Assessing menstrual cycles with urinary hormone assays


1Division of Reproductive Endocrinology, Department of Obstetrics, Gynecology and Women’s Health, Albert Einstein College of Medicine, Bronx, New York 10461; 2New England Research Institutes, Watertown, Massachusetts 02472; 3Departments of Epidemiology and Preventive Medicine and Population Health and Reproduction, University of California at Davis, Davis 95616; 4Division of Geriatrics and Center for Health Sciences, University of California at Los Angeles, Los Angeles, California 90095; 5Reproductive Sciences Program and Department of Epidemiology, University of Michigan, Ann Arbor, Michigan 48109; and 6Department of Obstetrics, Gynecology and Women’s Health, University of Medicine and Dentistry of New Jersey, Newark, New Jersey 07103

Submitted 29 August 2002; accepted in final form 15 November 2002

URINARY HORMONE DETERMINATIONS have been demonstrated to be useful in field studies of women with a variety of reproductive conditions (1–3, 5–6, 9–10, 13, 15, 17, 22–25, 31, 35–42). By use of urinary hormone metabolites as proxy markers of circulating gonadotropin and sex steroid hormones, it was possible to determine the window of fertility in a woman’s cycle (3, 41), as well as to describe the effects of exercise (6), smoking (14, 21, 42), weight (40), and reproductive aging (31) on hormone patterns.

The day-to-day patterns of secretion of LH, FSH, estradiol, and progesterone are duplicated reliably by use of an overnight urine specimen (24–25, 29). Correction of the urinary hormone value with creatinine excretion reduces within women variations in hormone concentrations, which can be considerable (25). Preservation of urine with glycerol is required to maintain gonadotropin activity in some (31) but not all (35) assays.

The Study of Women’s Health Across the Nation (SWAN) is a multisite, multiethnic, longitudinal study of midlife women (34). Among the goals of SWAN is the characterization of the reproductive hormone patterns as women approach and traverse the menopausal transition. The Daily Hormone Substudy within SWAN will collect and analyze cycles of daily urinary hormones annually, thus providing both cross-sectional and longitudinal data.

Unlike most previous studies in field settings, the women in SWAN are being observed as their cycles become irregular and eventually cease. We note that cycle patterns of women as they traverse the menopause are relatively unexplored. The application of algorithms designed to identify cycles with evidence of luteal activity among younger, regularly cycling women may be less reliable as reproductive aging occurs. We therefore sought to identify in a prospective fashion the characteristics of a presumably ovulatory

Address for reprint requests and other correspondence: N. Santoro, Division of Reproductive Endocrinology, Dept. of Obstetrics, Gynecology and Women’s Health, Albert Einstein College of Medicine, 1300 Morris Park Ave., Mazer 316, Bronx, NY 10461 (E-mail: glicktoro@aol.com).

http://www.ajpendo.org 0193-1849/03 $5.00 Copyright © 2003 the American Physiological Society
menstrual cycle that would appear to be most robust and to incorporate them into an algorithm with which to assess ovarian function in SWAN’s sample of women. We examined previously tested algorithms (2, 15, 40) and assessed their ability to provide a standardized, robust, and objective assessment of presumed ovarian status that would maintain predictive power as our participants proceed toward menopause.

MATERIALS AND METHODS

Participant Population

The Daily Hormone Substudy of SWAN enrolled women from all seven SWAN clinical sites (see citation in ACKNOWLEDGMENTS). Eligibility criteria were as follows: 1) an intact uterus and at least one ovary present at the time of recruitment; 2) at least one menstrual period in the 3 mo before recruitment; 3) no use of sex steroid hormones within 3 mo before recruitment; 4) not currently pregnant. 5) Since all women enrolled in the SWAN Study were asked to complete daily menstrual calendars that evaluated symptoms in addition to bleeding, compliance with the menstrual calendar protocol that was part of the parent SWAN study was initially considered a prerequisite to entry into the Daily Hormone Substudy (34). Women who had successfully completed at least four of the most recent six monthly menstrual calendars were initially asked to participate in the Daily Hormone Substudy. Eventually, this screening criterion was dropped to facilitate recruitment at all sites.

For the purpose of testing algorithms for menstrual cycle classification into those with and without evidence of luteal activity (ELA and NELA, respectively), the first 396 cycles at baseline were included in analyses. Five other cycles were examined but excluded from these analyses because they were double-ovulatory. Table 1 presents the distribution of baseline characteristics for all 848 Daily Hormone Substudy participants and for the subset of women included in the algorithm selection sample. The latter subset closely resembles the entire pool. Due to site differences in the timing of the fielding of the protocol, the algorithm selection sample included slightly more Caucasian women and fewer Hispanic women (P < 0.0074) and fewer participants at the Chicago SWAN site (P < 0.0001) compared with the remaining Daily Hormone Substudy participants. These small differences reflect the slightly later implementation of the study at the Chicago and New Jersey (Hispanic) sites. Although they are statistically significant because of large sample sizes, they should be inconsequential for purposes of algorithm testing.

Specimen Collection

Specimen collection kits, containing a supply of labeled polypropylene tubes [prefilled with glycerol to a final concentration of 7% (28)], an indelible marker pen, disposable plastic cups, and storage boxes, were delivered to the participants’ homes along with a miniature non-frost-free freezer (for those who wished to use one) to store the collected specimens. Women were instructed to collect their first morning voided urine into a cup, fill two tubes to the indicated fill line (5 ml), and place each tube into a box in the freezer within 2 h of collection. A specimen collection log was provided to allow participants to record any irregularities in the collection, such as a failure to remember to freeze the tube. Women were instructed to collect specimens on the first day of menstrual bleeding, if possible, and to end on the 1st day of bleeding in the subsequent cycle or after 50 days, whichever occurred first.

Weekly telephone calls were made to participants’ homes to encourage adherence to the protocol. At the end of the collection, urine kits were transported on ice from the participants’ homes and transported to the SWAN clinical sites. Specimens were then sent on dry ice to the Reproductive Sciences Program RSP-CLASS Laboratory at the University of Michigan for analysis.

Laboratory Analysis

LH, FSH, estrone conjugates (E1c), and pregnanediol glucuronide (PdG) were assayed using newly adapted chemiluminescent assays. Assays were configured to be compatible with the ACS-180 Autoanalyzer (CIBA-Corning). Specific assays are described as follows.

FSH. FSH was measured with a two-site chemiluminescent immunoassay, which uses constant amounts of two antibodies. The first antibody is an anti-human FSH antibody labeled with a dimethylacridinium ester (DMAE). The second antibody is an anti-human antibody that is covalently coupled to paramagnetic particles (PMP), a solid-phase reagent. The reporting range for the urine FSH assay is 0.3–136 mIU/ml, the minimum detectable concentration is 0.3 mIU/ml, and the inter- and intra-assay coefficients of variations (CVs) of the assay are 11.4 and 3.8%, respectively.

LH. LH was measured with a two-site chemiluminescent immunoassay utilizing constant amounts of two antibodies. The first antibody is a monoclonal mouse anti-LH antibody labeled with DMAE; the second antibody is a monoclonal mouse anti-LH antibody that is covalently coupled to PMP. The resulting luminescent signal is read on the ACS-180 Autoanalyzer, as for all assays yet to be described. The reporting range for the urine LH assay is 0.1–55.2 mIU/ml, the minimum detectable concentration is 0.1 mIU/ml, and the inter- and intra-assay CVs for the LH assay are 10.9 and 4.6%, respectively.
The E1c assay is a competitive immunoassay using direct, chemiluminescent technology. The first antibody is a rabbit anti-E1c antibody. The label is estrone-glucuronide labeled with DMAE. The second antibody is a goat anti-rabbit antibody that is covalently coupled to PMP and used as a solid-phase reagent. Urine samples and quality controls (QCs) are prediluted (1:51) in buffer. The reporting range for the urine E1c assay is 5.10 – 408.0 ng/ml, the minimum detectable concentration is 0.1 ng/ml, and the inter- and intra-assay CVs for the E1c assay are 11.5 and 8.1%, respectively.

The Pdg assay is a competitive immunoassay with direct, chemiluminescent technology. The first antibody is a rabbit anti-Pdg antibody. The label is Pdg labeled with DMAE. The second antibody is a goat anti-rabbit antibody that is covalently coupled to PMP and used as a solid-phase reagent. Urine samples and QCs are prediluted (1:51) in assay buffer. The reporting range for the urine Pdg assay is 0.005 – 25.5 μg/ml, the minimum detectable concentration is 0.0001 μg/ml, and the inter- and intra-assay CVs are 17.8 and 7.7%, respectively.

Glycerol-preserved specimens were used for all assays, because this has been reported to permit measurement of LH and FSH over long storage intervals (20) and does not interfere with E1c or Pdg assay (29, 31). Samples were all normalized for the amount of creatinine in each specimen and are expressed per milligram creatinine (36).

**Menstrual Cycle Parameters**

The characteristics of a regularly cycling midreproductive-aged woman’s cycle were considered the point of departure from which the raters constructed a common approach to the classification of cycles into those with evidence of luteal activity and those without evidence of luteal activity. In a normal midreproductive-aged woman’s menstrual cycle, an ovarian follicle matures within 10–20 days. The follicle produces increasing amounts of estradiol over the last 7–10 days of the follicular phase, with the peak estradiol triggering an ovulatory LH surge. The LH surge, in turn, triggers final maturation of the oocyte, ovulation, and transformation of the ovarian follicle into a corpus luteum. The corpus luteum produces progesterone for 12–16 days. Estradiol and progesterone production decreases substantially by the end of the cycle, unless pregnancy supervenes. The withdrawal of estradiol and progesterone results in menstruation, the event that defines the beginning and end of a menstrual cycle. These events have been summarized in greater detail (11) and are depicted in Figure 1.

In the cycles of women who are in their 40s, follicular phase shortening is often observed, along with increased FSH levels, especially in the early follicular phase of the cycle (16, 18, 31–32). The presumption that progesterone production follows ovulation and that corpus luteum function...
optimal may not be appropriate in these cycles. Because we did not directly recover an oocyte, did not perform studies of ovarian morphology on all women, and did not observe pregnancy, we cannot prove that ovulation occurred or did not occur in these cycles. For this reason, we do not describe cycles as “ovulatory” or “anovulatory,” because, strictly speaking, circulating hormone patterns cannot infallibly predict the ovarian event of ovulation. Thus we identified two key events of the menstrual cycle to be inferred by use of urinary hormones as proxy markers for ovulation and corpus luteum function:

Evidential of luteal activity. Progesterone production by the corpus luteum is a critical event that separates the menstrual cycle into the follicular and luteal phases. ELA was inferred by a rise in progesterone or a proportional increase in progesterone over follicular phase concentrations (relative threshold).

Day of the luteal transition. The timing of corpus luteum transformation [day of the luteal transition (DLT)] is usually observed with methods that pinpoint its initiation, such as a serum LH surge, or, when urinary hormones are analyzed, a change in the ratio of estrogen to progesterone.

Menstrual Cycle Algorithms

We examined several existing algorithms developed for assessment of menstrual cycle luteal phase activity and timing in midreproductive-aged women, as well as modifications to these algorithms.

ELA. All algorithms that we considered involved an increase in Pdg adjusted for creatinine (Cr). We denote Pdg/Cr by APdg, and the moving 5-day average of APdg by APdg5, as in Waller et al. (40).

The first algorithm employs an absolute threshold and requires Pdg to rise to a concentration of $\geq 3$ ng/mg Cr for three consecutive days.

A second group of algorithms, based on the method developed by Kassam et al. (15), uses a relative threshold. In the original algorithm proposed by Kassam et al., we will refer to as the Kassam method, a cycle-specific baseline is defined as the minimum APdg5, and a threshold for evidence of luteal activity as three times this baseline. Cycles with three successive values of APdg above the threshold are classified as ELA and all other cycles as NELA. This criterion assumes neither “normality” nor fecundability and is not a guarantee that ovulation has occurred. We note that Kassam et al. (15) validated this algorithm against a gold standard of weekly serum progesterone concentrations. Moreover, this method was developed explicitly for use with data that may not correspond to menstruation. Waller et al. (40) modified this algorithm by using a threshold equaling the baseline $+ 1$ $+ \sqrt{2}$ of the square root of the baseline. Cycles are defined as ELA if both the maximum APdg5 and $\geq 3$ of the 5 Pdg/Cr values in that 5-day sequence exceed the threshold. Cycles are classified as NELA if the maximum APdg5 is no more than the threshold minus 1. Remaining cycles are classified as questionable. We will refer to this algorithm as the Waller-ELA method.

A third group of algorithms, threshold/duration methods, is based on the method of Brown et al. (5), which identifies a Pdg rise if two consecutive measurements of APdg exceed the 5-day lagged APdg5 by at least three standard deviations (SDs).

Modifications to these algorithms included varying the number of days used to compute the moving average of APdg, the number of days required to be above the threshold, and the number of SDs for the threshold.

DLT. After omitting cycles classified as NELA by the best-performing ELA algorithm, we examined algorithms to detect the day of onset of luteal activity, i.e., day of luteal transition, or DLT. Existing algorithms (2, 40) require an increase followed by an immediate decrease in the daily E1c-to-Pdg ratio (E1c/Pdg).

The method of Baird et al. (2), a modification of work by Royston (28), examines 5-day sequences of E1c/Pdg, denoting the five consecutive values of E1c/Pdg by EP1 through EP5. The algorithm identifies sequences where EP1 is the maximum of EP1 through EP5 and EP4 and EP5 are at or below 40% of EP1; the 40% limit is known as the descent criterion. For cycles with one such sequence, the DLT is defined to occur on day 2 of the 5-day sequence. Cycles with no sequences meeting these criteria are classified as indeterminate regarding the DLT. For cycles with multiple nonoverlapping 5-day sequences meeting these criteria, the sequences are compared regarding the mean E1c/Pdg from the days before and after day 1 (i.e., the mean of EP0 and EP2). If one sequence’s mean is more than twice the corresponding mean from the other sequences, that sequence is selected for identification of the DLT. If no sequence is dominant according to this condition, the cycle is classified as indeterminate. We will refer to this algorithm as the Baird method.

Waller et al. (40) modified the method of Baird et al. (2) by using E1c/Pdg $+ 1$ instead of E1c/Pdg to handle very low Pdg values in their dataset, by using a descent criterion of 60%, and in cases of multiple qualifying sequences, selecting the 5-day sequence with the maximum mean of EP0 and EP2. This algorithm will be referred to as the Waller-DLT method. The data analyzed here, omitting the 1 from the denominator had no effect on the performance of the algorithm (results not shown).

Modifications to these algorithms included varying the descent criterion and removing the restriction that the DLT be on day 8 or later.

Additional modifications to the ELA and DLT algorithms included use of the LH midcycle surge (MCS). Data were evaluated using a 5-day moving average, with a 3-SD increase required to consider the rise in LH significant (5). In addition, the onset of menses within 17 days of the DLT, a feature of a “normal” menstrual cycle, could provide supporting evidence that ovulation had occurred. Finally, the mean LH and FSH from the DLT to luteal day 8 were considered normal if they were less than the follicular phase means of these hormones (excluding the MCS), indicating midluteal suppression of gonadotropins.

Testing of Algorithm Performance

Each menstrual cycle was randomly assigned to a pair of expert raters (from a pool of six) and classified by each rater regarding ELA. For those with ELA, the raters also specified a DLT. Two determinations that were identical were taken as the final result. In the event that all three raters disagreed (one ELA, one NELA, and one indeterminate), the cycle was rated as indeterminate. These ratings served as the criterion standard against which the algorithms were tested. Discrepancies were resolved by an independent third rater. Interrater reliability for ELA status was assessed in terms of percent agreement and the $\kappa$-statistic (12), and interrater reliability for DLT was assessed in terms of percent agreement $\geq 3$ days.

The rater pairs agreed on ELA status in 360 (91.9%) of the 396 cycles [Table 2; $\kappa = 0.70$, 95% confidence interval (CI) = 0.61–0.78], indicating good interrater agreement. Of these
Intrarater reliability was determined from blinded reratings of 25–26 randomly selected cycles by each rater. Of the 151 cycles rated twice by the same rater, 140 (92.7%) were given the same ELA rating both times (κ = 0.71, 95% CI = 0.56–0.86). This percentage ranged from 84.0% to 100% for individual raters. Among cycles rated twice by the same rater as ELA (n = 125), 98.4% agreed to ±3 days regarding the DLT. This percentage ranged from 95.5% to 100% for individual raters.

**Comparison of ratings with algorithm classifications.** Accuracy of each ELA/DTL algorithm was determined by comparison of the algorithm-based cycle classifications with the criterion standard expert ratings. Sensitivity was computed as the percentage of ELA-rated cycles that were classified by the algorithm as ELA. Similarly, specificity was computed as the percentage of NELA-rated cycles classified by the algorithm as NELA. For DLT algorithms, performance was assessed in the 328 cycles rated ELA by computing the percentage of cycles with no algorithm-assigned DLT, the percentage of cycles with an algorithm-assigned DLT that differed from the rater DLT by >3 days (considered discrepant), and the percentage of cycles with an algorithm-assigned DLT that differed from the rater DLT by ≤3 days (considered a match). Bootstrapped standard errors (SEs; 500 samples) were computed for all percentages (7).

**Comparison analyses in midreproductive-aged women.** A sample of 30 cycles from women between 19 and 34 yr of age was also assessed to determine algorithm performance in these “optimal” cycles. One subject of the 30 was eliminated from further analysis, because upon scrutiny of her cycle, she seemed to have initiated collection with the onset of a Pdg rise and may have experienced periovulatory bleeding that led to an incorrect day of collection. A similar assessment of algorithm performance was conducted in this group of women.

**RESULTS**

**ELA Performance**

Table 3 presents the comparison of Pdg-based ELA algorithms with the expert ratings. The constant cri-
terion had 100% specificity but only 81.4% sensitivity. Among the moving-average-based algorithms, use of a window of 3 or 4 days for the moving average gave slightly higher sensitivity than a 5-day window, but lower specificity. Increasing the length of the window to 6 or 7 days yielded no change in ELA classifications from a 5-day window (data not shown). Among the threshold/duration-based algorithms, shorter durations were more likely to classify cycles as ELA, thereby increasing sensitivity but decreasing specificity. None of the threshold/duration-based algorithms performed as well on both sensitivity and specificity as the original Kassam method. The Waller-ELA method was the only algorithm to include a questionable or indeterminate category. Its overall percent agreement with rater ELA classifications was 89.4% (κ = 0.67, 95% CI = 0.59–0.75). A relatively high percentage of rated NELA cycles were classified by this algorithm as indeterminate (12.9%), which lowered the specificity relative to other algorithms. In summary, the original method of Kassam et al. (15) yielded the best overall performance regarding both sensitivity and specificity.

**DLT Accuracy**

The performance of algorithms based on changes in E1c/Pdg was very similar across different modifications and is shown in Table 4 (columns labeled “Not using LH surge”). For all DLT algorithms, cycles classified as NELA by the Kassam method were considered to have no algorithm-assigned DLT. Among algorithms based on the Baird method, reducing the descent criterion from 40 to 30% yielded more cycles with no 5-day sequence satisfying the criterion, and thus a smaller proportion of cycles with a DLT match. Increasing the descent criterion to 60 or 70% produced more cycles with multiple 5-day sequences satisfying the criterion, with a corresponding increase in cycles with no dominant sequence. A descent criterion of 50% gave the highest probability of a match, at 91.8%. Among approaches based on the Waller-DLT method, omitting the requirement that the DLT be at least day 8 increased the probability of a match slightly, from 91.2 to 92.4%. Similar to the Baird-related methods, reducing the descent criterion from 60 to 50% increased the proportion of cycles with no algorithm-assigned DLT and decreased the probability of a match. In contrast, compared with a descent criterion of 60%, raising the descent criterion to 75% yielded a slightly lower percentage with no algorithm-assigned DLT, a slightly higher percentage with a discrepant DLT, and the same percentage of a match (92.4%). Thus use of the Waller-DLT method with the modification of removing the requirement that the DLT be day 8 or later, with a descent criterion of either 60 or 75%, yielded the highest probability of a match, although this percentage was only slightly higher than that of several other algorithms considered. We note the high amount of overlap of the 95% CIs across all algorithms.

**Other Criteria**

In the subset of cycles classified by the Kassam method as ELA but lacking an algorithm-assigned DLT, we modified each DLT algorithm by assigning the day of the LH surge (for cycles with multiple surges, using LH surge

---

**Table 4. Performance of E1c/Pdg-based algorithms for classification of luteal activity in detecting all 328 cycles rated ELA by raters**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>% in Category (95% CI)</th>
<th>No-algorithm DLT</th>
<th>DLT discrepant</th>
<th>DLT match</th>
</tr>
</thead>
<tbody>
<tr>
<td>30% Descent criterion</td>
<td>12.5 (9.6–15.4)</td>
<td>3.0 (1.5–4.7)</td>
<td>2.7 (1.5–4.3)</td>
<td>3.7 (2.1–5.4)</td>
</tr>
<tr>
<td>40% Descent criterion*</td>
<td>7.6 (5.4–10.0)</td>
<td>2.4 (1.0–3.9)</td>
<td>2.4 (1.0–3.9)</td>
<td>3.0 (1.5–4.8)</td>
</tr>
<tr>
<td>50% Descent criterion</td>
<td>5.5 (3.4–7.8)</td>
<td>1.8 (0.6–3.0)</td>
<td>2.7 (1.2–4.3)</td>
<td>3.0 (1.5–4.6)</td>
</tr>
<tr>
<td>60% Descent criterion</td>
<td>6.4 (4.3–8.8)</td>
<td>1.2 (0.3–2.2)</td>
<td>2.4 (1.1–3.9)</td>
<td>3.7 (1.9–5.4)</td>
</tr>
<tr>
<td>70% Descent criterion</td>
<td>7.9 (5.6–10.4)</td>
<td>1.2 (0.3–2.4)</td>
<td>2.4 (1.2–3.8)</td>
<td>3.7 (2.1–5.4)</td>
</tr>
<tr>
<td>60% Descent criterion, DLT = 8†</td>
<td>4.6 (2.7–6.7)</td>
<td>2.7 (1.2–4.3)</td>
<td>4.3 (2.4–6.2)</td>
<td>4.6 (2.7–6.5)</td>
</tr>
<tr>
<td>50% Descent criterion</td>
<td>7.3 (5.0–9.7)</td>
<td>3.0 (1.5–4.7)</td>
<td>4.6 (2.8–6.4)</td>
<td>5.2 (3.2–7.0)</td>
</tr>
<tr>
<td>60% Descent criterion</td>
<td>3.7 (2.0–5.5)</td>
<td>2.7 (1.2–4.3)</td>
<td>4.0 (2.2–5.8)</td>
<td>4.3 (2.5–6.2)</td>
</tr>
<tr>
<td>75% Descent criterion</td>
<td>3.0 (1.5–4.6)</td>
<td>2.7 (1.2–4.3)</td>
<td>4.6 (2.5–6.5)</td>
<td>4.9 (2.8–6.9)</td>
</tr>
</tbody>
</table>

E1c, estrone conjugates. The algorithms’ ability to detect these cycles with and without an additional criterion of a detectable LH surge is shown in split columns. Minor modifications of the algorithms are also shown in first column. *Method of Baird et al. (2); †method of Waller et al. (40).
the day of the maximum surge) as the algorithm-assigned DLT. As seen in Table 4 (columns labeled “Using LH surge”), this modification increased agreement of the Baird-related algorithms with the rater DLT, ranging from an increase of 3.3% for the 50% descent criterion to an increase of 8.5% for the 30% descent criterion. In general, the percentage of cycles with no algorithm-assigned DLT declined, and the increase in discrepant DLTs was more than compensated for by an increase in matching DLTs. Improvement in agreement with raters was smaller for the Waller-DLT-based algorithms, ranging from no change to 3.7% greater agreement, due to a smaller percentage of cycles that originally lacked an algorithm-assigned DLT.

Incorporating novel information in the algorithms, namely, menses within 17 days of the algorithm-generated DLT and midluteal gonadotropin suppression, did not increase the agreement between raters and the algorithms. For example, in cycles that would have had a Baird method-assigned DLT but were classified by the Kassam method as NELA, reclassifying as ELA those cycles that had menstrual bleeding within 17 days of the algorithm-assigned DLT created more mismatches than those that were corrected. Evidence of gonadotropin suppression also did not contribute further to algorithm performance, given the small number of cycles not rated ELA by the Kassam method but assigned a DLT by a Baird-related or Waller-DLT-related algorithm. Among cycles classified as ELA with an algorithm-assigned DLT, the percentage with FSH suppression ranged only from 73.1% for the Waller-DLT method with the restriction on a minimum DLT omitted to 82.9% for the Baird method with a 30% criterion. Percentages for LH suppression were lower, at 40.3% for the Baird method with a 70% descent criterion through 61.9% for the Waller-DLT method with a 50% descent criterion.

**Algorithm Performance in Midreproductive-Aged Non-SWAN Women**

Of the 29 cycles from midreproductive women included in these analyses, all of whom were rated as ELA, all 29 were classified as ELA when the Kassam method was used with modifications of the moving average window from 3 to 5 days (see Table 5). The constant criterion, requiring a rise to 3 μg/mg Cr for ≥3 days, yielded 28 cycles (96.6%) as ELA. The Waller-ELA algorithm classified 29 cycles as ELA, and the threshold/duration criteria, similar to its performance in the older women, became less robust as the duration of a significant rise in Pdg was lengthened from 2 to 5 days. By the time a 5-day rise was required, only 23 of the 29 cycles (79.3%) were considered to have ELA. When the DLT algorithms with their modifications were used (see Table 6), the Waller-DLT algorithm yielded a match in 27 cycles (93.1%), and the Baird algorithm matched in 28 cycles (96.6%) to within 3 days of the raters. Overall, tremendous consistency was noted across all descent criteria imposed.

### Table 5. Performance of pregnanediol-based algorithms for ELA classification of 29 midreproductive-aged women’s cycles (all rated ELA)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>% Algorithm-ELA (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant: Apdg &gt;3.0 for 3+ days</td>
<td>96.6(89.7–100.0)</td>
</tr>
<tr>
<td>Moving average</td>
<td></td>
</tr>
<tr>
<td>3-day window</td>
<td>100.0(100.0–100.0)</td>
</tr>
<tr>
<td>4-day window</td>
<td>100.0(100.0–100.0)</td>
</tr>
<tr>
<td>5-day window</td>
<td>100.0(100.0–100.0)</td>
</tr>
<tr>
<td>Modification by Waller et al.</td>
<td>100.0(100.0–100.0)</td>
</tr>
<tr>
<td>Threshold/duration</td>
<td></td>
</tr>
<tr>
<td>3 SD/2 days</td>
<td>100.0(100.0–100.0)</td>
</tr>
<tr>
<td>3 SD/3 days</td>
<td>93.1(84.5–100.0)</td>
</tr>
<tr>
<td>3 SD/4 days</td>
<td>89.7(79.3–96.6)</td>
</tr>
<tr>
<td>3 SD/5 days</td>
<td>79.3(65.5–89.7)</td>
</tr>
<tr>
<td>2 SD/2 days</td>
<td>100.0(100.0–100.0)</td>
</tr>
<tr>
<td>2 SD/3 days</td>
<td>93.1(84.5–100.0)</td>
</tr>
<tr>
<td>2 SD/4 days</td>
<td>89.7(79.3–96.6)</td>
</tr>
<tr>
<td>2 SD/5 days</td>
<td>89.7(79.3–96.6)</td>
</tr>
</tbody>
</table>

Note that although the number of cycles is much smaller, all moving average and two threshold/duration methods yielded 100% detection of luteal activity.

**DISCUSSION**

These data demonstrate the usefulness of objective, computer-based algorithms to describe the cardinal features of the menstrual cycle with reasonable accuracy. The algorithm-based cycle classifications agreed with our subjective, expert cycle raters in >90% of cycles. These data justify the ability of algorithms to assess probable ovulation by corpus luteum activity and cycle partition by day of luteal transition to help classify and elucidate the progress toward menopause among a population of reproductively aging women.

The degree of agreement between our raters and two simple, published criteria in a two-step algorithm (Kassam followed by either Waller-DLT or Baird) was 98% sensitivity for the detection of ELA and an 85–95% ability to predict the DLT to ±3 days. This degree of algorithm precision has not been published previously. Moreover, the prior use of algorithms has presumed that the cycles were all ovulatory. The fact that we have applied these results to a population that is becoming progressively more anovulatory suggests strongly that we will continue to be able to identify luteal function with reasonable reliability as the menopause transition proceeds. Attempts to improve or customize the existing algorithm did not yield significantly better agreement with our raters. Taken together, our findings suggest that the combined use of the Kassam method for detection of ELA and either the Waller or Baird method for determining the DLT will remain effective throughout the transition to menopause.

It is reassuring that the Kassam method had the greatest internal validity in testing with our expert raters. This method was validated against weekly serum sampling in the sample of women used to develop it. Therefore, there is reasonable assurance that the excursions of urinary Pdg we used to determine ELA are related to biologically meaningful increases in serum progesterone.
The sample of women we studied had a small proportion (4.3%) of cycles that were rated as indeterminate. This subgroup (n = 17) was too small to permit realistic modification and testing of ELA algorithms to add an indeterminate category. By use of the algorithm that performed optimally, the indeterminate cycles were rated approximately one-half ELA and one-half NELA; that is, there did not appear to be any systematic bias in the assignment of these cycles. Because we expect that luteal function will become increasingly abnormal as the women in SWAN traverse the menopause, we anticipate that more cycles of this type will be encountered. A top-performing algorithm for detection of luteal activity, the Waller-ELA method, includes a category that is “indeterminate.” It will be of interest to see if the proportion of women with cycles categorized in this manner increases over time when this algorithm is used to identify ELA status.

It was somewhat surprising to us that adding features to the two-step process (ELA determination, followed by DLT determination) did not further enhance the ability to detect cycles with presumed ovulation. A recent study of midreproductive-aged women (19) relied heavily on FSH elevations at midcycle to pinpoint the DLT. However, in this aging population of women, FSH elevations at midcycle demonstrated poor correspondence with the ELA determinations of the Kassam method. Use of an FSH midcycle peak as a criterion to identify ovulatory cycles resulted in the misclassification of 8 cycles as ELA that were determined NELA by the Kassam method and the exclusion of 56 cycles that had no detectable FSH peak but were ELA by the Kassam method. This implies that the monotropic rise in FSH that is a cardinal characteristic of the menopause transition obscures midcycle gonadotropin dynamics.

In the original publication by Kassam et al. (15), weekly serum progesterone levels were obtained on the participants to ensure that the criteria for luteal activity were reflective of change in serum progesterone. We did not independently validate the Kassam algorithm in these older women for several reasons. First, the SWAN study is essentially a field study of women at midlife, involving annual visits and blood draws. The additional participant burden imposed by weekly blood sampling was believed to be beyond what the participants would tolerate. Second, previous investigators have reported similar relationships between urinary pregnanediol glucuronide and serum progesterone in older women compared with younger women in the age groups we have studied herein (31). These data imply that there are not large changes in serum-to-urine steroid hormones that would render the Kassam algorithm less valid in the sample we have studied.

Performance of menstrual cycle algorithms was uniformly excellent in midreproductive-aged women. This was a small group obtained for comparison purposes and included women who had been prescreened for menstrual regularity and were therefore highly likely to have ovulatory cycles. Some differences in the hormone patterns of these women were noted when they were compared with the SWAN sample; however, the two groups cannot be considered strictly comparable.

Perimenopausal women experience irregular evidence of luteal activity, relative infertility, and great variability in cycle characteristics (22–23, 31, 33). Cycle parameters and cycle length are more predictable in younger women (18, 37–38). Younger women’s cycles are far more likely to be fertile (8). The basis for these changes in cyclicity is a diminished follicular pool, with concomitant changes in the key cycle parameters that are dependable features of the cycles of younger women (27). Overall, FSH is increased in older women’s cycles, and in the early perimenopause, estradiol is elevated (31, 33). Despite evidence of adequate circulating estradiol, perimenopausal women are less likely to produce an LH surge, consistent with other observations (31, 33) and dynamic studies (30). Older women also appear to suffer from relative luteal inadequacy. This finding has been reported in some (26, 31) but not all (4, 22, 32) previous studies. Thus the initial consequences of a dwindling supply of follicles include changes in all of the key characteristic hormones in terms of both relative quantity and pattern. Such
changes did not yet appear to make it unduly difficult to classify menstrual cycles in our SWAN Daily Hormone Substudy population. As the participants progress through the menopausal transition, the consequences of the physiological processes discussed above on hormonal patterns and the subsequent ability to interpret hormonal patterns appear to be able to be reliably assessed using algorithms based upon the more regular cycles of midreproductive women.

APPENDIX

We summarize the algorithms evaluated in the text, separately, for ELA status and for DLT.

ELA Status

First, we present some notation. Denote daily Pdg/Cr by APdg, and the moving 5-day average of APdg by APdg5 [see Waller et al. (40)]. Define a cycle-specific baseline (BASE) as the minimum APdg5. Define APdg5_5 as the 5-day lagged value of APdg5, and SD_5 as the SD of the 5 APdg values comprising APdg5_5.

**Absolute threshold**

ELA APdg = 3 ng/mg Cr for 3+ consecutive days

NELA otherwise

**Relative threshold**

**KASSAM METHOD**

ELA APdg = 3 × BASE for 3+ consecutive days

NELA otherwise

**WALLER-ELA METHOD**

ELA 1) Max APdg5 = BASE + 1 + BASE1/2

and 2) 3+ of 5 APdg values comprising max APdg5 = BASE + 1 + BASE1/2

NELA Max APdg5 = BASE + BASE1/2

**Questionable otherwise**

**Threshold/duration**

ELA APdg = APdg5_5 + 3 × SD_5 for 2+ consecutive days

NELA otherwise

**DLT**

Denote six consecutive daily values of E1/Pdg by EP0 through EP5, and the corresponding days by day 0 through day 5. Define AVG02 as the average of E0 and E2.

**Baird method.** Identify all 5-day sequences of E1 through E5 as days 1–5, satisfying both of the following equations

(a) EP1 = max{EP1, . . . , EP5}

(b) EP4 = 40% × EP1 and EP5 = 40% × EP1,

where 40% is called the descent criterion.

For overlapping 5-day sequences meeting both of the previous conditions, consider only the earliest. 1) If there is a single qualifying 5-day sequence, DLT = day 2. 2) If there are multiple nonoverlapping qualifying sequences, if one sequence’s AVG02 = 2 × AVG02 for all other qualifying sequences, then DLT = day 2 of that sequence. Otherwise, if no sequence is dominant, DLT = indeterminate. 3) If there are no qualifying sequences, DLT = indeterminate.

**Waller-DLT method.** Modifications to the Baird method include 1) substitute E1/Pdg + 1 for E1/Pdg; 2) use a 60% descent criterion instead of a 40% descent criterion; 3) for multiple nonoverlapping qualifying sequences, DLT = day 2 of the sequence with the maximum AVG02.

This manuscript was reviewed by the Publications and Presentations Committee of SWAN and has its endorsement. We gratefully acknowledge Drs. Kirsten Waller and Peter Meyer for advice on statistical analyses, and Edith Rodriguez for preparation of the manuscript.

The Study of Women’s Health Across the Nation (SWAN) was funded by the National Institute on Aging, the National Institute of Nursing Research, and the Office of Research on Women’s Health of the National Institutes of Health. Supplemental funding from the National Institute of Mental Health, the National Institute on Child Health and Human Development, the National Center on Complementary and Alternative Medicine, the Office of Minority Health, and the Office of AIDS Research is also gratefully acknowledged.

Clinical Centers: University of Michigan, Ann Arbor, MI [U01 NR-04061, Mary Fran Sowers, Principal Investigator (PI)]; Massachusetts General Hospital, Boston, MA (U01 AG-12531, Joel Finkelstein, PI); Rush University, Rush-Presbyterian-St. Luke’s Medical Center, Chicago, IL (U01 AG-12505, Lynda Powell, PI); University of California, Davis/Kaiser (U01 AG-12554, Ellen Gold, PI); University of California, Los Angeles (U01 AG-12539, Gail Greendale, PI); University of Medicine and Dentistry/New Jersey Medical School, Newark, NJ (U01 AG-12535, Gerson Weiss, PI); and the University of Pittsburgh, Pittsburgh, PA (U01 AG-12546, Karen Matthews, PI).

Laboratory: University of Michigan, Ann Arbor, MI (U01 AG-12495, Central Ligand Assay Satellite Services, Daniel McConnell, PI) and Medical Research Laboratories, Highland Heights, KY (subcontract of U01 AG-12553, Evan Stein, Director).

Coordinating Center: University of Pittsburgh, Pittsburgh, PA (U01 AG-12553, Kim Sutton-Tyrrell, PI) and New England Research Institutes, Watertown, MA (U01 AG-12553, Sonja McKinlay, PI, site of work for S. L. Crawford, J. E. Allsworth, and P. McGaffigan).

Project Officers: Taylor Harden, Carole Hudgings, Marcia Ory, and Sheryl Sherman.

Steering Committee Chair: Jennifer L. Kelsey

Current addresses: S. L. Crawford, Division of Preventive and Behavioral Medicine, Dept. of Medicine, University of Massachusetts Medical School, Worcester, MA 01655; J. E. Allsworth, Center for Gerontology and Health Care Research, Brown Medical School, Providence, RI 02912; P. McGaffigan, Harvard Injury Control Research Center, Harvard University, Boston, MA 02115.

REFERENCES


