Not within spitting distance: salivary immunoassays of estradiol have subpar validity for predicting cycle phase

Ruben C. Arslan[1,2], Khandis Blake[3], Laura J. Botzet[4], Paul-Christian Bürkner[5], Lisa DeBruine[6], Tom Fiers[7], Nicholas Grebe[8], Amanda Hahn[9], Ben C. Jones[10], Urszula M. Marcinkowska[11], Sunni L. Mumford[12], Lars Penke[4], James R. Roney[13], Enrique F. Schisterman[12], Julia Stern[14]

Abstract: Salivary steroid immunoassays are widely used in psychoneuroendocrinological studies of menstrual cycle phase, puberty, and menopause. Though manufacturers advertise their assays as suitable, they have not been rigorously validated for these purposes. We collated data from eight menstrual cycle studies across >1,200 women and >9,500 time points. Seven studies collected saliva and one collected serum. All assayed estradiol and progesterone and had an independent measure of cycle phase (LH-surge, menstrual onset). In serum, cycle phase measures strongly predicted steroid concentrations. In saliva, cycle phase poorly predicted estradiol values, which showed an upward bias compared to expectations from serum. For salivary progesterone, predictability from cycle phase was mixed. Widely used enzyme-linked assays performed poorly, while LC-MS/MS performed better. Imputing the population-average serum steroid changes from cycle phase may yield more valid values of hormonal changes for an independent person than directly assessing their hormone levels using salivary immunoassays.

Affiliations:

1. Personality Psychology and Psychological Assessment, University of Leipzig
2. Center for Adaptive Rationality, Max Planck Institute for Human Development
3. University of New South Wales, Sydney
4. University of Göttingen
5. University of Stuttgart
6. University of Glasgow
7. University of Gent
8. University of Michigan
9. California State Polytechnic University, Humboldt
10. University of Strathclyde, Glasgow
11. Institute of Public Health, Jagiellonian University Medical College
12. Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania
13. University of California, Santa Barbara
14. University of Bremen

OSF: osf.io/u9xad
Reproducible code documentation: https://rubenarslan.github.io/invalidity_on_steroids/

This is a peer-reviewed, accepted author manuscript of the following article: Arslan, R. C., Blake, K., Botzet, L. J., Bürkner, P. C., DeBruine, L., Fiers, T., Grebe, N., Hahn, A., Jones, B. C., Marcinkowska, U. M., Mumford, S. L., Penke, L., Roney, J. R., Schisterman, E. F., & Stern, J. (2023). Not within spitting distance: salivary immunoassays of estradiol have subpar validity for predicting cycle phase. Psychoneuroendocrinology, 149, [105994]. https://doi.org/10.1016/j.psyneuen.2022.105994

Introduction

Salivary immunoassays for estradiol and progesterone are widely used in psychoneuroendocrinology because they are cheap compared to other assays and easy to collect non-invasively. In research on the effects of the menstrual cycle phase, salivary measures of estradiol and progesterone are commonly used as indicators of cycle phase. Menstrual cycle phase effects are studied to test theories about sexual selection, to better understand sex differences, and to study cyclical changes in psychiatric and physical symptoms as well as cognitive abilities such as mental rotation. In recent years, there has been substantial controversy about replicability and problematic measures in menstrual cycle research (Harris et al., 2014; Jones et al., 2019). Since 2015, the NIH policy on "Sex as a biological variable" has directed researchers to include female research subjects and to consider female-unique factors such as cycle phase. Since 2021, "hormonal assessment for confirmation of cycle phase" was made a precondition for publication at the journal *Psychoneuroendocrinology (2022)* and this condition has mostly been fulfilled via steroid assays in saliva, as opposed to serum or urine (see Supplementary Note 1).

However, salivary immunoassays come with known issues (Granger et al., 2004; Schultheiss et al., 2018; Welker et al., 2016; Wood, 2009). Low concentrations of steroids, especially estradiol, are already very challenging to measure accurately in serum by immunoassay (Handelsman, 2017; Vesper et al., 2014a), mainly because of low specificity at lower concentrations (Garnett et al., 2020; Vesper et al., 2014b). The concentrations of estradiol and progesterone in saliva are only 1-2% of those in serum and reflect the free steroid concentration in serum, because only non-protein-bound forms can diffuse into saliva (Wood, 2009). Even though mean concentrations differ between free and total serum estradiol, the rank-order of total steroid concentration is largely preserved in free serum estradiol. By extension, the correlations of free estradiol with other variables such as cycle phase will also be similar (Dielen et al. 2017; Yeung et al. 2013). In saliva, contamination with small amounts of blood can substantially alter measured values , as can other errors in the pre-analytical phase (Celec and Ostatníková, 2012). The lower the concentration, the higher the specificity of the assay needs to be so that the signal is not overwhelmed by cross-reactivity or interference with other substances. Despite these challenges, Salimetrics, a widely used (see Supplementary Note 1) provider of salivary immunoassay kits and services, reports correlations of *r*s = .80/.87 between salivary and serum immunoassays of estradiol and progesterone, respectively (Salimetrics, 2020, 2019).

While serum measures of estradiol and progesterone show clear relationships with menstrual cycle phase and ovulation (Lynch et al., 2014), salivary measures have not been validated to the same extent. Manufacturers reported small-scale studies with mean salivary values and ranges grouped by cycle phase, with *N*s ranging from 18 to 20 for estradiol and *N*s from 27 to 202 for progesterone (IBL, 2019, 2015; Salimetrics, 2020, 2019). However, manufacturers do not report how they estimated cycle phase in these studies. Salimetrics (2020, 2019) additionally reported time series for one woman with daily progesterone and estradiol assays in serum and saliva for one cycle, which visually show some parallelism. Compared to Salimetrics' (2020, 2019) numbers, independent validations often show smaller saliva-serum correlations;

estimates vary widely, and are poorer at lower concentrations (Tivis et al. 2005; Shirtcliff et al. 2000; Dielen et al. 2017; Sun et al. 2019; Lu et al. 1999; Sakkas et al. 2021). The assay manufacturers IBL and Salimetrics do not report raw data of validation studies and only minimal information on the sample of women and their cycles and only rarely show scatter plots which would allow the assessment of heteroskedasticity and influential outliers. We are not aware of any study that directly validates multiple salivary estradiol and progesterone immunoassays against an independent measure of cycle phase within subjects.

In the current study, we aimed to close this gap. We obtained raw data from eight studies (Blake et al., 2017; Grebe et al., 2016; Jones et al., 2018; Jünger et al., 2018; Marcinkowska, 2020; Roney and Simmons, 2013; Stern et al., 2021; Wactawski-Wende et al., 2009) that collected repeated data from women across the menstrual cycle and measures of estradiol and progesterone, plus at least one independent measure of cycle phase (i.e., cycle day relative to the luteinising hormone surge or a menstrual onset). We compared the steroid measures across datasets with respect to averages, inter-individual differences and the strength of the association between hormones and our independent cycle phase measures, as well as the probability of being in the fertile window.

Methods

Datasets were obtained from public online repositories or first authors of the relevant publications. We attempted to pool data from multiple laboratories and assays, but used no systematic sampling strategy. Rather, we used eligible datasets that were either shared publicly or upon request by the first author within a reasonable timeframe (<1 year). All studies collected data only on adult women of reproductive age who were naturally cycling and not using hormonal contraception.[1] Because the datasets varied widely in how they were formatted, all data sets were first brought into the same standard format. This involved transforming all hormone measures to pg/ml, standardising cycle phase measures as described below, and restructuring the data so that cycle days were nested within women within studies (with cycle phase and hormones as columns). All datasets were analysed using an identical pipeline with allowance made for whether studies collected multiple cycles per woman or not. Each researcher checked the transformed version of their dataset for accuracy prior to analysis. Key features of the datasets are summarised in Table 1. All statistical code and intermediate results, as well as several of the datasets are on the OSF (osf.io/u9xad). The data for the BioCycle study can be obtained via NIH DASH. Several other studies shared their data on OSF, the relevant sources can be found in the references and on our OSF repository. All studies were subject to ethical review according to local regulations; details can be found in the respective publications (Blake et al. 2017; Grebe et al. 2016; Jones et al. 2018; Jünger et al. 2018; Marcinkowska 2020; Roney and Simmons 2013; Stern et al. 2021; Wactawski-Wende et al. 2009).

[1] By women in this context of menstrual cycle research, we are referring to biologically female persons. In studies where information on both gender identity and biological sex was collected, only cisgender women enrolled (see Supplementary Note 7).

Steroid assays

The BioCycle study collected serum, all others collected saliva, and most quantified hormones using enzyme-linked immunosorbent assays (ELISAs). Two studies quantified salivary progesterone using liquid chromatography tandem mass spectrometry (LC-MS/MS), one quantified salivary progesterone using a radioimmunoassay, and one quantified estradiol using a chemiluminescence immunoassay (see Table 1). Hormone values were log-transformed for the main analyses, but as a robustness check we also repeated central analyses with hormones untransformed, within-subject-centred raw hormones, and within-subject-centred after log-transformation. Measured hormone values can be left-censored, when values are at or below the limit of detection and not precisely quantifiable. A flag for left-censoring was added during data processing for all datasets based on laboratory notes where available or when values were at the limit of detection reported for the assay. For the BioCycle data, we applied a mass-action based algorithm to estimate the free estradiol level from the measured serum values for total estradiol, testosterone, sex-hormone binding globulin, and albumin (Dunn et al., 1981; Vermeulen et al., 1999).

Cycle phase

The menstrual cycle can be divided into the follicular phase, from menstrual onset to ovulation, and the luteal phase, from ovulation until menstrual onset. Often, the phases are further subdivided into early, mid, and late and some authors define a peri-ovulatory phase for the time of highest fertility or a peri-menstrual phase for the days surrounding the menstrual onset (Schmalenberger et al., 2021). The cycle phases vary in length from woman to woman and cycle to cycle and estradiol is highly variable within cycle phases. Here, we use the term *cycle phase* probabilistically, to indicate a day in the cycle relative to menstrual onset or urinary luteinizing hormone (LH) surge (see Table 1). We did not assign cycle days into phases, because in the absence of sonographic confirmation of ovulation measurement error and individual differences preclude a certain assignment to a single phase. In addition, a continuous approach better captures the signal in the data.

Studies differed in how they scheduled measurement time points. Two studies collected saliva every day for the whole cycle (Marcinkowska, 2020; Roney & Simmons, 2013), though they did not assay all samples. Two studies did not schedule appointments according to cycle phase, leading to a uniform distribution (Grebe et al. 2016; Jones et al. 2018). The other studies used a forecast of cycle phase to schedule appointments at specific times during the cycle (e.g., peri-ovulatory and luteal, see Table 1).

There were three approaches to estimate cycle phase independent of steroid hormones: counting *forwards* from the last recalled menstrual onset, counting *backwards* from the next observed menstrual onset, and counting from urinary measures of the day of the LH surge. Forward counting was possible for all datasets, but is known to provide the least valid estimate of the day of ovulation because of reporting errors for the last recalled menstrual onset and the high variability of the follicular phase's length (Blake et al., 2016; Gangestad et al., 2016; Schmalenberger et al., 2021). Backward counting was possible for all datasets, with the exception of Grebe et al. (2016), as the authors in this study did not follow up with participants

at their next menstrual onset. Because the luteal phase is less variable in length than the follicular phase and recall errors are reduced in prospective designs, backward-counting approximates the day of ovulation more precisely than forward-counting. However, anovulatory cycles cannot be identified using counting methods and variability remains substantial (Gangestad et al., 2016). Five studies additionally had women perform urinary LH tests at home. Such tests can detect the LH surge that precedes ovulation and are generally considered more valid than backward counting at a potential cost of improperly classifying cycles as anovulatory when the LH surge is borderline (Lynch et al., 2014; Marcinkowska, 2020). In summary, studies had between one and three measures of cycle phase that could be estimated independently from steroid hormones and each other.

For all three indicators, we first determined the day of the last menstrual onset and, if possible, of the next menstrual onset and the LH surge. Then, we estimated the relative position in the cycle of each day where steroid hormones were measured. We defined cycles as beginning on the day of menstrual onset and ending on the last day before the next menstrual onset. Therefore, the minimal value for forward-counted days was 0, the maximal value for backward-counted days was -1, and days relative to the LH surge ranged from -15 to 15 (observations further from the LH surge were discarded owing to their rarity). Counting in this way, the day of ovulation was expected to be on average on day 13 after the last reported menstrual onset, day -15 before the next observed menstrual onset or day 1 after the LH surge. Based on these cycle days, we were able to estimate the probability of being in the fertile window (i.e., when sex can result in conception) as outlined in Gangestad et al. (2016) and Stern et al. (2021) for each day (see also OSF merge files). We used this probability as another, more targeted measure of cycle phase, because many studies use salivary steroids to infer fertility status.

If cycle length was known, cycles shorter than 20 or longer than 35 days were excluded to reduce the odds of including irregular, anovulatory cycles (Magyar et al., 1979) and cycles in which a conception had occurred and was spontaneously aborted before detection. If cycle length was unknown, we excluded forward- and backward-counted days that exceeded the 35 day cutoff.

In addition to the steroid-independent cycle phase measures, we also computed a steroid-based measure of cycle phase, see Supplementary Note 2.

Main analyses

After hormone values had been log-transformed, we deemed no additional treatment of outliers necessary based on visual inspection.[2] Bayesian multilevel regressions were used to estimate the hormone's association with cycle phase separately for each hormone and dataset. To this end, log hormone values were modelled as Gaussian outcomes. Reported limits of detection (LODs), or analytical sensitivities, that is, the smallest values that could be distinguished from zero at 95% certainty, were used to model left-censoring, that is the fact that true values might

[2] We also report associations between fertile window probability and non-transformed hormones as a robustness check. Here, we did not exclude outliers either, because we know of no agreed-upon purely data-driven procedure to exclude values that are inconsistent with the assumed data-generating process that we could apply consistently across heterogeneous datasets.

be at the LOD or lower. The LODs are shown in all subsequent graphs as solid lines. Limits of quantitation, or functional sensitivities, that is, values at which the coefficient of variation reached 20%, are shown as dashed lines. Where only one line is shown, the other limit was not reported, or in the case of one IBL assay, both limits were reported as the same number. For all studies except Stern et al. (2021), censoring was rare (0-4%) and censored values were set to the LOD. For Stern et al.'s IBL estradiol ELISA (2021) censoring was common (12% of values) and we kept observed values below the LOD in subsequent analysis (setting them to LOD did not appreciably change any numbers or conclusions). All limits are reported exactly in Supplementary Note 7. Varying (random) intercepts for the woman and, if multiple cycles were covered, each cycle were added to estimate variance related to inter-individual and inter-cycle differences and to adjust standard errors for the data structure. For each available cycle phase measure, a cubic spline (Wood, 2003), a flexible piecewise polynomial function, was estimated across cycle day to continuously capture variation explained by cycle phase without discretizing the cycle *a-priori* into, for instance, follicular and luteal phase. Cubic splines allow us to smoothly interpolate hormone values over time without excessive oscillation, which high-degree polynomials can engender.

All analyses were computed with the statistical software R (R Core Team, 2021) and all multilevel models with the package brms (Bürkner, 2017) which implements an R interface to the probabilistic programming language Stan (Stan Development Team, 2022). We used default, minimally informative priors and checked convergence via the Rhat and effective sample statistics across four parallel chains. If chains did not converge or excessive divergences occurred, we increased the number of iterations or the adapt-delta parameter of the sampling algorithm.
We then estimated the variance explained by cycle phase with a Bayesian model-based R^2. As a safeguard against overfitting, which is likely when cubic splines are applied to small datasets, our main reported coefficient uses an approximative leave-one-out-adjustment (Vehtari et al., 2018), LOO-R^2. Where we use the coefficient to make claims about validity, we always report the square root of variance explained net of inter-individual and inter-cycle variation, i.e. LOO-R, not LOO-R^2, to make the coefficient more comparable to the correlations reported as evidence for validity in the literature. Where we use the coefficient to estimate the amount of variance explained by inter-individual or inter-cycle differences, we report LOO-R^2 to make it comparable to the intra-class correlations commonly reported in the literature.

Comparison to imputed serum values

Using the BioCycle data on cycle phase, we could predict serum values for estradiol and progesterone from cycle day (relative to the LH surge or a menstrual onset) for an average woman and cycle. The BioCycle data was used as up to 8 serum measures per cycle were available for 2 cycles, with visits well-timed to cycle phase using urinary fertility monitors which measured estradiol metabolites and luteinising hormone (Howards et al., 2009).
To more directly capture whether salivary measures performed similarly to serum measures, we then used the BioCycle models to impute serum hormone levels from the cycle phase estimates in all datasets. Three imputation models, one for each cycle phase measure, were estimated in

the BioCycle data, as described above, with cubic splines over cycle day. Average predicted mean hormone values for an average woman were generated for one cycle, i.e. one value per cycle day. These average predictions were merged on the other datasets by cycle day.
We then computed Pearson correlations at the individual level between the *measured* log hormones and the *imputed* log hormones for all datasets. In other words, for each cycle day, the imputed average hormone value was paired with each individual measured hormone value on that day and a correlation was then computed across all pairs for all cycle days. In addition, we took two steps to reduce the influence of differences in study design. The variance explained in our main models could be reduced or inflated depending on the sample characteristics, which might affect inter-individual differences, and depending on the scheduling procedure, which directly affects the variance of the cycle phase predictors. First, we subtracted the subject mean from the measured log hormones to account for the fact that imputations cannot recover interindividual differences and correlated the measured within-subject-centred log hormones with imputed log hormones. Second, we applied a correction for range restriction (Cohen et al. 2003; Fletcher 2010) to the correlation estimated in the first step. For the correction, we estimated the ratio of the observed standard deviation in the imputed hormone in each dataset to the standard deviation expected after daily measurement in a 29-day cycle. Some of the studies restricted hormone measures to the peri-ovulatory and luteal phase by design. As both progesterone and estradiol are at their lowest during menses, such designs restrict variation in hormones and attenuate the estimated correlations. In that case, correcting for range restriction implies that the correlation of imputed log hormones with within-subject differences in measured log hormones increases in proportion to the ratio. Thus, the correction should make the correlations more comparable across datasets that differed in scheduling (see also Supplementary Note 9 for a sanity check). In a final step, we now had estimates of the correlations between measured serum hormones and each of the three cycle-phase-based imputations in the BioCycle data ($r_{Imputation,\,Serum}$, i.e. the correlation between cycle-phase-based imputations and the measured serum values, closely related to the square root of the variance explained by the cycle phase predictor in this model) as well as estimates of the correlation between measured salivary hormones and the cycle-phase-based imputations, again based on the BioCycle data ($r_{Imputation,\,Saliva}$). We assumed linear associations and that there is no direct causal relationship between cycle phase and salivary steroid levels, so that their association would be fully accounted for by serum steroid levels. We could then use path tracing rules (Wright, 1934) to indirectly arrive at a rough expectation of the correlation between measured serum and measured salivary hormones, which we could not directly estimate (Eq. 1, see Supplementary Note 3 for all required assumptions and further explanation).

Eq. 1: $$r_{Serum,\,Saliva} = \frac{r_{Imputation,\,Saliva}}{r_{Imputation,\,Serum}}$$

Importantly, a comparison of $r_{Serum,\,Saliva}$ to $r_{Imputation,\,Serum}$ speaks to the question whether measured salivary hormone concentrations would be more highly correlated with individual serum measures than are the average serum measures for the same cycle days. A larger value for $r_{Imputation,\,Serum}$ than for $r_{Serum,\,Saliva}$ could therefore imply that individual serum hormone values can be better estimated by imputation than by salivary assay.

Prediction of fertile window probability

We also tested how well estradiol and progesterone could predict the estimated probability of being in the fertile window (Gangestad et al., 2016; Stern et al., 2021), either individually or jointly in the form of a ratio or as a flexible nonlinear interaction implemented as a thin-plate spline.

Results

We found large differences in variance explained by cycle phase, inter-individual and inter-cycle differences between assays in serum and tandem mass spectrometry in saliva on the one hand and immunoassays in saliva on the other hand (Figure 5).

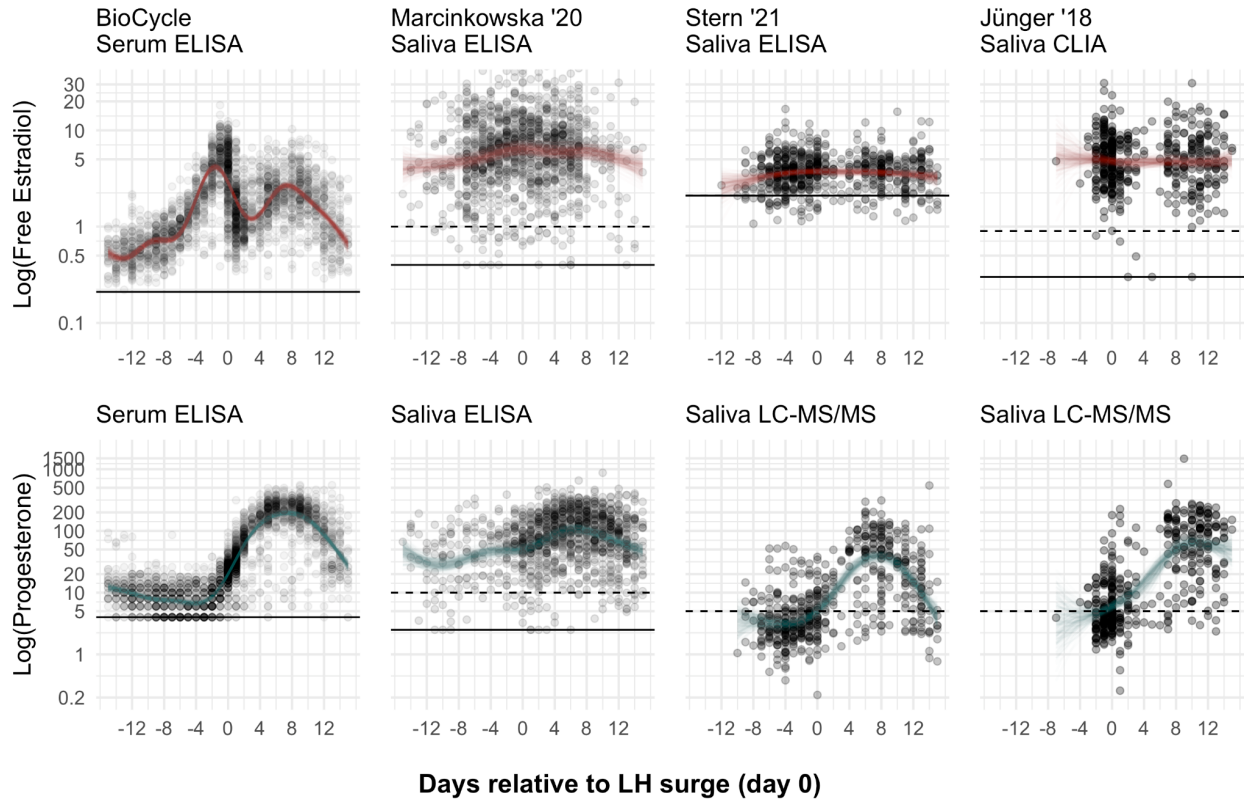

Figure 1. Associations with cycle day relative to the luteinising hormone surge (day 0) in the four largest datasets that tested urinary luteinising hormone. Dots show raw data. Coloured lines show two hundred random samples of the cubic spline fit using a Bayesian multilevel regression. Solid horizontal lines show the limit of detection; dashed the limit of quantitation. Progesterone values for BioCycle were multiplied by 2% as per Wood (2009) to make scales comparable.

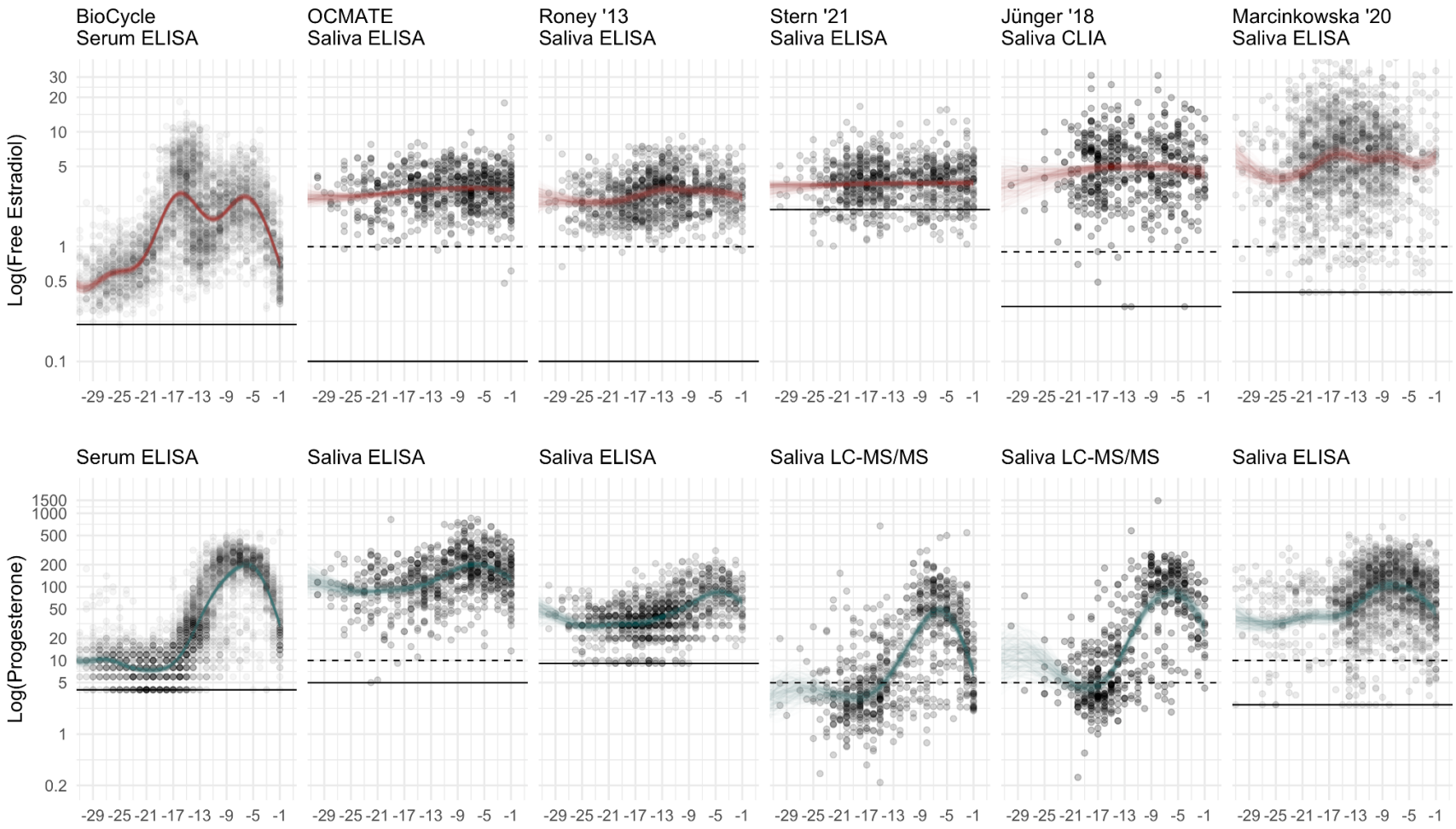

Figure 2. Associations with cycle day relative to the observed next menstrual onset in the six largest datasets. Dots show raw data. Coloured lines show two hundred random samples of the cubic spline fit using a Bayesian multilevel regression. Solid horizontal lines show the limit of detection; dashed the limit of quantitation. Progesterone values for BioCycle were multiplied by 2% as per Wood (2009) to make scales comparable.

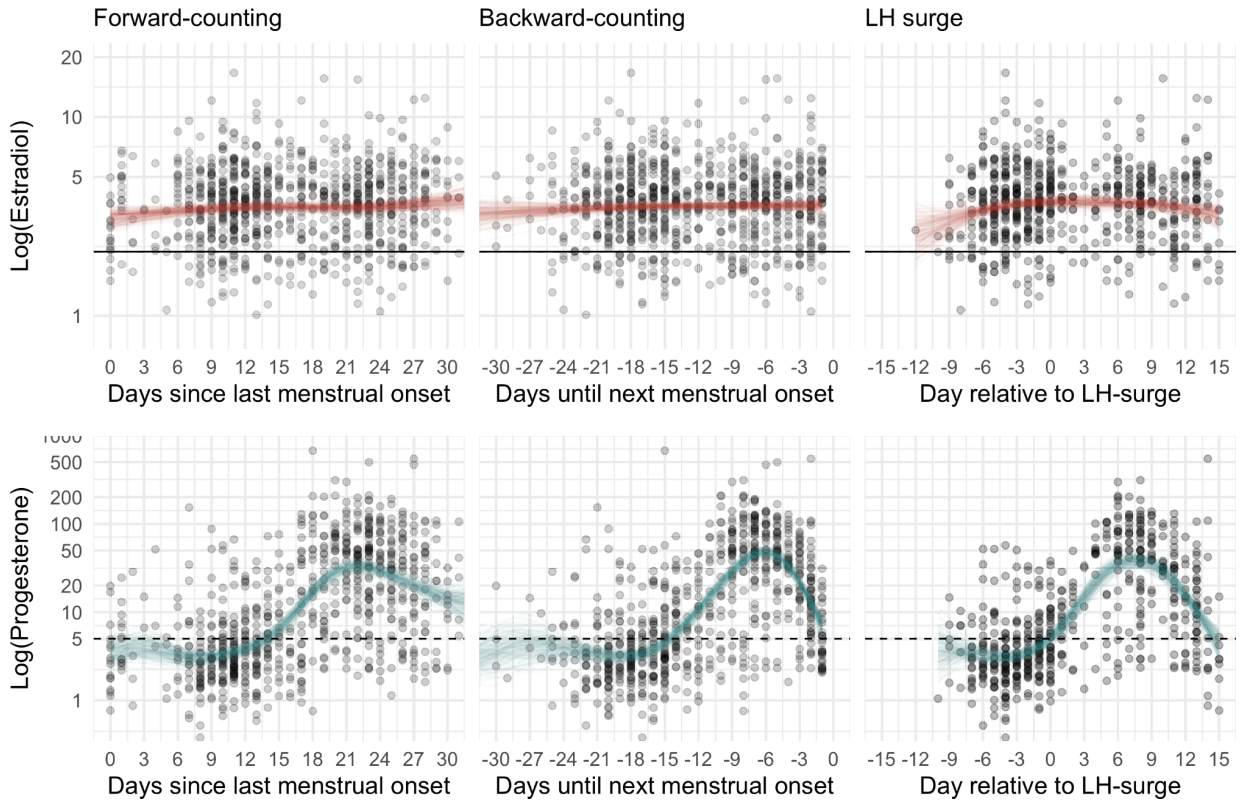

Figure 3. Associations with cycle day relative to the three different measures of cycle phase in Stern et al. (2021). Dots show raw data. Coloured lines show two hundred random samples of the cubic spline fit using a Bayesian multilevel regression. Solid horizontal lines show the limit of detection; dashed the limit of quantitation.

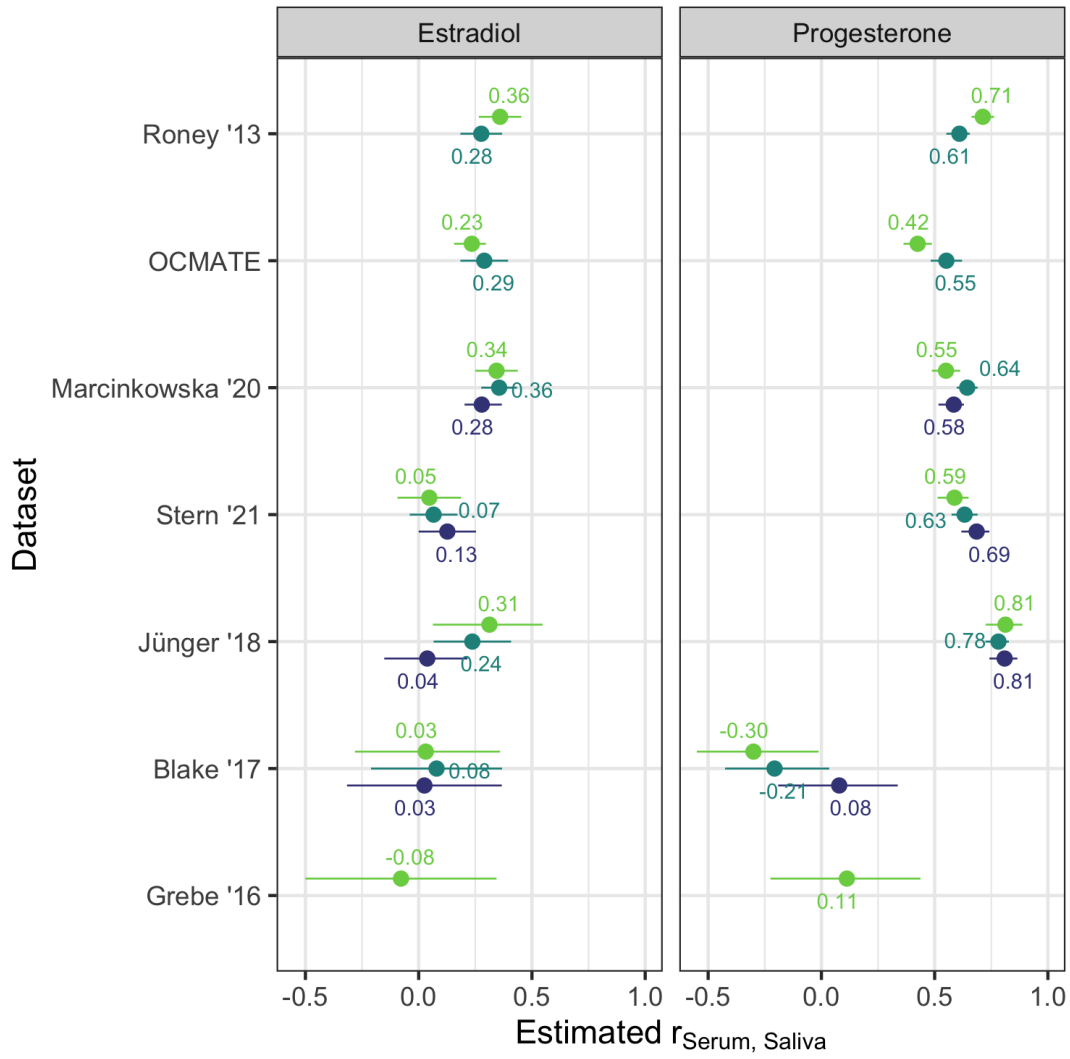

Figure 4. Correlations between serum and saliva, indirectly estimated from BioCycle cycle phase imputations as described in Eq. 1 and Supplementary Note 3. Colours reflect cycle phase measures. Green = forward-counted, blue = backward-counted, violet = relative to LH surge.

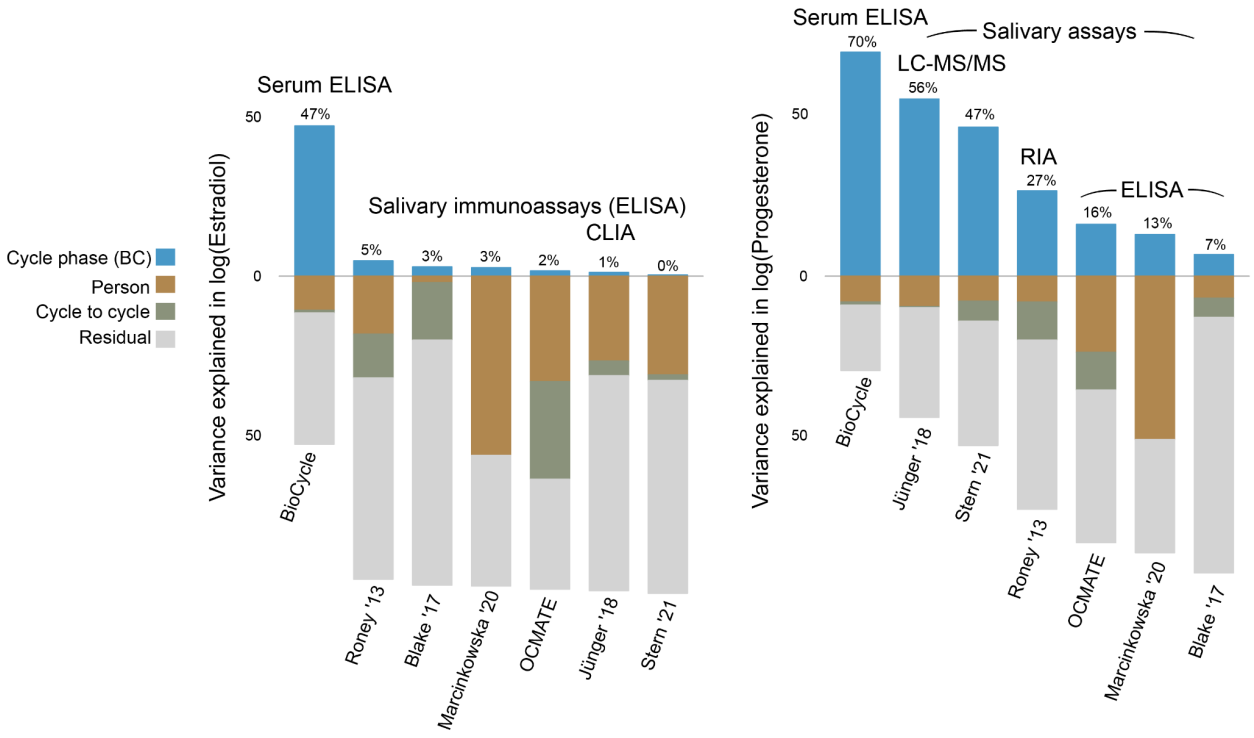

Figure 5. Variance explained in log estradiol and log progesterone, by dataset. Variance explained by backward-counted cycle day above the zero line, variance explained by inter-individual and inter-cycle differences, as well as residual variance below the line. LC-MS/MS = liquid chromatography tandem mass spectrometry; RIA = radioimmunoassay; CLIA = chemiluminescence immunoassay; ELISA = enzyme-linked immunosorbent assay. Variance components are shown without approximative leave-one-out adjustment, so that they sum to 100%, but can be inflated in the smaller datasets owing to overfitting.

Estradiol

In the BioCycle serum data, the urinary LH surge measure of cycle phase explained more than half the variance in estradiol (LOO-R = 0.72 95% credible interval [0.70;0.74]). Inter-individual differences accounted for a small percentage of the variance (LOO_R^2 = 0.06 [0.04;0.07]); additionally allowing for inter-cycle differences did not increase explained variance (LOO_R^2 = 0.05 [0.04;0.07]). With backward- and forward-counting the variance explained by cycle phase was somewhat reduced (LOO-R = 0.68 [0.66;0.69] and LOO-R = 0.57 [0.55;0.59]). Conditional effect plots of the cubic spline captured both the pre-ovulatory major peak of estradiol as well as the luteal minor peak, when predicted using backward-counting or LH (see Figures 1-2, and Supplementary Figure 2). The two peaks were less clearly separated when using forward-counting (see Supplementary Figure 1). In approximately the first week after the menstrual onset (days 0-6) and the first two days before the next menstrual onset (days -2 and -1), estimated mean values of free estradiol were below 1pg/ml.

In all salivary immunoassay datasets, the variance explained by cycle phase was much lower. The leave-one-out-adjusted *r* never exceeded .14, was indistinguishable from zero more often than not, and was not systematically larger for more valid measures of cycle phase. Inter-individual differences accounted for a larger percentage of the variance, on average (LOO-R^2s from negligible to 0.52); additionally allowing for inter-cycle differences occasionally substantially increased variance explained (LOO-R^2s from 0.04 to 0.51). The two estradiol peaks could not be discerned from conditional effect plots, and even the expected dip toward menstruation was not clearly apparent in all datasets (see Figures 1-3, and Supplementary Figures 1 and 2). The Salimetrics immunoassays have a reported limit of detection at 0.1pg/ml, but we observed very few values below 1pg/ml, the limit of quantitation, even in the days surrounding menstruation (see Figures 1-3). Censoring was never reported for Salimetrics assays.

When we compared the variance components in a model with backward-counted cycle phase as the predictor, differences were striking. For serum, cycle phase explained the most variance, whereas for saliva, inter-individual and inter-cycle differences dominated (Figure 5). These figures are not adjusted for differences in scheduling procedure, nor leave-one-out-adjusted. As such, they sum to 100%, but may be inflated by overfitting and affected by the study design.

When we divided the corrected saliva-imputed correlation by the serum-imputed correlation as per Eq. 1, the median value was 0.23 for the expected $r_{\text{Serum, Saliva}}$. Values ranged from -0.08 to 0.41 (see Figure 4). The highest values were seen for forward-counting rather than LH, and were largely a function of disattenuation for greater invalidity of the denominator rather than greater validity of the numerator. That is to say, $r_{\text{FC, Saliva}}$ was not higher than $r_{\text{LH, Saliva}}$, but because $r_{\text{FC, Serum}}$ was low, the estimated $r_{\text{Serum, Saliva}}$ was boosted for forward-counted cycle phase.

By contrast, the imputation models allowed us to generate estimates of serum estradiol from backward counting or LH tests that had a correlation ($r_{Imputation,\,Serum}$) of .68 or .72 with measured serum values and correlations of .76 and .79 with within-subject differences after correcting for range restriction. These imputed estimates easily exceed all our indirect estimates (see Eq. 1) of the true correlation of salivary with serum estradiol and come close to the correlation reported by Salimetrics (r=0.80).

Progesterone

In the BioCycle serum data, the LH measure of cycle phase explained three quarters of the variance in progesterone (LOO-R = 0.87 [0.85;0.88]). Inter-individual differences accounted for a small percentage of the variance (LOO_R^2 = 0.02 [0.01;0.02]); additionally allowing for inter-cycle differences did not increase explained variance (LOO_R^2 = 0.02 [0.01;0.02]). With backward- and forward-counting the variance explained by cycle phase was somewhat reduced (LOO-R = 0.83 [0.81;0.84] and 0.72 [0.70;0.74]). Conditional effect plots (Figures 1 and 2-) of the cubic spline captured the marked rise in progesterone around ovulation as well as a marked decrease towards the next menstrual onset. The expected pattern was clearest using LH or backward-counting, but still apparent using forward-counting (see Figures 1, 2, and Supplementary Figure 1). In the follicular phase, estimated mean values of total progesterone varied around a mean of approximately 500 pg/ml. Note that we multiplied serum progesterone by 2% in Figure 1 and 2 to approximate the concentrations seen in saliva (Wood, 2009).

In the two datasets that assayed progesterone using tandem mass spectrometry, findings were visually similar (Figures 1-2), but cycle phase explained less variance (e.g., LOO-Rs = 0.68 [0.62;0.73] and 0.69 [0.63;0.74] for LH as predictor). Inter-individual differences and inter-cycle differences accounted for a negligible portion of variance (i.e., LOO-R^2s veered negative). In the follicular phase, estimated mean values of free progesterone varied around a mean of 5pg/ml, the limit of quantitation for the assay.

In the salivary immunoassay datasets, the variance explained by cycle phase was lower, ranging from indistinguishable from zero using LOO-R to 0.48. Inter-individual differences accounted for a larger percentage of the variance, on average (LOO-R^2s from negligible to 0.39); additionally allowing for inter-cycle differences did not substantially increase variance explained (LOO-R^2s from negligible to 0.32). In some datasets (especially Marcinkowska 2020 for progesterone), the variance in cycle phase was severely restricted. In the larger datasets, the expected pattern was visible in the conditional effect plots (Figures 1 and 2) but weaker, with less clear separation between follicular and luteal phase. Interestingly, although the salivary immunoassays for progesterone report limits of quantitation and detection between 2.5-10pg/ml, the assays rarely called values below 10pg/ml. Even in the follicular phase, assays averaged between 20 and 100pg/ml across datasets (see Figures 1-3). Censoring was rare, but more frequent than for estradiol.

When we compared the variance components in a model with backward-counted cycle phase as the predictor, differences were striking. For serum and salivary tandem mass spectrometry, cycle phase explained the most variance and inter-individual and inter-cycle differences explained little, whereas for salivary enzyme-linked immunoassays, inter-individual and inter-cycle differences were larger or on par. The salivary radioimmunoassay fell between these two extremes (Figure 5).

When we divided the corrected saliva-imputed correlation by the serum-imputed correlation, the median value was 0.59 for the expected correlation between serum and saliva. Values ranged widely from -0.30 to 0.81, and were larger for more valid indicators of cycle phase (see Figure 4). Larger values were also found for the studies using tandem mass spectrography and two immunoassays (DRG ELISA and Siemens Health radioimmunoassay) than for studies using Salimetrics and IBL ELISAs. However, studies using different assays additionally differed in their cycle phase, scheduling procedure and age range. By way of comparison, the imputation models allowed us to generate estimates of serum P4 from backward counting or LH tests that had a correlation ($r_{Imputation,\,Serum}$) of .83 or .87 with measured serum values and correlations of .86 and .89 with within-subject differences after correcting for range restriction. These estimates exceeded our best indirect estimates (see Eq. 1) of the true correlation of salivary P4 with serum and matched the correlation reported by Salimetrics (r=0.87).

Ratio and probability of being in the fertile window

We also investigated different ways of jointly modelling estradiol and progesterone that have been discussed in the literature (Del Giudice & Gangestad, 2022; Roney, 2019). We found that, across all datasets, the logarithm of the ratio of estradiol over progesterone was much more strongly correlated with progesterone than with estradiol, because progesterone is more variable than estradiol on the log-scale. We then evaluated several models to predict the

estimated probability of being in the fertile window, with steroids as predictors. We compared a simple model with the log-transformed predictors estradiol and estradiol/progesterone ratio to a complex model allowing a nonlinear interaction between log-transformed estradiol and progesterone. In the BioCycle data, the complex model clearly outperformed the simple model for all cycle phase measures (e.g., for LH: loo-Rs=0.83 [0.81;0.85] and 0.69 [0.66;0.71]) and the correlation with the log-ratio (r=0.60 [0.57;0.62]). In the other datasets, these differences were much less marked: neither the ratio, nor the simple model, nor the complex model made a sizable improvement on prediction from log-transformed progesterone alone.

Robustness checks

Without taking a strong stance on the optimal approach, we estimated correlations between steroids and the estimated probability of being in the fertile window (PBFW) both with and without log-transformation and with and without within-subject centering. We used the PBFW as the criterion, as the probability is not itself a hormone that may or may not be log-transformed. On average, log transformation without subtracting the subject mean yielded the strongest correlations, but differences across transformations were small (at most 0.06 on average) and inconsistent across datasets (see Supplementary Figure 3).

We also investigated the predictive power of cycle phase by age to investigate the influence of anovulation rates, which vary by age (Supplementary Note 4), and the predictive power of cycle phase determined from serum LH in the BioCycle data (Supplementary Note 5).

Discussion

Salivary enzyme-linked immunoassays for estradiol and progesterone are widely used in psychoneuroendocrinology and hormonal assessment for confirmation of cycle phase is routinely recommended (Psychoneuroendocrinology Editorial Policy, 2022). Here, we show the most widely used assays exhibit subpar validity for predicting cycle phase using data from more than 1,200 women and 9,500 time points.

One potential reason for low validity is random error, which reduces power but can be compensated with larger sample sizes. The overall pattern we observe is inconsistent with this possibility: we see an upward bias compared to expectations from serum and salivary LC-MS/MS, especially in the early follicular phase when levels should be low. We therefore doubt that the problematic assays can serve as unbiased measures of menstrual cycle phase. Instead, the upward bias and low correlation with cycle phase could be well-explained by cross-reactions or other interferences with the immunoassays (Warade, 2017) when true steroid concentrations are low.

For estradiol, salivary immunoassays should be treated with extreme caution, especially when true levels are low (e.g., during puberty and menopause in women as well as in children and in

men), but may be appropriate when expected levels are high, for instance after ovarian hyperstimulation, conception, or estrogen treatment (Dielen et al., 2017; Sakkas et al., 2021; Sun et al., 2019; Tivis et al., 2005). For progesterone, our low indirect estimates of saliva-serum correlations for Salimetrics immunoassays are consistent in size with the correlations (from -0.02 to 0.22) reported in Sakkas et al. (2021). The Siemens radioimmunoassay and the DRG immunoassay for progesterone showed stronger associations with cycle phase than IBL and Salimetrics assays, but tandem mass spectrometry did best. However, comparisons of assay performance across studies can only be cautiously made and should be validated in future work because the included studies differed on other relevant variables as well, such as age range and assay scheduling.

Tandem mass spectrometry using the most recent generation of spectrometers may reduce the observed invalidity if the main problem is interference. Contamination with blood, short-term pulsatility of steroids in saliva (Bao et al., 2003), or a general higher error-proneness in analytic pipelines in psychological laboratories could also explain why the cycle phase relationships obtained in, for instance, the BioCycle data appear better. However, across laboratories we see a systematic upward bias in the follicular phase in saliva. This bias is difficult to explain without recourse to assay interference.Furthermore, the good performance of tandem mass spectrometry for salivary progesterone is inconsistent with error-prone pipelines and pulsatility. For estradiol, one previous attempt with mass spectrometry failed to detect salivary estradiol in a majority of cases and did not correlate with an IBL immunoassay (r = .06, Stern et al., 2019). Newer generation spectrometers may be sufficiently sensitive to be useful for salivary assays in premenopausal women (Fiers et al., 2017).

Counter to intuition, imputation from backward counting and LH surges may offer closer approximations to true serum steroid changes across an ovulatory cycle than measurements derived from salivary immunoassays, according to our indirect estimates and manufacturer-independent validations (Dielen et al., 2017; Sakkas et al., 2021; Sun et al., 2019; Tivis et al., 2005). Of course, imputation from cycle phase only speaks to average within-subject change whereas highly repeated salivary immunoassays could potentially reliably estimate inter-individual differences. We urge caution before generalising the findings from the BioCycle dataset to others (but see Supplementary Notes 5 and 9). We should also consider the possibility that more stringent screening criteria, as in BioCycle, could boost the validity of salivary immunoassays.

Limitations

We performed no direct comparison of matched samples in serum and saliva using multiple assays. Instead, we relied on a third variable, cycle phase, that was assessed in all studies. Cycle phase can only be determined with error using our methods. Error in the cycle phase phase measure deflates associations. However, measures perform as expected, given known uncertainties, in serum and for several measures of salivary progesterone: more valid cycle phase measures showed larger associations with ovarian hormone concentrations. In addition,

errors should cancel out in our indirect estimate of the saliva-serum correlation, if cycle phase measures have the same error level across studies.

To some extent, the cycle phase measures we deemed comparable across studies may differ depending on the sample, design, and urinary LH assay. Such differences could bias our estimates of variance explained by cycle phase and our indirect estimates of the saliva-serum correlation. Still, low correlations with salivary estradiol immunoassays were also observed in studies where the same cycle phase measure predicted LC-MS/MS progesterone well (Jünger et al., 2018; Stern et al., 2021). High rates of anovulation in some samples could explain low associations with counting-based cycle phase measures, but we would then expect substantial improvements when using the LH surge as a criterion (Lynch et al, 2014), which we did not observe. In addition, serum steroid measures consistently identify more cycles as ovulatory than urinary LH measures and salivary steroid measures (Lynch et al., 2014; Marcinkowska, 2020; Supplementary Note 4). Similarly, correcting for range restriction in cycle phase owing to the scheduling algorithm did not materially improve results. A sanity check supported the validity of our approach to adjusting for scheduling differences (Supplementary Note 9). We urge caution, however, before generalising our results to biologically female persons who are not cisgender, as all of our samples either included only cisgender women or did not inquire about gender apart from biological sex.

Our serum measure of free estradiol was determined via a mass action-based algorithm from total estradiol, not directly measured using, for instance, equilibrium dialysis. Free estradiol as determined by equilibrium dialysis correlates .92 with total estradiol in Dielen et al. 2017. The correlation with the algorithm-based estimate we used and total serum values was .97 in the BioCycle data. Given the strength of these associations, we doubt that there are major differences in cyclic patterns between free and total estradiol — the main difference is in the mean concentration.

Whether hormones should be treated as log-transformed and/or within-subject centred prior to analysis has been debated (Gangestad et al., 2019; Roney, 2019). In some cases, truncating outliers or within-subject centering has also been used to achieve an approximately normal distribution. In our robustness checks, we found that log transformation or centering did not materially improve associations with cycle phase (correlation coefficients differed by up to .06), though log transformation performed best on average.

In contrast to the concerning findings on salivary immunoassays, cycle phase strongly predicted measured serum values, potentially making imputation from cycle phase a low-cost alternative to salivary assays when within-cycle change is of interest. Our results are based on the BioCycle study, which applied rigorous screening criteria to exclude, among others, likely anovulatory women. The sample was older and more ethnically diverse than most other samples. Applied to other samples, the validity of our counting-based imputation models may be lower where anovulation and/or weak ovarian function are more common than in BioCycle. The validities should hence not be taken as given without further replication (see Supplementary

Notes 4 and 5). At least for progesterone, several of the salivary datasets provide encouraging evidence.

Implications

In combination, our results and several manufacturer-independent validation studies (Dielen et al. 2017; Sakkas et al. 2021; Sun et al., 2019; Tivis et al. 2005) call for caution when using salivary hormones as indicators of menstrual cycle phase. If salivary immunoassays of estradiol and progesterone have little validity for estimating cycle position, then questions arise as to whether they should be used to make confident inferences about the day of ovulation. Researchers who are interested in within-subject effects, such as the causal effects of hormonal change around ovulation, might question the reliability of previous results. Especially for estradiol, false negative results are likely to have occurred more frequently than expected. If cross-reactivity is the culprit, or if researchers engage in overfitting to noisy data, the chance for false positive results is also inflated.

Surprisingly, the correlations between day-specific average and day-specific individual serum hormone values were consistently larger than the inferred correlations between individual salivary and serum hormone values. This implies that salivary measures are so noisy that the actual serum hormone values on a given cycle day can be more accurately estimated by average serum values for that day than by measured salivary values. Although we acknowledge that definitive evidence for this conclusion requires collecting matched salivary and serum samples to directly measure their correlation, the inferential evidence presented here is nonetheless consistent with this conclusion. As such, existing studies that have a valid measure of cycle phase (e.g., LH surge day, prospective backward-counting) could be reanalysed in order to test whether their conclusions might differ if imputed serum hormones are substituted for measured salivary immunoassays of estradiol or progesterone, or if measured and imputed hormones are combined in, for instance, an overimputation model (Blackwell et al., 2017). To make this easier for future research, we have made tables with the imputed serum values by cycle phase available on OSF (osf.io/u9xad/). These files can simply be merged on the cycle day column, as explained on OSF.

In serum, cyclical variation was much larger than stable between-subject variation in estradiol and progesterone. Put simply, there are almost no women whose pre-ovulatory levels of estradiol or midluteal levels of progesterone are as low as any other woman's level during menses. The variance proportions seen in serum make sense considering the role of estradiol and progesterone as evolutionarily highly conserved signals which regulate reproduction, an essential function. Given this pattern, an important role for estradiol and progesterone for stable individual differences in e.g. personality or cognitive abilities in women seems less likely (see also Eisenlohr-Moul and Owens, 2016).
A focus on the variance proportions observed in salivary immunoassays (Ellison and Lipson, 1999) may have misdirected past theoretical debates, which operated under the assumption that between-subject variation was larger than cyclical variation (Havlíček et al., 2015). Studies that find substantial associations between inter-individual differences in estradiol and

psychological inter-individual differences (e.g., Marcinkowska et al., 2018) and studies estimating the heritability of inter-individual differences in estradiol (e.g., Grotzinger et al., 2017) based on salivary measures may capture covariation that is only partly related to stable interindividual differences in serum estradiol levels. The larger interindividual differences seen in saliva may, for example, instead result from unknown cross-reactants, into which further research is needed. . Direct comparison of multiple matched samples per person would be needed to substantiate or allay this concern (e.g., Stern et al., 2022).

Because assay details are normally only reported in the method section and in unstandardised form, it will be laborious to identify all studies that employed problematic assays. Standardisation and citation of protocols (Rosner et al., 2013) would help trace assays. If researchers choose to assay estradiol or progesterone using reagents classified as for "research use only", we advise to include a high-quality measure of cycle phase for internal validation, for example urinary LH surges. Authors should also report observed values together with the relevant reference ranges derived using gold standard methods, so that upward bias becomes apparent.

When saliva has been collected, we currently expect tandem mass spectrometry using the most recent generation of spectrometers to be superior to immunoassays of estradiol and progesterone (see also Fiers et al., 2017), but further validation should test our conjecture. To properly guide the choice of specimen and assay before sample collection, we encourage further empirical head-to-head comparisons between imputations, radio, luminescence, and enzyme immunoassays, and mass spectrometry assays, as well as comparisons across serum, blood spots, saliva, hair, sweat, and urine (see Supplementary Note 6) in the same women to clarify which specimens and assays are best suited for which research questions and populations (Stern et al., 2022; Sun et al., 2019).

Conclusions

Salivary immunoassays of estradiol have unacceptably low validity for the estimation of cycle phase. This problem is less marked for salivary immunoassays of progesterone, but for research on menstrual cycle change, imputing progesterone from a valid cycle phase measure appears to give a closer approximation of serum progesterone at a fraction of the cost. Tandem mass spectrometry combined with imputation on unassayed cycle days holds promise as a cost-effective approach for future work, but should be empirically validated. Substantial scientific resources may have been mis-allocated owing to the widespread use of assays with questionable validity, at the expense of sample size and number of measurement occasions. We are left with an underpowered literature and many questions about bias in need of answers.

Table 1.

	BioCycle	Roney 2013	OCMATE	Marcinkowska 2020	Stern 2021	Jünger 2018	Blake 2017	Grebe 2016
Sample								
Women	259	43	384	102	257	157	60	33
Cycles	509	79	907	102	454	398	109	33
Days	3911	2627	2394	2265	1028	628	120	66
Age ± SD	27.3 ± 8.21	18.8 ± 1.15	21.5 ± 3.29	28.8 ± 4.56	23.1 ± 3.28	23.2 ± 3.45	22.7 ± 4.87	20.8 ± 4.90
Partnered	25%[a]	33%	36%	65%	47%	48%	53%	100%
Body fluid	Serum	Saliva	Saliva	Saliva	Saliva	Saliva	Saliva	Saliva
Sampling time	routinely 7:00-8:30	morning ideally after waking	attempted to keep constant per woman	morning	12:00-16:00	11:30-16:00i	12:00-18:00	12:00-18:00 (S 1), Waking Time (S 2)
Cycle phase								
Cycle length	28.8 ± 4.10	27.8 ± 5.12	29.7 ± 6.73	28.2 ± 2.99	30.0 ± 4.75	29.5 ± 6.54	29.2 ± 2.50	28.8 ± 3.71[b]
Indicators	FC+BC+LH	FC+BC	FC+BC	FC+BC+LH	FC+BC+LH	FC+BC+LH	FC+BC+LH	FC
Scheduling	scheduled	each day	random	each day	scheduled	scheduled	scheduled	random
Estradiol								
Assay	E+MAA	ELISA	ELISA	ELISA	ELISA	CLIA	ELISA	ELISA
Women	257	42	360	100	243	157	58	31
Days	3682	1091	1664	1647	914	549	114	58
Geometric mean	1.49	2.83	3.10	5.47	3.63	4.81	6.30	2.27

Mean	2.13	3.10	3.36	7.61	4.00	5.88	7.42	2.45
SD	1.95	1.34	1.55	6.40	1.89	3.93	4.74	0.92
Range	0.21, 18.28	0.67, 9.17	0.48, 24.22	0.40, 46.52	1.01, 19.05	0.30, 31.00	2.10, 28.81	0.53, 5.62
Progesterone								
Assay	ELISA	RIA	ELISA	ELISA	LCMS/MS	LCMS/MS	ELISA	ELISA
Women	257	42	360	99	238	156	58	31
Days	3682	1121	1664	1550	778	537	114	57
Geometric mean	1394	42.95	122.73	70.81	9.58	17.64	117.92	48.42
Mean	3438	53.74	158.54	106.34	27.97	53.72	170.22	69.62
SD	4683	39.21	121.31	88.95	52.67	91.17	155.69	62.44
Range	200, 27700	9.14, 310.00	5.00, 1859.40	2.50, 875.96	0.22, 671.77	0.26, 1480.00	14.13, 748.71	5.00, 293.46

Note. Descriptive summary of the included datasets. All hormone values are in pg/ml. The sample sizes reported under sample are the whole sample before exclusions. Below each hormone, we again list the sample sizes after excluding cycles shorter than 20 or longer than 35 days, as well as days where the hormone value was missing. Note that the sample sizes for each cycle phase measure can be lower still, e.g. LH surges were not observed for all women. The specific sample size for each model can be found in Supplementary Note 8.
[a] Married or cohabiting.
[b] self-reported, not observed.
RIA = radioimmunoassay. ELISA=enzyme-linked immunosorbent assay, CLIA = chemiluminescence immunoassay, RIA = radioimmunoassay. LCMS/MS = tandem mass spectrometry, E+MAA = ELISA + mass-action algorithm. FC = forward-counting. BC = backward-counting. LH = counting relative to urinary luteinising hormone surge.

Acknowledgements

We thank all of the participants who provided their data. Ruben C. Arslan is supported by the German Research Foundation (#464488178). Urszula M. Marcinkowska's study was supported by the Polish National Science Centre (grant number 2014/12/S/NZ8/00722).

References

Bao, A.-M., Liu, R.-Y., van Someren, E.J.W., Hofman, M.A., Cao, Y.-X., Zhou, J.-N., 2003. Diurnal rhythm of free estradiol during the menstrual cycle. Eur. J. Endocrinol. 148, 227–232. https://doi.org/10.1530/eje.0.1480227

Blackwell, M., Honaker, J., King, G., 2017. A Unified Approach to Measurement Error and Missing Data: Overview and Applications. Sociol. Methods Res. 46, 303–341. https://doi.org/10.1177/0049124115585360

Blake, K.R., Bastian, B., O'Dean, S.M., Denson, T.F., 2017. High estradiol and low progesterone are associated with high assertiveness in women. Psychoneuroendocrinology 75, 91–99.

Blake, K.R., Dixson, B.J.W., O'Dean, S.M., Denson, T.F., 2016. Standardized protocols for characterizing women's fertility: A data-driven approach. Horm. Behav. 81, 74–83. https://doi.org/10.1016/j.yhbeh.2016.03.004

Bürkner, P.-C., 2017. brms: An R package for Bayesian multilevel models using Stan. J. Stat. Softw. 80. https://doi.org/10.18637/jss.v080.i01

Celec, P., Ostatníková, D., 2012. Saliva collection devices affect sex steroid concentrations. Clin. Chim. Acta 413, 1625–1628. https://doi.org/10.1016/j.cca.2012.04.035

Cohen, J., Cohen, P., West, S.G., Aiken, L.S., 2003. Applied multiple regression/correlation analysis for the behavioral sciences. Lawrence Erlbaum, Mahwah, NJ.

Dielen, C., Fiers, T., Somers, S., Deschepper, E., Gerris, J., 2017. Correlation between saliva and serum concentrations of estradiol in women undergoing ovarian hyperstimulation with gonadotropins for IVF/ICSI. Facts Views Vis Obgyn 9, 85–91.

Dunn, J.F., Nisula, B.C., Rodbard, D., 1981. Transport of steroid hormones: binding of 21 endogenous steroids to both testosterone-binding globulin and corticosteroid-binding globulin in human plasma. J. Clin. Endocrinol. Metab. 53, 58–68. https://doi.org/10.1210/jcem-53-1-58

Editorial Policy of Psychoneuroendocrinology [WWW Document], n.d. URL https://www.elsevier.com/journals/psychoneuroendocrinology/0306-4530/guide-for-authors (accessed 1.27.22).

Eisenlohr-Moul, T.A., Owens, S.A., 2016. Hormones and Personality, in: Zeigler-Hill, V., Shackelford, T.K. (Eds.), Encyclopedia of Personality and Individual Differences. Springer International Publishing, Cham, pp. 1–23. https://doi.org/10.1007/978-3-319-28099-8_762-1

Ellison, P.T., Lipson, S.F., 1999. Salivary estradiol--a viable alternative? Fertil. Steril. https://doi.org/10.1016/s0015-0282(99)00344-1

Fiers, T., Dielen, C., Somers, S., Kaufman, J.-M., Gerris, J., 2017. Salivary estradiol as a surrogate marker for serum estradiol in assisted reproduction treatment. Clin. Biochem. 50, 145–149. https://doi.org/10.1016/j.clinbiochem.2016.09.016

Fletcher, T.D., 2010. psychometric: Applied Psychometric Theory.

Gangestad, S.W., Dinh, T., Grebe, N.M., Del Giudice, M., Emery Thompson, M., 2019. Psychological cycle shifts redux, once again: response to Stern et al., Roney, Jones et al., and Higham. Evol. Hum. Behav. https://doi.org/10.1016/j.evolhumbehav.2019.08.008

Gangestad, S.W., Haselton, M.G., Welling, L.L.M., Gildersleeve, K., Pillsworth, E.G., Burriss, R.P., Larson, C.M., Puts, D.A., 2016. How valid are assessments of conception probability in ovulatory cycle research? Evaluations, recommendations, and theoretical implications. Evol. Hum. Behav. 37, 85–96. https://doi.org/10.1016/j.evolhumbehav.2015.09.001

Garnett, E., Bruno-Gaston, J., Cao, J., Zarutskie, P., Devaraj, S., 2020. The importance of estradiol measurement in patients undergoing in vitro fertilization. Clin. Chim. Acta 501, 60–65. https://doi.org/10.1016/j.cca.2019.09.021

Granger, D.A., Shirtcliff, E.A., Booth, A., 2004. The "trouble" with salivary testosterone. Psychoneuroendocrinology 29, 1229–40.

Grebe, N.M., Emery Thompson, M., Gangestad, S.W., 2016. Hormonal predictors of women's extra-pair vs. in-pair sexual attraction in natural cycles: Implications for extended sexuality. Horm. Behav. 78, 211–219. https://doi.org/10.1016/j.yhbeh.2015.11.008

Grotzinger, A.D., Mann, F.D., Patterson, M.W., Herzhoff, K., Tackett, J.L., Tucker-Drob, E.M., Harden, K.P., 2017. Twin models of environmental and genetic influences on pubertal development, salivary testosterone, and estradiol in adolescence. Clin. Endocrinol. https://doi.org/10.1111/cen.13522

Handelsman, D.J., 2017. Mass spectrometry, immunoassay and valid steroid measurements in reproductive medicine and science. Hum. Reprod. 32, 1147–1150. https://doi.org/10.1093/humrep/dex078

Harris, C.R., Pashler, H., Mickes, L., 2014. Elastic analysis procedures: An incurable (but preventable) problem in the fertility effect literature. Comment on Gildersleeve, Haselton, and Fales (2014). Psychol. Bull. 140, 1260–1264. https://doi.org/10.1037/a0036478

Havlíček, J., Cobey, K.D., Barrett, L., Klapilová, K., Roberts, S.C., 2015. The spandrels of Santa Barbara? A new perspective on the peri-ovulation paradigm. Behav. Ecol. 26, 1249–1260. https://doi.org/10.1093/beheco/arv064

IBL, 2019. 17beta-Estradiol Saliva ELISA 30121045.

IBL, 2015. Progesterone Luminescence Immunoassay RE62021 / RE62029.

Jones, B.C., Hahn, A.C., DeBruine, L.M., 2019. Ovulation, Sex Hormones, and Women's Mating Psychology. Trends Cogn. Sci. 23, 51–62. https://doi.org/10.1016/j.tics.2018.10.008

Jones, B.C., Hahn, A.C., Fisher, C.I., Wang, H., Kandrik, M., Han, C., Fasolt, V., Morrison, D., Lee, A.J., Holzleitner, I.J., O'Shea, K.J., Roberts, S.C., Little, A.C., DeBruine, L.M., 2018. No Compelling Evidence that Preferences for Facial Masculinity Track Changes in Women's Hormonal Status. Psychol. Sci. 29, 996–1005. https://doi.org/10.1177/0956797618760197

Jünger, J., Kordsmeyer, T.L., Gerlach, T.M., Penke, L., 2018. Fertile women evaluate male bodies as more attractive, regardless of masculinity. Evol. Hum. Behav. 39, 412–423. https://doi.org/10.1016/j.evolhumbehav.2018.03.007

Lynch, K.E., Mumford, S.L., Schliep, K.C., Whitcomb, B.W., Zarek, S.M., Pollack, A.Z., Bertone-Johnson, E.R., Danaher, M., Wactawski-Wende, J., Gaskins, A.J., Schisterman, E.F., 2014. Assessment of anovulation in eumenorrheic women: comparison of ovulation detection algorithms. Fertil. Steril. 102, 511–518.e2. https://doi.org/10.1016/j.fertnstert.2014.04.035

Magyar, D.M., Boyers, S.P., Marshall, J.R., Abraham, G.E., 1979. Regular menstrual cycles and premenstrual molimina as indicators of ovulation. Obstet. Gynecol. 53, 411–414.

Marcinkowska, U.M., 2020. Importance of Daily Sex Hormone Measurements Within the Menstrual Cycle for Fertility Estimates in Cyclical Shifts Studies. Evol. Psychol. 18, 1474704919897913. https://doi.org/10.1177/1474704919897913

Marcinkowska, U.M., Kaminski, G., Little, A.C., Jasienska, G., 2018. Average ovarian hormone levels, rather than daily values and their fluctuations, are related to facial preferences among women. Horm. Behav. 102, 114–119. https://doi.org/10.1016/j.yhbeh.2018.05.013

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Roney, J.R., 2019. On the use of log transformations when testing hormonal predictors of cycle phase shifts: Commentary on. Evol. Hum. Behav.

Roney, J.R., Simmons, Z.L., 2013. Hormonal predictors of sexual motivation in natural menstrual cycles. Horm. Behav. 63, 636–645. https://doi.org/10.1016/j.yhbeh.2013.02.013

Rosner, W., Hankinson, S.E., Sluss, P.M., Vesper, H.W., Wierman, M.E., 2013. Challenges to the measurement of estradiol: an endocrine society position statement. J. Clin. Endocrinol. Metab. 98, 1376–1387. https://doi.org/10.1210/jc.2012-3780

Sakkas, D., Howles, C.M., Atkinson, L., Borini, A., Bosch, E.A., Bryce, C., Cattoli, M.,

Copperman, A.B., de Bantel, A.F., French, B., Gerris, J., Granger, S.W., Grzegorczyk-Martin, V., Lee, J.A., Levy, M.J., Matin, M.J., Somers, S., Widra, E.A., Alper, M.M., 2021. A multi-centre international study of salivary hormone oestradiol and progesterone measurements in ART monitoring. Reprod. Biomed. Online 42, 421–428. https://doi.org/10.1016/j.rbmo.2020.10.012

Salimetrics, 2020. Salivary progesterone enzyme immunoassay kit.

Salimetrics, 2019. High sensitivity salivary 17β-estradiol enzyme immunoassay kit.

Schmalenberger, K.M., Tauseef, H.A., Barone, J.C., Owens, S.A., Lieberman, L., Jarczok, M.N., Girdler, S.S., Kiesner, J., Ditzen, B., Eisenlohr-Moul, T.A., 2021. How to study the menstrual cycle: Practical tools and recommendations. Psychoneuroendocrinology 123, 104895. https://doi.org/10.1016/j.psyneuen.2020.104895

Schultheiss, O.C., Dlugash, G., Mehta, P.H., 2018. Hormone measurement in social neuroendocrinology: A comparison of immunoassay and mass spectrometry methods, in: Schultheiss, O.C., Mehta, P.H. (Eds.), Routledge International Handbook of Social Neuroendocrinology. Routledge, Abingdon, UK.

Shirtcliff, E.A., Granger, D.A., Schwartz, E.B., Curran, M.J., Booth, A., Overman, W.H., 2000. Assessing estradiol in biobehavioral studies using saliva and blood spots: simple radioimmunoassay protocols, reliability, and comparative validity. Horm. Behav. 38, 137–147. https://doi.org/10.1006/hbeh.2000.1614

Stan Development Team, 2022. Stan Modeling Language Users Guide and Reference Manual.

Stern, J., Arslan, R.C., Gerlach, T.M., Penke, L., 2019. No robust evidence for cycle shifts in preferences for men's bodies in a multiverse analysis: A response to Gangestad et al. Evolution and Human Behavior 40, 517–525. https://doi.org/10.1016/j.evolhumbehav.2019.08.005

Stern, J., Arslan, R.C., Penke, L., 2022. Stability and validity of steroid hormones in hair and saliva across two ovulatory cycles. Comprehensive Psychoneuroendocrinology 9, 100114. https://doi.org/10.1016/j.cpnec.2022.100114

Stern, J., Kordsmeyer, T.L., Penke, L., 2021. A longitudinal evaluation of ovulatory cycle shifts in women's mate attraction and preferences. Horm. Behav. 128, 104916. https://doi.org/10.1016/j.yhbeh.2020.104916

Sun, B.Z., Kangarloo, T., Adams, J.M., Sluss, P.M., Welt, C.K., Chandler, D.W., Zava, D.T., McGrath, J.A., Umbach, D.M., Hall, J.E., Shaw, N.D., 2019. Healthy Post-Menarchal Adolescent Girls Demonstrate Multi-Level Reproductive Axis Immaturity. J. Clin. Endocrinol. Metab. 104, 613–623. https://doi.org/10.1210/jc.2018-00595

Tivis, L.J., Richardson, M.D., Peddi, E., Arjmandi, B., 2005. Saliva versus serum estradiol: implications for research studies using postmenopausal women. Prog. Neuropsychopharmacol. Biol. Psychiatry 29, 727–732. https://doi.org/10.1016/j.pnpbp.2005.04.029

Vehtari, A., Gelman, A., Gabry, J., Yao, Y., 2018. loo: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models.

Vermeulen, A., Verdonck, L., Kaufman, J.M., 1999. A critical evaluation of simple methods for the estimation of free testosterone in serum. J. Clin. Endocrinol. Metab. 84, 3666–3672. https://doi.org/10.1210/jcem.84.10.6079

Vesper, H.W., Botelho, J.C., Vidal, M.L., Rahmani, Y., Thienpont, L.M., Caudill, S.P., 2014a. High variability in serum estradiol measurements in men and women. Steroids 82, 7–13. https://doi.org/10.1016/j.steroids.2013.12.005

Vesper, H.W., Botelho, J.C., Wang, Y., 2014b. Challenges and improvements in testosterone and estradiol testing. Asian J. Androl. 16, 178–184. https://doi.org/10.4103/1008-682X.122338

Wactawski-Wende, J., Schisterman, E.F., Hovey, K.M., Howards, P.P., Browne, R.W., Hediger, M., Liu, A., Trevisan, M., BioCycle Study Group, 2009. BioCycle study: design of the

longitudinal study of the oxidative stress and hormone variation during the menstrual cycle. Paediatr. Perinat. Epidemiol. 23, 171–184. https://doi.org/10.1111/j.1365-3016.2008.00985.x

Warade, J., 2017. Retrospective Approach to Evaluate Interferences in Immunoassay. EJIFCC 28, 224–232.

Welker, K.M., Lassetter, B., Brandes, C.M., Prasad, S., Koop, D.R., Mehta, P.H., 2016. A comparison of salivary testosterone measurement using immunoassays and tandem mass spectrometry. Psychoneuroendocrinology 71, 180–188. https://doi.org/10.1016/j.psyneuen.2016.05.022

Wood, P., 2009. Salivary steroid assays - research or routine? Ann. Clin. Biochem. 46, 183–196. https://doi.org/10.1258/acb.2008.008208

Wood, S.N., 2003. Thin plate regression splines. J. R. Stat. Soc. Series B Stat. Methodol. 65, 95–114. https://doi.org/10.1111/1467-9868.00374

Wright, S., 1934. The Method of Path Coefficients. Annals of Mathematical Statistics 5, 161–215. https://doi.org/10.1214/aoms/1177732676