AMERICAN JOURNAL OF HUMAN BIOLOGY 24:515–525 (2012)

Original Research Article

Who Is Stressed? Comparing Cortisol Levels Between Individuals

PABLO A. NEPOMNASCHY,[1,2*] TERRY C.K. LEE,[3,4] LEILEI ZENG,[5] AND C.B. DEAN[6]
[1]*Faculty of Health Sciences, Simon Fraser University, Burnaby, British Columbia, Canada*
[2]*Human Evolutionary Studies Program, Simon Fraser University, Burnaby, British Columbia, Canada*
[3]*Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada*
[4]*CIHR Canadian HIV Trials Network, Vancouver, British Columbia, Canada*
[5]*Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada*
[6]*Faculty of Science, University of Western Ontario, London, ON, Canada*

Cortisol is the most commonly used biomarker to compare physiological stress between individuals. Its use, however, is frequently inappropriate. Basal cortisol production varies markedly between individuals. Yet, in naturalistic studies that variation is often ignored, potentially leading to important biases.

Objectives: Identify appropriate analytical tools to compare cortisol across individuals and outline simple simulation procedures for determining the number of measurements required to apply those methods.

Methods: We evaluate and compare three alternative methods (raw values, *Z*-scores, and sample percentiles) to rank individuals according to their cortisol levels. We apply each of these methods to first morning urinary cortisol data collected thrice weekly from 14 cycling Mayan Kaqchiquel women. We also outline a simple simulation to estimate appropriate sample sizes.

Results: Cortisol values varied substantially across women (ranges: means: 1.9–2.7; medians: 1.9–2.8; SD: 0.26–0.49) as did their individual distributions. Cortisol values within women were uncorrelated. The accuracy of the rankings obtained using the *Z*-scores and sample percentiles was similar, and both were superior to those obtained using the cross-sectional cortisol values. Given the interindividual variation observed in our population, 10–15 cortisol measurements per participant provide an acceptable degree of accuracy for across-women comparisons.

Conclusions: The use of single raw cortisol values is inadequate to compare physiological stress levels across individuals. If the distributions of individuals' cortisol values are approximately normal, then the standardized ranking method is most appropriate; otherwise, the sample percentile method is advised. These methods may be applied to compare stress levels across individuals in other populations and species. Am. J. Hum. Biol. 24:515–525, 2012.

The hypothalamic-pituitary-adrenal axis (HPAA) acts as a mediator of the interface between the individual and its environment, allowing the organism to respond and temporarily adapt to ecological challenges including psychosocial, energetic, and health stressors. Responding and adapting to said stressors involve important costs. As metabolic energy is limited, the organism has to reallocate energetic resources away from other metabolic tasks that can be postponed to attend to the challenges at hand. As a consequence of this energy reallocation, critical metabolic tasks such as immune response, cardiovascular function, and reproductive physiology may be affected, leading to negative health outcomes. Consequently, stress has become one of the most important fields of research in contemporary ecological, psychological, and health sciences (Hruschka et al., 2005).

The reallocation of energy needed to respond to ecological challenges is mediated by increases in circulating levels of cortisol (Kanaley and Hartman, 2002; Kudielka and Kirschbaum, 2003; Negrão et al., 2000). Variations in cortisol levels are, therefore, frequently used to monitor HPAA function and activation and to compare physiological stress levels between individuals (Hruschka et al., 2005; Pollard, 1995, 1997).

The use of glucocorticoids to compare stress levels between individuals is not, however, without complications (Hruschka et al., 2005). Experimental designs are commonly used to assess an individual's ability to respond to stress challenges (Adam and Kumari, 2009; Golden et al., 2011). Stress responsivity may be affected by that individual's genome, ontogenetic history, and chronic exposure to stress. Thus, by itself, cortisol increases in response to experimental challenges cannot be used to assess the current level of physiological stress of an individual.

Researchers interested in comparing physiological stress levels between individuals in naturalistic settings commonly compare cortisol levels assessed in a variety of matrices including blood, saliva, urine, hair, and nails (D'Anna-Hernandez et al., 2011; Gunnar et al., 2001a,b; Nepomnaschy et al., 2004, 2006, 2011; Pruessner et al., 1997; Warnock et al., 2010). Hair and nails are used to assess chronic stress levels over prolonged time periods on the order of weeks. These two matrices do not provide information on day-to-day variations in cortisol levels. Blood and specially saliva are often used to assess cortisol levels at the time of sample collection. These matrices, however, present several problems. Some participants may, for example, experience cortisol increases triggered by the anticipatory anxiety generated by the impending challenge

Leilei Zeng is currently at Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada.

C.B. Dean is currently at Faculty of Science, University of Western Ontario, London, ON, Canada.

Financial support for the research involved and preparation of the article was provided by a CIHR IGH Operating Grant (CIHR #106705) and a Simon Fraser University President's Start-up Grant to PAN, an NSERC Operating Grant to CBD and an NSERC Discovery Grant, #327107 to LZ.

*Correspondence to: Pablo A. Nepomnaschy, PhD, Assistant Professor, Faculty of Health Sciences, Simon Fraser University, Blusson Hall, Room 9417, 8888 University Drive, Burnaby, BC V5A 1S6, Canada.
E-mail: pablo_nepomnaschy@sfu.ca

Received 5 October 2011; Revision received 24 December 2011; Accepted 5 February 2012

DOI 10.1002/ajhb.22259

Published online 21 March 2012 in Wiley Online Library (wileyonlinelibrary.com).

or the prick that precedes blood collection, which affects the "basal cortisol reading." In addition to being invasive and uncomfortable, blood collection carries a risk of infection, which is increased by protocols that require repetitive sampling (Miller et al., 2004).

Saliva collection is less invasive and comparatively easier to collect than blood and has, therefore, been favored. Nonetheless, the use of saliva in studies in naturalistic designs is still complicated by two important factors. First, there is a variety of exposures other than stressors that can affect cortisol levels in the short term, including normal physical activity and the consumption of food, caffeine, or alcohol (Adam and Kumari, 2009; Bonen, 1976; Flinn and England, 1995; Meulenberg and Hofman, 1990; Pruessner et al., 1997; Weitzman et al., 1971). Controlling these exposures either statistically or by design in a natural setting is extremely difficult (Kirschbaum et al., 1990; Pollard, 1995; Shirtcliff et al., 2005). Second, cortisol is secreted following a complex pattern based on two overlapped frequencies occurring on different time scales. The overall 24-h circadian rhythm is "built" by ultradian oscillations taking place every 20–120 min (Markovic et al., 2011). To control for the effect of circadian effects, researchers tend to assess cortisol either at the peak and nadir of the cortisol circadian rhythm, or both. It is assumed that stressed individuals will have elevated cortisol levels at both (Golden et al., 2011). The problem with this method is that timing of the true peaks and nadirs for each individual are unknown, so researchers rely on assumptions with regard to the occurrence of these two time points that are usually applied to all the participants. Thus, this protocol is likely to result in samples being collected before or after the true peak and nadir for each individual, thereby affecting the levels of cortisol measured and potentially biasing the final results. Furthermore, even after controlling for between-individual variation in circadian profiles, researchers have to control for between-individual differences in basal cortisol production as well as their temporal variability around that baseline (Hruschka et al., 2005; Kirschbaum et al., 1990, 1999; Nepomnaschy et al., 2004; Williams, 2008). Differences in basal HPAA function muddle the interpretation of single glucocorticoid measurements outside of the context of the basic parameters of each individual's distribution of glucocorticoid values and complicate comparisons between individuals (Golden et al., 2011; Markovic et al., 2011). The same levels of glucocorticoids may reflect different stress levels for individuals with different baseline cortisol values and standard deviations (Gunnar et al., 2001a,b; Hruschka et al., 2005; Nepomnaschy et al., 2004; Pollard, 1995).

Human ecologists and health scientists have considered a variety of alternatives to control for interindividual variation in basal cortisol production (Adam and Kumari, 2009; Golden et al., 2011) including the assessment of cortisol awakening response (Pruessner et al., 1997), diurnal cortisol slope (Adam and Gunnar, 2001), daily average cortisol (Nicolson, 2004), and area under the curve (D'Anna-Hernandez et al., 2011). These methods are still complicated to implement, as they rely heavily on the proper timing of sample collection, which is difficult to monitor and require access to participants multiple times a day, logistically difficult, and costly as they involve the collection and evaluation of multiple salivary specimens per day (Nepomnaschy et al., 2011).

A valid alternative for naturalistic studies is to monitor changes in first morning urinary cortisol levels. First morning urine presents several advantages over other matrices for naturalistic longitudinal designs involving repetitive sampling from each participant (Miller et al., 2004). First, the nocturnal integrative nature of the measurement helps to reduce the confounding effect of ephemeral changes in circulating cortisol levels due to nonstress factors such as food consumption or physical activity. As those confounders are most often diurnal, they are more likely to affect daytime serum or salivary cortisol levels than to affect overnight cortisol secretion. Mild challenges such as those regularly faced over the course of a day can activate the HPAA resulting in transient cortisol increases. These types of challenges, however, seem to not have a major impact on overnight HPAA functioning. More intense and longer term challenges, the type that interest most researcher studying physiological stress, do appear to affect overnight cortisol (Dahlgren et al., 2005, 2009; Haus, 2007; Markovic et al., 2011; Nepomnaschy et al., 2004). Second, being the first urinary void of the day, the timing of sample collection is easier to monitor than in the case of saliva and compliance is high (Collins et al., 1979; Kesner et al., 1992; Lasley et al., 1994; Miller et al., 2004). The specimens can be self-collected by the participants reducing the involvement of research assistants and, therefore, the invasiveness of the protocol (Miller et al., 2004). Finally, only a single urinary specimen per day needs to be collected and analyzed, which reduces the costs of research.

These advantages make first morning urinary cortisol levels an appropriate tool to monitor longitudinal changes in physiological stress levels and to draw comparisons between individuals. One main remaining challenge, however, subsists. To make sense of single cortisol measurements and to compare stress levels across individuals, those measurements still have to be interpreted in the context of each participant's basic distribution parameters (i.e., baseline cortisol and standard deviation around the baseline).

In this article, we evaluate alternative methods to standardize cortisol using each individual's cortisol distribution (a proxy for their HPA function during a particular period) to make meaningful comparisons between individuals. We also outline simple simulation procedures for determining the number of measurements required to apply those methods. Specifically, we evaluate two alternative methods to rank individuals according to their first morning urinary cortisol levels: (i) *Z*-scores calculated based on each individual's arithmetic mean and standard deviation, and (ii) sample percentiles within each woman's cortisol profile and compared them with the use of raw cortisol values. We apply these alternative methods to a data set containing first morning urinary cortisol values obtained thrice weekly for up to 1 year from 14 cycling Mayan Kaqchiquel women living in Guatemala.

The nonparametric methods proposed here to rank individuals according to their cortisol levels are simple and do not require a priori assumptions or time-dependent linear models. We expect these methods to be useful to compare stress levels between individuals across diverse populations and species.

MATERIALS AND METHODS

Ethics statement

This research was approved by the Research Ethics Board of Simon Fraser University. As most individuals in

the study population were illiterate, informed consent from the participants was obtained orally. The consent document was read in Kakchiquel Mayan by a female research assistant (a native Kakchiquel speaker) and signed by each volunteer with a cross, finger print, or name initials, according to her preference. This research is in compliance with the national legislation and the Code of Ethical Principles for Medical Research Involving Human Subject of the World Medical Association (Declaration of Helsinki).

Study population and criteria for participant inclusion

This article is based on data collected in the context of the Society, Environment and Reproduction study. Fieldwork took place over 12 months between the years 2000 and 2001 in a rural Kaqchikel Mayan community located in the southwest highlands of Guatemala. This community was composed at the time of 1,159 inhabitants, who were almost exclusively Kaqchikel Mayan. All women within this population who fit the following profile were invited to participate: living with a coresident male partner, not pregnant, not using any form of chemical contraceptive method, had given birth at least once in the past, and their last birth had taken place at least 6 months before the onset of the study (Nepomnaschy et al., 2004, 2006).

Sample

Throughout the year, 61 women (about three-quarters of those eligible) volunteered to participate via written consent. We restricted our analysis to data collected from a subset of 14 women who had resumed ovarian function after their last birth, were not pregnant, and who had provided 20 or more first morning urine specimens. In this population, it is tradition for women to continue nursing their infants until a new pregnancy is conceived. Accordingly, all women in our sample were breastfeeding during the period studied. A minimum of 20 urine specimens would ensure that we obtained meaningful Kernel density estimates for our analysis. The number of specimens for each woman ranged from 22 to 119. The ages of these 14 women ranged from 18 to 31 years (SD = 4.6 years, median = 25.5 years).

Data and specimen collection

Data and urine specimen collection were performed by trained local female field assistants. Every other day, for a total of three times each week, assistants visited participants in their homes and gathered first morning urine samples. The urine specimens were self-collected by each participant following standard urine collection protocols in clean, dry, nonreactive plastic containers that we provided the night before. Samples were kept on ice until assistants returned to the laboratory (<2 h from the urinary void). Two-milliliter aliquots from the original specimens were stored frozen at -10°C in the field. Samples were shipped on dry ice to the CLASS laboratory at the University of Michigan, where they were stored at -80°C until analysis.

Hormone assays

Concentrations of urinary free cortisol and reproductive hormones used to assess menstrual cycle day [estrone conjugates (E_1C), pregnandiol glucuronide (PdG), luteinizing hormone (LH), and follicle stimulating hormone (FSH)] were determined using immunoassays developed in the CLASS laboratory for use on the Bayer Automated Chemiluminescence System (ACS-180) immunoassay analyzer (Nepomnaschy et al., 2004). Creatinine was assayed using a spectrophotometric assay. All samples from a single participant were run in duplicate in the same assay. Outliers were identified and the samples rerun. Ranges and intraassay and interassay coefficients of variation (IACV and IECV, respectively) were within acceptable ranges (creatinine: range = 0.05–1.4 mg/ml, IACV = 5.4%, IECV = 9.8%; cortisol: range = 0.2–75 mg/dl, IACV = 2.0%, IECV = 6.5%; E_1C: range = 5.10–408.0 ng/ml, IACV = 3.8%, IECV = 6.5%; PdG: range = 0.005–25.5 mg/ml, IACV = 3.6%, IECV = 11.6%; LH: range = 0.1–53.1 mIU/ml, IACV = 3.5%, IECV = 5.4%; and FSH: range = 0.3–144.0 mIU/ml, IACV = 2.3%, IECV = 5.8%) (Nepomnaschy et al., 2004).

DATA ANALYSIS

Characterization of menstrual cycles and cycle day attribution

Following standard procedures, to control for urine dilution we divided the concentration of each hormone by the concentration of creatinine in the same sample (Miller et al., 2004). We then log transformed the creatinine-corrected hormone measurements in an attempt to normalize the distribution of the data. All log-transformed, creatinine-corrected measurements are simply referred to by the name of the hormone (e.g., log-transformed, creatinine-corrected cortisol is hereafter referred to as cortisol).

Menstrual cycles were considered to begin on the first day of vaginal bleeding and end the day before the next bleeding. If reports of vaginal bleeding were missing or confusing, we imputed the last day of the cycle to the day in which PdG levels fell to 40% of its luteal peak and remained low for $\geq$2 days. Cycles presenting a threefold rise in PdG levels above baseline were considered ovulatory (Nepomnaschy et al., 2004, 2006). The time of ovulation was inferred using an algorithm based on the urinary ratio of E_1C/PdG (Baird et al., 1991) and verified using the presence of LH and FSH surges. Menstrual cycles were aligned to the estimated day of ovulation, which was designated "day 0." Follicular days were given negative numbers and luteal days were given positive numbers.

Confounding factors.

Cortisol secretion can be affected by circadian rhythms, physical activity, food consumption, smoking, caffeine, alcohol, and steroid medications (Bonen, 1976; Meulenberg and Hofman, 1990; Pruessner et al., 1997; Weitzman et al., 1971). First morning specimens provide a proxy for overnight cortisol secretion. Working with overnight cortisol secretion minimizes the effects that circadian rhythms have on this metabolite's profile. None of the participants smoked or consumed alcohol. Urine specimens were collected as soon as the participants woke up each morning, before they consumed food or performed any major physical activity, thereby eliminating the influence of those confounders. Our sample of women was relatively homogeneous in terms of age. Thus, we did not evaluate the possible

impact of age on cortisol levels in our models (Nepomnaschy et al., 2004).

Cortisol across the menstrual cycle.

Several reports suggest that basal cortisol profiles do not vary across the menstrual cycle (Kanaley et al., 1992; Kirschbaum et al., 1999; Stewart et al., 1993). In our sample, we find that menstrual cycle day is not a significant predictor of urinary cortisol levels when we restrict our analysis to days between day −14 (follicular) and day 14 (luteal) (P-value > 0.05). Yet, day of the menstrual cycle becomes a statistically significant predictor of cortisol levels when we extend our analyses beyond that central 28-day period (Nepomnaschy et al., 2011). Thus, as no proper characterization of day-to-day variation in first morning urinary cortisol across long menstrual cycles is currently available, for this initial model we restricted our analysis to days −14 to +14 of the menstrual cycle.

Distribution of cortisol.

Across the 14 women, the individual cortisol means and medians ranged from around 1.9 to 2.8. The sample standard deviation also varied substantially across the women, with a minimum value of 0.26 and a maximum value of 0.49. Anderson-Darling normality test suggested that more than half of the women have cortisol distributions that were non-normal (not shown).

Kernel density estimates, nonparametric estimates for the probability density function of the true probability distribution for cortisol based on all the cortisol measurements (range: 22–119) of each of the 14 participants, are given in Figure 1. These estimates exhibit marked variation in distributions across women. Most of the estimates are somewhat different from the bell-shaped curve one would expect from a normal distribution. Some participants' cortisol profiles also exhibit two peaks, which might represent the modes (i.e., the most frequently observed value) of two stress states: normal versus high. Also, as pointed out above, the sample mean, median, and standard deviation can differ substantially across women. This evidence suggests that cortisol is not normally distributed and that the probability distribution varies across women in our sample. Nonetheless, the log transformation was helpful in alleviating the challenges posed by severe skewness in the data and made the data distribution for many of our participants much closer to normal, and in some cases, approximately normal.

Dependence of cortisol measurements within woman.

To assess the potential dependence of cortisol measurements within woman in our sample, we examined the autocovariance function (ACF) of the cortisol values. The ACF measures the linear dependence between two points on the same series observed at different times. In our data set, cortisol measurements were taken either on a "Monday-Wednesday-Friday" or a "Tuesday-Thursday-Saturday" schedule. Therefore, the time between measurements is either 2 or 3 days. However, because we have censored the data (by restricting our analyses to days ±14 days around ovulation), the time lag between consecutive data points used in this analysis can vary substantially. Nevertheless, the majority (79%) of the time lags was 2 or 3 days. Thus, for the sake of simplicity, we assumed that the time lag between consecutive measurements was similar for our calculations of the sample's ACF. Most of the sample ACFs were within the 95% confidence bounds for no linear dependence (with the exception of a few time lags from four women whose ACFs were marginally outside the confidence bound). These results suggest that it is reasonable to assume that within-women cortisol values were uncorrelated and independent in our data set. The various statistical procedures presented in the following sections are based on that independence.

Ranking physiological stress levels based on cortisol.

If the true cortisol distribution for each woman is known, an intuitive way of ranking physiological stress levels across women is to compare cortisol values' true percentiles within each woman's cortisol distribution. The true percentile indicates the percentage of cortisol values that are smaller than each particular cortisol measurement within each individual's profile. Within the context of each woman's cortisol distribution, a larger percentile can be interpreted as a higher level of physiological stress. The "true" cortisol distributions are, however, not known for the individuals. It is, therefore, not feasible to rank participants based on true percentiles. Thus, here we evaluate three alternative methods for ranking stress levels between women. These methods will equal or approach the true percentile method under certain conditions as explained below (see also Table 1). The three methods are as follows:

i. Comparing raw values:
   If true cortisol distributions were the same across the women being compared, then physiological stress level rankings based on raw values (whether log transformed or not) would be the same as ranking by true percentiles. We found, however, that the cortisol distributions across women varied significantly in our data set and, thus, ranking based on raw cortisol values may not be appropriate. Nonetheless, we use this method for comparison purposes.
ii. Comparing standardized values:
   To account for the differences in cortisol distribution between women, Nepomnaschy et al. standardized the observed cortisol values relative to each woman's baseline and overall variability before conducting their analyses (Nepomnaschy et al., 2004). The jth standardized cortisol measurement for woman i is defined as Eq. (1): $z_{ij} = \frac{\text{cort}_{ij} - \mu_i}{\sigma_i}$, where cort_{ij} is the observed cortisol value, μ_i and σ_i are the arithmetic sample mean and sample standard deviation of cortisol for woman i, respectively. The standardized value indicates how many standard deviations an observation is away from the mean. If the true mean and standard deviation for each woman's cortisol value are known and the underlying cortisol distribution is normal, then ranking based on the standardized values will be the same as ranking based on the true percentiles. Comparing standardized values across women, however, is only optimal if the distribution for cortisol is normal for all the women under consideration. Because we found evidence against normality in our data set, we propose an alternative method.
iii. Comparing sample percentiles:
   We ranked women based on sample percentiles. The sample percentile for the jth cortisol measurement of woman i is defined as Eq. (2): $p_{ij} = 100 \times \frac{r_{ij} - 0.5}{n_i}$, where r_{ij}

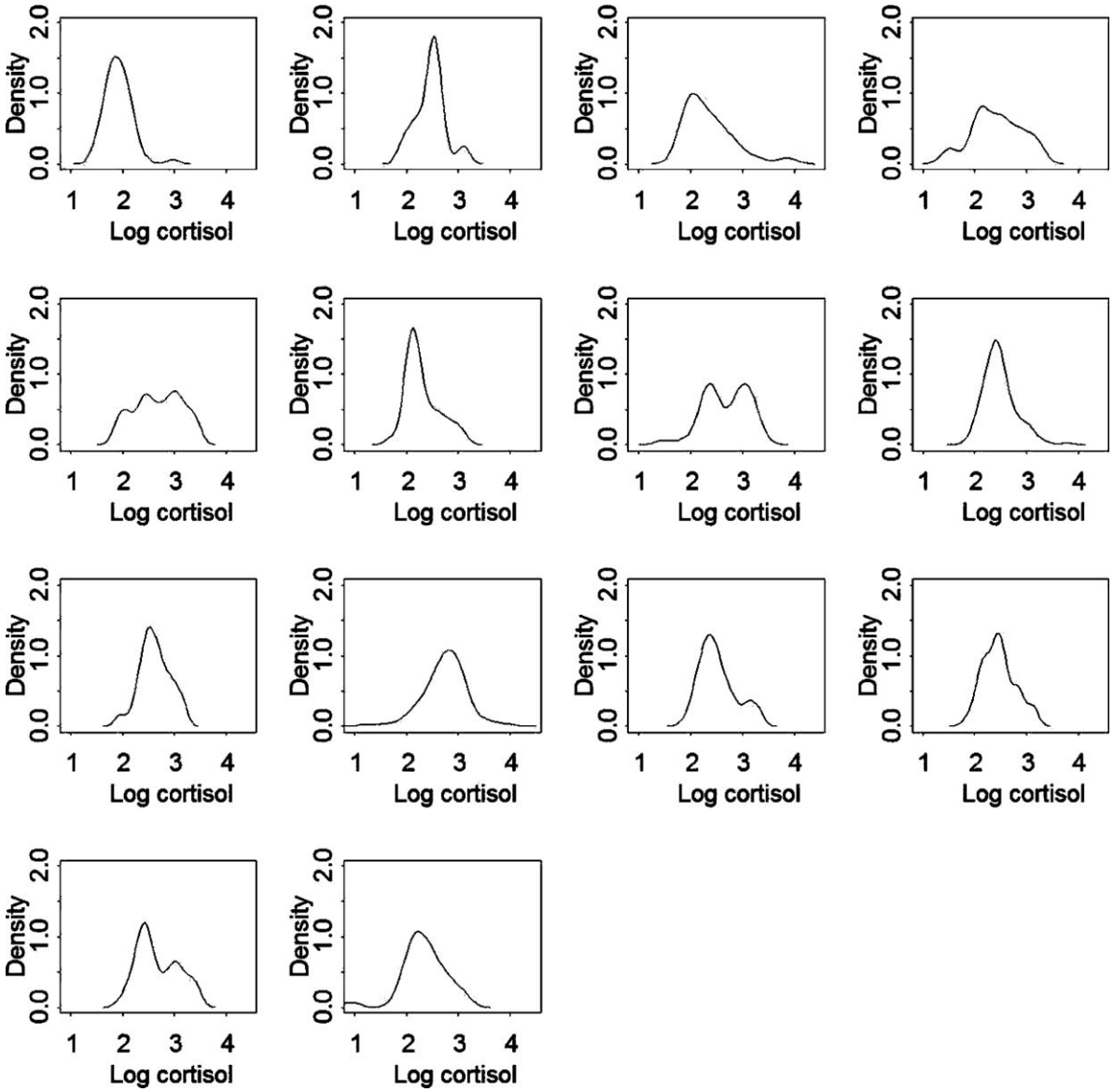

Fig. 1. Kernel density estimates of the distribution of cortisol values for the 14 participants.

is the rank (from smallest to largest) within woman i for the jth measurement and n_i is the total number of measurements for woman i. Ranking based on sample percentiles will approach the ranking based on true percentiles when all the n_i approach infinity.

Assessing the accuracy of ranking methods.

We conducted a simulation study to assess the performance of the three ranking methods described above. We used our participants' cortisol values to generate data series and then used those data series to evaluate the three ranking methods described above. Kernel density provided appropriate nonparametric estimates of the actual distributions of our participants' cortisol data; thus, we used them to generate data series in a simulated group of women. The kernel density estimate for each participant in the simulated group was randomly chosen (with replacement) from the 14 kernel density estimates obtained from our data. We then applied the three methods described above to rank the last simulated values of the participants in the simulated group (using the last simulated values was an arbitrary decision, we could have chosen any other values and the conclusions presented here would be the same).

To assess the accuracy of the ranking methods, we compared the mean of the squared difference (denoted as mean square error, MSE) between the estimated ranks and the true ranks. Because the underlying cortisol distributions were known in the simulation, we can obtain the true ranks based on the true percentiles.

Evaluating appropriate sample sizes. To examine the impact of sample size on the accuracy of the ranking methods, we compared the MSEs obtained when the num-

TABLE 1. Conditions under which ranking based on raw values, standardized values, and sample percentiles will equal or approach the ranking based on true percentiles

Ranking method	Conditions
Raw values	Cortisol distribution is the same for all participants
Standardized values	Cortisol distribution is normal for all participants and true mean and variance of the distributions are known. If one of these parameters is unknown, the sample size per participant will need to be large
Sample percentiles	The sample size per participant is large

ber of participants being ranked varied from 5 to 30 and the number of cortisol values per participant varied from 3 to 30. We started our simulations at three cortisol values because the standard deviations needed to calculate standardized values cannot be computed with less than three values per participant.

RESULTS

In our sample, rankings based on raw values differ substantially from both those based on standardized values and sample percentiles (Fig. 2). In 10 of the 14 participants (71%), rankings based on raw values differed from those based on standardized values and sample percentiles. Rankings based on raw values were similar to rankings based on standardized (Z-score) values in four occasions (29%) and to those based on sample percentiles in only two instances (14%). In contrast, the rankings based on standardized values and sample percentiles were exactly the same for eight women (57%) (Fig. 2). Furthermore, these eight ranks represented the top and bottom four of our rank distribution, which suggests an important level of consistency. To evaluate the accuracy of the three ranking methods, we conducted simulations based on the actual distributions of our participants' cortisol data as described in the "Materials and Methods" section.

Figure 3 illustrates the ratios of the average MSE for the standardized and percentile methods relative to the average MSE based on ranks obtained using raw values. Ratios are presented as a function of the number of measurements per participant, and average MSEs are obtained by averaging the MSEs over 10,000 simulation runs. Regardless of the total number of participants being ranked, the MSE ratio for rankings based on standardized values are somewhat similar to those based on sample percentile rankings (Fig. 3). We attribute this similarity to the fact that, for many of our participants, the log transformation made cortisol distribution approximate normality. Without the log transformation we would have observed a stronger deviation from a Gaussian distribution, which, in turn, would translate into poorer performance for the standardized values method. In contrast, as log transformation is monotonic, the performance of the raw value and sample percentile methods would remain the same without the transformation. The average MSE of the raw value rankings, on the other hand, was higher than those obtained by the other two methods, i.e., MSE ratio < 1 (when n, the number of measurements per participant, was larger than five). Note that the raw value ranking method only utilizes the last measurements of each participant, so the average MSE is constant regardless of the total number of measurements available per participant. When the number of specimens per participant

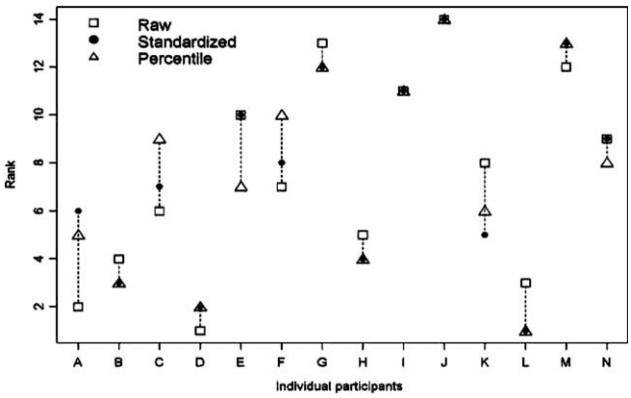

Fig. 2. Participants stress levels ranked based on raw values, standardized values, and sample percentiles for their last cortisol measurements. The ranks from the three methods are joined by a dotted line to visualize the difference between the ranks.

was low (<5), ranks based on raw values yield smaller average MSEs than the other two methods; when the sample sizes are very small, all average MSEs are fairly large, which implies that under those conditions the accuracy of the ranks is poor, independently of the method used. When the number of measurements per participant increases, the average MSEs for the ranks based on standardized values and sample percentiles decrease quite sharply while accuracy increase (Fig. 4). However, after about 10 measurements per participant, there is deceleration in the reduction of the average MSE, and after about 15 measurements that deceleration decreases even further.

DISCUSSION

As is the case with various other biomarkers, interindividual variation in glucocorticoid production is an "underutilized resource" by all: ecological physiologists, psychologists, and health scientists (Bennett, 1987). In a recent review, Williams revisited Bennett's (Bennett, 1987) "tyranny of the Golden Mean," the extended practice of focusing statistical analyses on average values, to emphasize the importance of incorporating interindividual variation into the study of dynamic endocrinological systems (Williams, 2008). Interindividual variation in endocrine regulation is a critical mediator of phenotypic plasticity. Plastic responses allow organisms to face challenges that take place across a wide variety of time frames ranging from those imposed by their ontogenetic trajectories, which may involve modification of basal function, to short term, acute challenges that only require ephemorous changes in physiological status. Cortisol is secreted following a complex pattern based on the overlap between ultradian oscillations and a 24-h circadian rhythm (Markovic et al., 2011). These patterns of cortisol secretion vary between individuals and within individuals across time. Thus, to make sense of an individual's current physiological status, it is critical to know and take into account the "starting point" for that individual at the time of the evaluation (i.e., basal function and normative response range). Only then meaningful comparisons between individuals can be drawn.

We examined three alternative methods to compare physiological stress levels across individuals based on single first morning urinary cortisol measurements: (i) raw

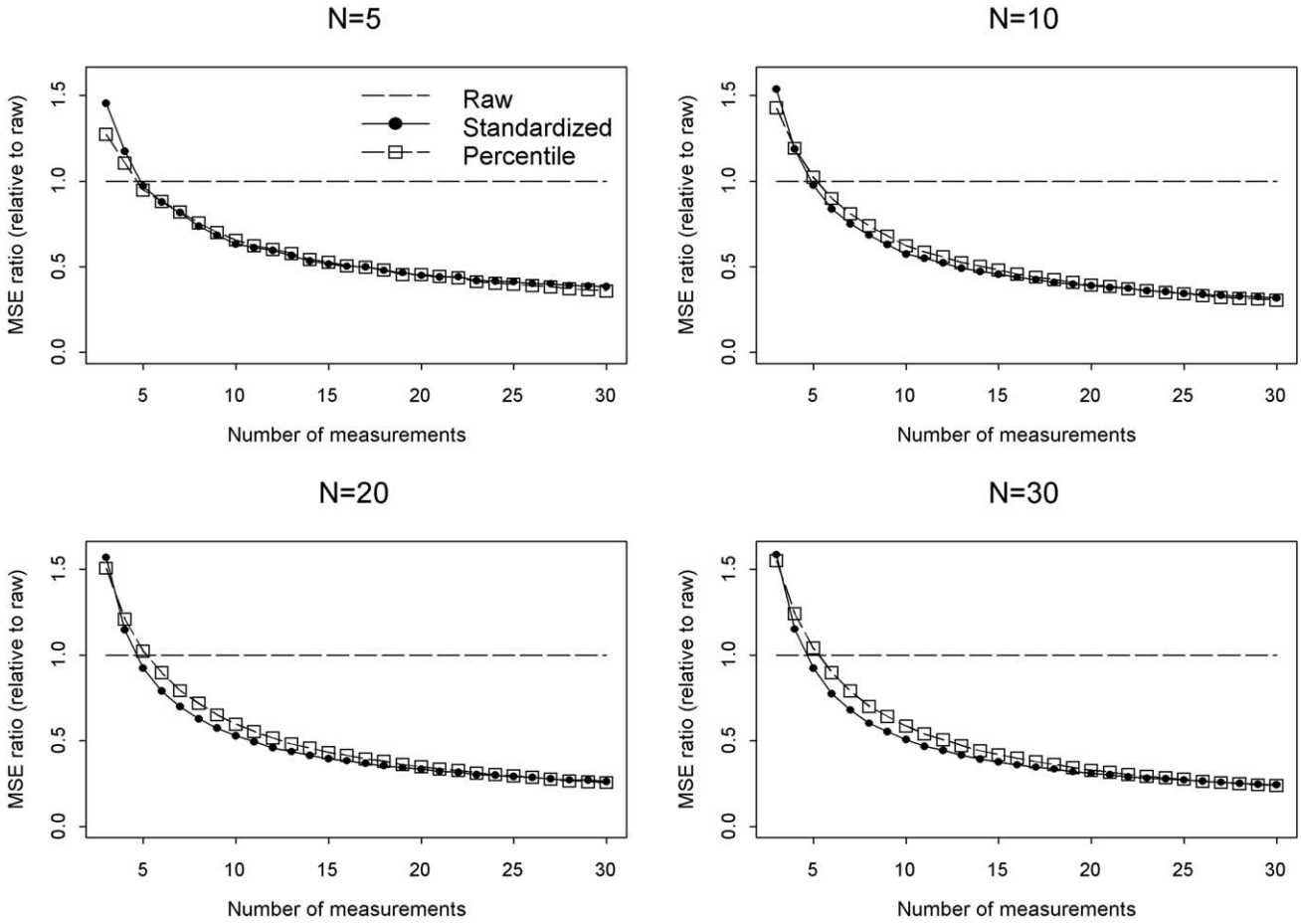

Fig. 3. Ratio of average MSE between true and estimated ranks based on the standardized and percentile methods relative to the average MSE based on ranks obtained using raw values. Plots show the results of 10,000 simulation runs based on kernel density estimates. Each plot corresponds to different values of N (the total number of subjects being ranked). MSE ratios are presented as a function of the number of measurements per subject. A MSE ratio > 1 suggests that the method being considered has larger average MSE than ranking by raw values.

values, (ii) standardized (Z-scores), and (iii) sample percentiles. The results of simulations based on data from a group of 14 women experiencing regular menstrual cycles suggest that the rankings obtained using standardized cortisol values and sample percentiles are similar, and that those two methods are more accurate than rankings obtained using raw cortisol values.

Our simulations suggest that to obtain an accurate approximation to each individual's cortisol distribution in our population it is necessary to collect between 10 and 15 first morning urinary specimens per participant (based on a sample of between 5 and 30 participants). The deceleration observed in the reduction of the average MSE after 15 measurements suggests that 15 would be the point of diminishing returns in terms of the gain in accuracy and statistical power provided by collecting extra measurements per participant. However, as none of the three methods is highly accurate when number of measurements per participant is small, we strongly advise researchers to collect a minimum of 10 specimens and as close to 15 as possible to make comparisons across individuals.

For comparative purposes, we performed two other simulation studies (see Appendix) using two different types of distributions to generate true ranks: (a) normal and (b) a mixture of two different normal distributions. These simulations suggest that when cortisol distributions are normal, rankings based on standardized cortisol values are more accurate (i.e., present lower average MSE) than those based on sample percentiles. This is not surprising because standardized value rankings are based on the assumption of normality. The ranking method based on sample percentile values, on the other hand, performs better than the standardized method when individual cortisol distributions deviate from normality more than our kernel density estimates. This is because sample percentile rankings do not place any assumption on the cortisol distributions and thus are not affected by deviations from normality.

In terms of their practical consequences, our findings suggest that if cortisol distributions can be normalized using, for example, a logarithmic transformation, then rankings based on standardized cortisol values are the most appropriate comparison method. Otherwise, it is most advisable to use sample percentiles to compare physiological stress levels between individuals.

There are important caveats to take into account when considering our results: first, the estimated cortisol distributions used in the analysis are based on only 14 women. This small sample may not represent the full range of var-

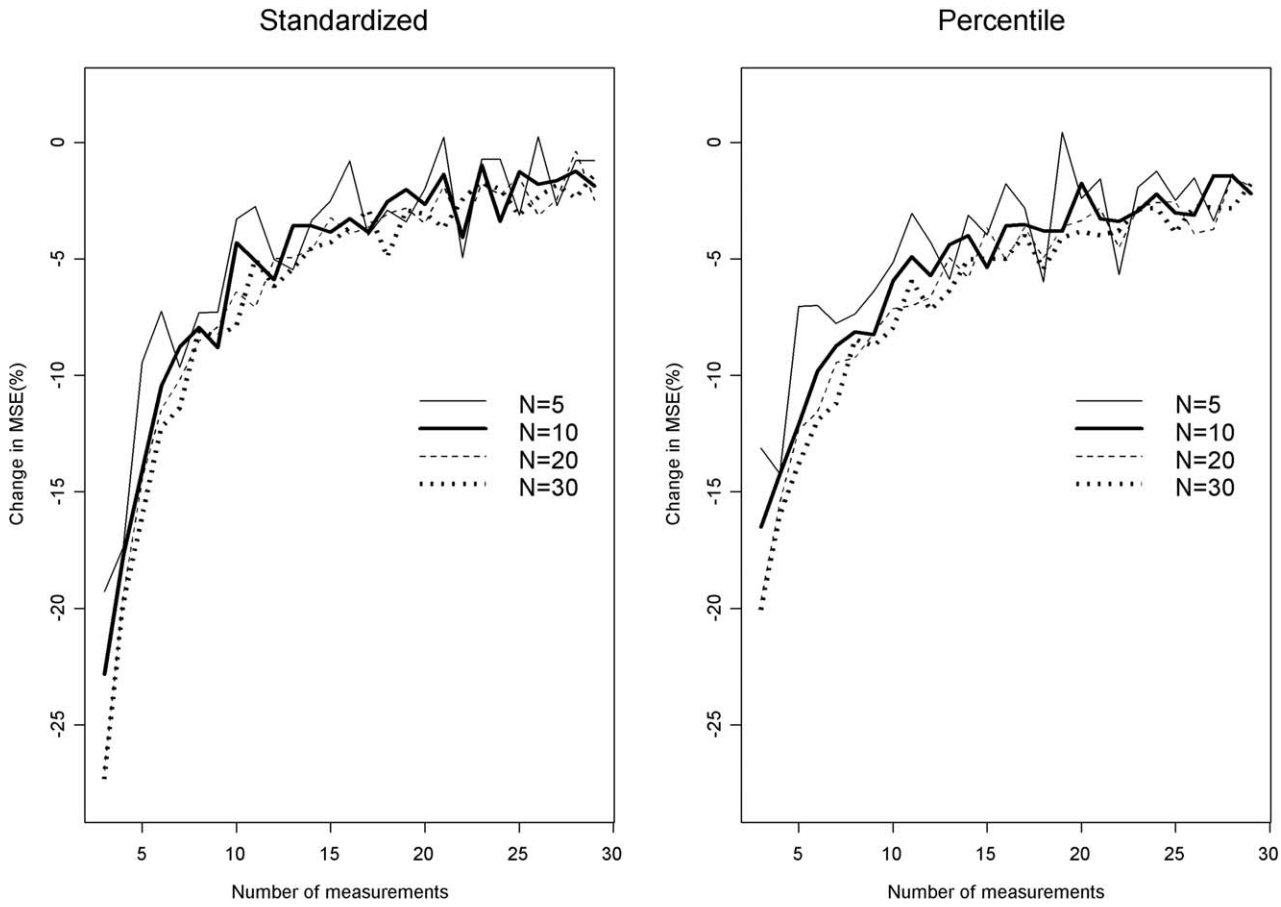

Fig. 4. Change in average mean square error of the estimated ranks based on standardized and percentile ranking methods for various values of N.

iation present in the study population. Second, the distribution of cortisol values amongst the Kakchiquel Mayan women living in a rural village may differ from that of other populations. Although the general nature of HPAA function is universal for all women, small differences in baseline cortisol and variability may exist between ethnic groups exposed to particular environmental pressures and, thus, different ontogenetic trajectories and life styles. Third, for the sake of simplicity, we restricted our analyses to data collected within ± 14 days of ovulation; yet, we have found evidence suggesting that cortisol levels may vary with the day of the menstrual cycle outside that period (Nepomnaschy et al., 2011). Therefore, further studies conducted with larger samples in a variety of populations are necessary to characterize normal ranges of interindividual variation in cortisol secretion. These studies will provide the data needed for an accurate assessment of the standard errors associated with the estimated ranks required for proper statistical inferences across different populations and throughout the entirety of the menstrual cycle. Nonetheless, these results represent a critical step forward in the development of standard methods needed to compare physiological stress levels across women using cortisol levels.

CONCLUSIONS

To summarize, consistent with previous studies, our results suggest that it is inadequate to use single raw cortisol values to compare physiological stress levels across individuals (Hruschka et al., 2005). Our findings imply that when using naturalistic designs, it is necessary to collect multiple specimens per individual. In our particular case, given the between-women variability in cortisol observed in our sample, a minimum of 10 specimens are needed to rank women according to their cortisol levels. Furthermore, our simulations provide an idea of the level of accuracy gained by increasing the number of participants and of specimens per participant. This information is paramount for the conception and design of naturalistic longitudinal studies of which the evaluation of stress is a part.

On the basis of our results, we recommend that future studies interested in comparing cortisol levels begin by collecting pilot cortisol data to assess the between-individual variability in the population of interest and then use the simulation procedure we outline in this article to determine the appropriate number of measurements per participant needed. If a preliminary assessment of the between-individual variability in the population of interest is not possible, then collecting around 15 specimens per participant in as many days would represent a conservative approach. The work of Hruschka et al. (2005) provides a complementary analysis on a different population reaching similar conclusions.

In terms of statistical analyses to compare cortisol levels between women, the most appropriate method appears to be

the standardized ranking method evaluated here. If individual cortisol distributions are unknown, however, then it is advisable to use the sample percentile method, which is more robustness when distributions deviate from normality.

Glucocorticoids are widely used to assess physiological stress across disciplines ranging from ecological physiology to psychology and public health. The HPA axis function is dynamic and complex as it depends on each individual's genome, ontogenetic environment, as well as the time of the day, physical activity, food consumption, and other ephemerous daily stimuli (Adam and Kumari, 2009). Therefore, it is difficult to compare the HPA axis activity between different individuals or to draw unequivocal conclusions about the overall physiological stress status of an individual using single time-point measurements of cortisol levels. Nonetheless, single cortisol measurements are frequently used for that purpose (Cagnacci et al., 2011; McBurnett et al., 2000; Phillips et al., 2011; Reynolds et al., 2010; Soriano-Rodriguez et al., 2010) as they are useful to compare physiological stress levels between individuals in a given group at a given time point or in response to a particular stressor. Our results, however, demonstrate that they can only be used appropriately in the context of each individual's cortisol distribution. The standardization methods we evaluated are appropriate for that use as well as for comparisons of cortisol profiles between individuals across time (Nepomnaschy et al., 2004).

These methods are simple, nonparametric, and properly account for the sizable amount of variation observed in cortisol and other glucocorticoids across individuals. These methods do not require a priori assumptions, are not time dependent, and are not based on linear models. Thus, they should be useful to all colleagues interested in comparing stress, or other physiological states, between individuals using naturalistic designs in human and nonhuman species. Future studies describing normative HPAA variation across life history (e.g., adrenarche, menarche, or menopause) and reproductive (e.g., postpartum, resumption of ovarian function, and gestation) transitions are needed, so that said information can be incorporated into standardization procedures.

ACKNOWLEDGMENTS

The authors thank the members of their Guatemalan research team for their assistance during fieldwork as well as Guatemala's Ministry of Health for permits and logistical collaboration. They thank the personnel of CLASS Laboratory at The University of Michigan for assistance in the hormonal analyses. They further thank Saranee Fernando and Rachel Williamson for their assistance with the literature review and Dr. Katrina Salvante and Rachel Williamson for their help in editing earlier versions of this manuscript. The funding sources had no involvement in the study design, data collection, analysis and interpretation of data, writing of the report, or in the decision to submit the article for publication.

LITERATURE CITED

Adam EK, Gunnar MR. 2001. Relationship functioning and home and work demands predict individual differences in diurnal cortisol patterns in women. Psychoneuroendocrinology 26:189–208.

Adam EK, Kumari M. 2009. Assessing salivary cortisol in large-scale, epidemiological research. Psychoneuroendocrinology 34:1423–1436.

Baird DD, Weinberg CR, Wilcox AJ, McConnaughey DR, Musey PI. 1991. Using the ratio of urinary oestrogen and progesterone metabolites to estimate day of ovulation. Stat Med 10:255–266.

Bennett AF. 1987. Interindividual variability: an under-utilized resource. In: Feder ME, Bennett AF, Burggren WW, Huey RB, editors. New directions in ecological physiology. Cambridge: Cambridge University Press. p 147–169.

Bonen A. 1976. Effects of exercise on excretion rates of urinary free cortisol. J Appl Physiol 40:155–158.

Cagnacci A, Connoletta M, Caretto S, Zanin R, Xholli A, Volpe A. 2011. Increased cortisol level: a possible link between climacteric symptoms and cardiovascular risk factors. Menopause 18:273–278.

Collins WP, Collins PO, Kilpatrick MJ, Manning PA, Pike JM, Tyler JP. 1979. The concentrations of urinary oestrone-3-glucuronide, LH and pregnanediol-3-glucuronide as indices of ovarian function. Acta Endocrinol 90:336–348.

Dahlgren A, Kecklund G, Akerstedt T. 2005. Different levels of work-related stress and the effects on sleep, fatigue and cortisol. Scand J Work Environ Health 31:277–285.

Dahlgren A, Kecklund G, Theorell T, Akerstedt T. 2009. Day-to-day variation in saliva cortisol—relation with sleep, stress and self-rated health. Biol Psychol 82:149–155.

D'Anna-Hernandez KL, Rossa RG, Natviga CL, Laudenslager ML. 2011. Hair cortisol levels as a retrospective marker of hypothalamic–pituitary axis activity throughout pregnancy: comparison to salivary cortisol. Physiol Behav 104:348–353.

Flinn MV, England BG. 1995. Childhood stress and family environment. Curr Anthropol 36:854–866.

Golden SH, Wand GS, Malhotra S, Kamel I, Horton K. 2011. Reliability of HPA axis assessment methods for use in population-based studies. Eur J Epidemiol 26:511–525.

Gunnar MR, Bruce J, Hickman SE. 2001a. Salivary cortisol response to stress in children. Adv Psychosom Med 22:52–60.

Gunnar MR, Morison SJ, Chisholm K, Schuder M. 2001b. Salivary cortisol levels in children adopted from Romanian orphanages. Dev Psychopathol 13:611–628.

Haus E. 2007. Chronobiology in the endocrine system. Adv Drug Deliv Rev 59:985–1014.

Hruschka DJ, Kohrt BA, Worthman CM. 2005. Estimating between- and within-individual variation in cortisol levels using multilevel models. Psychoneuroendocrinology 30:698–714.

Kanaley JA, Boileau RA, Bahr JA, Misner JE, Nelson RA. 1992. Cortisol-levels during prolonged exercise: the influence of menstrual phase and menstrual status. Int J Sports Med 13:332–336.

Kanaley JA, Hartman ML. 2002. Cortisol and growth hormone responses to exercise. Endocrinologist 12:421–432.

Kesner JS, Wright DM, Schrader SM, Chin NW, Kreig EF. 1992. Methods of monitoring menstrual function in field studies: efficacy of methods. Reprod Toxicol 6:385–400.

Kirschbaum C, Kudielka BM, Gaab J, Schommer NC, Hellhammer DH. 1999. Impact of gender, menstrual cycle phase, and oral contraceptives on the activity of the hypothalamus-pituitary-adrenal axis. Psychosom Med 61:154–162.

Kirschbaum C, Steyer R, Eid M, Patalla U, Schwenkmezger P, Hellhammer DH. 1990. Cortisol and behavior. II. Application of a latent state-trait model to salivary cortisol. Psychoneuroendocrinology 15:297–307.

Kudielka BM, Kirschbaum C. 2003. Awakening cortisol responses are influenced by health status and awakening time but not by menstrual cycle phase. Psychoneuroendocrinology 28:35–47.

Lasley BL, Mobed K, Gold EB. 1994. The use of urinary hormonal assessments in human studies. Ann N Y Acad Sci 709:229–311.

Markovic VM, Cupic Z, Vukojevic V, Kolar-Anic L. 2011. Predictive modeling of the hypothalamic-pituitary-adrenal (HPA) axis response to acute and chronic stress. Endocr J 58:889–904.

McBurnett K, Lahey BB, Rathouz PJ, Loeber R. 2000. Low salivary cortisol and persistent aggression in boys referred for disruptive behavior. Arch Gen Psychiatry 57:38–43.

Meulenberg EP, Hofman JA. 1990. The effect of pretreatment of saliva on steroid hormone concentrations. J Clin Chem Clin Biochem 28:923–928.

Miller RC, Brindle E, Holman DJ, Shofer J, Klein NA, Soules MR, O'Connor KA. 2004. Comparison of specific gravity and creatinine for normalizing urinary reproductive hormone concentrations. Clin Chem 50:924–932.

Negrão AB, Deuster PA, Gold PW, Singh A, Chrousos GP. 2000. Individual reactivity and physiology of the stress response. Biomed Pharmacother 54:122–128.

Nepomnaschy PA, Altman RM, Watterson R, Co C, McConnell D, England BG. 2011. Is cortisol excretion independent of menstrual cycle day? PLoS One 6:e18242.

Nepomnaschy PA, Welch K, McConnell D, Strassmann BI, England BG. 2004. Stress and female reproductive function: a study of daily variations in cortisol, gonadotrophins, and gonadal steroids in a rural Mayan population. Am J Hum Biol 16:523–532.

Nepomnaschy PA, Welch KB, McConnell DS, Low BS, Strassmann BI, England BG. 2006. Cortisol levels and very early pregnancy loss in humans. Proc Natl Acad Sci USA 103:3938–3942.
Nicolson NA. 2004. Childhood parental loss and cortisol levels in adult men. Psychoneuroendocrinology 29:1012–1018.
Phillips AC, Batty GD, Gale CR, Lord JM, Arit W, Carroll D. 2011. Major depressive disorder, generalized anxiety disorder, and their comorbidity: associations with cortisol in the Vietnam Experience Study. Psychoneuroendocrinology 36:682–690.
Pollard TM. 1995. Use of cortisol as a stress marker: practical and theoretical problems. Am J Hum Biol 7:265–274.
Pollard TM. 1997. Physiological consequences of everyday psychosocial stress. Coll Anthropol 21:17–28.
Pruessner JC, Wolf OT, Hellhammer DH, Buske-Kirschbaum A, von Auer K, Jobst S, Kaspers F, Kirschbaum C. 1997. Free cortisol levels after awakening: a reliable biological marker for the assessment of adrenocortical activity. Life Sci 61:2539–2549.
Reynolds RM, Labad J, Strachan MWJ, Braun A, Fowkes FGR, Lee AJ, Frier BM, Seckl JR, Walker BR and Price JF, on behalf of the Edinburgh Type 2 Diabetes Study (ET2DS) Investigators 2010. Elevated fasting plasma cortisol is associated with ischemic hearth disease and its risk factors in people with Type 2 diabetes: the Edinburgh Type 2 Diabetes Study. J Clin Endocrinol Metab 95:1602–1608.
Shirtcliff EA, Granger DA, Booth A, Johnson D. 2005. Low salivary cortisol levels and externalizing behavior problems in youth. Dev Psychopathol 17:167–184.
Soriano-Rodriguez P, Osiniri I, Grau-Cabrera P, Riera-Perez E, Prats-Puig A, Carbonell-Alferez M, Schneider S, Mora-Maruny C, De Zegher F, Ibanez L, Bassols J, López-Bermejo A. 2010. Physiological concentrations of serum cortisol are related to vascular risk markers in prepubertal children. Pediatr Res 68:452–455.
Stewart PM, Penn R, Holder R, Parton A, Ratcliffe JG, London DR. 1993. The hypothalamo-pituitary-adrenal axis across the normal menstrual cycle and in polycystic ovary syndrome. Clin Endocrinol 38:387–392.
Warnock F, McIsaac S, Macritchie KAN, Ramirez-Aponte T, Young AH, McElwee K, Seo RJ, Seim D. 2010. Measuring cortisol and DHEA in fingernails: a pilot study. Neuropsychiatr Dis Treat 6:1–7.
Weitzman ED, Fukushima D, Nogeire C, Roffwarg H, Gallagher TF, Leon H. 1971. Twenty-four hour pattern of the episodic secretion of cortisol in normal participants. J Clin Endocrinol Metab 33:14–22.
Williams TD. 2008. Individual variation in endocrine systems: moving beyond the 'tyranny of the Golden Mean'. Philos Trans R Soc Lond B 363:1687–1698.

APPENDIX

Here, we present results from two additional simulation studies for which we used distributions other than the kernel density estimates from our data to represent the individual cortisol distributions of each participant. In the first set of simulations, we used a normal distribution to generate data and to obtain the true rank. The underlying mean and variance for each subject's cortisol distribution are equal to the sample mean and sample variance for the corresponding subject in our data. The cortisol distribution for subject i is $N(\mu_{1i}, \sigma_{i2})$, where μ_{1i} is the sample mean and σ_{i2} is the sample variance for subject i in our data.

Our data's kernel density estimates suggest that the cortisol profiles of some of our participants can be classified into two states: a normal stress state and a high stress state. The mixture of normal distributions is a possible parametric model to describe this phenomenon. Thus, we conducted a second set of simulations by generating the data based on a mixture of two normal distributions. Under this mixture distribution framework, the cortisol distributions for these two stress states are allowed to follow two different normal distributions. We hypothesize that the cortisol distributions under these two states are both normal, but the mean cortisol value for the high stress state will be higher than that in the normal stress state. Thus, the cortisol distribution for subject i is defined as Eq. (A1): $p_i\,\mathrm{N}(\mu_{1i}, \sigma_i^2) + (1 - p_i)\,\mathrm{N}(\mu_{2i}, \sigma_i^2)$, where p_i is the probability that the subject is at the normal stress

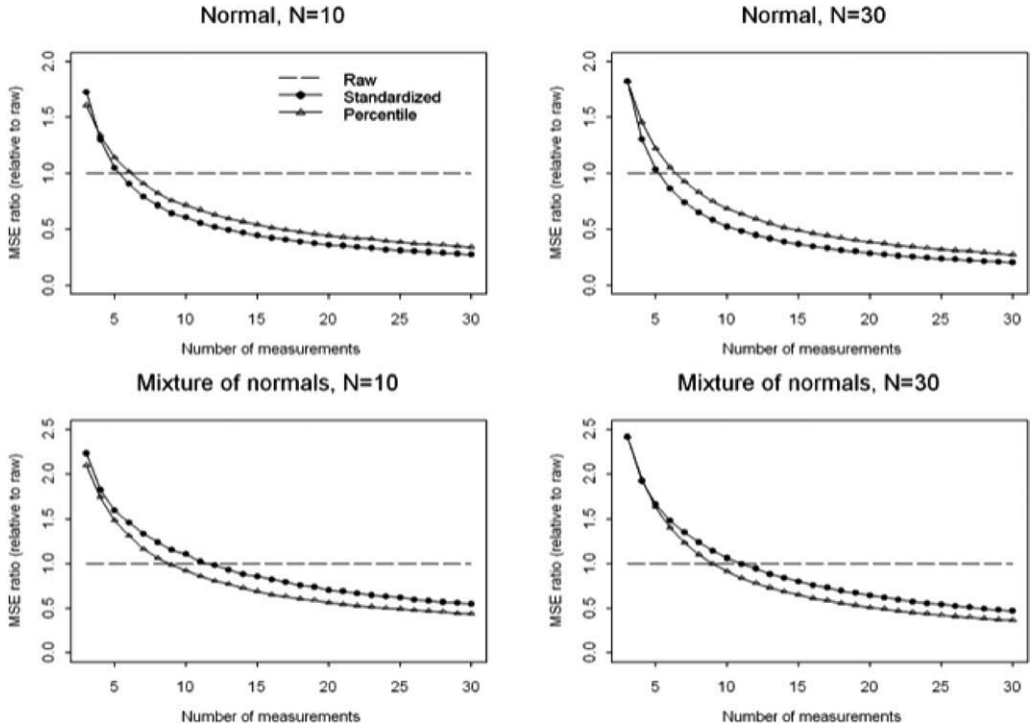

Fig. A1. Ratio of average MSE between true and estimated ranks based on the standardized and percentile methods relative to the average MSE based on ranks obtained using raw values. Plots show the results of 10,000 simulation runs from two additional simulation studies. MSE ratios are presented as a function of the number of measurements per subject. A MSE ratio > 1 suggests that the method being considered has larger average MSE than ranking by raw values.

state, $N(.)$ denotes a normal distribution, and $\mu_{1i} < \mu_{2i}$. In our simulation, we used $p_i = 0.9$ and $\mu_{2i} = 2\mu_{1i}$. The parameters μ_{1i} and σ_i^2 are set to be the sample mean and variance of subject i in our data, respectively.

Figure A1 shows the ratio of the average MSE of the estimated ranks based on the standardized or percentile method relative to the average MSE based on ranks obtained from the raw values; the averaging is over 10,000 simulation runs with the results displayed as a function of the number of measurements per subject. Our results are similar to those obtained using a kernel distribution in that the standardized values and sample percentiles ranking methods generally provide more accurate rankings than those obtained using raw values. For the first simulation, the MSE ratio based on standardized value rankings was most often smaller than those resulting from sample percentile ranking. This suggests that the standardized method is more accurate when the distributions approach normality. For the second simulation, the MSE ratio resulting from sample percentile ranking was smaller than those resulted from the standardized value ranking method. These results are explained by the fact that the sample percentile ranking method makes no assumption on the distribution of the data so when the distributions are not normal this method is more accurate.