 100:631–637. [PubMed: 23993665]
- Proust-Lima C, Taylor JMG. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: a joint modeling approach. *Biostatistics*. 2009; 10:535–549. [PubMed: 19369642]

- Pulkstenis EP, Ten Have TR, Landis JR. Model for the analysis of binary longitudinal pain data subject to informative dropout through remedication. *J Am Statist Ass.* 1998; 93:438–450.
- Pyper C, Bromhall L, Dummett S, Altman DG, Brownbill P, Murphy M. The Oxford conception study design and recruitment experience. *Paed Perntl Epidem.* 2006; 20:51–59.
- R Development Core Team. R: a Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2012.
- Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics.* 2011; 67:819–829. [PubMed: 21306352]
- Rizopoulos D, Ghosh P. A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statist Med.* 2011; 30:1366–1380.
- Rizopoulos D, Verbeke G, Molenberghs G. Shared parameter models under random effects misspecification. *Biometrika.* 2008; 95:63–74.
- Rothman, K.; Greenland, S.; Lash, T. *Modern Epidemiology.* Boston: Wolters Kluwer Health; 2008.
- Scarpa B, Dunson DB. Bayesian methods for searching for optimal rules for timing intercourse to achieve pregnancy. *Statist Med.* 2007; 26:1920–1936.
- Scheike TH, Jensen TK. A discrete survival model with random effects: an application to time to pregnancy. *Biometrics.* 1997; 53:318–329. [PubMed: 9147597]
- Scheike TH, Petersen JH, Martinussen T. Retrospective ascertainment of recurrent events: an application to time to pregnancy. *J Am Statist Ass.* 1999; 94:713–725.
- Small CM, Manatunga AK, Klein M, Dominguez CE, Feigelson HS, McChesney R, Marcus M. Menstrual cycle variability and the likelihood of achieving pregnancy. *Rev Environ Hlth.* 2010; 25:369–378.
- Song X, Davidian M, Tsiatis AA. A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics.* 2002; 58:742–753. [PubMed: 12495128]
- Stanford JB, Dunson DB. Effects of sexual intercourse patterns in time to pregnancy studies. *Am J Epidem.* 2007; 165:1088–1095.
- Sundaram R, McLain AC, Buck Louis GM. A survival analysis approach to modeling human fecundity. *Biostatistics.* 2012; 13:4–17. [PubMed: 21697247]
- Ten Have TR, Kunselman AR, Pulkstenis EP, Landis JR. Mixed effects logistic regression models for longitudinal binary response data with informative drop-out. *Biometrics.* 1998; 54:367–383. [PubMed: 9544529]
- Wang Y, Ke C, Brown MB. Shape-invariant modeling of circadian rhythms with random effects and smoothing spline anova decompositions. *Biometrics.* 2003; 59:804–812. [PubMed: 14969458]
- Wilcox A, Day Baird D, Dunson DB, Mc Connaughey DR, Kesner JS, Weinberg CR. On the frequency of intercourse around ovulation: evidence for biological influences. *Hum Reproduct.* 2004; 19:1539–1543.
- Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics.* 1988; 44:175–188.

Appendix A: Shared parameter models

In this section, we provide details concerning our joint model formulation and provide justification for expression (3.3) in our framework of shared parameter models. Observe that N , as defined in point (f) of Section 2, depends both on the TTP and on the menstrual cycle lengths. We look to simplify the dependence structure by using the current biological understanding of the dependence structure of intercourse, TTP and menstrual cycle length. More precisely, we assume that there are unobserved latent variables $\mathbf{b}$ (e.g. unmeasured factors such as libido and stress level) that relate intercourse and TTP, and unobserved latent variables $\mathbf{c}$ (e.g. unmeasured female reproductive hormones such as luteinizing hormone and oestrogen) that relate the TTP and menstrual cycle length, whereas intercourse and

menstrual cycle length may be indirectly related through the TTP (see Fig. 5 for a visual representation of these dependences). In view of this, we make the following assumptions.

Assumption 1. Given $\mathbf{b}$ and the TTP, intercourse acts are independent of the menstrual cycle lengths.

Assumption 2. Given the TTP, the menstrual cycle lengths are independent of $\mathbf{b}$.

In Fig. 5, we include a latent variable $\mathbf{c}$ (e.g. unmeasured female reproductive hormones) that induces dependence between $\mathbf{K}$ and T . It should be noted that conditioning on T in the distribution of $\mathbf{K}$ could create dependence between $\mathbf{K}$ and $\mathbf{b}$. Note that T is an outcome of $\mathbf{b}$ (directly) and menstrual cycle length (indirectly through $\mathbf{c}$), which is commonly referred to as a collider. It is well known in the epidemiological literature (Rothman *et al.*, 2008) that, when a collider is conditioned on, dependence between two otherwise independent variables is generated. For example, there may be dependence between menstrual cycle length and $\mathbf{b}$ after conditioning on TTP. However, this dependence would be attenuated by the strength of the relationships between T and $\mathbf{c}$, followed by $\mathbf{K}$ and $\mathbf{c}$. Consequently, we assume that the relationship between $\mathbf{K}$ and $\mathbf{b}$ given T is negligible. Thus, assumptions 1 and 2 are reasonable in our context.

We shall now establish expression (3.3), i.e.

$$f(\mathbf{X}_1, \dots, X_N, N, T | \mathbf{b}; \boldsymbol{\eta}) \propto f_X(X_1, \dots, X_N | \mathbf{b}, T; \theta) f_T(T | \mathbf{b}; \kappa),$$

where θ and κ are the unknown parameters in the models for intercourse and TTP respectively. For this, note that the joint distribution of the longitudinal process $\mathbf{X}$, N and T under the shared parameter framework can be expressed as

$$f(X_1, \dots, X_N, N, T | \mathbf{b}; \boldsymbol{\eta}) = f(X_1, \dots, X_N, N | \mathbf{b}, T; \theta) f_T(T | \mathbf{b}; \kappa).$$

Note that $N = \sum_{j=1}^T K_j$ is a function of two random processes T and $\mathbf{K}_T$. So, we have

$$f(X_1, \dots, X_{N_O}, N = N_O | \mathbf{b}, T; \theta) = \int_{\sum_{j=1}^T K_j = N_O} f_X(X_1, \dots, X_{N_O} | T, \mathbf{K}_T, \mathbf{b}; \theta) f(\mathbf{K}_T | \mathbf{b}, T; \omega) d\mathbf{K}_T \quad (\text{A.1})$$

$$\begin{aligned} &= f_X(X_1, \dots, X_{N_O} | \mathbf{b}, T; \theta) \int_{\sum_{j=1}^T K_j = N_O} f(\mathbf{K}_T | T; \omega) d\mathbf{K}_T \\ &\propto f_X(X_1, \dots, X_{N_O} | \mathbf{b}, T; \theta). \end{aligned} \quad (\text{A.2})$$

Note that we derived equation (A.2) from equation (A.1) as a consequence of assumptions 1 and 2. Consequently,

$$f(X_1, \dots, X_N, N, T | \mathbf{b}; \theta, \kappa) \propto f_X(X_1, \dots, X_N | \mathbf{b}, T; \theta) f_T(T | \mathbf{b}; \kappa).$$

Appendix B: Inference method

In this section, we present a maximum likelihood method using adaptive Gaussian quadrature to estimate the parameter vector $\psi = \{\theta, \kappa, \mathbf{D}, \nu, \sigma\}$. We use the formulation of the data structure that was introduced in Sections 2 and 4. For subject i we observe $\mathcal{D}_i = \{\mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i, \mathbf{U}_i, \mathbf{Y}_i, \mathbf{R}_i, T_i\}$ where $\mathbf{X}_i$ is a vector of the longitudinal binary outcome, $\mathbf{Z}_i, \mathbf{W}_i$ and $\mathbf{U}_i$ are the design matrices and $\mathbf{R}_i = \{R_{i1}, \dots, R_{iT_i}\}$ is a vector of indicators that delineates $\mathbf{Y}_i$ into observed and missing portions. To estimate ψ , we use expression (3.3). Under these assumptions the likelihood for subject i is

$$L(\psi | \mathcal{D}) \propto \prod_{i=1}^n \int \int \left(f_T(T_i | \mathbf{b}; \kappa)^{\Delta_i} \bar{F}_T(T_i | \mathbf{b}; \kappa)^{1-\Delta_i} \left[\int \prod_{j=1}^{T_i} \left\{ \int f_{Y_i}(Y_{ij}^* | \nu + \sigma_{b_3} b_3; \sigma) \prod_{k=1}^{K_{ij}} p_{ijk}^{x_{ijk}} (1-p_{ijk})^{1-x_{ijk}} dY_{ij\text{mis}} f_b(\mathbf{b}; \mathbf{D}) db_3 db_1 db_2 \right\} \right] \right) d\mathbf{b} \quad (\text{B.1})$$

where $Y_{ij}^* = (1 - R_{ij})Y_{ij} + R_{ij}Y_{ij\text{mis}}$ and $\mathcal{D} = \{\mathcal{D}_1, \dots, \mathcal{D}_n\}$. We use adaptive Gaussian quadrature to evaluate expression (B.1), which is based on the conditional likelihood for subject i :

$$L(\psi | \mathcal{D}_i, \mathbf{a}) \propto f_T(T_i | \mathbf{b}; \kappa)^{\Delta_i} \bar{F}_T(T_i | \mathbf{b}; \kappa)^{1-\Delta_i} \prod_{j=1}^{T_i} \int f_{Y_i}(Y_{ij}^* | \nu; \sigma) \prod_{k=1}^{K_{ij}} p_{ijk}^{x_{ijk}} (1-p_{ijk})^{1-x_{ijk}} f_b(\mathbf{b}; \mathbf{D}),$$

where $\mathbf{a} = \{\mathbf{b}, \mathbf{Y}_{i\text{mis}}\}$. For each candidate value of ψ , say $\psi = \psi^c$, we estimate

$$\tilde{\mathbf{a}}_i \equiv \arg \max_{\mathbf{a} \in \mathcal{A}_i} \log\{L(\psi^c | \mathcal{D}_i, \mathbf{a})\}, \quad (\text{B.2})$$

and

$$\tilde{H}_i = \frac{\partial^2}{\partial \mathbf{a}' \partial \mathbf{a}} \log\{L(\psi^c | \mathcal{D}_i, \mathbf{a})\} |_{\mathbf{a}=\tilde{\mathbf{a}}_i},$$

where $\mathcal{A}_i = \mathbb{R}^{3+\sum_{j=1}^{T_i} R_{ij}}$, for all $i = 1, \dots, n$. Following Pinheiro and Bates (1995) the integral in expression (B.1) is approximated by a weighted sum of $L(\psi | \mathcal{D}_i, \mathbf{a})$ evaluated at Gaussian quadrature nodes centred and scaled by $\tilde{\mathbf{a}}_i$ and $\tilde{H}_i$ respectively, for $i = 1, 2, \dots, n$. To calculate $\hat{\psi} = \arg \max_{\psi} \log\{L(\psi | \mathcal{D})\}$ we used the numerical optimization routine `optim` in R (R Development Core Team, 2012). Section E of the on-line supplementary materials contains the R code.

In the on-line supplementary materials we present further details on our inference method and results of a simulation study. In our simulation study we tested the performance of a model that is similar to that presented here. Our adaptive Gaussian quadrature method was

fitted with five, 10 and 15 quadrature nodes. We found that when 10 quadrature nodes were used the estimates of the parameters had little bias, and a numerical approximation to the Hessian matrix provided standard error estimates that were relatively similar to the Monte Carlo standard error. Wald-type confidence intervals yielded converged probabilities that were close to the nominal 0.95-level. We did not find that 15 quadrature nodes were significantly superior to 10 quadrature nodes. As a result, we used 10 quadrature nodes in our analysis.

Fig. 1.
Average number of acts of intercourse on non-bleeding days by the day relative to ovulation (—), along with the longitudinal hormonal patterns of luteinizing hormone (-----) and oestrone-3-glucuronide (-.-.-.-) scaled to (0,1)

Fig. 2.

(a) Calibration curve of PP/ *versus* OP/ given in expression (4.3) and (b) ROC curves: both compare a GLM (· · · · ·), a GNLMM (-----) and the proposed model (——) for the test data set; AUC for the proposed model was 0.796 for the GNLMM 0.725 and for the GLM 0.679

Fig. 3.

Dynamic individualized predictive profiles by using the proposed model (—) with 95% pointwise confidence intervals (-----) and the true realizations (●) (empirical Bayes estimates of unknowns utilized $\mathcal{D}_{i1}$ for cycle 2, and $\mathcal{D}_{i2}$ for cycle 3): (a) cycle 2, woman 1135; (b) cycle 3, woman 1135, (c) cycle 2, woman 1468; (d) cycle 3, woman 1468

Fig. 4. ROC curve for predicting the probability $T_i > 6$ by (a) $\pi_{\hat{\lambda}}(6|1)$ and by (b) $\pi_{\hat{\lambda}}(6|2)$ with the proposed method (—) and the standard approach (-----): AUC for the proposed method and standard approach is (0.707, 0.714) and (0.565, 0.588) respectively

Fig. 5. Schematic representation of the underlying random processes of interest: intercourse $\mathbf{X}$, observed intercourse $\mathbf{X}_{\text{obs}}$, menstrual cycle length K and TTP T : $\circ$, latent variables representing the unobserved shared random effect $\mathbf{b}$, and the unobserved covariates $\mathbf{c}$ that relate K and T

Table 1

Definitions of fixed effect covariates

Effect	Definition	Type
X_{ijk-1} Lag 1	intercourse indicator	Day varying
ALC	Indicator of alcohol consumption	Day varying
BLD	Indicator of menstrual bleeding	Day varying
SM	K Indicator of smoking	Day varying
WKND	Indicator of a Friday, Saturday or Sunday	Day varying
PAR	Indicator whether the women had at least 1 live birth	Baseline
AGE	Age of the female partner minus 31.5 years	Baseline

Table 2

Estimates $\hat{\theta}$ and 95% confidence intervals for intercourse outcome with the proposed method, GNLM and GLM

	Results for proposed model		Results for GNLM		Results for GLM	
	$\hat{\theta}$	95% confidence interval	$\hat{\theta}$	95% confidence interval	$\hat{\theta}$	95% confidence interval
Intercept	-1.02	(-1.39, -0.65)	-1.51	(-2.16, -0.86)	-1.41	(-2.04, -0.78)
Lag 1	-0.21	(-0.3, -0.12)	-0.12	(-0.19, -0.03)	0.26	(0.17, 0.33)
SMK	-0.22	(-0.47, 0.02)	-0.18	(-0.40, 0.04)	0.01	(-0.10, 0.13)
ALC	0.18	(0.08, 0.28)	0.18	(0.09, 0.26)	0.05	(-0.03, 0.14)
BLD	-2.20	(-2.41, -1.99)	-2.06	(-2.26, -1.86)	-1.88	(-2.08, -1.68)
WKND	0.07	(-0.002, 0.15)	0.09	(0.011, 0.17)	0.09	(0.021, 0.16)
PAR	-0.15	(-0.35, 0.05)	-0.12	(-0.31, 0.06)	-0.10	(-0.17, -0.02)
AGE	-0.02	(-0.05, 0.01)	-0.02	(-0.04, 0.00)	-0.01	(-0.02, 0.00)
AGE*PAR	-0.07	(-0.11, -0.02)	-0.06	(-0.10, -0.02)	-0.05	(-0.07, -0.04)
f^*	0.09	(-0.04, 0.22)	—	—	—	—
f^{*2}	-0.02	(-0.04, 0.01)	—	—	—	—

Table 3

Estimates $\hat{\theta}$ and 95% confidence intervals for the proposed model[†]

	D-parameters		TTP parameters	
	$\hat{\theta}$	95% confidence interval	$\hat{\theta}$	95% confidence interval
σ_{b1}	0.84	(0.67, 1.00)	γ_{PAR}	0.57 (0.22, 0.92)
ρ_{12}	-0.64	(-0.84, -0.44)	γ_{AGE}	-0.26 (-0.43, -0.09)
σ_{b2}	0.62	(0.51, 0.73)	ξ_1	0.24 (0.03, 0.44)
σ_{b3}	2.49	(2.12, 2.86)	ξ_2	0.30 (0.05, 0.55)

[†]The GNLM estimated σ_{b1} as 0.71 with 95% confidence interval (0.56, 0.86).