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

Ruben C. Arslan[1,2], Khandis Blake[3], Laura J. Botzet[4], Paul-Christian Bürkner[5], Lisa DeBruine[6], Tom Fiers[7], Nick 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. Though manufacturers advertise their assays as suitable, they have not been rigorously validated. We collated data from eight studies across >1,200 women and >9,500 time points. Seven studies collected saliva; one study collected serum. All assayed estradiol and progesterone and had an independent measure of cycle phase (e.g., day in cycle relative to the luteinising hormone surge or a menstrual onset). In serum, all independent cycle phase measures strongly predicted steroids. By contrast, salivary immunoassays of estradiol were only weakly predictable from cycle phase and showed an upward bias compared to expectations from serum. For salivary immunoassays of progesterone, predictability from cycle phase was more mixed, but two widely used assays performed poorly. Imputing average serum steroid levels from cycle phase may yield more valid values than several widely used 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
Online supplement: https://rubenarslan.github.io/invalidity_on_steroids/

Introduction

Salivary immunoassays for estradiol and progesterone are widely used in psychoneuroendocrinology because they are comparatively cheap and easy to collect non-invasively. In research on the effects of the menstrual cycle phase, they are commonly used as indicators of cycle phase. In recent years, "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).
Salivary immunoassays come with known issues (Granger et al., 2004; Schultheiss et al., 2018; Welker et al., 2016; Wood, 2009). Low values 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, because only non-protein-bound forms can diffuse into saliva. Hence, even contamination with small amounts of blood can substantially alter values measured in saliva, 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. In spite of these difficulties, Salimetrics, a widely used (see Supplementary Note 1) provider of salivary immunoassay kits and services, reports correlations of *r*s=0.80/0.87 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, notwithstanding small-scale studies reported by manufacturers, with Ns ranging from 1 to 202 (IBL, 2019, 2015; Salimetrics, 2020, 2019). Independent validations often show smaller saliva-serum correlations; estimates are heterogeneous, and poorer at lower concentrations (Dielen et al., 2017; Shirtcliff et al., 2000; Sun et al., 2019; Tivis et al., 2005). 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. We are not aware of any study that directly validates multiple salivary estradiol and progesterone immunoassays for use as indicators 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 collected measures of estradiol and progesterone plus at least one independent measure of cycle phase. Here, we use the term *cycle phase* loosely, or probabilistically, to indicate a day in the cycle relative to menstrual onset or urinary luteinizing hormone (LH) surge (see Table 1). We did not explicitly assign cycle days into phases, because measurement error and individual differences preclude certain assignments and a continuous approach better captures the signal in the data. 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) and one quantified salivary progesterone using a radioimmunoassay (see Table 1). 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.

Methods

Datasets were obtained from public online repositories or first authors of the relevant publications. All studies collected data only on adult women of reproductive age who were naturally cycling and not using hormonal contraception. 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. 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 (Wactawski-Wende et al. 2009; Jones et al. 2018; Grebe et al. 2016; Jünger et al. 2018; Stern et al. 2021; Marcinkowska 2020; Roney and Simmons 2013; Blake et al. 2017).

Steroid assays

Steroids were measured either in serum or saliva using radioimmunoassays, ELISAs, or tandem mass spectrometry. 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, 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

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 random distribution (Jones et al. 2018, Grebe et al. 2016). 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 except Grebe et al. (2016), who did not follow up until the next menstrual onset. Because the luteal phase is less variable in length 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 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).
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 as outlined in Gangestad et al. (2016) and Stern et al. (2021) for each day (see also OSF merge files). 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 aborted. 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.[1] Bayesian multilevel regressions were used to estimate the hormone's association with cycle phase separately for each hormone and dataset. To this

[1] 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.

end, log hormone values were modelled as Gaussian outcomes. Reported limits of detection (LODs) were used to model left-censoring. The LODs are shown in all subsequent graphs as solid lines. Limits of quantitation are shown as dashed lines. All limits are exactly reported in the online supplement. Varying intercepts for the woman and, if multiple cycles were covered, each cycle were added to absorb variance related to inter-individual and inter-cycle differences. For each available cycle phase measure, a thin-plate spline (Wood, 2003), a flexible smoothing function, was estimated across cycle day to continuously capture variation explained by cycle phase without segmenting the cycle *a-priori* into, for instance, follicular and luteal phase.

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 thin-plate 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 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 fertility monitors. To more directly capture whether salivary measures performed similarly to serum measures, we 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. Average predicted values for an average woman were generated for one cycle. These average predictions were merged on the other datasets by cycle day.
We then computed Pearson correlations between the measured log hormones and the imputed log hormones for all datasets. In addition, we subtracted the subject mean from the measured log hormones to account for the fact that imputations cannot recover interindividual differences and correlated the result with imputed log hormones. 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. To account for this, we applied a correction for range

restriction in cycle phase based on comparing the observed standard deviation in the imputed hormone in each dataset with 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 and attenuate correlations. After correcting for range restriction, the correlation of imputed hormones with within-subject differences should be more comparable across datasets. We can then divide the estimated saliva-imputed correlation by the serum-imputed correlation to arrive at a rough expectation of the (unmeasured) serum-saliva correlation via path tracing rules (see Supplementary Note 3 for required assumptions and further explanation; Wright, 1934).

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

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.

Results

We found large differences between assays in serum and tandem mass spectrometry in saliva on the one hand and immunoassays in saliva on the other hand.

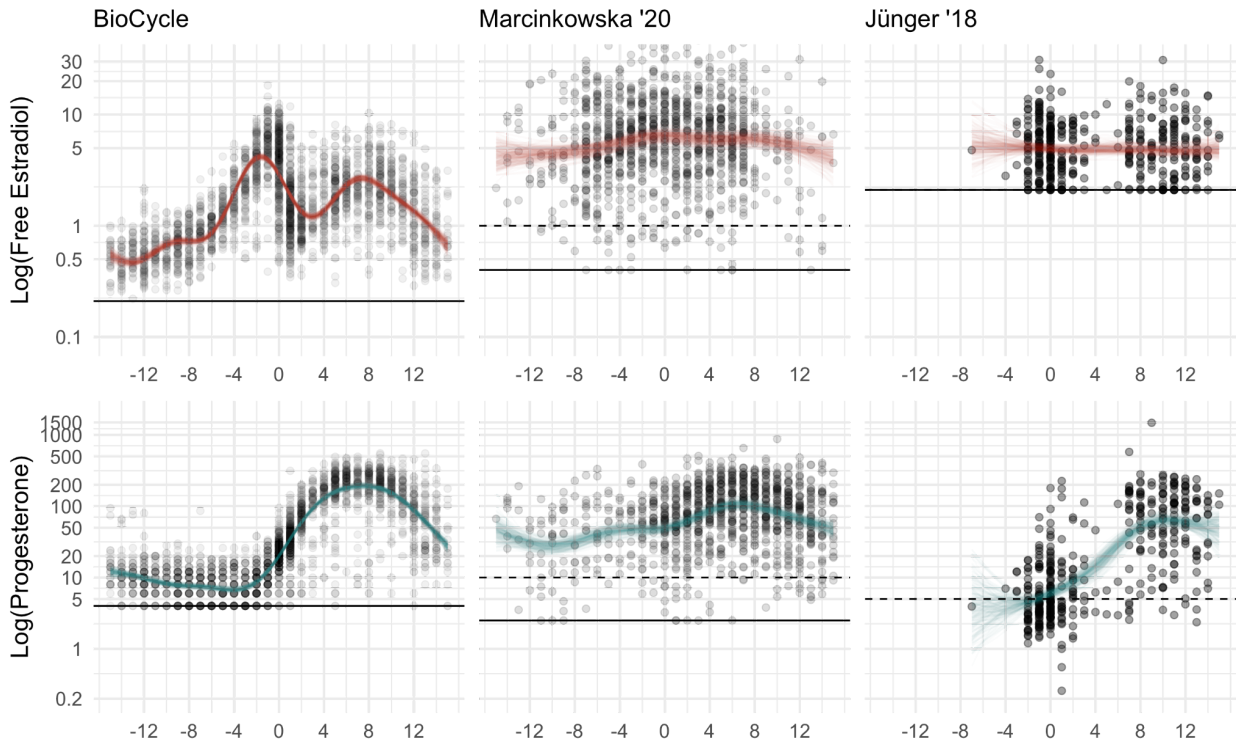


Figure 1. Associations with cycle day relative to the luteinising hormone surge in three selected datasets. Dots show raw data. Coloured lines show two hundred random samples of the thin-plate spline fit using a Bayesian multilevel regression. Solid horizontal lines show the limit of detection; dashed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

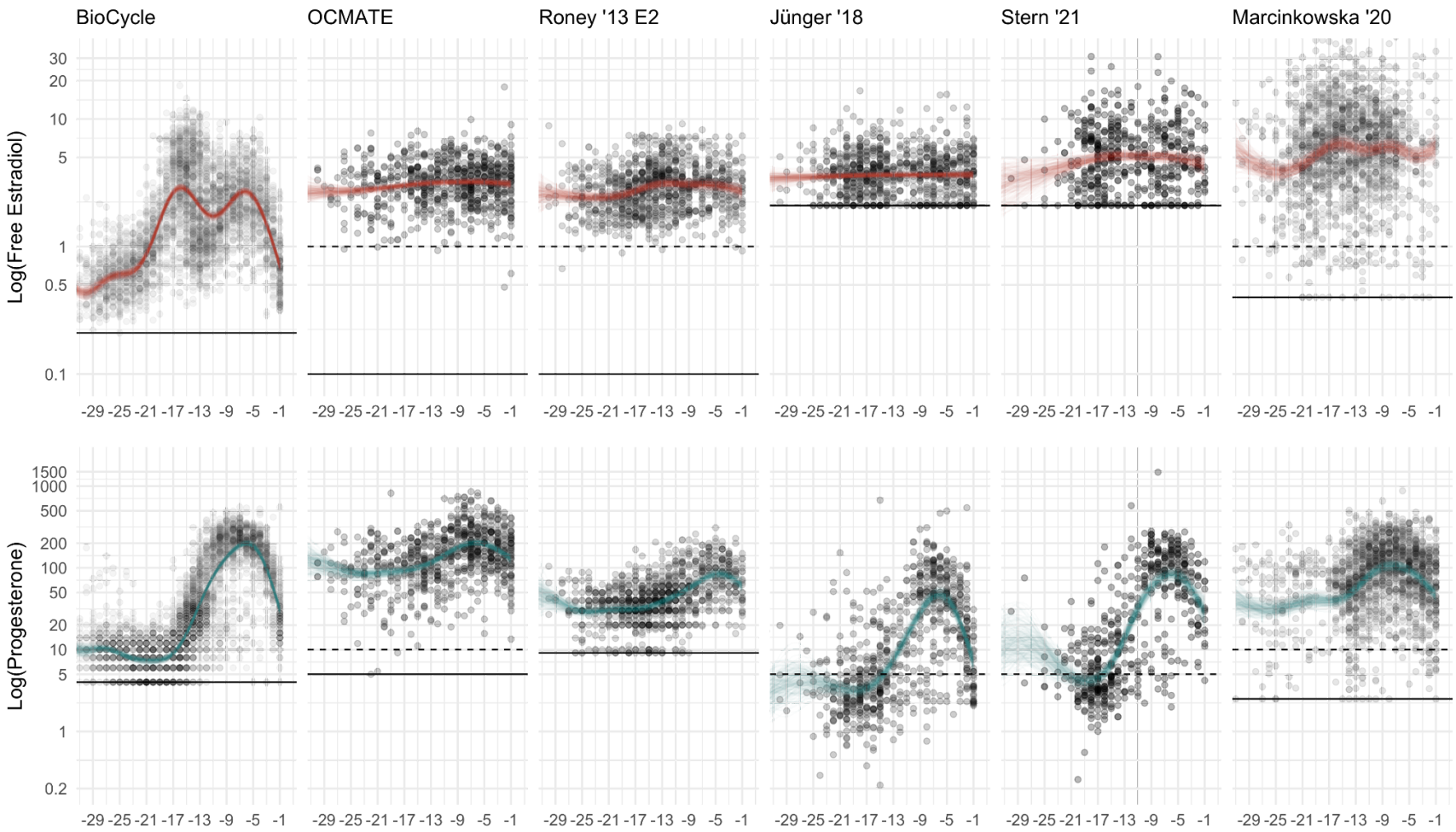


Figure 2. Associations with cycle day relative to the observed next menstrual onset in the six largest datasets. Dots show raw data. Coloured lines show two hundred random samples of the thin-plate spline fit using a Bayesian multilevel regression. Solid horizontal lines show the limit of detection;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

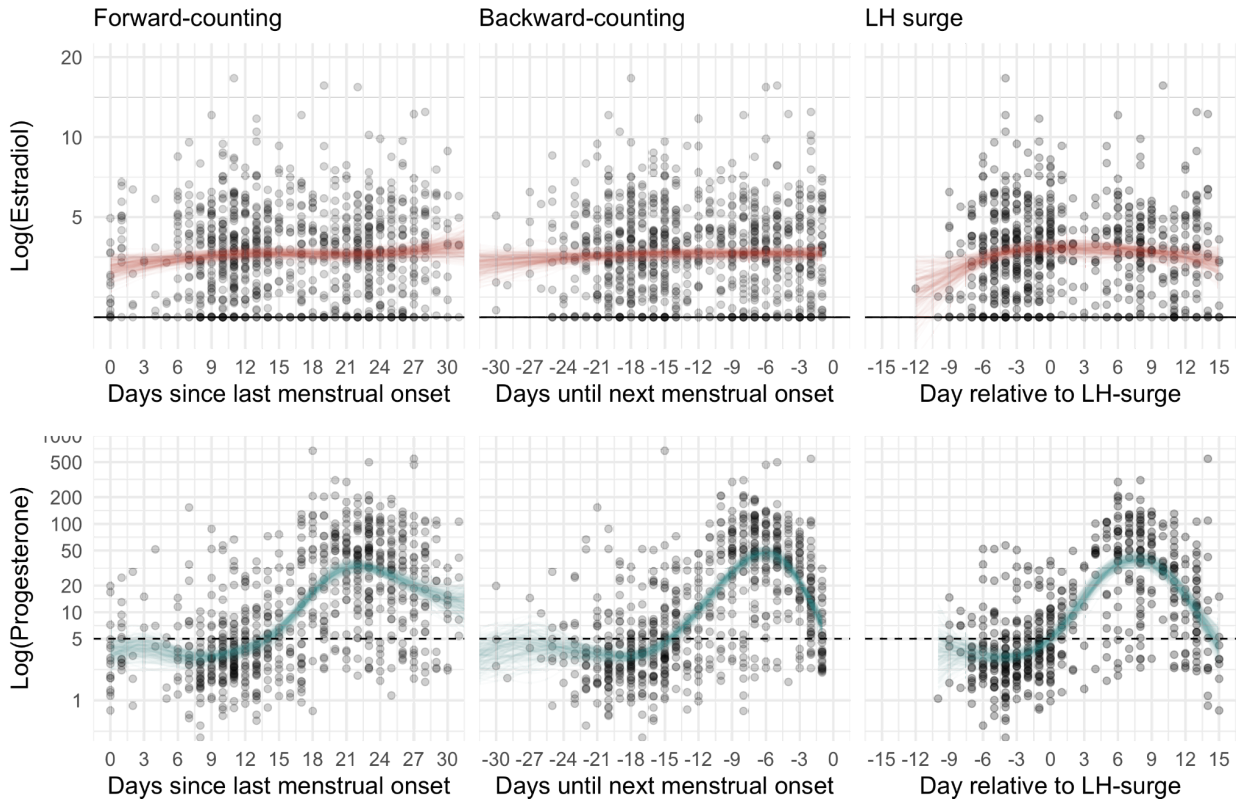


Figure 3. Associations with cycle day relative to the three different measures of cycle phase in Stern et al. (2021). Dots show raw data. Coloured lines show two hundred random samples of the thin-plate spline fit using a Bayesian multilevel regression. Solid horizontal lines show the limit of detection;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

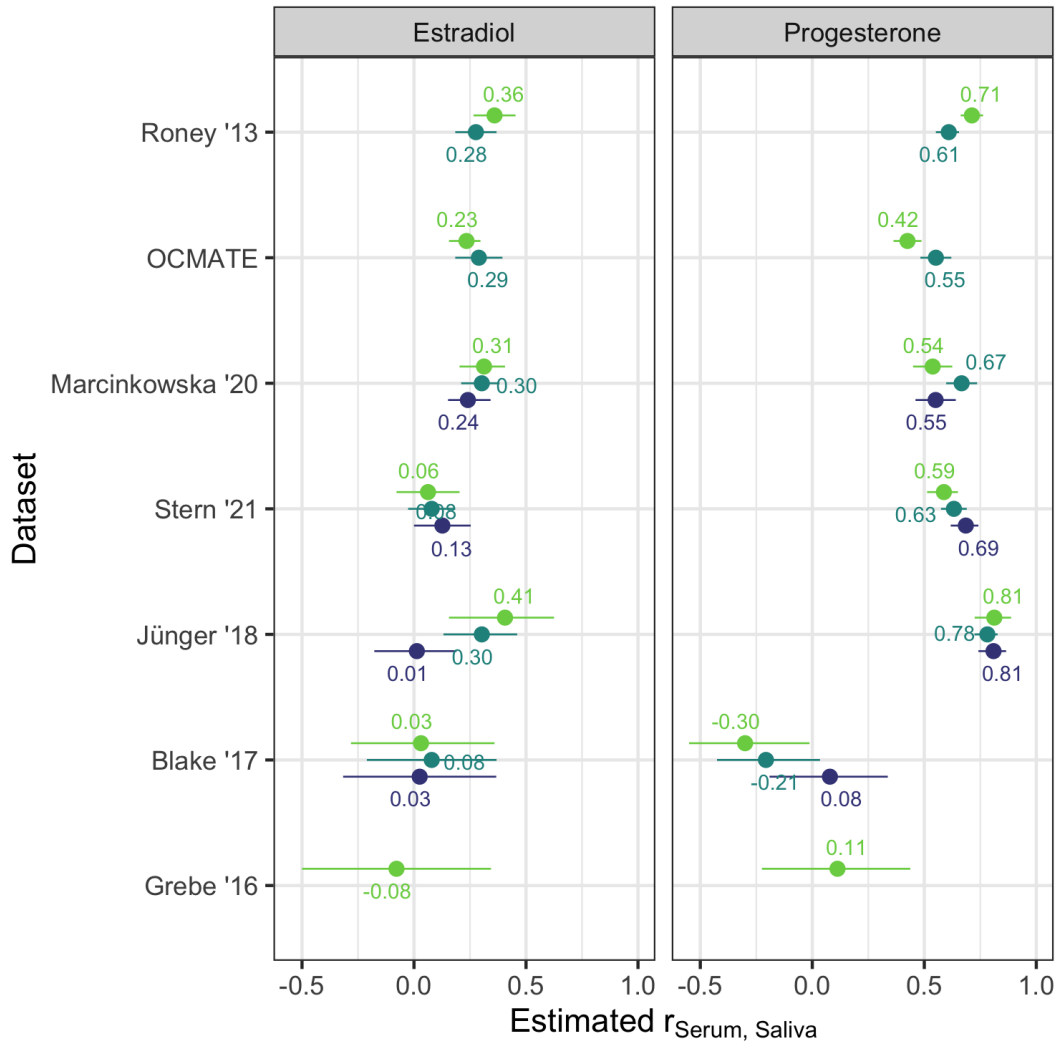


Figure 4. Correlations between serum and saliva, estimated as described in Eq. 1 and Supplementary Note 3. Colours reflect cycle phase measures. Green = forward-counted, blue = backward-counted, violet = relative to LH surge.

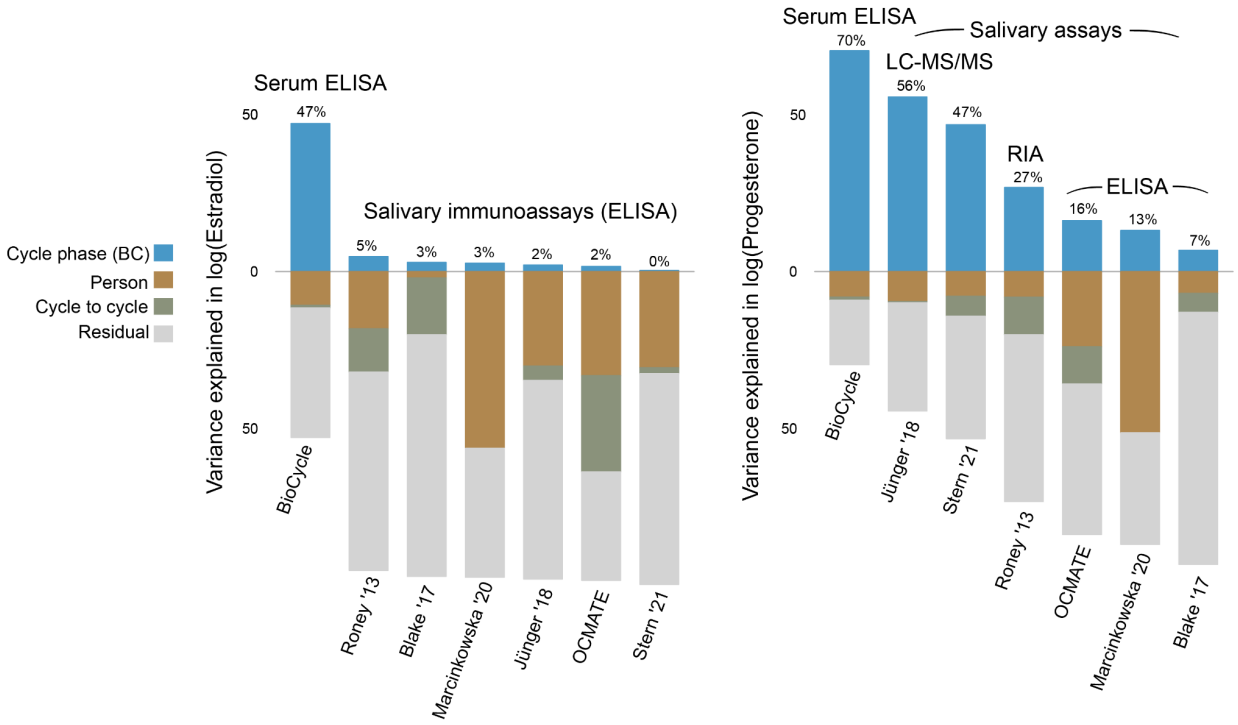


Figure 5. Variance explained in log estradiol and log progesterone, by dataset. Variance explained by backward-counted cycle day above the zero line, variance explained by inter-individual and inter-cycle differences, as well as residual variance below the line. LC-MS/MS = liquid chromatography tandem mass spectrometry; RIA = radioimmunoassay; 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 thin-plate 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 Figure 2, 3, 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 and the first two days before the next menstrual onset, 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 2, 3, and Supplementary Figures 1 and 2). The IBL immunoassays have a limit of detection at 2.1pg/ml. Consequently, censored observations were common (2-12%) and obscured the peri-menstrual variation of free estradiol seen in the serum data. The Salimetrics immunoassays have a reported limit of detection at 0.1pg/ml, but we observed very few values below 1pg/ml, even in the days surrounding menstruation (see Figure 2, 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 saliva-imputed correlation by the serum-imputed correlation as per Eq. 1, the median value was 0.23 for the expected $r_{Serum,\ Saliva}$. Values ranged from -0.08 to 0.41. 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, $r_{FC,\ Saliva}$ was not higher than $r_{LH,\ Saliva}$, but because $r_{FC,\ Serum}$ was low, the estimated $r_{Serum,\ Saliva}$ 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 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 estimates 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 2 and 3) of the thin-plate 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 Figure 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 2 and 3 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 2 and 3), 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), the variance in cycle phase was severely restricted. In the larger datasets, the expected pattern was visible in the conditional effect plots (Figures 2 and 3) 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 2 and 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 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 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. 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. 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 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 estimates 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. 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).

A complete table of all results can be found in our online supplement.

Discussion

Despite known difficulties (Granger et al., 2004; Schultheiss et al., 2018; Sun et al., 2019; Tivis et al., 2005; Welker et al., 2016; Wood, 2009), salivary immunoassays for estradiol and progesterone are widely used in psychoneuroendocrinology and hormonal assessment for confirmation of cycle phase is routinely recommended. Here we examined the validity of salivary versus serum immunoassays for estradiol and progesterone using data from more than 1,200 women and 9,500 time points. Consistent with expectations, salivary immunoassays showed smaller associations with independent measures of cycle phase than serum immunoassays. The extent of the difference, however, was much larger than expected, especially for estradiol. The weak association between estradiol and cycle phase raises doubts about the validity of commonly used salivary immunoassays, especially those offered by Salimetrics (2019) and IBL (2019), the two most commonly employed assays (Supplementary Note 1). Progesterone immunoassays performed comparatively better. Compared to serum immunoassays and salivary tandem mass spectrography, salivary immunoassays also showed a less clear separation between the follicular and luteal phase, larger inter-individual differences, and a lack of values in the low range expected for the follicular phase.

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: Even with a large sample size, we still 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 studied salivary immunoassays of estradiol can be interpreted as unbiased measures of menstrual cycle phase or proximity to ovulation. Instead, the observed patterns are consistent with cross-reactions or other interferences with the immunoassays (Warade, 2017) when true steroid concentrations are low. However, Salimetrics (2020, 2019) and IBL (2019, 2015) note no strong cross-reactivity except for the rare anti-cancer drug fulvestrant.

For estradiol, salivary enzyme immunoassays should be treated with extreme caution pending further validation, especially in populations where levels are expected to be low (e.g., children, adolescent women, men), but may be appropriate when levels are expected to be 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), with further validation needed for intermediate levels, for instance, to identify fecund cycles (Lipson and Ellison, 1996). For progesterone, our low estimates of saliva-serum correlations for Salimetrics immunoassays are consistent in size with the correlations reported in Sakkas et al. (2021). The Siemens radioimmunoassay and the DRG immunoassay performed better, but tandem mass spectrometry performed best.

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 (Wactawski-Wende et al., 2009) and the internal studies of saliva-serum associations reported by Salimetrics (2019, 2020) appear better. However, the systematic upward bias in the follicular phase across laboratories is difficult to explain without recourse to some form of assay interference. The good performance of tandem mass spectrometry for salivary progesterone (Jünger et al., 2018, Stern et al., 2021) is inconsistent with heavily error-prone pipelines. 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 ($r = 0.06$, (Stern et al., 2019). Newer generation spectrometers may be able to remedy this problem (Fiers et al., 2017). Given the higher cost of LC-MS/MS compared to immunoassays, there is a tradeoff between sample size and validity. We considered conducting a cost-benefit analysis, but given the very low estimated validity of all considered salivary estradiol immunoassays the conclusion is foregone for estradiol. If estradiol is measured with LC-MS/MS, the additional cost for a second LC-MS/MS analyte, such as progesterone, is not higher than the cost for an additional immunoassay.

Counter to intuition and frequently repeated advice, imputation from backward counting and LH surges may offer more valid approximations to true serum steroid changes across an ovulatory cycle than measurements derived from salivary immunoassays. Of course, imputation from

cycle phase only speaks to average within-subject change, and we urge caution before generalising the findings from the BioCycle dataset to others.

Limitations

We performed no direct comparison of yoked samples in serum and saliva using multiple assays. Instead, we relied on a third variable, cycle phase, that was commonly assessed across studies. Given our indirect approach, we should ask whether instead the cycle phase measures are flawed. Indeed, none of our cycle phase measures can determine the day of ovulation without error. Still, they perform as expected in serum and for several measures of salivary progesterone: more valid cycle phase measures show larger associations with ovarian hormone concentrations. Hence, it seems unlikely that the low correlations for salivary estradiol immunoassays result only from flawed measures of cycle phase.

To some extent, the cycle phase measures we deemed comparable across studies may differ depending on the sample, design, and urinary LH assay. Still, low correlations with salivary estradiol immunoassays are also observed in studies where the same cycle phase measure predicted 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 based 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.

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 in the literature (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. The debate is complicated by the fact that the observed distributions and intra-class correlations differ between assays, that accuracy decreases with concentration, and that the subject mean itself is estimated with varying accuracy. In our robustness checks, we found that this decision did not materially affect our conclusions (correlation coefficients differed by up to .06), though log transformation performed best on average.

In contrast to the concerning findings on salivary immunoassays, imputation from cycle phase performed well. The true validity may be even higher if, for instance, quantitation inaccuracy in the serum assay, such as censoring, attenuated the variance explained. Our results are based on the BioCycle study, which applied rigorous screening criteria to exclude, among others, likely anovulatory women and women in poor health, for instance women with ovulatory disorders or irregular cycles. 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. We should also consider the possibility that more stringent screening criteria, as in BioCycle, could boost the validity of salivary immunoassays.

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, we cannot use them to make confident inferences about the day of ovulation, and the effects of being fertile. Researchers who are interested in within-subject effects, such as the causal effects of hormonal change around ovulation, should question 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, our favourable results for imputation imply that existing studies of hormonal effects across the menstrual cycle that have a valid measure of cycle phase (e.g., LH surge day, prospective backward-counting) could be fruitfully reanalyzed. They might yield more trustworthy results if imputed 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, 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.

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 of protocols (Rosner et al., 2013) and citation 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.

Based on the results from serum, it is implausible that estradiol and progesterone have strong links to stable individual differences in e.g. personality or cognitive abilities in women, because cyclical variation is much larger than stable between-subject variation (see also Eisenlohr-Moul and Owens, 2016). For traits that are stable and do not clearly vary cyclically, only small effects

are plausible. 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., Kurath and Mata, 2018; Marcinkowska et al., 2018) and studies estimating the heritability of inter-individual differences (e.g., Grotzinger et al., 2017) based on salivary measures may capture covariation that is only partly related to variation in the steroid of interest in serum. Direct comparison of multiple yoked samples per person would be needed to substantiate this concern (see e.g. Stern et al., 2022).

When saliva has been collected, we expect tandem mass spectrometry using the most recent generation of spectrometers to be superior to immunoassays of estradiol and progesterone. To properly guide choices about methods before sample collection, we encourage further empirical head-to-head comparisons between imputations, radio 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 (Stern et al., 2022; Sun et al., 2019).

Conclusions

Salivary immunoassays of estradiol have unacceptably low validity for the estimation of cycle phase. Some salivary immunoassays of progesterone perform better, but imputing progesterone from a valid cycle phase measure performs just as well at a fraction of the cost. Too few existing approaches have been independently validated against a gold standard, but tandem mass spectrometry combined with imputation on unassayed cycle days holds promise as a worthy approach for future work. 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
In rel.	25%[1]	33%	36%	65%	47%	48%	53%	100%
Body fluid	Serum	Saliva	Saliva	Saliva	Saliva	Saliva	Saliva	Saliva
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[a]
Indicators	FC+BC+LH	FC+BC	FC+BC	FC+BC+LH	FC+BC+LH	FC+BC+LH	FC+BC+LH	FC
Scheduling	distributed	each day	random	each day	distributed	distributed	distributed	random
Estradiol								
Assay	E+MAA	ELISA	ELISA	ELISA	ELISA	ELISA	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	3.71	5.00	6.30	2.27
Mean	2.13	3.10	3.36	7.61	4.04	5.94	7.42	2.45
SD	1.95	1.34	1.55	6.40	1.84	3.86	4.74	0.92
Range	0.21, 18.28	0.67, 9.17	0.48, 24.22	0.40, 46.52	2.10, 19.05	2.10, 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. RIA = radioimmunoassay. ELISA=enzyme-linked immunosorbent assay, 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. [a] self-reported, not observed. 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 sizes for each model can be found in our online supplement.[1] Married or cohabiting.

Author contributions

RCA planned the study, collated and analysed the data, and wrote the first manuscript draft. LJB provided code review. PCB provided assistance with Bayesian modelling. TF provided the spreadsheet for the mass action algorithm and expertise on assay technology. All other authors provided data and assisted in data preparation and interpretation. All authors revised the manuscript critically.

Acknowledgements

We thank all of the participants who provided their data. Ruben Arslan is supported by the German Research Foundation (#464488178).

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

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

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
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., 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
Kurath, J., Mata, R., 2018. Individual differences in risk taking and endogeneous levels of testosterone, estradiol, and cortisol: A systematic literature search and three independent meta-analyses. Neurosci. Biobehav. Rev. https://doi.org/10.1016/j.neubiorev.2018.05.003
Lipson, S.F., Ellison, P.T., 1996. Comparison of salivary steroid profiles in naturally occurring conception and non-conception cycles. Human Reproduction 11, 2090–2096. https://doi.org/10.1093/oxfordjournals.humrep.a019055
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

Supplementary

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

Ruben C. Arslan[1,2], Khandis Blake[3], Laura J. Botzet[4], Paul-Christian Bürkner[5], Lisa DeBruine[6], Tom Fiers[7], Nick 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]

Additional online supplement: https://rubenarslan.github.io/invalidity_on_steroids/

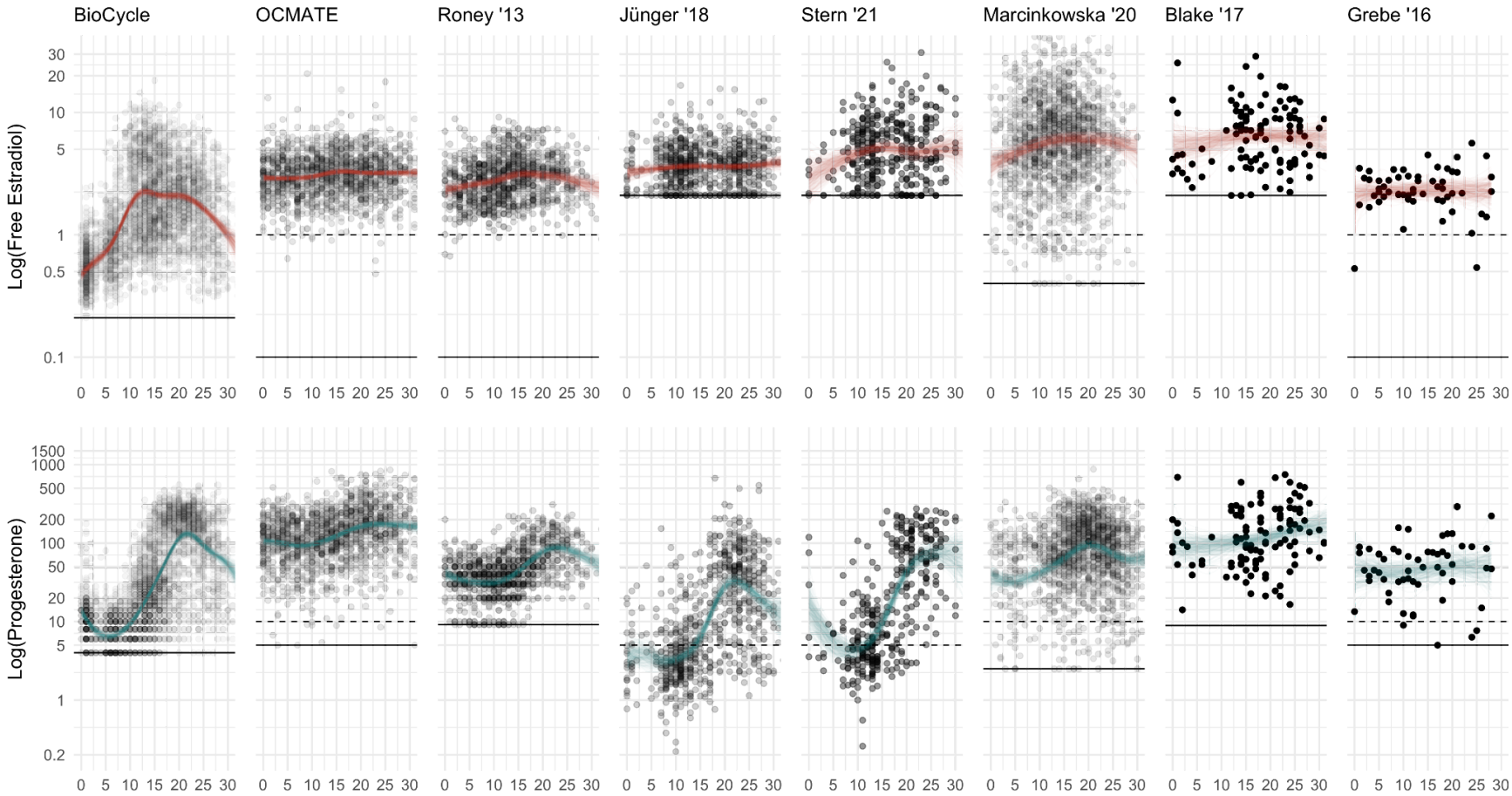
Supplementary Note 1: Literature search on assays

We performed the following search on SCOPUS.
Journal: Psychoneuroendocrinology or Hormones and Behavior
In abstract, title or keywords: Menstrual Cycle, Estradiol
Limited to articles published 2017 or later.

RCA then examined the full text of each article to code the following information:
- Is the article an empirical primary study related to menstrual cycle effects?
- Did researchers collect saliva, serum, urine, or something else?
- How was estradiol assayed?

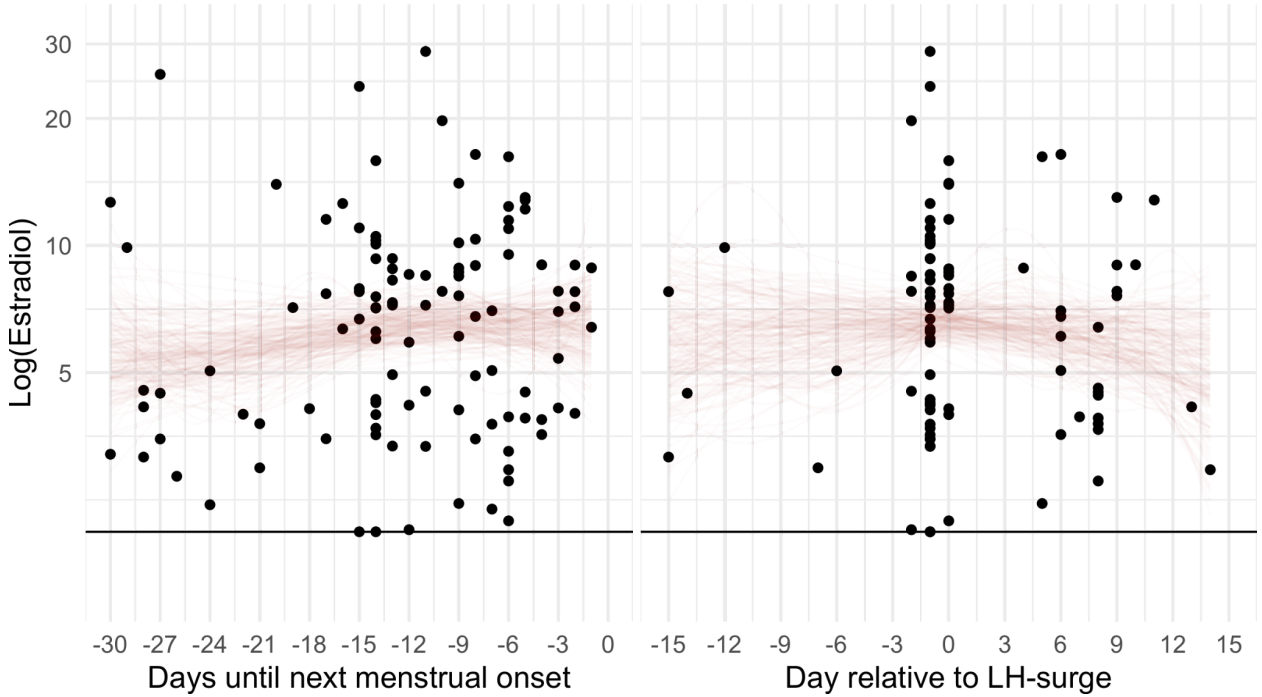
The search terms yielded N=58 papers. Of those, n=42 papers were empirical primary studies with at least one research question related to menstrual cycle changes within-women. Of those, n=1 used an app-based symptom tracking approach, n=1 collected urine, n=2 collected serum, and n=37 collected saliva and assayed estradiol. The estradiol immunoassays studied here, i.e. Salimetrics, IBL, and DRG immunoassays accounted for 78% of the assays performed on salivary estradiol (*n*s=12, 11, 6). In most cases, the progesterone assay was done using kits from the same company. Of studies using salivary estradiol assays, n=22 were published in Psychoneuroendocrinology, n=15 in Hormones and Behavior.

Supplementary Figure 1



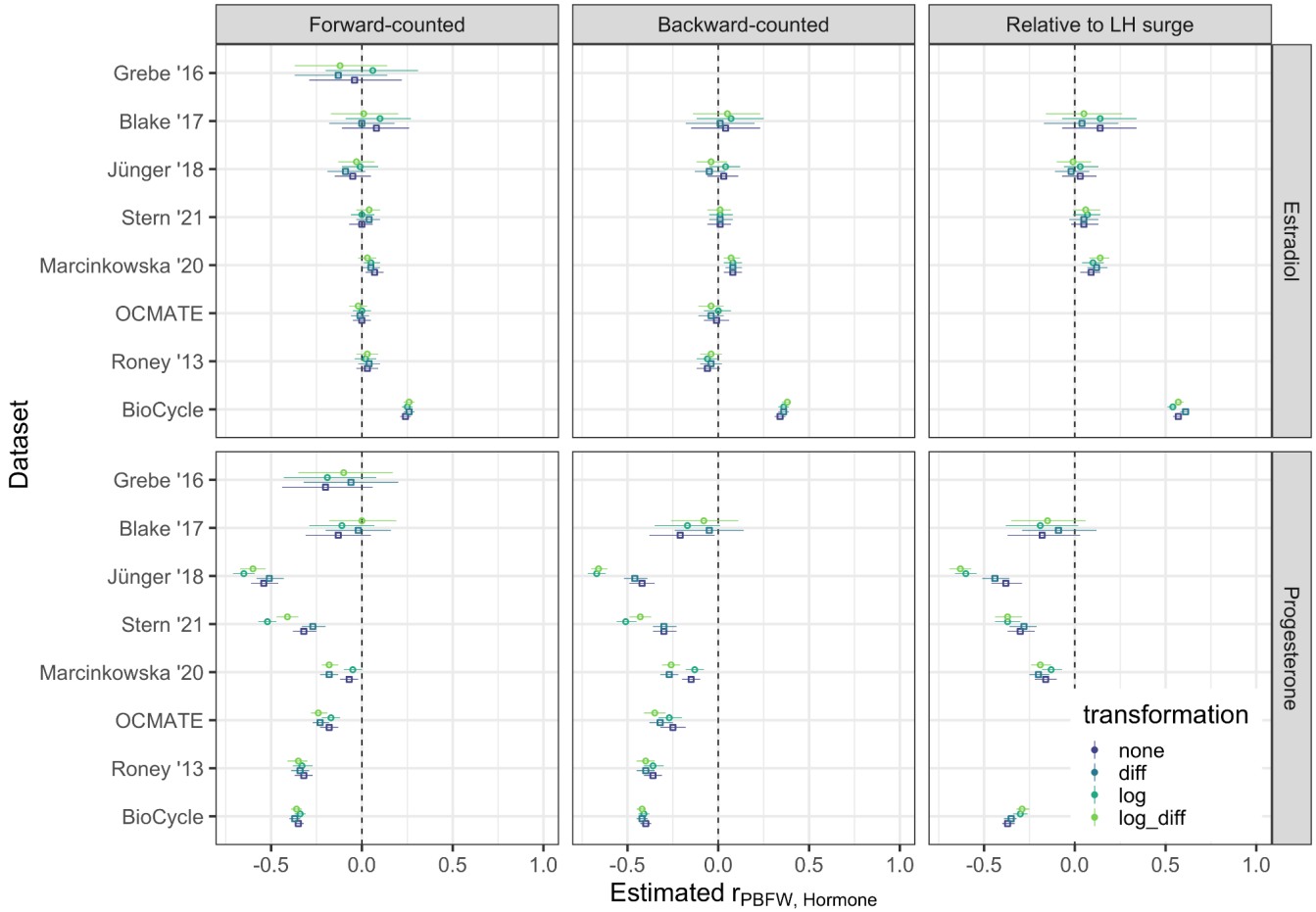
Supplementary Figure 1. Associations with cycle day relative to the recalled or observed last menstrual onset in all datasets. Dots show raw data. Coloured lines show two hundred random samples of the thin-plate spline fit using a Bayesian multilevel regression. Solid horizontal lines show the limit of detection; dashed the limit of quantitation. Progesterone values for BioCycle were multiplied by 2% as per Wood (2009) to make scales comparable.

Supplementary Figure 2



Supplementary Figure 2. Associations with cycle day relative to the observed next menstrual onset and relative to the LH surge in Blake et al. (2017). Dots show raw data. Coloured lines show two hundred random samples of the thin-plate spline fit using a Bayesian multilevel regression. Solid horizontal lines show the limit of detection.

Supplementary Figure 3

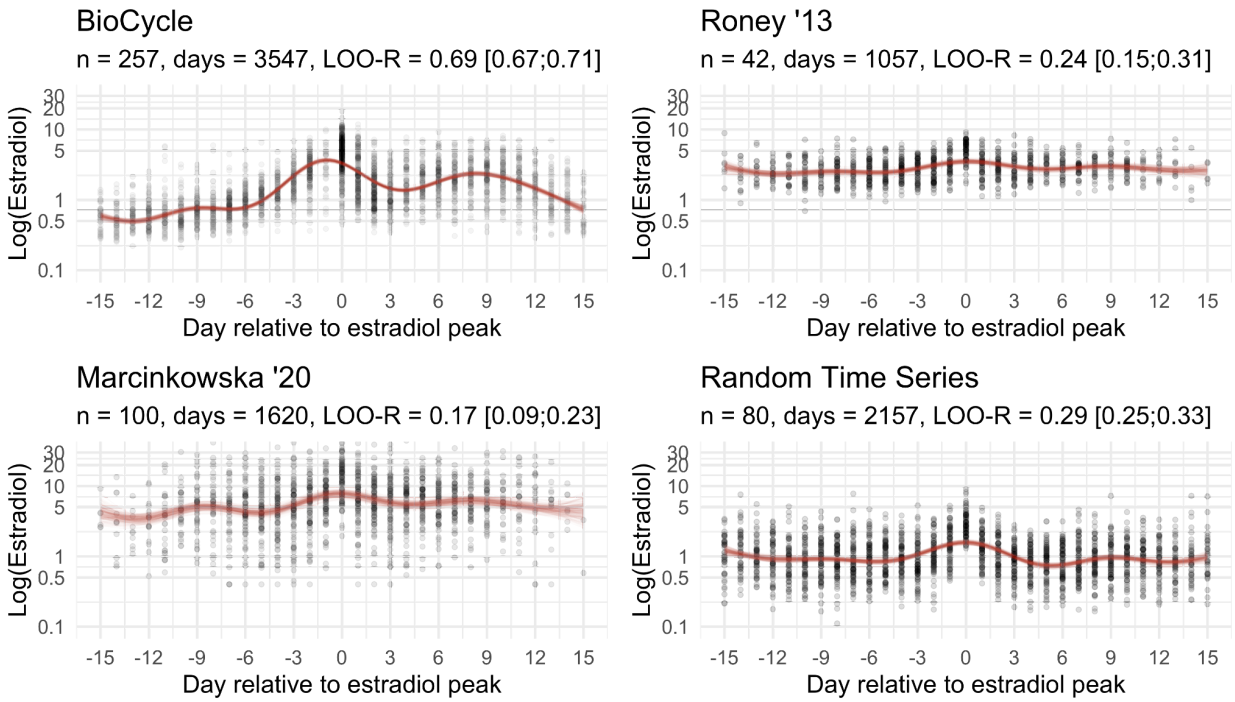


Supplementary Figure 3. Correlations between probability of being in the fertile window, according to forward-counting, backward-counting, and the LH surge, and hormone measures in various transformations (none, within-subject differences, log, within-subject differences from log).

Supplementary Note 2: Estradiol peaks as cycle phase measure

Several studies have employed salivary steroids themselves to estimate the day of ovulation. Algorithms vary and are sometimes not well-described, but one common approach is to identify the day of the highest E2 peak. If our steroid-independent measures are more noisy than we think, a steroid-based measure could show that the cycle phase - steroid relationship is stronger in reality. However, we also have to be mindful of circular reasoning. Previous studies have displayed a curve of the averages of estradiol when centred on the estradiol peak or the day of the largest estradiol drop. If this curve shows a peak in the middle, this is, by itself, not evidence of validity. Any random, stationary time series, when centred on its peak or drop, will show a peak or drop. It would be more informative to ask whether we can also see the secondary peak of the estradiol curve and how much variance this cycle phase measure explains compared to what a similar measure explains in a random time series. For three datasets that had sufficient repeated midcycle measurements per woman to approximately identify the estradiol peak, we applied the following algorithm: consider only forward-counted cycle days 7-20, identify the day of the maximum estradiol value in this set, compute the relative position of all other days in the cycle. We also simulated a dataset to contain 80 women with 30 days each, where the "estradiol" values were simply random, stationary time series with a small degree of autocorrelation. We then estimated multilevel models comparable with the LH surge models described above. We graphed the conditional means and computed the LOO-R. Only in the BioCycle data was the secondary estradiol peak clearly visible. LOO-R was also highest in the BioCycle data, whereas it was comparable in the Roney '13 and Marcinkowska '20 datasets and the random time series (Supplementary Figure 3). We also computed how well different algorithms identified the day of the urinary LH surge in the BioCycle and Marcinkowska data, and in both datasets simple backward counting had fairly similar accuracy to computing the day of the largest estradiol drop or peak.

Supplementary Figure 4



Supplementary Figure 4. Log (free) estradiol according to the day of the estradiol peak. The figure shows the three datasets with a sufficient number of measurement time points per woman to find the midcycle peak per cycle, as well as one simulated dataset that was set up to contain ~30 days per woman as stationary, autocorrelated time series (see OSF for detailed code).

Supplementary Note 3: Calculating the expected correlation between serum and saliva

We assumed that there is no direct causal relationship between cycle phase and saliva steroid levels, so that the association is fully accounted for by serum steroid levels (i.e. the causal graph resembles one of, CP -> Serum -> Saliva, or Saliva <- Serum -> CP). If this assumption holds, i.e. there is no other path from cycle phase to saliva steroid levels, we can derive the correlation between serum and saliva levels from their respective correlations with cycle phase (in this case, our cycle-phase-imputed expectations for each steroid) by standard path tracing rules (Wright, 1934). We can obtain the expected correlation between salivary levels and cycle phase by multiplying the cycle phase serum correlation with the saliva serum correlation.

$$r_{Cycle\ phase, Saliva} = r_{Serum, Saliva} \cdot r_{Cycle\ phase, Serum}$$

$$r_{Serum, Saliva} = \frac{r_{Imputation, Saliva}}{r_{Imputation, Serum}}$$

Solving the resulting equation for $r_{Serum, Saliva}$ and using our cycle-phase-based-imputation as a predictor gives us Eq. 1, a way to estimate the expected correlation between serum and salivary steroid levels. Our estimate of $r_{Serum, Saliva}$ will be biased downward if the cycle phase measure is measured with more error in the saliva studies than in the serum study, and vice versa. Concretely, if there were more anovulatory cycles in a saliva study than in the serum study, a counting-based cycle phase measure, which ignores anovulation, would lead to lower estimated validities (see e.g. Supplementary Note 4). This form of bias is likely less important for urinary luteinising hormone based measures. However, these also differ in their sensitivity across studies, with the BioCycle fertility monitor being the only one to use a digital readout of both urinary estrogen metabolites and luteinising hormone.

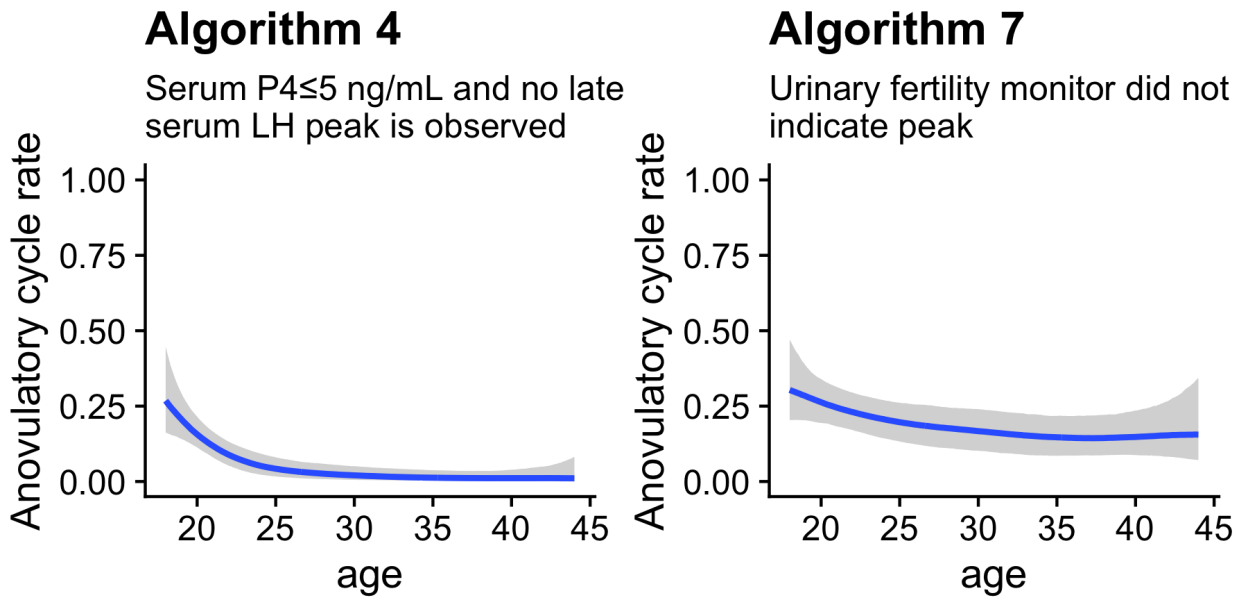
Our estimate would also be biased downward, if variation in cycle phase was more restricted in a saliva study (as, e.g., in Marcinkowska, 2020), but we adjust for range restriction to bring the correlations to a common denominator and avoid this bias.

We show 95% confidence intervals for these estimated correlations in Figure 4. These only include uncertainty in the numerator (i.e., we divided the lower and upper 95% interval bound of $r_{Imputation, Saliva}$ by the mean estimate for $r_{Imputation, Serum}$). Uncertainty in the denominator (explained variance in the BioCycle study) was quite low (forward-counted: [0.62;0.66], backward-counted: [0.74;0.77], relative to LH surge: [0.78;0.81]), but would of course additionally increase the uncertainty of the expected correlations.

Supplementary Note 4: Anovulation rates and imputation accuracy by age

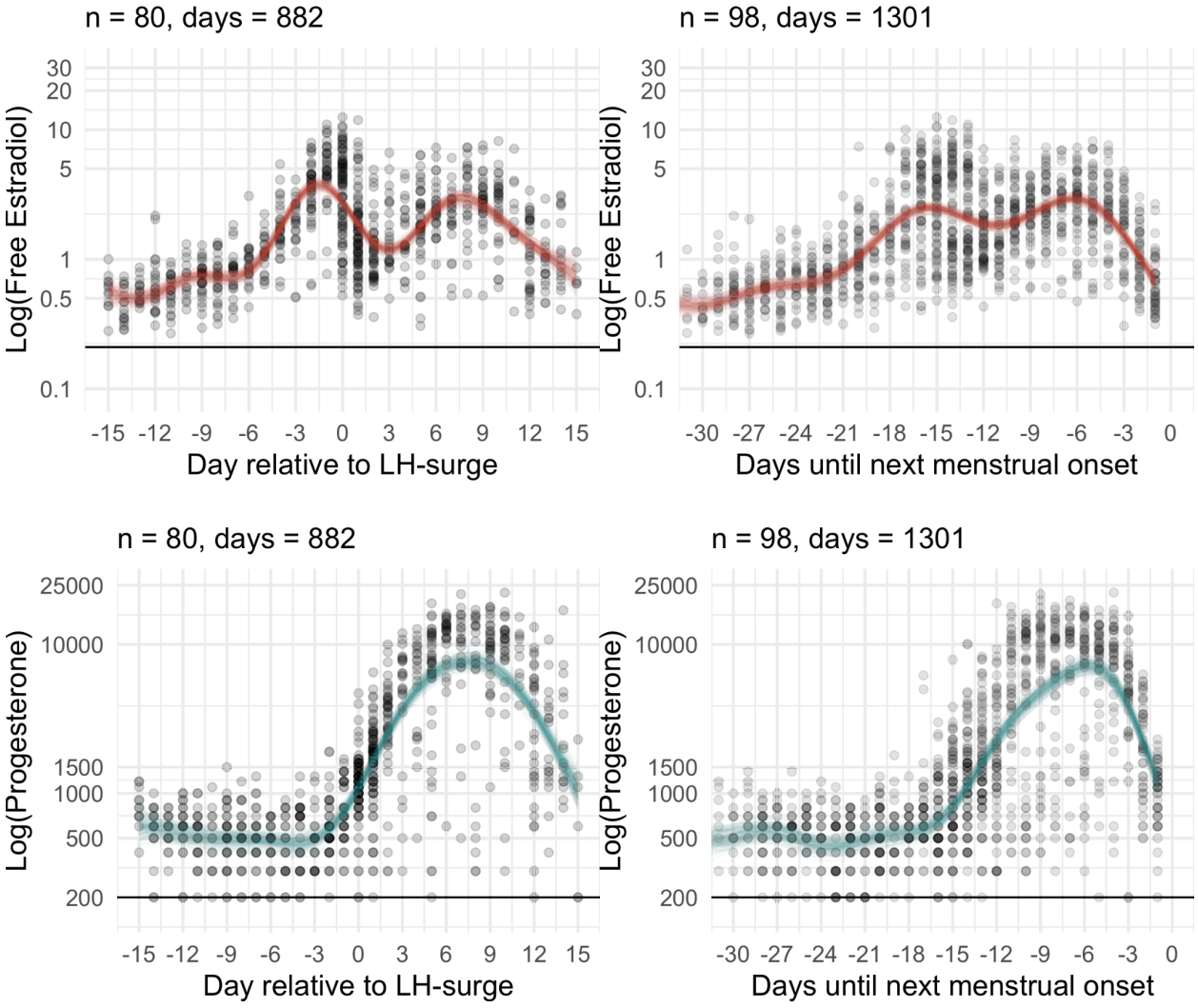
We used algorithms 4 and 7 from Lynch et al. (2014) to determine likely anovulatory cycles in the BioCycle data. We then estimated a model using several demographic factors (age, ethnicity, partnered, ever used birth control, BMI, high school education, parity, smoker) to predict anovulation. Only the 95% CI for age excluded zero. At age 18, 26.8% of cycles were predicted to be anovulatory. This rate reduced nonlinearly to 8.9% at age 22, 2.0% at age 30, and 1.1% at age 44 according to a Bayesian Poisson regression with only a thin-plate spline over age as the predictor to estimate the rate of anovulation per cycle continuously. However, the urinary fertility monitor classified far more cycles as anovulatory (see Supplementary Figure 5; Lynch et al., 2014). Based on this relationship with age, some of the included studies (especially Roney '13, OCMATE, Grebe '16) may have included substantially more anovulatory cycles than the BioCycle data. One might expect the accuracy of imputed hormones to be deflated in younger samples as a result. To quantify this deflation, we restricted the BioCycle data to the n=101 women aged 18-22. LOO-R using days relative to the urinary LH surge as the predictor of log free estradiol was 0.70 [0.66;0.73] and using backward-counted days, it was 0.66 [0.63;0.69], compared to 0.72 and 0.68 for the whole sample. For progesterone, the respective values were LOO-R 0.81 [0.77;0.84] and 0.75 [0.71;0.78], compared to 0.87 and 0.83 (see Supplementary Figure 6). In addition, the Marcinkowska '20 sample was older than the BioCycle sample on average, yet showed comparable results to other datasets. In summary, age and anovulation cannot fully explain the discrepancy in variance explained by cycle phase across saliva and serum studies.

Supplementary Figure 5



Supplementary Figure 5. Model-estimated rate of cycles classified as anovulatory by age in the BioCycle data, according to one serum-based and one urine-based algorithm.

Supplementary Figure 6



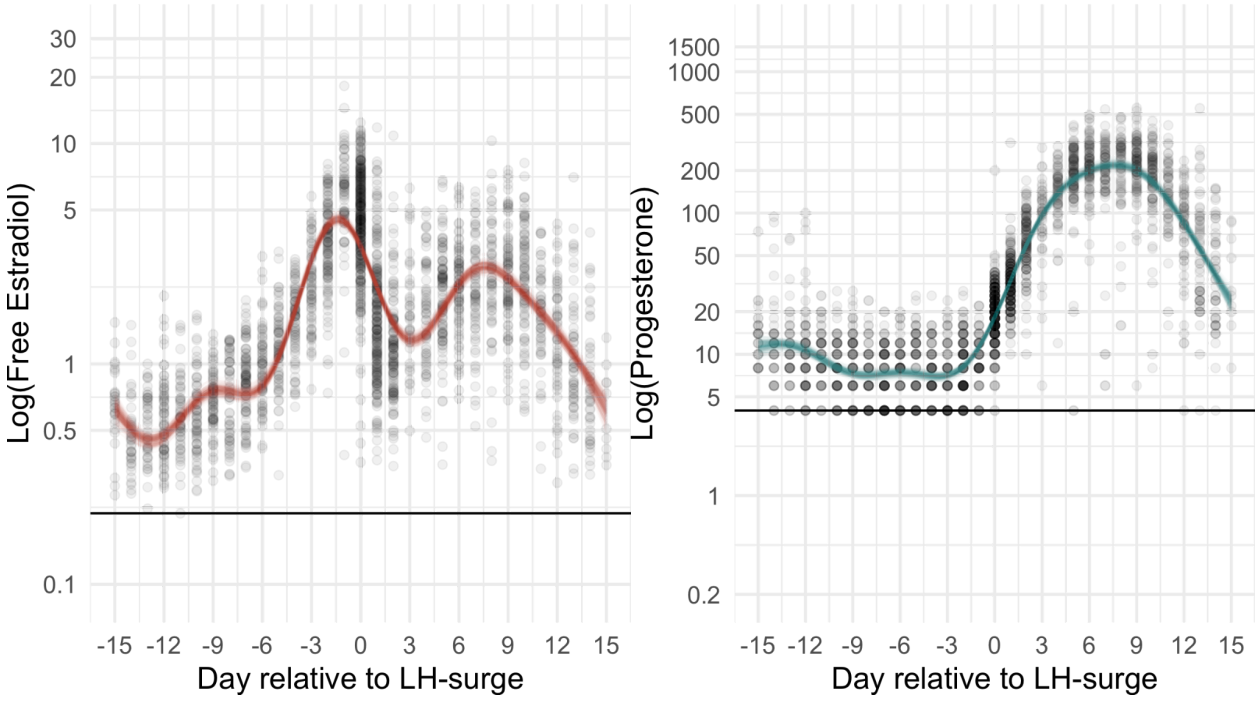
Supplementary Figure 6. Log free estradiol and log progesterone according to the urinary LH surge and observed next menstrual onset in all women 18-22 in the BioCycle data.

Supplementary Note 5: Serum steroids according to the serum luteinizing hormone surge

The BioCycle study measured LH both using a serum immunoassay and a ClearBlue home fertility monitor based on urine test strips. For imputation and for comparison with the salivary steroid studies, the urinary measure was more interesting for us. However, to facilitate comparisons with the literature (Stricker et al., 2006) and to understand how well steroids can be imputed from LH surges if they are measured with less error in studies collecting serum, we also computed the relationship between steroids and the cycle day relative to the serum LH surge (the first day of the cycle where LH > 20 ng/ml). For log estradiol, the LOO-R was 0.76 (95% CI [0.74;0.78]). For log progesterone, the LOO-R was 0.91 [0.90;0.92], see also Supplementary Figure 4.

We also computed correlations between average hormones and the 5% and 95% percentiles reported in Stricker et al. (2006) according to the LH peak, to see whether our imputations were generalizable. The means computed by Stricker et al. (2006) correlated .99 with the means computed by us in the BioCycle data for both estradiol and progesterone. The 5% and 95% percentiles were correlated between .90 and .97 across studies. In a plot, absolute levels also were highly similar. Thus, both means and estimates of variability across women were highly similar across two high-quality studies based on serum.

Supplementary Figure 7



Supplementary Figure 7. Log free estradiol and log 2% progesterone according to the serum LH surge in the BioCycle data.

Supplementary Note 6: Urinary validity data.

Urinary immunoassays of estrogen and progestogen metabolites are an infrequently used non-invasive alternative to salivary immunoassays (see Supplementary Note 1). Roos et al. (2015) report correlations of r=0.54 and r=0.82 between estradiol and progesterone in serum and their respective urinary metabolites one day later in a sample of 40 women who participated for an entire cycle (see also Denari et al., 1981). These validities, while suggesting an improvement over salivary immunoassays for estradiol, are still lower than those for our values imputed from cycle phase (backward-counting and LH) in the BioCycle data. But unlike imputation, urinary assays additionally permit examination of between-cycle and between-woman variation. Yet, if within-cycle effects are of interest and budgets constrained, researchers could achieve higher statistical power through imputation from inexpensive LH tests or backward counting methods.

We cite the Roos et al. (2015) coefficients for the urine-serum correlation. They are interpretable as coefficients of validity, as they reflect correlations between individual yoked samples. Other work (e.g. Gordon et al., 2021) has on occasion cited O'Connor et al. (2003) to show validity of urinary immunoassays. O'Connor et al. (2003) report r=.93 for estradiol and r=.97 for progesterone at a one day lag. However, O'Connor et al. (2003) aggregated values by cycle day before computing correlations. Therefore, the correlations are extremely inflated if taken as indicators of the validity of individual measurements. For comparison, we merged the aggregated serum values by cycle day reported in Roos et al. (2015) with the urine values reported in Johnson et al. (2015) for the same dataset. Correlations at a one day lag were r=0.90 for estradiol and r=0.98 for progesterone and their respective metabolites.

Supplementary References

Denari, J.H., Farinati, Z., Casas, P.R., Oliva, A., 1981. Determination of ovarian function using first morning urine steroid assays. Obstet. Gynecol. 58, 5–9.

Gordon, J.L., Halleran, M., Beshai, S., Eisenlohr-Moul, T.A., Frederick, J., Campbell, T.S., 2021. Endocrine and psychosocial moderators of mindfulness-based stress reduction for the prevention of perimenopausal depressive symptoms: A randomized controlled trial. Psychoneuroendocrinology 130, 105277. https://doi.org/10.1016/j.psyneuen.2021.105277

Johnson, S., Weddell, S., Godbert, S., Freundl, G., Roos, J., Gnoth, C., 2015. Development of the first urinary reproductive hormone ranges referenced to independently determined ovulation day. Clin. Chem. Lab. Med. 53, 1099–1108. https://doi.org/10.1515/cclm-2014-1087

O'Connor, K.A., Brindle, E., Holman, D.J., Klein, N.A., Soules, M.R., Campbell, K.L., Kohen, F., Munro, C.J., Shofer, J.B., Lasley, B.L., Wood, J.W., 2003. Urinary estrone conjugate and pregnanediol 3-glucuronide enzyme immunoassays for population research. Clin. Chem. 49, 1139–1148. https://doi.org/10.1373/49.7.1139

Roos, J., Johnson, S., Weddell, S., Godehardt, E., Schiffner, J., Freundl, G., Gnoth, C., 2015. Monitoring the menstrual cycle: Comparison of urinary and serum reproductive hormones referenced to true ovulation. Eur. J. Contracept. Reprod. Health Care 20, 438–450. https://doi.org/10.3109/13625187.2015.1048331

Stricker, R., Eberhart, R., Chevailler, M.-C., Quinn, F.A., Bischof, P., Stricker, R., 2006. Establishment of detailed reference values for luteinizing hormone, follicle stimulating hormone, estradiol, and progesterone during different phases of the menstrual cycle on the Abbott ARCHITECT analyzer. Clin. Chem. Lab. Med. 44, 883–887. https://doi.org/10.1515/CCLM.2006.160

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