Biometrics 67, 1573–1582
December 2011
DOI: 10.1111/j.1541-0420.2011.01558.x

A Hot-Deck Multiple Imputation Procedure for Gaps in Longitudinal Recurrent Event Histories

Chia-Ning Wang,[1,*] Roderick Little,[1] Bin Nan,[1] and Siobán D. Harlow[2]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
[2]Department of Epidemiology, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
[*]*email:* cnwang@umich.edu

Summary. We propose a regression-based hot-deck multiple imputation method for gaps of missing data in longitudinal studies, where subjects experience a recurrent event process and a terminal event. Examples are repeated asthma episodes and death, or menstrual periods and menopause, as in our motivating application. Research interest concerns the onset time of a marker event, defined by the recurrent event process, or the duration from this marker event to the final event. Gaps in the recorded event history make it difficult to determine the onset time of the marker event, and hence, the duration from onset to the final event. Simple approaches such as jumping gap times or dropping cases with gaps have obvious limitations. We propose a procedure for imputing information in the gaps by substituting information in the gap from a matched individual with a completely recorded history in the corresponding interval. Predictive mean matching is used to incorporate information on longitudinal characteristics of the repeated process and the final event time. Multiple imputation is used to propagate imputation uncertainty. The procedure is applied to an important data set for assessing the timing and duration of the menopausal transition. The performance of the proposed method is assessed by a simulation study.

Key words: Hormone treatment; Menopause; Menstrual periods; Missing data; Predictive mean matching; Terminal event.

1. Introduction

Data on recurrent events are recorded in prospective longitudinal studies where participants experience a particular event repeatedly. Examples includes epileptic seizures, asthma, migraine episodes, and menstruation. In many studies, the intensity or frequency of the recurrent events changes as a disease progresses, or in response to physiological changes. Such changes in recurrent events are often used to define markers that indicate different phases of disease progression or different life stages. Investigators are typically interested in estimating the occurrence time of events that mark onset of new phases or stages. In many biomedical applications, this recurrent event process can be stopped permanently by a terminal event such as death, or as in our motivating application, the final menstrual period (FMP); and the phase/stage marker events may be useful in predicting the final or terminal event. Research interest often concerns the duration time from the event marking onset of a phase/stage to the final event.

We consider here the problem of gaps in events histories when information is not recorded, which create uncertainty about the times of marker events. We address the problem by a hot-deck multiple imputation procedure in which the missing data in the gap is multiply imputed based on information obtained from donor records, matched to the recipient on relevant longitudinal characteristics by an extension of a predictive mean matching (PMM; Rubin, 1986; Little, 1988).

1.1 *Motivating Application*

Our methods are motivated by analyses for ReSTAGE, a multi-study collaboration to evaluate the association between ages at onset of the bleeding marker events that define the onset of the menopausal transition and age at menopause, the end of reproductive life (Harlow et al., 2006, 2008). Menopause is defined as the FMP, which is confirmed by at least 12 months of amenorrhea. Several marker events for the menopausal transition based on menstrual bleeding criteria have been proposed (Mitchell, Woods, and Mariella, 2000; Soules et al., 2001; Taffe and Dennerstein, 2002a, 2002b; Lisabeth et al., 2004). ReSTAGE compared the reliability, reproducibility, and representativeness of these markers and assessed how well they predict time to FMP (Harlow et al., 2007). Definitions of these bleeding markers are based on the intervals between menstruation times, which are called menstrual cycles. An example is the first occurrence of a menstrual cycle longer than 60 days, which is a marker for onset of the late menopausal transition (Lisabeth et al., 2004; Harlow et al., 2006).

We apply our proposed method to the TREMIN study, one of four large cohort studies that provide the data for ReSTAGE. TREMIN enrolled 1997 female students at the University of Minnesota between 1935 and 1939 and followed them up to 40 years throughout their reproductive life (Treloar et al., 1967). The women used menstrual diary cards to record the days when they experienced menstrual bleeding or spotting. The first day of menstrual bleeding/spotting was coded as an event in the menstrual histories and the interval between adjacent events defines the length of a menstrual cycle. Other information such as pregnancies and medical treatments including hormone therapy (HT) was collected via annual questionnaires. Gaps occur in the menstrual record when

women failed to record their menstrual data or when they used HT, because HT masks the menstrual cycles that would have occurred in the absence of HT use. Incorporation of the observed data for individuals with gaps requires a strategy for addressing large gaps in the menstrual record—when a marker has not occurred prior to a gap, it is often unclear whether or not a marker occurred during a gap.

1.2 *Limitations of Existing Approaches*

Existing approaches to the gap problem are (a) to discard cases that have any gaps in their recurrent event records; (b) to treat women with gaps as censored at the time of gap initiation; and (c) to ignore, or jump, the gap and compute the time of the marker based on the information recorded before and after the gap. Approach (a) results in considerable information loss, and could lead to biased estimation; (b) discards information recorded after gaps, and is potentially biased if the HT censoring mechanism is informative; and (c) retains all the recorded information but implicitly assumes that the marker did not occur during the gap and hence tends to bias results by pushing dates of markers into the future. Which of these would you choose? Our analyses provide rather unexpected information on the relative biases of these approaches.

1.3 *Outline of Proposed Approach*

Our approach is to fill in missing information in the gaps using a form of hot-deck imputation (Andridge and Little, 2010) where the case with a gap (called the recipient) is matched to a similar case with no gaps (called the donor), and the number and times of recurrent events within the recipient's gap are imputed using information from the donor. This imputation approach has several advantages. First, it is relatively nonparametric and avoids the need to build a parametric model to estimate the times of recurrent events. In the case of menstrual cycle histories, such a model is difficult to develop because the length and variability of cycles change across time in varied and complex ways. Second, as imputed event times are based on real rather than simulated data, the imputed event times within the gap are likely to be realistic. Third, once the gaps are filled in, standard complete-data methods can be applied to answer a wide spectrum of research questions, with the missing information in the gaps being handled in a consistent way. Imputation uncertainty is incorporated by creating multiply imputed data sets with different matches of donors to recipients, and then applying simple multiple imputation (MI) combining rules (Little and Rubin, 2002).

An earlier version of this approach matched donors and recipients with respect to the time (i.e., age) at the start and end of the gap (Little et al., 2008). We refine that method by an extension of PMM that allows an extensive set of relevant recurrent event characteristics and covariates to be included in the match between donors and recipients. This has two important advantages. First, the method assumes the data are missing at random (MAR), that is after conditioning on the variables used to define donor-recipient matches, the distribution of events within the gap is the same for recipients as for donors. This assumption is weaker and more defensible when an extensive set of covariate information is included in the matching metric. Second, failure to condition on covariate information leads to attenuation of the estimated relationship between the covariates and the variables being imputed. In our example, women who have experienced more variable and more unstable menstrual patterns are more likely to have gaps in their menstrual calendars, either because they fail to record information or because they use HT. Menstrual cycle variability increases as women approach FMP, so failure to condition on information about FMP in imputing the gaps may bias estimated relationships between markers and FMP.

We now describe the proposed method in more detail. In Section 3, we apply the method to the TREMIN data, and compare results for a particular menopausal marker with alternative methods for handling the missing data. In Section 4, we describe a simulation study that examines the statistical properties of the proposed method. Section 5 concludes with discussion, including other potential applications.

2. The Imputation Procedure

We propose an extension of PMM (see Rubin, 1986; Little, 1988; Heitjan and Little, 1991) to incorporate longitudinal covariate information in the match between donors and recipients. To match simultaneously on multiple covariates, PMM defines a metric to measure the "closeness" (or distance) between the subjects based on the predictive mean. Suppose values of a variable Y are missing and for a subject i with y_i missing, $x_i = (x_{i1}, \ldots, x_{iK})$ are the values of K covariates for that subject i. The distance between subject i and a potential donor subject j is defined as

$$d(i, j) = (\hat{Y}(x_i) - \hat{Y}(x_j))^2,$$

where $\hat{Y}(x_j)$ is the predicted value of Y from the regression of Y on X, estimated from the complete cases. After the distances are computed, all subjects j with y_j observed and $d(i, j) < \delta$, a prespecified maximum distance δ, are selected to constitute the donor pool for nonrespondent i. From this donor pool, one donor j' is randomly drawn, and the observed value $y_{j'}$ from the donor is used to replace the missing value y_i for nonrespondent i.

We cannot apply PMM directly to our problem because we need to impute a set of recurrent events within gaps of varying lengths. Also, the information used for selecting matched donors includes the longitudinal event characteristics before and after the gap. To adapt PMM for this setting, we use summary statistics for data reduction and multivariate regressions to compute predictive means for each summary statistic. We then compute distances between donors and recipients based on the multivariate predictive means. For ease of presentation, we describe the procedure in the context of the TREMIN application. In Section 5, we discuss applications to other settings.

Consider a gap for a woman i, and let $start_i$ and end_i be the age at the start and the end of the gap—in our application, these ages coincide with times of menstruation. The gap is imputed as follows:

(1) A set S_i of potential donors j is selected who have menstrual bleeds at age $start_j$ close to $start_i$ and end_j

close to end_i, with the $ratio = \frac{end_i - start_i}{end_j - start_j}$ between L = 1/(1 + *a*) and U = 1/(1 − *a*), and completely recorded menstrual histories between $start_j$ and end_j. The choice of *a*, which we call the bound ratio, is a trade-off between the number of potential donors and similarity to the recipient on the interval of interest. We chose a value of $a = 0.2$, which is a reasonable choice illustrated by a plot of *a* against the distribution of number of potential donors, as shown in Web Figure 1.

(2) For each potential donor *j* in S_i, we calculate summary statistics *Y* of the cycle lengths within the gap (from age $start_j$ to end_j). In our application, we use the median and interquartile range (IQR) of the cycle lengths, $Y.median_j$ and $Y.IQR_j$, to summarize the longitudinal recurrent event patterns within chosen intervals. These measures, rather than the mean and standard deviation, are chosen to limit the influence of outliers.

(3) For each potential donor *j* in S_i, we calculate the predictors X_j that are input into our predictive mean model. In our application, these consisted of four adjacent running medians and IQRs before and after the gaps, age at FMP and the censoring indicator for whether FMP is or is not observed. For ease of computation burden, running medians and IQRs of the cycle lengths are calculated using a one-year bandwidth with a half-year step size. Four adjacent running medians and IQRs before and after the gaps, which cover 2.5 years of menstrual cycles, respectively, can capture the menstrual patterns well for the period before and after the gaps.

(4) Estimate the multivariate regression with outcomes ($Y.median$, $Y.IQR$) and covariates *X*, based on the cases in S_i.

(5) Let $\hat{Y}_i$ and $\hat{Y}_j$ be the predicted value of ($Y.median$, $Y.IQR$) for subjects *i* and *j* from the multivariate regression, and *G* is the residual covariance matrix of ($Y.median$, $Y.IQR$). The distance between subject *i* and subject *j* in S_i is calculated as

$$d(i,j) = (\hat{Y}_i - \hat{Y}_j)G^{-1}(\hat{Y}_i - \hat{Y}_j),$$

which is the difference in predicted values of ($Y.Median, Y.IQR$), scaled by the inverse of the residual covariance *G*. We select $D = 10$ donors in S_i who have the smallest distance from woman *i* as the donor pool. As discussed below, the impact of the choice of *D* is assessed by comparing key results for different values of *D*, and it appears to be minor in our application.

(6) For each recipient *i*, select a donor (say j') randomly from the *D* closest donors found in Step 5.

(7) The cycles from $start_{j'}$ to $end_{j'}$ are used to impute the cycles from $start_i$ to end_i for each recipient *i*. Specifically, let n_j be the number of events between $start_j$ and end_j, and $l_{j1}, l_{j2}, \ldots, l_{jn_j}$ be the lengths of the n_j cycles between $start_j$ and end_j. The imputed lengths of cycles for subject *i* are $l_{i1} = l_{j1}{}^{*}ratio, l_{i2} = l_{j2}{}^{*}ratio, \ldots, l_{in_j} = l_{jn_j}{}^{*}ratio$, where $ratio = \frac{end_i - start_i}{end_j - start_j}$ adjusts for the difference in the lengths of the gap and ranges from 0.83 to 1.25 by construction of S_i.

Multiply imputed data sets are obtained by repeating steps (6) and (7). Note that the ages at the start and the end of the gaps are not included in the PMM model as predictors, but these ages are implicitly conditioned because the predictions are focused on information for that interval.

3. Applications

We applied the proposed imputation approach to the TREMIN data. In the interest of space, we present here findings for just one marker of the late menopausal transition, the first occurrence of a cycle of more than 60 days. The age at onset of the 60-day marker and the duration of the late transition, that is, the time to FMP, were estimated after missing gaps had been imputed by our proposed imputation procedure. We compared these estimates to three other approaches for handling missingness: censoring at the start of the first gap, jumping the gaps, and multiply imputing the gaps conditional only on age.

Data for this analysis include the subset of 735 women in the original cohort who were still participating at age 35, the baseline for our analysis of the menopausal transition, and who subsequently provided a minimum of 10 consecutive menstrual cycles. After age 35, women participated from 0.84 years to 24.28 years with a median of 14.27 years, and contributed 12 to 322 segments (median = 169). Among the 735 women, 331 (45.03%) women reached their FMP and the rest were treated as censored when they withdrew (n = 150, 20.41%), had hysterectomy (n = 99, 13.47%), or initiated HT use at the end of their menstrual record (n = 155, 21.09%).

More than 60% of women had gaps in their menstrual histories (n = 475, 64.63%) in TREMIN either because they failed to record menstrual information (missing gaps, n = 305) or because they used HT but stopped treatment before the end of their menstrual record (HT gaps, n = 269). A total of 1331 gaps (861 missing gaps and 470 HT gaps) were listed in the menstrual histories with a median length of 0.16 years for missing gaps and 1.03 years for HT gaps. Figure 1 illustrates the times of menstrual bleeds as well as missing gaps and HT gaps, for a systematic sample of 1 in 10 women. Missing gaps or HT gaps were treated indistinguishably and were imputed by the same procedure, an approach that is valid under MAR. Women may have had more than one gap, and for simplicity, each gap was imputed independently using the proposed PMM approach. In this application, we constrained the imputation procedure to permit a maximum gap length of 4 years. If a woman had a longer gap, she was censored at the start of the gap and treated as a withdrawal (n = 33).

3.1 *Imputation and Analysis*

The PMM algorithm of the previous section was applied to create five MI data sets. Donor sets of size $D = 10$ were created for each gap. For comparison purposes, gaps were also imputed conditional only on ages at the start and the end of the gaps (Little et al., 2008).

We then estimated age at onset of the markers using imputed data sets and compared results from imputations conditional only on age and imputations that used the PMM approach to condition on age, menstrual characteristics

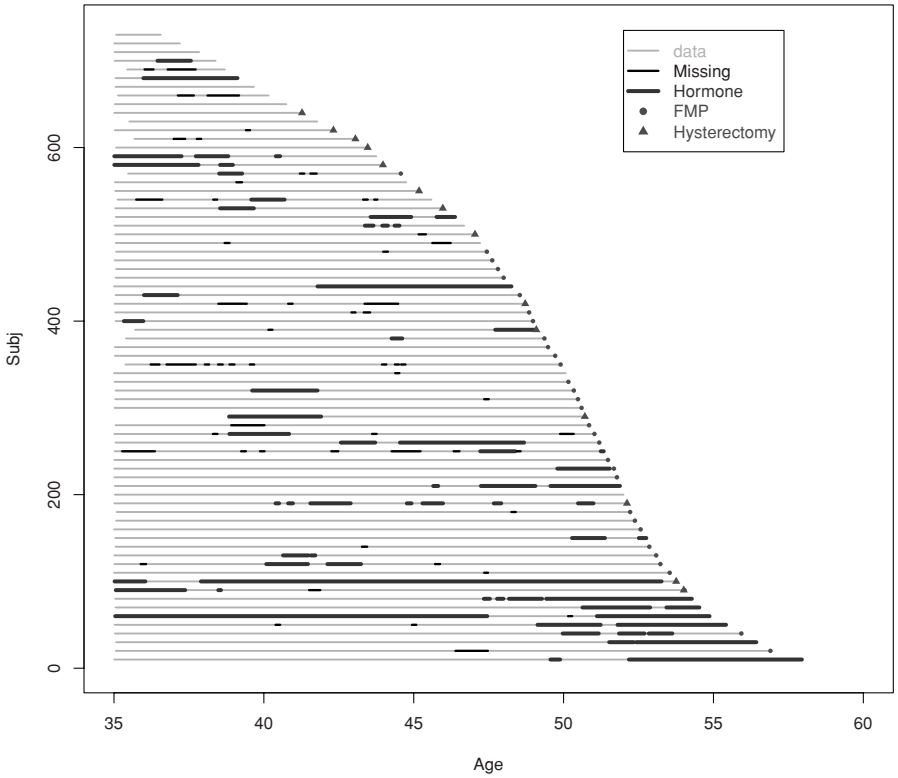

Figure 1. Observation pattern for TREMIN data. Horizontal lines indicate where menstrual bleeds were recorded, missing or during HT gaps, for every 10th woman.

before and after the gaps and FMP status. These results also were compared with the results when we censored at the first gap and when we jumped gaps in the data. Kaplan–Meier (KM) curves were calculated for describing the distributions of marker ages.

Table 1 presents the median age of the 60-day marker and associated 95% confidence limits for

(a) COMPLETE: the complete cases with no gaps;
(b) CENSOR: the data censored at the first gap;
(c) JUMP: the data jumping over the gaps;
(d) AGEMI: the "gap age" MI procedure that conditions only on the ages at the start and end of the gaps; and
(e) PMMI: the MI procedure with PMM.

The median from PMMI is slightly higher than the median from AGEMI, 0.4 years less than the median from JUMP, and 0.7 years less than the median from CENSOR. If the PMMI estimate is not seriously biased, as is supported by the simulations in the next section, then both JUMP and CENSOR have a substantial bias, with the CENSOR estimate even more biased than the JUMP estimate. This finding is unexpected, and evidence that the censoring mechanism is informative. The 95% confidence interval from PMMI is slightly narrower than the corresponding interval from AGEMI, consistent with a gain in efficiency from conditioning on the covariate information.

Figure 2 shows the KM curves of the age of the 60-day marker for the five methods. The KM curves for the two MI approaches are very close, and slightly lower than the curve from JUMP. Because JUMP assumes the marker has not occurred during the gap, it tends to overestimate the age of the marker. The KM curve from CENSOR is substantially higher than the other curves, suggesting evidence of bias from informative censoring if the MI methods are valid. One possible explanation for informative censoring is that women are more likely to take HT or fail to record their bleeding cycles when they begin to experience changes in menstrual cycles or symptoms such as hot flashes, which often emerge as menopause approaches. Treating women as censored at the time of gap initiation and mistakenly assuming independent censoring potentially introduces bias and overestimate the age of the 60-day marker.

The median time from 60-day marker to FMP was estimated using the method of Lin, Sun, and Ying (1999), which accounts for the fact that the probability that the duration is censored increases with the age of marker. Table 1 presents the median years of the duration and associated 95% confidence limits. The medians for AGEMI and PMMI are 2.76 and 2.81 years, respectively, compared with 2.60 years for complete cases, 2.59 years from jumping the gap, and 4.11 years from censoring at first gap. Again, CENSOR is divergent from the other methods, evidence of informative censoring.

The two MI methods do not differ much with respect to these aggregate results, but the value of PMM in generating better matches becomes apparent when we examine imputations for individual cases. For example, Figure 3 shows the

Table 1
Medians and 95% CIs of the 60-day marker and the duration time from the marker to FMP

				Marker onset age		Duration	
	N	Marker	FMP	Median	95% CI	Median	95% CI
COMPLETE	260	147	103	48.32	(47.19, 48.94)	2.60	(2.10, 3.07)
CENSOR	735	203	132	48.76	(48.48, 49.25)	4.11	(2.96, 4.91)
JUMP	735	487	317	48.48	(48.02, 48.77)	2.59	(2.28, 2.83)
AGEMI				48.04	(47.66, 48.59)	2.76	(2.52, 3.03)
Imputed no. 1	735	478	310	48.16	(47.82, 48.65)	2.74	(2.48, 2.98)
Imputed no. 2	735	482	313	48.04	(47.62, 48.59)	2.77	(2.48, 3.03)
Imputed no. 3	735	478	312	48.02	(47.62, 48.53)	2.82	(2.56, 3.05)
Imputed no. 4	735	480	312	48.04	(47.68, 48.52)	2.77	(2.48, 2.98)
Imputed no. 5	735	478	311	48.13	(47.80, 48.65)	2.76	(2.48, 3.05)
PMMI				48.11	(47.77,48.53)	2.81	(2.53, 3.08)
Imputed no. 1	735	475	308	48.11	(47.80, 48.50)	2.79	(2.53, 3.08)
Imputed no. 2	735	470	308	48.13	(47.80, 48.60)	2.83	(2.58, 3.08)
Imputed no. 3	735	474	308	48.13	(47.78, 48.52)	2.76	(2.48, 3.04)
Imputed no. 4	735	474	310	48.04	(47.68, 48.59)	2.87	(2.59, 3.12)
Imputed no. 5	735	478	313	48.13	(47.82, 48.56)	2.80	(2.53, 3.05)

sequences of menstrual lengths for a selected woman with a missing gap between ages 42.64 and 45.75 (ID 4613, first row). Note that this woman had very variable menstrual cycles after the gap, and reached FMP at age 52.07. Rows 2–4 show imputed sequences for the three closest donors from AGEMI (left column) and PMMI (right column). The imputations from PMMI match the characteristics of the individual after the gap much more closely than the imputations from

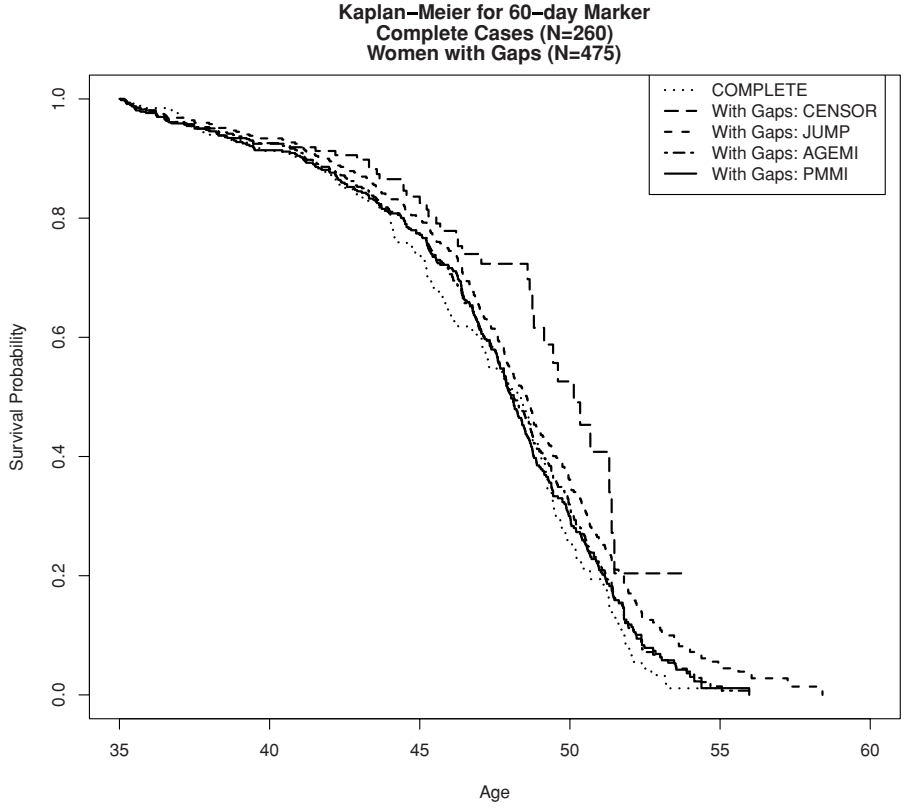

Figure 2. KM estimates for the age of the 60-day marker.

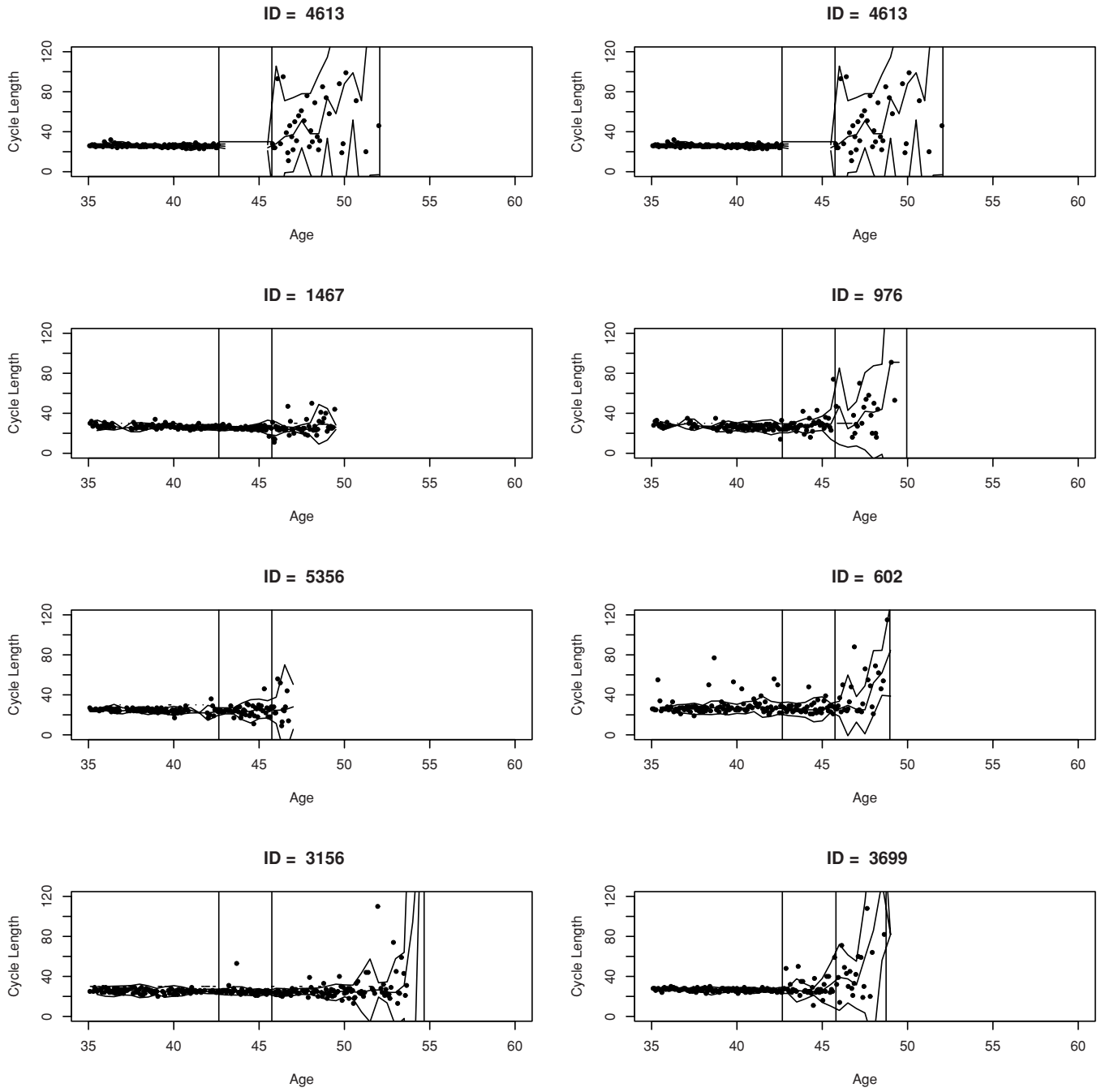

Figure 3. The sequences of menstrual lengths for a selected woman with a missing gaps (first row, ID=4613) and for three closest donors from AGEMI (left column) and from PMMI (right column). The vertical lines between age 42.64 and age 45.75 demarcate the missing gap and the corresponding age period and FMP is also indicated by the vertical line if FMP is observed.

AGEMI. This individual is chosen to illustrate the point, but PMMI produces better matches for other cases as well.

3.2 *Sensitivity Analysis*

We also performed sensitivity analyses to evaluate the effects of the bound ratio a and size of donor pool D. Web Tables 1 and 2 show that the estimated median age of the 60-day marker ranges from 48.10 to 48.13, when a is chosen from 10% to 50% or D is chosen from 10 to 50. The 95% CI widens slightly as D or a increase. Web Figure 1 shows the distribution of number of potential donors for different choices of a, and Web Figure 2 shows the average of D smallest distances with D ranging from 10 to 50. From the results, it can been seen that $a = 0.2$ results in most of the donors being included in the imputation set, and $D = 10$ results in the donors within the donor pools all having reasonably small distances. These

Table 2
Estimates of marker onset age and its biases

	Scenario 1			Scenario 2			Scenario 3a			Scenario 3b		
Marker time	Bias	CI Width	Coverage (%)	Bias	CI Width	Coverage (%)	Bias	CI Width	Coverage (%)	Bias	CI Width	Coverage (%)
Based on subjects whose marker happened after the start of the gap												
Before deletion	52.23	1.35		51.34	0.78		50.45	0.91		49.92	0.87	
JUMP	0.64	1.20		0.58	0.59		0.60	0.84		0.00	0.87	
AGEMI	−0.16	1.29		−0.38	0.87		0.00	0.87		−0.08	0.89	
PMMI	−0.01	1.36		−0.04	0.83		0.00	0.94		−0.00	0.87	
Based on all subjects												
Before deletion	49.94	0.48	94.4	49.94	0.48	94.4	49.94	0.48	94.4	49.94	0.48	94.4
COMPLETE	−0.46	0.50	5.6	−0.56	0.56	3.4	−0.11	0.59	87.8	0.01	0.58	95.4
CENSOR	−0.07	0.48	90.8	−0.11	0.51	86.0	0.25	0.56	59.6	0.01	0.58	95.2
JUMP	0.15	0.49	76.8	0.22	0.53	63.6	0.19	0.47	64.2	0.00	0.48	94.6
AGEMI	0.03	0.48	93.2	−0.09	0.48	85.0	0.03	0.48	93.8	−0.02	0.49	94.6
PMMI	0.01	0.48	95.4	−0.01	0.49	95.8	0.00	0.49	95.6	−0.00	0.48	95.4

two figures and the results from the sensitivity analysis provide the evidence that our choices of $a = 0.2$ and $D = 10$ are appropriate.

4. Simulation Study

4.1 *Data Generation and Missing Gap Mechanism*

We conducted a simulation study to assess the performance of PMMI and the other methods. We simulated recurrent cycle length data for a population of 100,000 women. From this population, a total of 500 complete random sample data sets of size 1000 were generated, and gaps with missing cycles were created under different missing data mechanisms. These missing gaps were imputed and the differences in the estimates based on the imputed data sets and the full data sets before deletion were assessed. All simulations and data analysis were performed using the software package `R`.

The population data were generated using the model for longitudinal menstrual cycles in Huang, Harlow, and Elliott (2010) to model the TREMIN data. To our knowledge, this is the most sophisticated model for menstrual data in the literature to date that specifically models changes in length and variability that precede the FMP. It includes subject-specific change points for the mean and variance of the menstrual cycle lengths as a function of age. Like any parametric model it may not be entirely realistic, but it allows us to simulate a large and plausible population of menstrual cycles for testing the various imputation methods. (Details about the generating procedure are described in Web Appendix B.) Among the 100,000 generated subjects, 87% of them reached FMP at an average age of 53.94 years. Subjects who did not reach FMP were treated as censored after their last cycle, and the mean censoring age is 60.13.

Gaps with missing cycles were created by four missing gap mechanisms for representing different scenarios. In scenario 1, missing gaps become more likely as the subjects approached FMP, whereas in scenario 2, the probability of a missing gap depends on age alone, and increases with age. In scenario 3a and scenario 3b, missing gaps occur with constant probability, but in scenario 3a, missing gaps occur only after age 45, whereas in scenario 3b, gaps occur before age 45, when most subjects have not yet been censored or reached the final event. (Details about missing gap mechanisms are provided in Web Appendix B.)

In all these scenarios, because missing gaps can only happen before subjects reach the final events or are censored, the final event time or the censoring time needs to be included into the imputation model to avoid violating the MAR assumption. These models led to approximately 30% subjects with a missing gap uniformly distributed between 2 and 3 years in their event history. The number of subjects with missing gaps are 287.3, 297.0, 290.9, and 309.7 on average in each scenario.

4.2 *Analysis and Results*

The five missing data adjustment methods—COMPLETE, CENSOR, JUMP, AGEMI, and PMMI were applied to each data set, and compared with the results applied to the data set without gaps. For the two MI methods, $M = 5$ imputed data sets were created, and estimates and standard errors computed using the MI combining rules. The marker "first cycle of longer than 60 days" was chosen for these comparisons.

We first compared estimates of the distribution of age at marker. In the generated population, 99% of subjects had this marker event observed, with a median age of 49.95 and a mean age of 50.03. Table 2 shows the estimates of marker age and biases of the various methods, based on all subjects and restricted to subjects for whom the gap preceded the marker, because these are the subjects affected by the missing data method. In scenario 3b, where missing gaps were generated before age 45, that is before the marker tends to occur, all methods perform fine. In other scenarios, COMPLETE underestimates and JUMP overestimates the age at the marker onset, and both these methods suffer from confidence interval undercoverage. CENSOR is biased and undercovers in Scenarios 2 and 3a. The MI methods have uniformly smaller empirical bias and are closer to nominal coverage. The PMMI is consistently the best, with smallest bias and close to nominal confidence coverage in all scenarios.

We also compared estimates of the duration from the marker to the final event. (Details are described in Web

Table 3
Bias of the characteristics of imputed cycles

	AGEMI		PMMI	
	Mean	SD	Mean	SD
Bias in median of the cycle lengths				
Scenario 1	−4.15	15.56	0.22	8.76
Scenario 2	−2.57	11.85	0.17	5.10
Scenario 3a	−7.13	15.76	−0.29	8.52
Scenario 3b	−1.06	4.66	0.00	1.56
Bias in IQR of the cycle lengths				
Scenario 1	−4.35	16.38	−1.25	11.23
Scenario 2	−0.45	11.32	−0.42	6.51
Scenario 3a	−4.94	15.05	−1.41	10.07
Scenario 3b	−0.20	3.17	−0.12	1.86
Bias in the number of cycles				
Scenario 1	5.09	12.11	0.14	4.89
Scenario 2	3.16	12.73	−0.22	4.66
Scenario 3a	8.23	12.73	0.42	4.83
Scenario 3b	1.56	6.51	0.01	1.84

Appendix C.) Web Table 3 presents the median duration and empirical biases when missing gaps are treated by different approaches, for all subjects and restricted to the subjects with a gap before the marker. In scenario 3b, empirical bias is small and confidence coverage is satisfactory for all the methods. In other scenarios, JUMP underestimates the median duration time for subjects where the gap preceded the marker by 0.5 to 0.9 years, and has poor confidence coverage. COMPLETE also underestimates the duration and suffers from undercoverage. CENSOR performs better, but has less than nominal confidence coverage for scenario 2. The MI methods again have small empirical bias and close to nominal coverage, with PMMI being superior to AGEMI.

The quality of imputed cycles in the missing gaps is the foundation for a wide spectrum of other statistical analyses that may be applied to the data. Accordingly, we also compared the number, median, and IQR of imputed and actual cycles in the gap, for the two MI methods. The empirical bias of estimates of these three quantities are shown for the two MI methods in Table 3. AGEMI tends to impute too many cycles in the missing gaps, and these imputed cycles are shorter and less variable than the actual ones. This tendency is minor for scenario 3b where all missing gaps were generated before age 45; at these ages, subjects have more stable cycle patterns and are very similar to each other, so conditioning only on age at the beginning and end of the gap yields satisfactory matches. AGEMI imputations are less satisfactory in scenario 3a, where missing gaps were generated after age 45, and cycles tend to be longer and more variable. In this and other scenarios, the PMMI method yields imputed cycles that match the number and characteristics of the actual cycles much more closely.

4.3 *Summary*

For the estimates of age at marker and for the estimates of the duration, PMMI and AGEMI both perform very well, whereas PMMI is slightly better than AGEMI with less bias and close nominal confidence converge. Comparisons of individual imputed cycles show that PMMI imputes cycles that resemble the simulated cycles more closely. That is, PMMI provides more realistic imputations of the missing cycles than AGEMI. This increases our confidence in statistical analyses applied to the data multiply imputed by the PMMI method.

Concerning the performance of other approaches, JUMP overestimates the age at marker and underestimates the duration, which is intuitive, because JUMP forces the later onset of marker. COMPLETE underestimates both the age at marker and the duration. By missing gap mechanisms, we can see that people with longer event histories (therefore, later marker onset or longer duration) are more likely to have gaps in their event histories. Consequently, due to the violation of missing completely at random, COMPLETE induces biases for the estimates. The effects of CENSOR are different in different scenarios; in some scenarios, CENSOR overestimates but in other scenarios it underestimates the parameters.

5. Discussion

We have proposed a regression-based hot-deck MI method for imputing gaps in longitudinal histories involving recurrent events and the final event. To the best of our knowledge, the only prior study dealing with missing gaps in recurrent event histories is Little et al. (2008), and in that article the missing gaps are imputed conditional only on the time, that is age, at the start and end of the gaps. We use an extension of PMM to incorporate information on the longitudinal recurrent event patterns before and after the gap and other important covariates, such as age at the final event. Simulations suggest that the proposed approach has good statistical properties.

The proposed methodology is developed in the context of data on menstrual bleeding, but the idea of using a hot deck with PMM, with suitable modification for special features, has potential applications in other situations where the times of recurrent events are recorded longitudinally, and gaps exist in the records of individual subjects. Possible applications include diary or panel studies of asthmas, seizures, or migraines. Key features of our proposed approach are (i) choice of appropriate summary statistics, such as the median and IQR in our application, to capture characteristics of the missing events within the gaps; (ii) the estimation of regression models relating these characteristics to covariates that are predictive of the recurrent event patterns within the gaps, including summaries of the recurrent events adjacent to the gap; (iii) the selection of donor sets based on PMM, with parameters estimated from the regression; and (iv) multiple imputation of information within the gaps based on the information from randomly selected donors within the donor set. The resulting imputed data sets can be analyzed as complete longitudinal recurrent event data, and estimates combined using MI combining rules. We recommend that future users perform sensitivity analyses with different sets of covariates, interval bounds or potential donor sizes. Because a separate regression is fitted for each gap, the procedure is computation intensive, but manageable with current computing power, particularly

if the regression models are linear. The proposed method is easily programmed using `R` or `SAS` macro algorithms.

One possible application of the proposed imputation approach is to surveillance data collected to monitor the extensive sanitation program implemented by the Bahia state government (Brazil) since 1997. One data set focuses on the incidence and prevalence of diarrhea in children up to 3 years or age. Daily data on diarrhea episodes are available on 926 children from home visits over 455 days from 2000 to 2002. Information gaps occur in these daily-recorded data, mainly when the data collector was not available (Strina et al., 2003; Borgan et al., 2007). Another possible application is to analyses of the Copenhagen Studies on Asthma in Childhood, which include two longitudinal cohort studies with 411 and 800 mothers and their children, respectively. Parents prospectively reported respiratory and skin-related symptoms, treatment, and history of common childhood infections in daily diaries, and gaps occur from lapses of recording.

An alternative to our hot-deck approach to handling gaps in event histories data is to posit a multivariate model for the recurrent event times conditional on the covariate information, and apply methods such as maximum likelihood or Bayesian inference based on the observed data (Little and Rubin, 2002). An example in our motivating application is the Bayesian model of Huang et al. (2010) which was used to generate the population for our simulation study. The complexity of the recurrent event models and resulting likelihoods are drawbacks to this approach, and sensitivity to model misspecification is a concern. Our proposed hot-deck MI method is simpler and less parametric, in that the regression models used to construct predictive means are only used to develop matches between donors and recipients, and imputed cycles are based on existing donor cycle information.

In the PMM approach, matched donors are selected based on the predicted means from regression models, where the outcomes are summary statistics of the events in the gap, and the predictors are covariates and summary statistics before and after the gap that are predictive of these summary statistics. The strength of the predictive mean metric is that covariates that are more predictive of the recurrent events are given more weight than covariates that are less predictive. These weights are determined empirically for each gap, using regression methods. Other metrics could be used for selecting matched donors, such as the Mahalanobis distance or the maximum deviation (Little and Rubin, 2002, Chapter 4), but these are inferior in our view because they do not distinguish between covariates that are highly predictive of the missing values and variables that are weakly predictive.

As with other hot-deck approaches, our PMM approach is more appropriate in large samples, where there is a substantial pool of donors, than in small samples, where more parametric imputation approaches may be more effective. A hybrid approach may be attractive when the sample size is small and the donor pool is limited. One such approach would be to create a pseudosample using a parametric model, such as the model used to simulate data in our simulation study, with the parameters estimated from the original sample. This pseudosample is combined with the original sample to increase the pool of potential donors, and the proposed PMM method applied to the enlarged sample. Because the parametric model is used only for generating possible donors, and the imputation is still based on the proposed PMM, this hybrid approach is likely to be more robust to model misspecification than a purely parametric imputation method. The details and assessment of such hybrid methods need further study.

Our proposed imputation approach did well in our simulations, where it outperformed other approaches such as discarding the event histories with gaps, jumping gaps, or censoring at the beginning of the gaps, which are potentially biased and generally involve some loss of observed information. Useful further analyses of our method would be to conduct sensitivity analyses to assess the impact of different choices of covariates and summary statistics before and after gaps as predictors in the PMM model. Sensitivity analyses could also be developed to capture and assess deviations from MAR, by introducing offsets between the predictive means of donors and recipients to reflect differences in their predictive distributions.

The PMM imputation in this study uses multiple imputation to reflect the major component of imputation uncertainty, but it is improper (Rubin, 1987) in that it fails to reflect all the uncertainty in the use of sample estimates to create the donor sets. The simulation study suggests that this is not a major problem here, perhaps because the fraction of missing information is relatively small. However, the procedure could be made proper by applying a method such as the approximate Bayesian bootstrap (Rubin and Schenker, 1986) when selecting the donors for imputation. For each MI data set, we would first create a random sample of these subjects by drawing subjects at random with replacement, and then, apply the proposed PMM approach to each random sample and create the imputations within the gap. Then appropriate estimates and associated variances could be acquired using MI combining rule.

In this article, a hot-deck MI method has been proposed for imputing missing gaps in longitudinal recurrent event data. The gaps are imputed based on observed information from matched donors conditional on longitudinal patterns and important covariates. This proposed method without involving complicated models can be easily implemented and makes good use of observed information with gaps. We applied this proposed method to menstrual bleeding data for assessing the menopausal transition and FMP. And the simulation study is also performed. The simulation findings show that this proposed method provides substantial gain in terms of reduced bias and increased efficiency over other approaches.

6. Supplementary Materials

Web Appendixes, Figures, and Tables are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

ReSTAGE is supported by grant AG 021543 (SDH, PI) from the National Institute on Aging. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Aging or the National Institutes of Health. We appreciate the helpful comments of two referees.

References

Andridge, R. R. and Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review* **78,** 40–64.

Borgan, O., Fiaccone, R. L., Henderson, R., and Barreto, M. L. (2007). Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in Brazil. *Scandinavian Journal of Statistics* **34,** 53–69.

Harlow, S. D., Cain, K., Crawford, S., Dennerstein, L., Little, R. J. A., Mitchell, E. S., Nan, B., Randolph, John, F. J., Taffe, J., and Yosef, M. (2006). Evaluation of four proposed bleeding criteria for the onset of late menopausal transition. *Journal of Clinical Endocrinology and Metabolism* **91,** 3432–3438.

Harlow, S. D., Crawford, S., Dennerstein, L., Burger, H., Mitchell, E. S., and Sower, M. F. (2007). Recommendations from a multi-study evaluation of proposed criteria for staging reproductive aging. *Climacteric* **10,** 112–119.

Harlow, S. D., Mitchell, E. S., Crawford, S., Nan, B., Little, R. J. A., and Taffe, J. (2008). The ReSTAGE collaboration: Defining optimal bleeding criteria for onset of early menopausal transition. *Fertility and Sterility* **89,** 129–140.

Heitjan, D. F. and Little, R. J. A. (1991). Multiple imputation for the fatal accident reporting system. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **40,** 13–29.

Huang, X., Harlow, S. D., and Elliott, M. R. (2010). Modelling menstrual cycle length at the approach of menopause using Bayesian changepoint models. University of Michigan Working Paper 87, Ann Arbor, Michigan.

Lin, D. Y., Sun, W., and Ying, Z. (1999). Nonparametric estimation of the gap time distribution for serial events with censored data. *Biometrika* **86,** 59–70.

Lisabeth, L. D., Harlow, S. D., Gillespie, B., Lin, X., and Sowers, M. F. (2004). Staging reproductive aging: A comparison of proposed bleeding criteria for the menopausal transition. *Menopause* **11,** 186–197.

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics* **6,** 287–296.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edition. New York: Wiley.

Little, R. J. A., Yosef, M., Cain, K. C., Nan, B., and Harlow, S. D. (2008). A hot-deck multiple imputation procedure for gaps in longitudinal data on recurrent events. *Statistics in Medicine* **27,** 103–120.

Mitchell, E. S., Woods, N., and Mariella, A. (2000). Three stages of the menopausal transition from the Seattle Midlife Women's Health Study: Toward a more precise definition. *Menopause* **7,** 334–349.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics* **4,** 87–94.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* **81,** 366–374.

Soules, M. R., Sherman, S., Parrott, E., Rebar, R., Santoro, N., Utian, W., and Woods, N. (2001). Executive summary: Stages of reproductive aging workshop (STRAW). *Fertility and Sterility* **76,** 874–878.

Strina, A., Cairncross, S., Barreto, M. L., Larrea, C., and Prado, M. S. (2003). Childhood diarrhea and observed hygiene behavior in Salvador, Brazil. *American Journal of Epidemiology* **157,** 1032–1038.

Taffe, J. R. and Dennerstein, L. (2002a). Menstrual patterns leading to the final menstrual period. *Menopause* **9,** 32–40.

Taffe, J. R. and Dennerstein, L. (2002b). Time to the final menstrual period. *Fertility and Sterility* **78,** 397–403.

Treloar, A. E., Boynton, R. E., Behn, B. G., and Brown, B. W. (1967). Variation of the human menstrual cycle through reproductive life. *International Journal of Fertility* **12,** 77–126.

Received June 2010. *Revised November* 2010.
Accepted November 2010.