STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:965–978

Assessing human fertility using several markers of ovulation‡

D. B. Dunson[1,*,†] C. R. Weinberg[1], D. D. Baird[2], J. S. Kesner[3] and A. J. Wilcox[2]

[1] *Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC, U.S.A.*
[2] *Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC, U.S.A.*
[3] *Experimental Toxicology Branch, Division of Biomedical and Behavioral Science, National Institute for Occupational Safety and Health*

SUMMARY

In modelling human fertility one ideally accounts for timing of intercourse relative to ovulation. Measurement error in identifying the day of ovulation can bias estimates of fecundability parameters and attenuate estimates of covariate effects. In the absence of a single perfect marker of ovulation, several error prone markers are sometimes obtained. In this paper we propose a semi-parametric mixture model that uses multiple independent markers of ovulation to account for measurement error. The model assigns each method of assessing ovulation a distinct non-parametric error distribution, and corrects bias in estimates of day-specific fecundability. We use a Monte Carlo EM algorithm for joint estimation of (i) the error distribution for the markers, (ii) the error-corrected fertility parameters, and (iii) the couple-specific random effects. We apply the methods to data from a North Carolina fertility study to assess the magnitude of error in measures of ovulation based on urinary luteinizing hormone and metabolites of ovarian hormones, and estimate the corrected day-specific probabilities of clinical pregnancy. Published in 2001 by John Wiley & Sons, Ltd.

1. INTRODUCTION

Studies that collect information on intercourse and ovulation can be used to address fundamental questions about human fertility. These questions include: (i) What is the length of the fertile interval and on which days around ovulation can a woman potentially become pregnant? (ii) What pattern do the day-specific conception probabilities follow within this interval? (iii) What factors affect fecundability? (iv) How much variability is there from couple to couple in fecundability? (v) How error prone are commonly used markers of ovulation? In order to address these questions properly, new statistical methods are needed.

In fertile women of reproductive age, a single ovum is typically released during each menstrual cycle in a process called ovulation. Ovulation is associated with alterations in basal

* Correspondence to: David B. Dunson, Biostatistics Branch, MD A3-03, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, NC 27709, U.S.A.
† E-mail: dunson1@niehs.nih.gov

Received June 1999
Published in 2001 by John Wiley & Sons, Ltd. *Accepted April 2000*

body temperature, cervical mucus and the production of ovarian hormones. The probability that an act of intercourse results in a pregnancy is dependent on a number of factors including the timing of intercourse relative to ovulation, the viability of the sperm and ovum, and survival of the embryo to detection of the pregnancy [1]. The day-specific probability of pregnancy is theoretically zero outside of a fertile interval around ovulation. The determinants of this fertile interval are unknown, but may include sperm and egg survival [2, 3], cervical mucus penetrability, ovarian hormones and contraceptive glycoproteins [4]. The ovum remains viable for only a short time after ovulation, while sperm can survive for several days in the normal female reproductive tract [5]

The day of ovulation can be estimated from serial physiological measurements including body temperature (last day of hypothermia), cervical mucus (maximum viscosity), luteinizing hormone (LH) (surge above baseline), and the ratio of oestrogen to progesterone (abrupt decline) [6]. Schwartz *et al.* [1] proposed a model to account for both timing of intercourse relative to ovulation and cycle viability, which includes viability of the ovum and survival of the embryo to detection. Their model was later extended to allow for covariate effects [7, 8], heterogeneity among couples [9], missing intercourse data [10], a sterile subpopulation of couples [11] and measurement error in a single marker of ovulation [12].

If there is error in identifying the day of ovulation, then for some proportion of cycles the estimated day will be shifted by one or more days from the true day. This creates bias in the fecundability parameters [12], artefactually extends the fertile interval [13] and attenuates estimates of covariate effects. In previous work we developed a model and a Bayesian Markov chain Monte Carlo estimation procedure for accounting for measurement error when a single marker of ovulation is available [12]. In the absence of validation data containing both the error-prone marker and a direct measure of ovulation, such as ultrasound documentation of follicular rupture, the identifiability of the measurement error distribution relies on the assumed model for the fertilization probability. Similar identifiability problems exist in modelling of dietary data [14]. Conditional on a given model relating the intercourse pattern to the fertilization probability, the measurement error distribution is identifiable based on menstrual cycle data provided there are a variety of intercourse patterns and conception outcomes. As in other measurement error applications where validation data are not available [15, 16], repeated measurements can provide additional information about the true underlying variable, in this case the timing of ovulation.

Several recent studies have obtained multiple estimates of ovulation. The European Study of Daily Fecundability independently estimated the last day of hypothermia and the cervical mucus peak [17]. The estimated day-specific probabilities of clinical pregnancy relative to the last day of hypothermia were substantially different from the estimates relative to the mucus peak, suggesting that at least one of the markers must be error prone; a fact verified by the low concordance between the ovulation days estimated by the two markers. The North Carolina Early Pregnancy Study [18] originally estimated ovulation using the rapid decline in the ratio of oestrogen to progesterone metabolites in the urine (day of luteal transition, DLT) [6]. Recently, a second measure of ovulation, LH surge (LHS), was independently estimated using an ultrasensitive immunofluorometric assay of LH [19] applied to the original urine samples. New statistical methods are needed to reconcile differences between estimates based on different markers of ovulation and to take full advantage of the information in the multiple markers.

Dunson and Zhou [11] developed a latent normal mixture model that accounts for the pattern of intercourse relative to ovulation, heterogeneity among couples, and general covariate

effects on fecundability and sterility. In this paper we generalize the Dunson and Zhou model to account for multiple markers of ovulation, and we assign each marker a distinct non-parametric measurement error distribution. As in conventional epidemiology studies which incorporate repeated measures of exposure, we make the simplifying assumption that the measurement errors in the multiple markers of ovulation are independent.

There are important advantages to this approach over methods based on a single estimate of ovulation [12]. First, under an assumed model relating the intercourse pattern to the fertilization probability in each menstrual cycle, the relative values of independent estimates of ovulation are informative about both the true day of ovulation and the measurement error distribution for each marker. By incorporating information from multiple markers, we can greatly improve the accuracy of estimates of the error distributions and the fertility parameters. Second, when there are multiple markers, one marker will often provide an estimated day of ovulation in a cycle for which the other marker is missing (for example, due to incomplete data collection). The proposed method requires only that cycles have at least one estimate of ovulation. Thus cycles can be included in the analysis that otherwise would have been discarded.

Section 2 introduces the model and proposes a Monte Carlo implementation of the EM algorithm for maximum likelihood estimation [20, 21]. Section 3 applies the method to data from the North Carolina Early Pregnancy Study (EPS) to assess the magnitude of error in the DLT and LHS measures of ovulation, and to estimate the day-specific conception probabilities allowing for measurement error in DLT and LHS. Section 4 discusses the results.

2. A MIXTURE MODEL FOR FECUNDABILITY

2.1. Notation

Suppose there are N couples in the study and that couple i is followed for n_i cycles ($i = 1, \ldots, N$). If conception occurs in cycle j for couple i then $Y_{ij} = 1$, otherwise $Y_{ij} = 0$ ($j = 1, \ldots, n_i$). The outcome for each cycle depends on a susceptibility factor (cycle viability) and a timing factor representing the effects of intercourse on various days of the cycle relative to ovulation. We assume that there are K days on which intercourse can potentially result in a conception. Let $X_{ijk} = 1$ if there was intercourse on day k, $k = 1, 2, \ldots, K$, of cycle j for couple i, and 0 otherwise. The intercourse indicators $\mathbf{X}_{ij} = (\mathbf{X}_{ij1}, \ldots, \mathbf{X}_{ijK})$ are indexed relative to the estimated day of ovulation.

When ovulation is estimated incorrectly, X_{ijk} is misindexed by l days, where l is the number of days between the estimated day and the true day. Often multiple measures of ovulation are available for some or all of the cycles. Let W_{ijm} denote the cycle day (where the onset of menstruation occurs on day 1) estimated using the mth marker of ovulation, where $m = 1, 2, \ldots, M$. In the next section we propose a mixture model that uses information available in Y_{ij}, $\mathbf{X}_{ij}$, and $\mathbf{W}_{ij} = (W_{ij1}, \ldots, W_{ijM})$ to account for measurement error in modelling day-specific fecundability.

2.2. Intercourse-based models

Schwartz *et al.* [1] proposed a biologically plausible model to relate the intercourse pattern $\mathbf{X}_{ij}$ to the probability of conception,

$$\Pr\,(Y_{ij} = 1 \mid \mathbf{X}_{ij}, \omega, \mathbf{p}) = \omega \left\{ 1 - \prod_{k=1}^{K} (1 - p_k)^{X_{ijk}} \right\} = \omega\ g(\mathbf{X}_{ij}, \mathbf{p}) \tag{1}$$

where ω is the probability that the cycle is viable (that is, capable of supporting a pregnancy), $\mathbf{p}=(p_1,\ldots,p_K)$, p_k is the conditional probability that conception would have resulted from intercourse on day k if the cycle were viable, and ωp_k is the probability that conception would result from intercourse on only day k. Model (1) assumes that sperm deposited into the reproductive tract on different days commingle and then compete independently for the chance to fertilize the ovum. The ω parameter is the probability that all factors unrelated to the timing of intercourse are favourable for conception to occur and survive to the point of clinical recognition, and $g(\mathbf{X}_{ij},\mathbf{p})$ is the probability that factors related to the timing of intercourse are favourable. In fact, ω may vary from couple to couple and will be 0 for couples who are sterile.

Dunson and Zhou [11] generalized model (1) to allow for heterogeneity among couples and general covariate effects on fecundability and sterility

$$\Pr\,(Y_{ij}=1\,|\,\mathbf{R}_i,\mathbf{U}_{ij},\mathbf{X}_{ij},\boldsymbol{\phi},\boldsymbol{\beta},\mathbf{p},b_i)=\{1-\Phi(\mathbf{R}_i\boldsymbol{\phi})\}\,\Phi(b_i+\mathbf{U}_{ij}\boldsymbol{\beta})\,g(\mathbf{X}_{ij},\mathbf{p}) \tag{2}$$

where $\mathbf{R}_i$ is a $1\times r$ vector of covariates for couple i, $\boldsymbol{\phi}$ is a $r\times 1$ parameter vector linking $\mathbf{R}_i$ to the probability that couple i is sterile, $b_i\sim \mathrm{N}\,(0,\tau^{-1})$ is a couple-specific random effect, $\mathbf{U}_{ij}$ is a $1\times q$ vector of covariates for cycle j from couple i, $\boldsymbol{\beta}$ is a $q\times 1$ parameter vector linking $\mathbf{U}_{ij}$ to the probability that the cycle is viable, and $\Phi(t)=(2\pi)^{-1/2}\int_{-\infty}^{t}\exp(-s^2/2)\,\mathrm{d}s$. Model (2) is built on the assumption that conception in a particular cycle requires the success of three independent events, namely, that the couple is not sterile, the cycle is viable, and the timing of intercourse is favourable for conception.

We extend model (2) to allow for measurement error in ovulation in the case where multiple imperfect markers of ovulation are available. By convention, the intercourse pattern $\mathbf{X}_{ij}$ is indexed relative to the measured day of ovulation based on marker one. In practice we have M markers ($m=1,\ldots,M$) that estimate ovulation to be on days $W_{ij1},\ldots,W_{ijM}$ for cycle j from couple i. Let $\pi_l^{(m)}=\Pr\,(T_{ij}-W_{ijm}=l)$, where T_{ij} is the unknown true day of ovulation. To ensure identifiability we assume that for the first marker of ovulation $\pi_0^{(1)}\geqslant\pi_l^{(1)}$ for all $l\neq 0$. This is equivalent to assuming that marker one has its mode at the true day of ovulation, an assumption which is consistent with validation data for commonly used markers such as the last day of hypothermia and the luteinizing hormone surge [22–25]. We do not constrain the measurement error parameters for the other markers ($m>1$), as it is not necessary for identifiability.

If the true error shift $l=T_{ij}-W_{ij1}$ were known then the intercourse pattern could easily be corrected by letting the new X_{ijk} be the old $X_{ij,k+l}$. We can extend (2) to account for error in estimating ovulation by mixing across the measurement error distribution

$$\begin{aligned}&\Pr\,(Y_{ij}=1\,|\,\mathbf{W}_{ij},\mathbf{R}_i,\mathbf{U}_{ij},\mathbf{X}_{ij},\boldsymbol{\pi},\boldsymbol{\phi},\boldsymbol{\beta},\mathbf{p},b_i)\\&\quad=\{1-\Phi(\mathbf{R}_i\boldsymbol{\phi})\}\Phi(b_i+\mathbf{U}_{ij}\boldsymbol{\beta})\sum_{l=-L_1}^{L_1}p(l\,|\,\mathbf{W}_{ij},\boldsymbol{\pi})\;g(\mathbf{X}_{ij(l)},\mathbf{p})\end{aligned} \tag{3}$$

where $\mathbf{X}_{ij(l)}=(X_{ij,1+l},X_{ij,2+l},\ldots,X_{ij,K+l})$ and the probability of an error shift of l in marker $m=1$ is

$$p(l\,|\,\mathbf{W}_{ij},\boldsymbol{\pi})=\frac{\pi_l^{(1)}\prod_{m=2}^{M}\pi^{(m)}_{l+W_{ij1}-W_{ijm}}}{\sum_{s=-L_1}^{L_1}\pi_s^{(1)}\prod_{m=2}^{M}\pi^{(m)}_{s+W_{ij1}-W_{ijm}}}$$

where $\pi_l^{(1)} = 0$ for $l < -L_1$ and $l > L_1$, and $T_{ij} - W_{ijm}$ is assumed to be independent of $T_{ij} - W_{ijm'}$ for all $m' \neq m$.

Under the assumed form for $g(\cdot)$, the measurement error parameters $\boldsymbol{\pi}$ and the day-specific fertilization probabilities $\mathbf{p}$ are separately identifiable as long as the observed data consist of menstrual cycles with a variety of intercourse patterns, including cycles with multiple intercourse acts in the fertile window. However, if each couple in the study restricted their intercourse to one act in the fertile window per cycle, the measurement error parameters would not be identifiable based on a single marker of ovulation. In this case, multiple markers would allow estimation of a common set of measurement error parameters for all the markers under the assumption of independent measurement errors, but distinct measurement error parameters for each marker would not be estimable. Fortunately, multiple intercourse acts per cycle are common in most studies.

Though the data are informative about the correlation between the measurement errors for the different markers under an assumed form for $g(\cdot)$, we make the simplifying assumption that the measurement errors are uncorrelated. This assumption is reasonable for commonly used markers such as the last day of hypothermia, the cervical mucus peak, LHS and DLT. For basal body temperature, errors are known to occur due to insomnia, rising during the night, or mild febrile illness, such as a cold. The mucus-based methods can be disrupted by medications, recent intercourse or improper assessment of mucus quality. The hormone-based measures are subject to laboratory error. These distinct error mechanisms should generally act separately (independently) on the corresponding marker for ovulation. If correlation does exist, it is likely to be positive. In this case, our method will tend to underestimate the magnitude of measurement error in the markers, and our estimates for the fertility parameters may not be fully corrected for measurement error.

2.3. *EM algorithm*

The EM algorithm can be used to fit model (3). Implementation of the EM is complicated by the need to integrate across the random-effects density. However, Monte Carlo [21] or numerical approximations [26] can be used. To simplify the presentation we assume that there is no sterile subpopulation (that is, $\boldsymbol{\phi} = -\infty$) and that no observed covariates affect cycle viability ($\mathbf{U}_{ij} = 1$). It is straightforward to extend the algorithm to the general case.

2.3.1. Latent data. Weinberg *et al.* [7] used the EM algorithm to fit model (1) by defining two unobservable indicator variables Z_{ij} and S_{ij}, where Z_{ij} indicates that the cycle is viable and S_{ij} indicates the hypothetical success of the sperm, that is, that conception would have occurred if the cycle had been viable. Dunson and Weinberg [12] defined an additional unobservable indicator E_{ijl} that indicates a shift of l days in the true day of ovulation relative to the estimated day in a setting where only one marker of ovulation was available.

We let E_{ijlm} correspond to the indicator of an error of l days for the mth marker of ovulation, where $m = 1, \ldots, M$. If the error shifts were known for any one of the markers then E_{ijlm} could easily be calculated for each of the other markers using $\mathbf{W}_{ij}$. Following Dunson and Zhou [11], we define a latent normal random variable V_{ij} that has mean $b_i + \beta$ and variance 1, and let $Z_{ij} = 1$ if $V_{ij} > 0$ and $Z_{ij} = 0$ if $V_{ij} < 0$. It follows that the cycle viability probability for the jth cycle from couple i is $\Phi(b_i + \beta)$. We propose an EM algorithm to estimate the parameters in model (3) using the latent data $\{S_{ij}, E_{ijlm}, V_{ij}, b_i\}$.

2.3.2. Maximization step. If the latent data were directly observable we would simply maximize the complete data log-likelihood to estimate the unknown parameters. In the case where marker one is observed for all cycles the complete data log-likelihood is proportional to

$$\sum_{i=1}^{N}\left[\log\,\tau-\tau\,b_i^2+\sum_{j=1}^{n_i}\left\{-(V_{ij}-b_i-\beta)^2+\sum_{m=1}^{M}\sum_{l=-L_1}^{L_1}E_{ijlm}\log\,\pi_l^{(m)}\right.\right.$$
$$\left.\left.+\sum_{l=-L_1}^{L_1}S_{ij}E_{ijl1}\log\,g(\mathbf{X}_{ij(l)},\boldsymbol{p})+(1-S_{ij})E_{ijl1}\,\log(1-g(\mathbf{X}_{ij(l)},\boldsymbol{p}))\right\}\right] \tag{4}$$

which is the sum of four components that each have a distinct parameter vector: $l_1(\tau;\mathbf{b})$; $l_2(\beta;\mathbf{V},\mathbf{b})$; $l_3(\boldsymbol{\pi};\mathbf{E})$, and $l_4(\boldsymbol{p};\mathbf{S},\mathbf{E},\mathbf{X})$. This separation is possible due to conditional independence assumptions implicit in model (3). The complete data likelihood can easily be extended to allow marker one to be missing for some of the cycles.

We can maximize (4) by separately maximizing the four components. There are closed form expressions for the conditional maximum likelihood estimates of τ, β, and $\boldsymbol{\pi}$

$$\hat{\tau}=N\Big/\sum_{i=1}^{N}b_i^2,\qquad \hat{\beta}=\sum_{i=1}^{N}\sum_{j=1}^{n_i}(V_{ij}-b_i)\Big/\sum_{i=1}^{N}n_i,\qquad \hat{\pi}_l^{(m)}=\left(\sum_{i=1}^{N}\sum_{j=1}^{n_i}E_{ijlm}\right)\Big/\sum_{i=1}^{N}n_i$$

for $m=1,\ldots,M$. There is no closed form solution for the conditional MLE of $\boldsymbol{p}$. However, l_4 can be expressed as a Bernoulli log-linear model since $\log\{1-g(\mathbf{X}_{ij(l)},\boldsymbol{p})\}=\sum_{k=1}^{K}X_{ij,k+l}\log(1-p_k)$. It is biologically reasonable to assume that the day-specific probabilities of conception increase to a peak and then decrease. Therefore, one may want to estimate $\mathbf{p}$ under the inequality constraint: $p_1\leqslant p_2\leqslant\cdots\leqslant p_*\geqslant\cdots\geqslant p_K$, where the peak is unknown. The pooled-adjacent-violators algorithm [27] can be used to estimate $\mathbf{p}$ under this constraint, and the peak day can be chosen to maximize the observed data likelihood.

2.3.3. Expectation step. To carry out the E-step we calculate expectations of sufficient statistics of the parameters $\boldsymbol{\Theta}=(\tau,\beta,\boldsymbol{\pi},\boldsymbol{p})$ with respect to the posterior distribution of the latent data. From Section 2.3.2 we know that sufficient statistics for β, $\boldsymbol{\pi}$, and $\boldsymbol{p}$ are linear in V_{ij}, E_{ijlm} and $S_{ij}E_{ijl1}$, respectively, and $\sum_{i=1}^{N}b_i^2$ is sufficient for τ. If we could generate samples $\{b_i^{(t)},V_{ij}^{(t)}\}$, $t=1,\ldots,T$, from the joint posterior distribution of b_i and V_{ij}, we could approximate conditional expectations of the sufficient statistics from

$$E\left(\sum_{i=1}^{N}b_i^2\,|\,\mathbf{D},\hat{\boldsymbol{\Theta}}\right)=\int\sum_{i=1}^{N}b_i^2\,\mathrm{d}f(b_1,\ldots,b_N\,|\,\mathbf{D},\hat{\boldsymbol{\Theta}})\approx\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}b_i^{(t)2}$$
$$E(V_{ij}\,|\,\mathbf{D},\hat{\boldsymbol{\Theta}})=\int V_{ij}\,\mathrm{d}f(V_{ij}\,|\,\mathbf{D},\hat{\boldsymbol{\Theta}})\approx\frac{1}{T}\sum_{t=1}^{T}V_{ij}^{(t)}$$
$$E(E_{ijlm}\,|\,\mathbf{D},\hat{\boldsymbol{\Theta}})=\int\Pr\,(E_{ijlm}=1\,|\,b_i,\mathbf{D},\hat{\boldsymbol{\Theta}})\,\mathrm{d}f(b_i\,|\,\mathbf{D},\hat{\boldsymbol{\Theta}})$$
$$\approx\frac{1}{T}\sum_{t=1}^{T}\Pr(E_{ijlm}=1\,|\,b_i^{(t)},\mathbf{D},\hat{\boldsymbol{\Theta}})$$

$$E(S_{ij}E_{ijl1} \mid \mathbf{D}, \hat{\boldsymbol{\Theta}}) = \int \Pr\ (S_{ij}=1, E_{ijl1}=1 \mid b_i, \mathbf{D}, \hat{\boldsymbol{\Theta}})\, \mathrm{d}f(b_i \mid \mathbf{D}, \hat{\boldsymbol{\Theta}})$$
$$\approx \frac{1}{T}\sum_{t=1}^{T} \Pr\ (S_{ij}=1 \mid E_{ijl1}=1, b_i^{(t)}, \mathbf{D}, \hat{\boldsymbol{\Theta}})\ \Pr\ (E_{ijl1}=1 \mid b_i^{(t)}, \mathbf{D}, \hat{\boldsymbol{\Theta}})$$

where $\mathbf{D} = \{Y_{ij}, \mathbf{W}_{ij}, \mathbf{X}_{ij}\}$ represents the observed data. Unfortunately, it is difficult to sample efficiently from the joint posterior distribution of V_{ij} and b_i.

We instead alternately sample from the full conditional distributions of V_{ij} and b_i, which follow simple forms. According to Gibbs sampling theory these samples will converge to the joint posterior density under mild regularity conditions [28–30]. The posterior distribution of b_i conditional on the observed data $\mathbf{D}$, the current parameter estimates $\hat{\boldsymbol{\Theta}}$, and the latent $V_{i1}, \ldots, V_{in_i}$ for couple i is

$$\mathrm{N}\left((\tau + n_i)^{-1}\sum_{j=1}^{n_i}(V_{ij}-\beta), (\tau+n_i)^{-1}\right)$$

For conception cycles the full posterior distribution of V_{ij} is

$$\frac{I_{[V_{ij}>0]}\ \exp\{(V_{ij}-b_i-\beta)^2/2\}}{\sqrt{(2\pi)}\Phi(b_i+\beta)}$$

which is $\mathrm{N}(b_i+\beta, 1)$ truncated below by 0. For non-conception cycles the posterior distribution of V_{ij} is

$$\left[\frac{I_{[V_{ij}<0]} + \widehat{\Pr}(S_{ij}=0) I_{[V_{ij}>0]}}{1-\Phi(b_i+\beta)\widehat{\Pr}(S_{ij}=1)}\right] \times \left[\frac{1}{\sqrt{(2\pi)}}\ \exp\{(V_{ij}-b_i-\beta)^2/2\}\right] \tag{5}$$

where $\widehat{\Pr}(S_{ij}=1) = \sum_{l=-L_1}^{L_1} p(l \mid \mathbf{W}_{ij}, \hat{\boldsymbol{\pi}})\, g(\mathbf{X}_{ij(l)}, \hat{\boldsymbol{p}})$ is the probability that conception would have occurred had the cycle been viable. To sample from (5) let $s \sim \mathrm{Bernoulli}\{\widehat{\Pr}(S_{ij}=1)\}$. If $s=0$, $V_{ij} \sim \mathrm{N}(b_i+\hat{\beta}, 1)$, otherwise $V_{ij} \sim \mathrm{N}(b_i+\hat{\beta}, 1)$ truncated above by 0.

We require the probability that marker m is shifted by l days from the true day of ovulation conditional on the observed data, the random-effect b_i, and $\hat{\boldsymbol{\Theta}}$. For conception cycles

$$\Pr(E_{ijlm}=1 \mid b_i, \mathbf{D}, \hat{\boldsymbol{\Theta}}) = p(l + W_{ijm} - W_{ij1} \mid \mathbf{W}_{ij}, \hat{\boldsymbol{\pi}}) \frac{g(\mathbf{X}_{ij(l+W_{ijm}-W_{ij1})}, \hat{\boldsymbol{p}})}{\widehat{\Pr}(S_{ij}=1)}$$

For non-conception cycles

$$\Pr(E_{ijlm}=1 \mid b_i, \mathbf{D}, \hat{\boldsymbol{\Theta}}) = p(l + W_{ijm} - W_{ij1} \mid \mathbf{W}_{ij}, \hat{\boldsymbol{\pi}}) \frac{1-\Phi(b_i+\hat{\beta})\, g(\mathbf{X}_{ij(l+W_{ijm}-W_{ij1})}, \hat{\boldsymbol{p}})}{1-\Phi(b_i+\hat{\beta})\widehat{\Pr}(S_{ij}=1)}$$

We also require $\Pr(S_{ij}=1, E_{ijl1}=1 \mid b_i, \mathbf{D}_{ij}, \hat{\boldsymbol{\Theta}})$, which equals the conditional probability of an error shift of l in the first marker when $Y_{ij}=1$ and equals

$$p(l + W_{ijm} - W_{ij1} \mid \mathbf{W}_{ij}, \hat{\boldsymbol{\pi}}) \left[\frac{\{1-\Phi(\hat{\mu}_{ij})\} g(\mathbf{X}_{ij(l+W_{ijm}-W_{ij1})}, \hat{\boldsymbol{p}})}{1-\Phi(\hat{\mu}_{ij})\widehat{\Pr}(S_{ij}=1)}\right]$$

when $Y_{ij} = 0$. This is the joint probability that $Y_{ij} = 0$, $S_{ij} = 1$ and $E_{ijl1} = 1$ divided by the probability that $Y_{ij} = 0$.

Beginning with initial estimates of the parameters, the algorithm alternates between approximating the conditional expectations of each component of the log-likelihood (Monte Carlo E-step) and estimating the conditional MLEs of each of the parameters (M-step). It is straightforward to extend this algorithm to allow for general covariate effects, multiple random effects, and a sterile subpopulation of couples.

3. EARLY PREGNANCY STUDY

3.1. Data

We applied the methods to data from the North Carolina Early Pregnancy Study of 221 healthy women who enrolled at the time they discontinued contraception in order to conceive [18]. Women collected daily urine samples and kept daily records of intercourse during a six-month follow-up period. Ovulation was later estimated for each menstrual cycle by (i) measures of oestrogen and progesterone metabolites in the urine, the ratio of which decreases abruptly with luteinization of the ovarian follicle [6], and (ii) an immunofluorometric assay of urinary lutenizing hormone (LH), which rises abruptly around the time of ovulation [19]. We refer to measure one as the day of luteal transition (DLT) and measure two as the LH surge (LHS). There is evidence to suggest that DLT and the day of LHS do not vary systematically from each other or from the time of rupture of the ovarian follicle [22, 31]. DLT was taken to be marker one, that is, we assume that the true day of ovulation is the day most likely to be identified using DLT.

Implantation (defined here as a measurable rise in urinary excretion of human chorionic gonadotropin) occurred in 199 out of 740 menstrual cycles. There were 151 clinical pregnancies and 48 early losses (lost within 6 weeks of last menstrual period). We were able to estimate ovulation using either DLT or LHS for 708 cycles and 219 women. DLT was available for 696 cycles, LHS for 676 cycles, and both for 654 cycles. The distribution of DLT minus LHS is plotted in Figure 1. In 80 per cent of the cycles the two measures differ by one day or less. We use the methods described in Section 2 to evaluate the performance of DLT and investigate the relationship between timing of intercourse and the probability of pregnancy.

3.2. Analysis

We model the probability of a clinically-recognized conception, treating early losses as non-events. We fit the data using: (a) model 1 (the Schwartz *et al.* model); (b) model 3 (the model accounting for error in measurement of ovulation) with a fixed probability of cycle viability for all couples; (c) model 3 with a normal density for the latent cycle viability variable. We assume that the day-specific probabilities increase to a peak and then decrease. It is straightforward to fit (a) and (b) using a standard EM algorithm, and we use the Monte Carlo algorithm outlined in Section 2.3 to fit (c).

Based on Figure 1, it appears very unlikely that the error in DLT or LHS exceeds three days. Therefore, we chose a $[-3, 3]$ support for both measurement error distributions. Without correcting for measurement error, Wilcox *et al.* [33] observed that the day-specific probabilities of clinical pregnancy were positive in a five day interval ending on the day of ovulation. Based

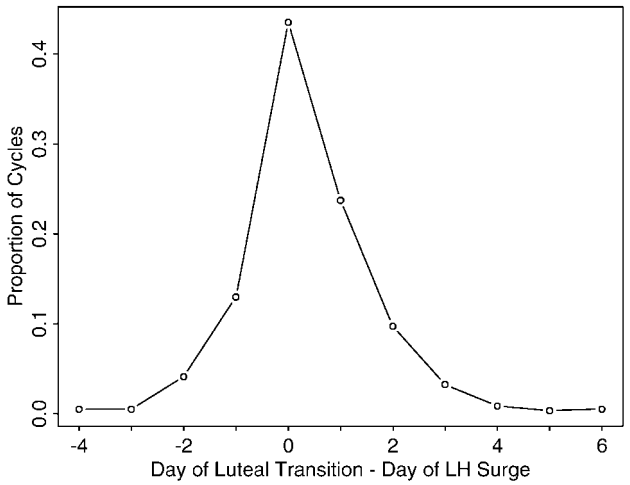

Figure 1. Distribution of day of luteal transition (DLT)-day of LH surge (LHS).

on their results, we assume that the day-specific probabilities are zero outside of a seven-day window (-5, $+1$ around ovulation).

We assessed convergence of the Monte Carlo EM algorithm by monitoring the observed data log-likelihood, which we calculated at each iteration, and by examining diagnostic plots of the sampled parameters. We also repeated each analysis under a variety of widely different starting values. In each case, convergence was apparently achieved by 10 000 iterations. To further verify convergence, we collected an additional 100 000 iterates for one set of starting values. The resulting estimates were essentially identical to the pooled estimates from the other chains. The estimates are not sensitive to the initial values and plots of the parameter values at each iteration behave as expected (that is, the parameters change rapidly at first then more slowly as the maximum is approached).

The estimated day-specific conception probabilities under each model are shown in Figure 2. Comparing (a) and (b), we can see that the estimates that account for measurement error are lower on the edge of the fertile interval (that is, days $-4, -3, -1, 0$) and are higher in the centre of the fertile interval (that is, day -2). As expected, the random-effects estimates that account for both measurement error and heterogeneity among couples are uniformly higher than the fixed-effects estimates. Numerical values and standard errors for the estimated cycle viability probability and the conditional day-specific conception probabilities within the $[-4, 0]$ fertile interval are presented in Table I. As expected, the standard errors are slightly higher for the models that account for measurement error. In each case, the estimated standard errors are low for the cycle viability probability and are high for the conditional day-specific conception probabilities. Thus, if covariates such as age or smoking exposure were incorporated into the model for the cycle viability probability we would expect reasonable power to detect effects. However, we would expect low power to detect small to moderate differences between groups in the pattern of day-specific probabilities. Such differences could be better assessed in a larger study, such as the European Study of Daily Fecundability [17].

The estimated cycle viability density under the random-effects model is shown in Figure 3. The mean cycle viability is 0.30 and there is considerable heterogeneity among

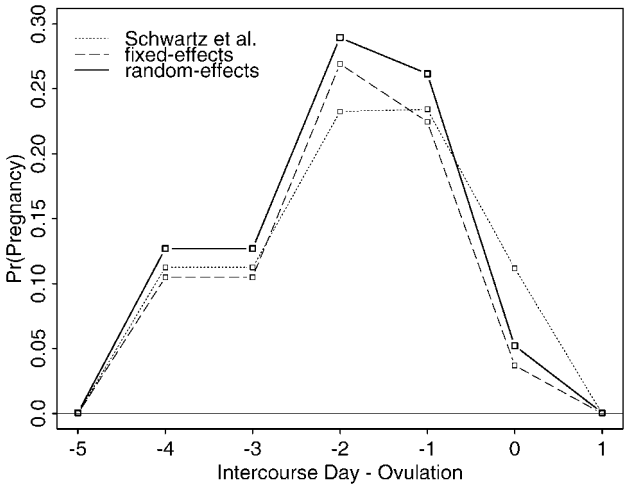

Figure 2. Estimated day-specific probabilities of clinical pregnancy under: the order-restricted Schwartz *et al*. model; the fixed-effects model allowing for measurement error in both markers; the random-effects model allowing for measurement error in both markers.

Table I. Estimated fertility parameters for the North Carolina Early Pregnancy Study under the order-restricted Schwartz *et al.* model, the fixed-effects model that allows for measurement error in both markers, and the random-effects model that allows for measurement error in both markers.

Parameter	Schwartz		Fixed-effects		Random-effects	
	MLE	SE	MLE	SE	MLE	SE
ω	0.258	0.033	0.275	0.040	0.271	0.037
p_{-4}	0.443	0.192	0.382	0.208	0.433	0.221
p_{-3}	0.443	0.236	0.382	0.285	0.433	0.309
p_{-2}	0.905	0.510	0.979	0.612	0.988	0.432
p_{-1}	0.905	0.442	0.817	0.309	0.892	0.392
p_0	0.428	0.230	0.135	0.353	0.177	0.393

The estimate for ω under model (c) is for a typical couple (that is, $b_i = 0$).

couples. The estimated measurement error distributions under the random-effects model are plotted in Figure 4. These estimates are very similar to the estimates under the fixed-effects model, suggesting that the estimated measurement error distributions may be robust to specification of the cycle viability model. Based on Figure 4, it appears that both measures have a high probability of correctly identifying the true day of ovulation, and both error distributions are skewed; an error in DLT or LHS appears more likely to result from identifying a day that occurs before the true day of ovulation.

Model (3) is based on the assumption of independent measurement errors. If the error in DLT, $T_{ij} - W_{ij1}$, and the error in LHS, $T_{ij} - W_{ij2}$, are positively correlated, model (3) will underestimate the measurement error probabilities $\pi_l^{(1)}$ and $\pi_l^{(2)}$ for $l \neq 0$. However, we expect that the correlation between these measurement errors is small. The day of luteal transition (DLT) is estimated based on changes in oestrogen and progesterone, whereas the day of the

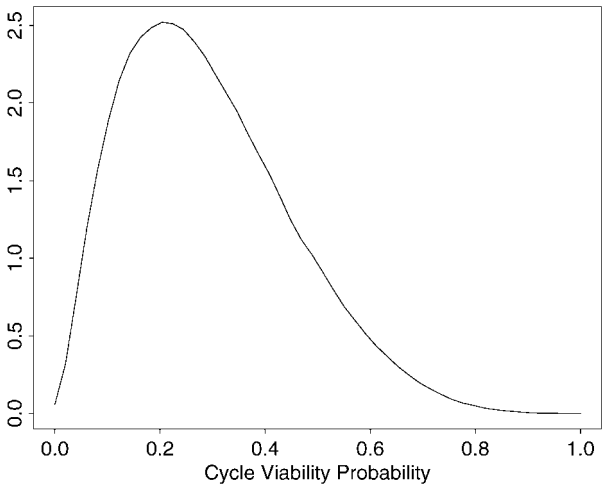

Figure 3. Estimated cycle viability density function under the random-effects model.

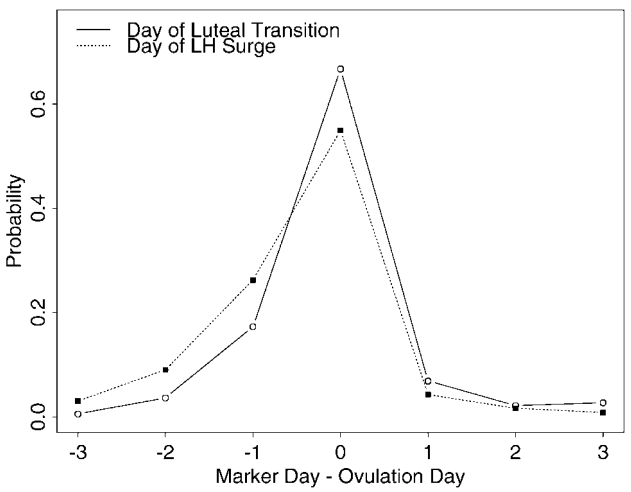

Figure 4. Measurement error distribution for DLT and day of LH surge (LHS).

LH surge (LHS) is based on the rapid increase in luteinizing hormone production which accompanies ovulation. In hormonally aberant cycles the error in these measures may be correlated, for example, if a weak LHS is associated with a late rise in progesterone, both markers may identify ovulation day late. However, the major contributor to error is expected to be assay error which should be uncorrelated for the two separate assays conducted at different laboratories.

4. DISCUSSION

We have proposed a method of accounting for measurement error in identifying ovulation within a general mixture model that allows multiple covariates to affect human fertility. The model assigns a non-parametric error distribution to each marker of ovulation. Under independence of the measurement errors and an assumed model for the fertilization probability, the set of measurement error distributions can be estimated jointly with the error-corrected fertility parameters using a Monte Carlo EM algorithm.

Our results agree with earlier observations that clinical pregnancy in the North Carolina Early Pregnancy Study cohort occurred only with intercourse in a five-day interval ending on the day of ovulation [30]. Although the measurement error correction did not change the estimate of the fertile interval, the estimated day-specific probabilities within the interval were modified. In particular, we estimate that the probability a pregnancy results from intercourse on only the day of ovulation is slightly less than 5 per cent, which is lower than previously estimated.

An important advantage of the proposed method is that it provides estimates of the measurement error distribution for each marker of ovulation without requiring validation data. Direct biological validation of a marker requires data on the time of follicular rupture, which is difficult and expensive to obtain. In fact, direct validation may not be possible using the conventional method of regular ultrasound monitoring for evidence of follicular rupture, since ultrasound-related heat stimulation can potentially induce ovulation. In our analysis of the North Carolina study the probability that the DLT marker correctly identified the true day of ovulation was 0.65, suggesting that the decline in the ratio of oestrogen to progesterone metabolites in the urine corresponds closely to ovulation. This provides the most direct assessment of the performance of the DLT marker to date [6]. However, care should be taken to avoid overinterpretation of this estimate since it is based on assumptions that, though biologically reasonable, have not been directly validated.

In order to make meaningful comparisons across fertility studies that utilize different markers of ovulation, it is crucial to account for measurement error in identifying ovulation [34]. Otherwise, the conception probabilities may be biased due to differences in the markers instead of biological differences between cohorts. The North Carolina study used two measures of ovulation based on hormonal measures. Both markers appear to occur on or slightly before the true day of ovulation. However, more commonly used measures such as the last day of hypothermia or the cervical mucus peak should be much more error-prone; a conjecture supported by the much wider fertile intervals estimated in studies using these markers [17, 35]. Differences in the markers may explain at least some of the differences in the day-specific conception probabilities between the North Carolina study, the Barrett and Marshall study [35] and the European Study of Daily Fecundability [17]. In a previous analysis [34] we compared the day-specific fecundabilities for North Carolina and British couples using a method that adjusts for measurement error when a single marker of ovulation is available for each cycle [12]. In future work we plan to apply the methodology presented here to jointly analyse data from the North Carolina study and the large multinational European study [17], which utilizes basal body temperature and cervical mucus measurements to identify ovulation.

REFERENCES

1. Schwartz D, MacDonald PDM, Heuchel V. Basal body temperature, ovulation and the risk of conception, with special reference to the lifetime of sperm and egg. *Population Studies* 1980; **34**:397–400.
2. Royston JP. Basal body temperature, ovulation and the risk of conception, with special reference to the lifetime of sperm and egg. *Biometrics* 1982; **38**:397–406.
3. Weinberg CR, Wilcox AJ. A model for estimating the potency and survival of human gametes *in vivo*. *Biometrics* 1995; **51**:405–412.
4. Seppala M, Koistinen H, Mandelin E *et al.* Glycodelins: role in regulation, potential for contraceptive development and diagnosis of male infertility. *Human Reproduction* 1998; **13** Suppl. **3**:262–269.
5. Perloff WH, Steinberger E. In vivo survival of spermatozoa in cervical mucus. *American Journal of Obstetrics and Gynecology* 1964; **88**:439–442.
6. Baird DD, Weinberg CR, Wilcox AJ, McConnaughey DR, Musey PI. Using the ratio of urinary oestrogen and progesterone metabolites to estimate day of ovulation. *Statistics in Medicine* 1991; **10**:255–266.
7. Weinberg CR, Gladen BC, Wilcox AJ. Models relating the timing of intercourse to the probability of conception and the sex of the baby. *Biometrics* 1994; **50**:358–367.
8. Zhou H, Weinberg CR. Modelling conception as an aggregated Bernoulli outcome with latent variables via the EM algorithm. *Biometrics* 1996; **52**:945–954.
9. Zhou H, Weinberg CR, Wilcox AJ, Baird DD. A random-effects model for cycle viability in fertility studies. *Journal of the American Statistical Association* 1996; **91**:1413–1422.
10. Dunson DB, Weinberg CR. Accounting for unreported and missing intercourse in human fertility studies. *Statistics in Medicine* 2000; **19**:665–679.
11. Dunson DB, Zhou H. A Bayesian model for fecundability and sterility. *Journal of the American Statistical Association* 2000; (under revision).
12. Dunson DB, Weinberg CR. Modelling human fertility in the presence of measurement error. *Biometrics* 2000; **56**:288–292.
13. Bongaarts J. The proximate determinants of natural marital fertility. In: *Determinants of fertility in developing countries. Vol. 1*, Bulatao RA, Lee RD, Hollerbach PE, Bongaarts J (eds). Academic Press: New York, 1983; 103–138.
14. Carroll RJ, Freeman LS, Kipnis V, Li L. A new class of measurement-error models, with applications to dietary data. *Canadian Journal of Statistics* 1998; **26**:467–477.
15. Freedman LS, Carroll RJ, Wax Y. Estimating the relation between dietary-intake obtained from a food frequency questionnaire and true average intake. *American Journal of Epidemiology* 1991; **134**:310–320.
16. Landin R, Freedman LS, Carroll RJ. Adjusting for time trends when estimating the relationship between dietary-intake obtained from a food frequency questionnaire and true average intake. *Biometrics* 1995; **51**:169–181.
17. Masarotto G, Romualdi C. Probability of conception on different days of the menstrual cycle: an ongoing exercise. *Advances in Contraception* 1997; **13**:105–115.
18. Wilcox AJ, Weinberg CR, Baird DD. Timing of sexual intercourse in relation to ovulation. *New England Journal of Medicine* 1995; **319**:189–194.
19. Kesner JS, Knecht EA, Krieg EF, Wilcox AJ, O'Connor JF. Detecting pre-ovulatory lutenizing hormone surges in urine. *Human Reproduction* 1998; **13**:15–21.
20. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, *Series B* 1977; **39**:1–38.
21. Wei G, Tanner MA. A Monte Carlo implementation of the EM algorithm and the poor mans data augmentation algorithms. *Journal of the American Statistical Association* 1991; **85**:699–704.
22. Collins WP, Branch CM, Collins PO, Sallam HM. Biochemical indices of the fertile period in women. *International Journal of Fertility*, 1983; **26**:196.
23. France JT, Graham FM, Gosling L, Hair P, Knox BS. Characteristics of natural conception cycles occurring in a prospective study of sex preselection: fertility awareness symptoms, hormone levels, sperm survival, and pregnancy outcome. *International Journal of Fertility*, 1994; **37**:244–255.
24. Lloyd R, Coulam CB. The accuracy of urinary luteinizing hormone testing in predicting ovulation. *American Journal of Obstetrics and Gynecology* 1989; **160**:1370–1375.
25. Guida N, Tommaselli GA, Palomba S, Pellicano M, Moccia G, Di Carlo C, Nappi C. Efficacy of methods for determining ovulation in a natural family planning program. *Fertility and Sterility* 1999; **72**:900–904.
26. Abramowitz M, Stegun IA. *Handbook of Mathematical Functions* Dover Publications: New York, 1987.
27. Barlow RE, Bartholomew DJ, Bremner JM, Brunk HD. *Statistical Inference Under Order Restrictions: the Theory and Application of Isotonic Regression*. Wiley: London, 1972.
28. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Patern Analysis and Machine Intelligence* 1984; **6**:721–741.
29. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 1987; **82**:528–550.

30. Tierney L. Markov chains for exploring posterior distributions. *Annals of Statistics* 1994; **22**:1701–1762.
31. Collins WP, Branch CM, Collins PO, Sallam HM. Biochemical indices of the fertile period in women. *International Journal of Fertility* 1983; **26**:196.
32. Baird DD, McConnaughey DR, Weinberg CR, Musey PI, Collins DC, Kesner JS, Knecht EA, Wilcox AJ. Application of a method for estimating day of ovulation using urinary estrogen and progesterone metabolites. *Epidemiology* 1995; **6**:547–550.
33. Wilcox AJ, Weinberg CR, Baird DD. Post-ovulatory aging of the human oocyte and embryo failure. *Human Reproduction* 1998; **13**:394–397.
34. Dunson DB, Baird DD, Wilcox AJ, Weinberg CR. Day-specific probabilities of clinical pregnancy based on two studies with imperfect measures of ovulation. *Human Reproduction* 1999; **14**:1835–1839.
35. Barrett JC, Marshall J. The risk of conception on different days of the menstrual cycle. *Population Studies* 1969; **23**:455–461.