npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Characterizing physiological and symptomatic variation in menstrual cycles using self-tracked mobile-health data

Kathy Li[1,2], Iñigo Urteaga[1,2], Chris H. Wiggins[1,2], Anna Druet[3], Amanda Shea[3], Virginia J. Vitzthum[3,4] and Noémie Elhadad[2,5]✉

The menstrual cycle is a key indicator of overall health for women of reproductive age. Previously, menstruation was primarily studied through survey results; however, as menstrual tracking mobile apps become more widely adopted, they provide an increasingly large, content-rich source of menstrual health experiences and behaviors over time. By exploring a database of user-tracked observations from the Clue app by BioWink GmbH of over 378,000 users and 4.9 million natural cycles, we show that self-reported menstrual tracker data can reveal statistically significant relationships between per-person cycle length variability and self-reported qualitative symptoms. A concern for self-tracked data is that they reflect not only physiological behaviors, but also the engagement dynamics of app users. To mitigate such potential artifacts, we develop a procedure to exclude cycles lacking user engagement, thereby allowing us to better distinguish true menstrual patterns from tracking anomalies. We uncover that women located at different ends of the menstrual variability spectrum, based on the consistency of their cycle length statistics, exhibit statistically significant differences in their cycle characteristics and symptom tracking patterns. We also find that cycle and period length statistics are stationary over the app usage timeline across the variability spectrum. The symptoms that we identify as showing statistically significant association with timing data can be useful to clinicians and users for predicting cycle variability from symptoms, or as potential health indicators for conditions like endometriosis. Our findings showcase the potential of longitudinal, high-resolution self-tracked data to improve understanding of menstruation and women's health as a whole.

npj Digital Medicine (2020)3:79 ; https://doi.org/10.1038/s41746-020-0269-8

INTRODUCTION

Menstruation is an important indicator of overall health and quality of life in women: the reproductive endocrine system is associated with sexual and reproductive health, bone and heart health, and cancers[1–9]; it affects fertility[10,11], menopause[12–14], exercise[15], and diet[16]. Seminal work on variation of menstrual cycle length throughout the reproductive lifespan[17,18] has concluded that "complete regularity in menstruation through extended time is a myth," and recent empirical studies[19–21] have confirmed that variation between cycles, women, and populations is the norm[22–29]. Establishing a clear, informative, and quantitative characterization of the patterns and the underlying female physiology of what has been hypothesized as "the fifth vital sign"[30–33] has been a long-explored issue in women's health, but remains an open research question[27,34,35], in part due to limited access to large, reliable datasets concerning menstruation.

With the rise of data-powered health, we now have the ability to identify menstrual patterns at scale and explore their relationships with a broad set of symptoms. Observational health data sources have shed light on individual clinical trajectories[36], increased self-awareness about individual health[37], and helped deliver on the promise of precision medicine[38]. Mobile-health solutions enable a high-resolution view of a large, highly diverse range of individuals over time[39–42], and can provide insights into chronic diseases and behaviors[43–52]. Menstrual trackers in particular have become increasingly common: they are the second most popular app for adolescent girls and the fourth most popular for adult women[53,54]. Millions of women around the world routinely track their menstrual cycles and a variety of contextual factors and symptoms, accumulating high volumes of temporal, heterogeneous data via many different apps[55–59]. As exemplified by studies connecting the menstrual cycle to variations in women's mood, behavior, and vital signs[60], self-tracked data can provide insights into cycle characteristics[61], ovulation timing, and the evolution of reproductive health for large populations[62], as well as empower informed decision-making through increased self-awareness[63].

We utilize deidentified user-tracked data from Clue by BioWink GmbH[55], one of the most popular and accurate menstrual trackers worldwide[64]. In addition to period data, Clue users can track symptom information in categories like exercise, pain, and sexual activity (see Fig. 1). Note that Clue users are not required to specify gender—in this paper, we refer to Clue users or menstruators as 'women,' but we acknowledge that not all menstruators are women and vice versa. This large-scale dataset provides a high-resolution, long-term view of variation in both physiology (period and cycle duration) and symptoms (e.g., pain and mood) across menstrual cycles, enabling us to study the shared information between quantitative, temporal attributes and qualitative, symptomatic attributes of menstrual experiences.

While previous work has examined how menstrual cycle characteristics like cycle and phase length vary with age and body mass index (BMI)[61], we aim instead to use the observed variability in cycle length statistics to investigate differences in symptomatic behavior between those who exhibit more or less variable cycle lengths. Namely, we seek to answer two research questions: (1) how do cycle length characteristics for a large, self-tracked user population differ among groups of users? and (2)

[1]Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027, USA. [2]Data Science Institute, Columbia University, New York, NY 10027, USA. [3]Clue by BioWink GmbH, Adalbertstraße 7-8, 10999 Berlin, Germany. [4]Kinsey Institute & Department of Anthropology, Indiana University, Bloomington, IN 47405, USA. [5]Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA. ✉email: noemie.elhadad@columbia.edu

Fig. 1 Clue app screenshot. Sample screenshots of the Clue app. Users can track daily symptoms across 20 categories; Table 3 provides a description of the available Clue categories and their corresponding symptoms. On the left for example, the app displays what day the user is currently on in their cycle. On the right, a user can choose from 'cramps,' 'headache,' 'ovulation,' or 'tender breasts' symptoms for the category 'pain' (the third most tracked category in our dataset, see Table 3). Screenshots produced by and used with permission from Clue by BioWink GmbH.

how do users who fall at different ends of the cycle length variability spectrum self-track their menstrual symptoms?

To this end, we select users from the Clue dataset aged 21–33 (because menstrual cycle lengths are relatively less variable, and cycles are more likely to be ovulatory during this age interval[18,22,29,65,66]) with natural menstrual cycles (i.e., no hormonal birth control or intrauterine device (IUD)). We define a menstrual cycle as the span of days from the first day of a period through to and including the day before the first day of the next period[29]. A period consists of sequential days of bleeding (greater than spotting and within 10 days after the first greater-than-spotting bleeding event) unbroken by no more than 1 day on which only spotting or no bleeding occurred.

In this paper, we take symptom tracking behavior to be a proxy for true physiological behavior. The Clue tracking categories (summarized in Table 3) encompass a wide range of experiences (subject to user interpretation of the category rather than based on specific validated scales), enabling broad usage of the app to meet individual user needs. Self-tracked data reliability is dependent on consistent and accurate user tracking; for instance, cycle length can be arbitrarily long if a user forgets to track their period, which would skew the analysis of menstrual patterns by misrepresenting a long cycle as due to physiological behavior rather than tracking behavior. We propose a procedure to mitigate such potential engagement artifacts by quantifying engagement with cycle tracking and identifying cycles lacking engagement, allowing us to separate true menstrual patterns from tracking anomalies. To investigate the spectrum of variability in women's menstrual health experiences, we propose cycle length difference or CLD—the absolute difference between subsequent cycle lengths—as a robust metric for quantifying cycle variability, and we examine users who fall at opposite ends of the variability spectrum.

RESULTS

Study population

The cohort for this study comprises 378,694 users located on all continents, aged 21–33 years old (see Table 1 for detailed summary statistics). The average user is 25.49 (median of 25) years old (per-country and per-age detailed statistics are provided in the Supplementary Information). As reported in Table 2, the average number of cycles tracked per user is 12.89 (median of 11), with an average cycle length of 29.73 (median of 29) days and mean period length of 4.08 (median of 4) days.

Cycle length difference (CLD) as a robust metric for quantifying cycle variability

We propose cycle length difference, or CLD—the absolute difference between subsequent cycle lengths—as a powerful metric to characterize the spectrum of menstrual variability. We examine each user's CLDs to identify those who are 'consistently highly variable' in their cycle lengths. We find that a median CLD of 9 days splits consistently highly variable and consistently not highly variable cycle behavior, and we use this threshold to separate the menstrual experiences and symptom reporting of those at different ends of the variability spectrum. Tables 1–3 showcase the summary statistics, high-level cycle characteristics,

Table 1. Summary statistics of the full cohort.

Variable	Full cohort	Consistently not highly variable	Consistently highly variable
Number of users	378,694 (100.00%)	349,606 (92.32%)	29,088 (7.68%)
Number of observations	117,014,597 (100.00%)	112,093,683 (95.79%)	4,920,914 (4.21%)
Number of days of observation	34,056,343 (100.00%)	32,699,312 (96.02%)	1,357,031 (3.98%)
Number of cycles	4,881,697 (100.00%)	4,701,694 (96.31%)	180,003 (3.69%)

Summary statistics of the full cohort, as well as for the consistently not highly variable and consistently highly variable user groups. We utilize a greater than 9 day median cycle length difference threshold to place users in each group—those in the consistently highly variable group represent the far end of a cycle variability spectrum.

Table 2. Per-user cycle characteristics.

Variable	Full cohort's mean ± sd (95% CI)	Full cohort's median	Consistently not highly variable group's mean ± sd (95% CI)	Consistently not highly variable group's median	Consistently highly variable group's mean ± sd (95% CI)	Consistently highly variable group's median
Number of cycles	12.89 ± 9.11 (3.00,36.00)	11.00	13.45 ± 9.19 (3.00,37.00)	11.00	6.19 ± 3.87 (2.00,17.00)	5.00
Cycle length	29.73 ± 5.73 (21.00,43.00)	29.00	29.45 ± 4.98 (21.00,41.00)	29.00	37.04 ± 13.71 (13.00,69.00)	34.00
Period length	4.08 ± 1.76 (1.00,7.00)	4.00	4.07 ± 1.72 (1.00,7.00)	4.00	4.28 ± 2.54 (1.00,9.00)	4.00
Median CLD	4.15 ± 4.94 (1.00,18.00)	3.00	3.04 ± 1.86 (1.00,8.00)	2.50	17.48 ± 9.15 (9.50,43.00)	14.00
Maximum CLD	10.07 ± 7.49 (2.00,31.00)	8.00	8.82 ± 5.65 (2.00,23.00)	8.00	25.15 ± 10.10 (12.00,53.00)	23.00

Per-user high-level cycle characteristics for the full cohort, as well as for the consistently not highly variable and consistently highly variable user groups. We utilize a greater than 9 day median cycle length difference threshold to place users in each group—those in the consistently highly variable group represent the far end of a cycle variability spectrum. The 'cycle length difference' (CLD) refers to the absolute difference between two consecutive cycles.

Table 3. Description of the Clue app tracking categories and symptoms.

Category	Description	Symptoms	Number of tracking events (%) for the consistently not highly variable group	Number of tracking events (%) for the consistently highly variable group
Period	Period flow	Spotting, light, medium, and heavy	22,096,884 (19.71%)	913,403 (18.56%)
Emotion	Emotional state	Happy, sensitive, sad, and PMS	11,377,997 (10.15%)	501,610 (10.19%)
Pain	Type of pain experienced	Cramps, tender breasts, headache, and ovulation pain	9,730,958 (8.68%)	406,710 (8.26%)
Energy	Energy level	Low, high, exhausted, and energized	8,710,403 (7.77%)	410,216 (8.34%)
Sleep	Hours of sleep	0–3, 3–6, 6–9, and >9	8,597,769 (7.67%)	405,726 (8.24%)
Skin	Skin health	Acne, good, oily, and dry	5,896,540 (5.26%)	263,258 (5.35%)
Mental	Mental state	Calm, distracted, focused, and stressed	5,871,137 (5.24%)	252,621 (5.13%)
Sex	Sexual health	Unprotected sex, high sex drive, protected sex, and withdrawal sex	5,813,292 (5.19%)	271,540 (5.52%)
Motivation	Motivation level	Motivated, unmotivated, productive, and unproductive	5,467,728 (4.88%)	236,052 (4.80%)
Craving	Food cravings	Sweet, salty, carbs, and chocolate	4,867,777 (4.34%)	224,751 (4.57%)
Digestion	Digestive health	Great, bloated, gassy, and nauseated	4,825,627 (4.30%)	209,651 (4.26%)
Social	Social behavior	Sociable, withdrawn, supportive, and conflict	4,178,744 (3.73%)	186,110 (3.78%)
Poop	Stool health	Normal, constipated, great, and diarrhea	3,889,471 (3.47%)	172,716 (3.51%)
Hair	Hair health	Good, bad, oily, and dry	3,128,384 (2.79%)	147,844 (3.00%)
Fluid	Vaginal discharge type	Creamy, egg white, sticky, and atypical	2,378,211 (2.12%)	106,782 (2.17%)
Collection method	Method for period collection	Pad, tampon, panty liner, and menstrual cup	2,027,258 (1.81%)	84,270 (1.71%)
Exercise	Physical exercise	Running, yoga, biking, and swimming	1,222,568 (1.09%)	44,946 (0.91%)
Party	Party-related experiences	Drinks, cigarettes, big night, and hangover	900,444 (0.8%)	40,779 (0.83%)
Medication	Type of medication taken	Pain, cold/flu, antihistamine, and antibiotic	561,540 (0.5%)	21,030 (0.43%)
Ailment	Physical maladies	Cold/flu, allergy, injury, and fever	550,951 (0.49%)	20,899 (0.42%)

Description of tracking categories and the corresponding symptoms for the Clue app, along with the per-symptom number of tracking observations (and their corresponding proportion with respect to the total number of observations) for the consistently not highly variable and consistently highly variable user groups.

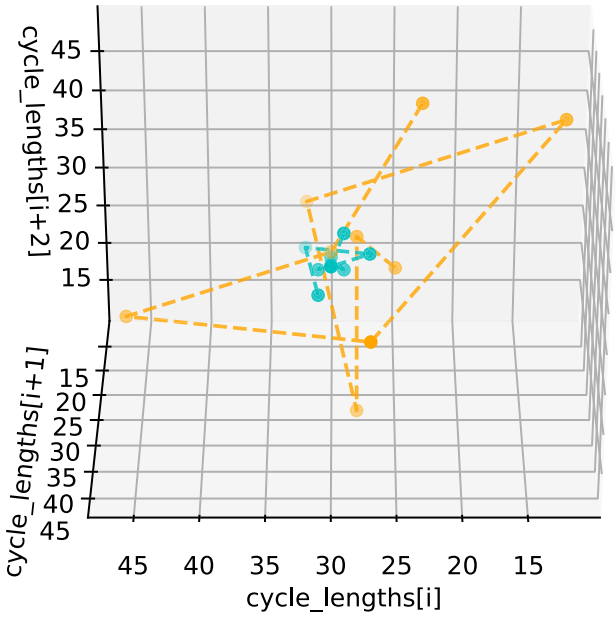

Fig. 2 Time series embedding for cycle length (one user, across groups). We sample one consistently highly variable and one consistently not highly variable user, each with the median number of cycles (11), from the user cohort and plot each set of three consecutive cycles on the *x*, *y*, and *z* axes, respectively. This allows us to visualize how much a user's cycle lengths change throughout their entire cycle tracking history—we would expect that a not consistently highly variable user would have points that cluster closer together in space. We see that the consistently not highly variable (teal) user occupies a small region, while the consistently highly variable (orange) user's points move through the space. This indicates that the teal user's cycle lengths are consistently very similar to one another, whereas the orange user experiences more consistent fluctuation in cycle lengths. Thus, we see that separating users into groups on the basis of median CLD identifies those who are more and less consistently highly variable.

and category tracking frequencies for the resulting two groups, respectively.

We note that the consistently highly variable group comprises about 7.68% of the user cohort (29,088 out of 378,694 users), and that their relative category tracking frequencies are similar to the larger, consistently not highly variable user group. Period flow, emotional state, and experienced pain are the most frequently tracked categories across both groups; they account for 38.54% of the events for the consistently not highly variable group and 37.01% for the consistently highly variable group. Below, we summarize the commonalities and differences in cycle and period length characteristics and symptom tracking behavior between these two populations.

Cycle variability characterization—women in the consistently highly variable group experience volatile cycle lengths

We examine the cycle length characteristics of the proposed user groups, both visually and statistically. At an individual level, we visualize for two randomly sampled users (one per group) a time series embedding of all of their consecutive cycle lengths in Fig. 2. We sample one consistently highly variable and one consistently not highly variable user with the median number of cycles (11) from the cohort and plot each set of three consecutive cycles on the *x*, *y*, and *z* axes, respectively. In Fig. 2, the consistently not highly variable (teal) user occupies a small region of the space, indicating that this user experiences similar cycle lengths throughout their history; however, the consistently highly variable (orange) user's points wander through the space, indicating that this user experiences consistently fluctuating cycle lengths throughout their cycle history.

In Fig. 3a, we plot the time series embeddings of cycle length for the entire cohort, where each point represents three consecutive cycles randomly sampled from each user's cycle history (for users with at least three cycles). In contrast to Fig. 2, each user is represented by one point, instead of plotting the whole cycle histories of two randomly sampled users. We visualize at a population level whether our median CLD metric successfully

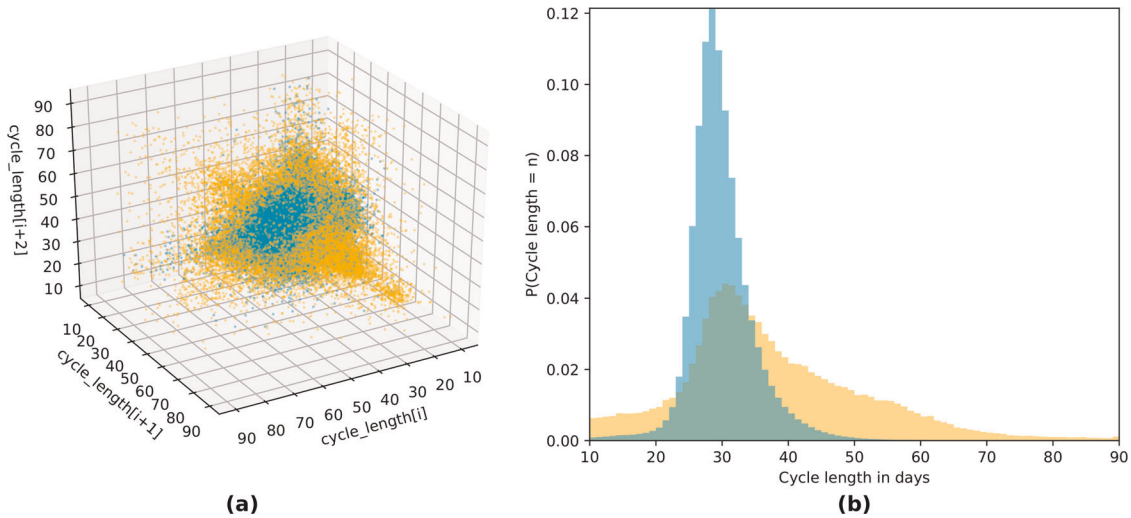

Fig. 3 Time series embedding and probability distribution for cycle length (all users, across groups). Time series embedding (**a**) and probability distributions (**b**) of cycle length for the consistently not highly variable (teal) and consistently highly variable (orange) groups. **a** The cycle lengths of three consecutive randomly sampled cycles from each user in the cohort are plotted on the *x*, *y*, and *z* axes. Each consistently not highly variable user is represented by a teal point, and each consistently highly variable user by an orange point. It is visually evident that the teal cluster of users occupies a tighter region of the space around the $x = y = z$ line, with the orange cluster fanning outward. **b** The cycle length probability distributions of the cohort, where we note that the orange group's distribution has a much wider spread and is less peaked than the teal group. Cycle lengths are more heterogeneous or widely distributed for the orange group, confirming that the consistently highly variable group represents those with more fluctuation in cycle length. The cumulative distributions per group differ significantly (as per a two-sample Kolmogorov–Smirnov test).

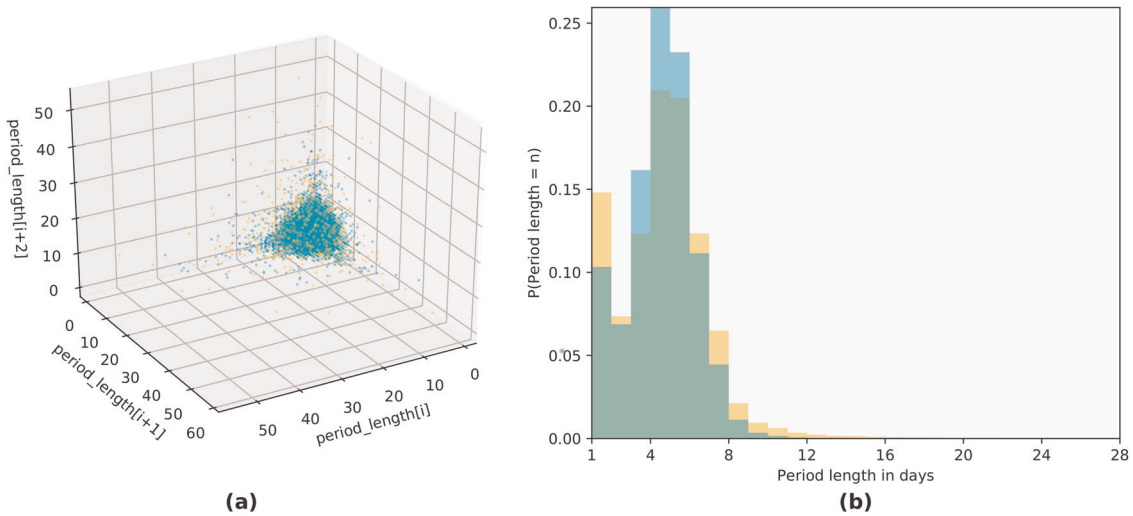

Fig. 4 Time series embedding and probability distribution for period length (all users, across groups). Time series embedding (**a**) and probability distributions (**b**) of period length for the consistently not highly variable (teal) and consistently highly variable (orange) groups. **a** The period lengths of three consecutive randomly sampled cycles from each user in the cohort are plotted on the *x*, *y*, and *z* axes. Visually, we observe that both groups occupy a very similar region of the period length space (few orange points are placed outside the region occupied by the teal cluster). **b** The period length probability distributions of the cohort, where we observe that the orange and teal distributions are largely overlapping, with the same median of 4 days and a similar shape, indicating that period lengths are distributed very similarly for the two groups. We notice a slight peak in single day period reports in both groups, which we argue is reminiscent of app usage behavior: some users are interested in knowing (approximately) when they had their period, not in tracking how long it was, so they may only track the day it occurred and not continue tracking after that.

separates out groups of users based on their cycle length fluctuations. If a user is perfectly consistently not highly variable, then its representative point would fall exactly on the $x = y = z$ line, since the three cycle lengths would be identical (i.e., not fluctuating at all). We observe a consistent phenomenon in Fig. 3a: the consistently not highly variable group (teal) occupies a tighter region of the space than the consistently highly variable one (orange). That is, a user in the consistently highly variable group experiences volatile menstrual patterns (i.e., highly varying cycle lengths).

Furthermore, we study the empirical cycle length distributions per group, and as seen in Fig. 3b, the cycle length distributions differ significantly between the two user groups. Observe that not only are cycle length statistics such as mean and median cycle length different, but that the shapes of the distributions are also distinct. Specifically, in addition to being centered at longer cycle lengths (median of 34 vs. 29 days), the cycle length distribution for the consistently highly variable group is less peaked with a wider spread (encompasses a more volatile range of cycle lengths), has much heavier tails, and is skewed toward longer cycle lengths.

Cycle variability characterization—period length statistics are homogeneous across the variability spectrum

We find that while women in different groups as separated by median CLD differ in their cycle length variability, their period length distributions are much less variable and fluctuate similarly between the groups. Period length is centered around the same median of 4 days for both groups and displays a similar length distribution. Figure 4 confirms that the variability in cycle length is not due to period length differences between the groups, as the period length varies the same amount across all women. These results show that our metric (median CLD) identifies two distinct groups of users based on their cycle (not period) length variability. Note that while the period length distributions do differ significantly by the two-sample Kolmogorov–Smirnov (KS) test, the KS statistic for the period length distributions is 0.066 with a 95% confidence interval of (0.064, 0.068) (details on computing the confidence interval using bootstrapping are presented in the Methods section). These numbers are nearly an order of magnitude smaller than those for the cycle length distributions (0.377 with a 95% confidence interval of (0.375, 0.378)). That is, the KS test identifies the cycle length distributions to differ more drastically and with much higher probability than the period length distributions.

Cycle variability characterization—cycle and period length statistics are stationary within groups over the app usage timeline

We study per-group cycle statistics over the app usage timeline (as represented by cycle ID) in Fig. 5 and find that cycle and period length statistics are stationary over time at the group level. Cycle ID enables us to align all users according to their subsequently tracked cycles (not absolute time), i.e., a cycle ID of 1 corresponds to the first cycle of a user, 2 to their second cycle, and so on. As reported in Table 2, the mean cycle length for the consistently not highly variable group is 29.45 days (median of 29), and the mean is 37.04 days (median of 34) for the consistently highly variable group. We observe that while the average cycle and period lengths are similar over subsequent reported cycles for both the entire user cohort and the consistently not highly variable user group, consistently highly variable users exhibit a wider spread (i.e., higher volatility). This volatility is maintained across cycles for users in the consistently highly variable group, showcasing that this group accounts for a large degree of the volatility in the data; this detail would be largely 'smoothed out' and lost if we considered the whole population rather than separating the users into two groups. Since cycle and period length statistics are constant within groups across app usage, we are confident that the proposed median CLD is not merely capturing spurious correlations that depend on how long the user stays with the app.

Reported symptom differences—women located at different ends of the spectrum of menstrual variability exhibit different symptom patterns

We find that there exists a relationship between median CLD and cycle level user symptom tracking behavior—despite CLD being a measure of cycle length variability, it is also correlated with

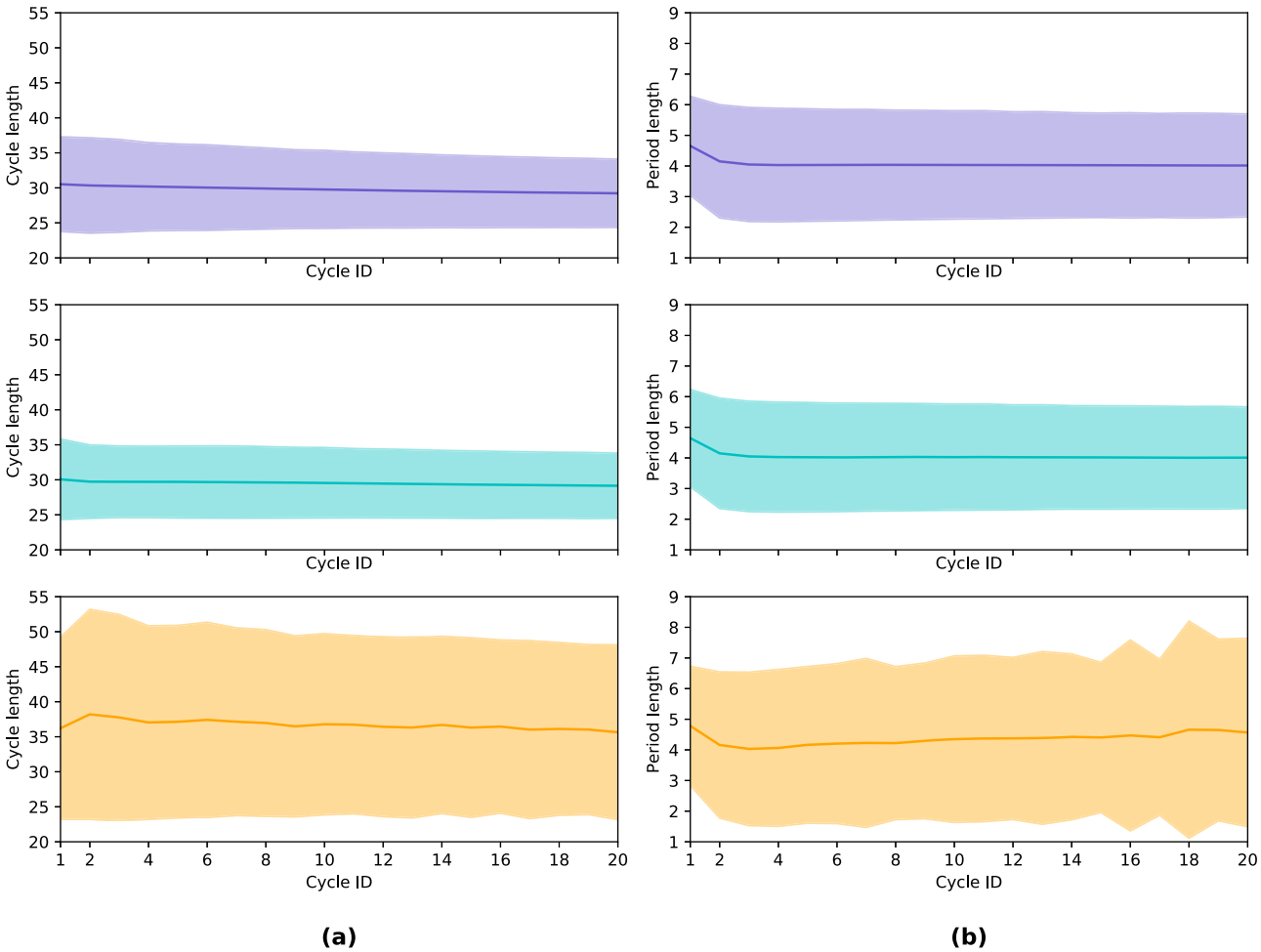

Fig. 5 Average cycle and period length by cycle ID. For each user's cycles (indexed by cycle ID), we average cycle (**a**) and period length (**b**) across three different groups: the entire user cohort (top, purple), the consistently not highly variable user cohort (middle, teal), and the consistently highly variable user cohort (bottom, orange). This allows us to visualize how cycle and period length vary over time for each group on average and in terms of standard deviation (for illustrative purposes, we restrict the cycle ID to 20). Cycle and period length statistics are stationary over the app usage timeline within each plot. We note that the top and middle plots look similar in each figure (i.e., the consistently not highly variable group looks similar to the overall population in terms of both cycle and period length), but the wider shaded orange spread of the bottom plot demonstrates the higher degree of variability in the consistently highly variable group. In addition, this spread is consistently wider for the orange plot over time. This showcases that the consistently highly variable group represents a large degree of the variability that we see in the data overall.

symptom tracking behavior. Specifically, our analysis of the symptoms tracked across the two variability groups showcases that while users exhibit similar tracking frequencies (i.e., the total number of times they track throughout history) per category (as in Table 3), there are notable differences among their symptom tracking patterns (i.e., how they track throughout history). The population-level distributions of our metric (i.e., 'proportion of cycles with symptom out of cycles with category' in Eq. (1)) differ between the user groups across most categories, with these differences being significant for all symptoms within the period, pain, and emotion categories, a result that may be clinically useful for assessing menstrual conditions and overall wellness. We present the KS test results for symptoms within those categories in Table 4.

Reported symptom differences—women in the consistently highly variable group display more heterogeneous period tracking behavior

We find that women in the consistently highly variable group are significantly more likely not to report heavy periods throughout their cycle history (odds ratio of 1.734 on the low extreme end of the proportion range in Table 5). In addition, the tracking pattern for spotting period flow is more heterogeneous for the consistently highly variable group, as shown by the higher odds ratios on both extremes of the proportion range (i.e., either in all or none of their cycle history) shown in Tables 5 and 6.

Reported symptom differences—women in the consistently highly variable group report pain-related symptoms more unpredictably

We observe generally more heterogeneous experiences for non-bleeding related symptoms like pain for the consistently highly variable group. Of particular interest is the finding that users in the consistently highly variable group are much more likely associated with tracking headaches and tender breasts in at least 95% of their cycles, with odds ratios of 1.663 and 1.715, respectively (see Table 6).

DISCUSSION

Characterization of menstrual patterns has been previously explored, though typically in relation to cycle and period lengths

only. While common knowledge refers to a 28-day cycle as "normal," this belief has been consistently disproved by clinical studies[18,22], as well as by recent analysis of high-level cycle characteristics via menstrual self-tracking apps[61,62].

Table 4. Kolmogorov–Smirnov test results for symptom tracking patterns.

Category	Symptom	Kolmogorov–Smirnov statistic (95% CI)
Period flow	Heavy	0.181 (0.178,0.183)
Period flow	Medium	0.134 (0.132,0.137)
Period flow	Light	0.121 (0.118,0.124)
Period flow	Spotting	0.089 (0.087,0.092)
Type of pain experienced	Cramps	0.101 (0.097,0.104)
Type of pain experienced	Ovulation pain	0.096 (0.093,0.099)
Type of pain experienced	Headache	0.089 (0.087,0.092)
Type of pain experienced	Tender breasts	0.082 (0.080,0.084)
Emotional state	Sensitive emotion	0.115 (0.112,0.118)
Emotional state	Happy	0.108 (0.105,0.111)
Emotional state	PMS	0.086 (0.083,0.089)
Emotional state	Sad	0.076 (0.073,0.079)

Kolmogorov–Smirnov test results for symptom tracking patterns that are significantly different (at a $p = 0.000001$ level) between users in the consistently not highly variable and consistently highly variable groups.

Overall, our results align with conclusions from these studies in that the cycle lengths have slightly higher values (median of 29 in our dataset) and wider ranges than what was previously commonly believed. While our study population demographics may differ slightly from other studies, we believe these still provide a reasonable basis for comparison. We show comparative summary statistics in the Supplementary Information, demonstrating the consistency of our cycle and period characteristics: (i) the average number of cycles tracked per user in this dataset (12.9) is bigger than in ref. [61] (8.6), while it matches those (12.8) of ref. [62]; (ii) the cycle length statistics are all similar: mean of 29.7 in this work, 29.3 in ref. [61]; median of 29 in this work, 28 in ref. [62]. Interestingly, both this work and ref. [61] report an overall variability of cycle length of around 5 days, and both this work and ref. [62] acknowledge the presence of "a heavy tail on longer cycles." The period length averages of this work and ref. [61] are in agreement as well (4.08 ± 1.76 vs. 4.0 ± 1.50, respectively). Besides, these high-level cycle statistics align well with the results of previous clinical studies[18,22,27].

Examining cycle length is often insufficient for capturing all fluctuations in menstrual patterns—studies regarding menstrual variability showcase that although average cycle length is associated with cycle length consistency, women still experience significant variability in cycle lengths, regardless of their average cycle length[27]. In this work, we address this limitation by utilizing our proposed metric, median CLD, to characterize menstrual cycle variability. Separating users according to their median CLD yields two distinct groups of users with statistically significant differences in cycle length, cycle length variations, and symptom tracking behaviors. We are unaware of any single figure of merit which so helpfully separates users into distinct segments. Clue uses the International Federation of Gynecology and Obstetrics (FIGO) definitions for clinically irregular cycles in the app[67], but has not found connections with differences in tracking.

Table 5. Pain and period tracking odds ratios for low extreme end of the proportion range.

Category	Symptom	Consistently highly variable group's likelihood for $\lambda_s < 0.05$ (95% CI)	Consistently not highly variable group's likelihood for $\lambda_s < 0.05$ (95% CI)	Odds ratio (95% CI) for $\lambda_s < 0.05$
Period flow	Heavy	0.170 (0.169,0.170)	0.098 (0.096,0.100)	1.734 (1.703,1.766)
Period flow	Spotting	0.314 (0.313,0.315)	0.239 (0.237,0.241)	1.314 (1.300,1.328)
Type of pain experienced	Headache	0.326 (0.325,0.327)	0.269 (0.266,0.272)	1.212 (1.199,1.225)
Type of pain experienced	Tender breasts	0.366 (0.365,0.367)	0.320 (0.317,0.322)	1.145 (1.134,1.156)

Likelihood of low λ_s per group, with the associated odds ratio (and 95% confidence intervals). The probability of not tracking 'heavy period' for users in the consistently highly variable group is 0.17 and 0.098 in the other, with an odds ratio of 1.734: the consistently highly variable group is more likely not to track 'heavy period'.

Table 6. Pain and period tracking odds ratios for high extreme end of the proportion range.

Category	Symptom	Consistently highly variable group's likelihood for $\lambda_s > 0.95$	Consistently not highly variable group's likelihood for $\lambda_s > 0.95$	Odds ratio (95% CI) for $\lambda_s > 0.95$
Period flow	Heavy	0.078 (0.077,0.079)	0.096 (0.094,0.097)	0.817 (0.802,0.833)
Period flow	Spotting	0.067 (0.066,0.067)	0.039 (0.037,0.040)	1.729 (1.679,1.782)
Type of pain experienced	Tender breasts	0.193 (0.192,0.194)	0.113 (0.111,0.115)	1.715 (1.684,1.746)
Type of pain experienced	Headache	0.218 (0.217,0.219)	0.131 (0.129,0.133)	1.663 (1.636,1.691)

Likelihood of high λ_s per group, with the associated odds ratio (and 95% confidence intervals). The probability of consistently tracking 'tender breast' pain for users in the consistently highly variable group is 0.193 and 0.113 in the other, with an odds ratio of 1.715: the consistently highly variable group is more likely to regularly track 'tender breast' pain.

While there has been ample work on hormone-level characterizations of the menstrual cycle[68–71], studies of the relationship between menstrual patterns and symptomatic variables are limited—recent work has explored this association using self-tracked data, but over a limited set of symptoms[72] and without discriminating over age or birth control usage[60]. A method for estimating ovulation timing based on Fertility Awareness Method observations (i.e., basal body temperature (BBT), cervical mucus, cervix position, and vaginal sensation) has been presented[62], but such data are inaccessible for this study due to the European Union's General Data Protection Regulation and other data-privacy concerns (sensitive fields such as appointments, ovulation and pregnancy tests, and BBT were not available in Clue's dataset). Nonetheless, such studies showcase the potential large-scale self-tracked data offer in exploring questions relating to menstruation.

This work further demonstrates that mobile self-tracked data provide an accessible option for clinicians and researchers investigating changes of a variable of interest across the menstrual cycle. Our dataset allows us to explore symptoms of interest like pain, types of bleeding, and emotions explicitly, and we are able to connect variability in cycle lengths to patterns in self-reported symptom tracking. In contrast to existing work, our methods allow us to comment quantitatively and qualitatively on the menstrual experience over a broad set of symptoms. While cycle length has been proposed as a biomarker of menstrual health (e.g., very long and very short cycles are associated with a higher risk of infertility), this work suggests that cycle variability may also be a useful biomarker.

We propose a definition of menstrual cycle variability and find that users in high and low variability groups showcase both statistically significant differences in their cycle statistics, as well as in their symptom tracking patterns. We argue that the discovery of such distinct forms of symptom expression allows for phenotype identification and the investigation of clinical associations. In particular, of the symptoms that show statistically significant association with timing data (as measured by median CLD), some of the most arguably unambiguous ones like period flow and pain are also the diagnostic symptoms frequently appearing in the assessment of menstrual health conditions like endometriosis and polycystic ovary syndrome (PCOS). Thus, cycle variability and these high-signal self-tracked symptom patterns can be potentially useful either for predicting each other (e.g., predicting cycle variability from symptoms) or health consequences (e.g., PCOS), insights that are useful to both clinicians and users.

Our perspective on how users experience their menstruation enables development of data-driven models to predict multiple aspects of the menstrual cycle based on self-tracked history, ranging from modeling time to the next cycle, to forecasting the occurrence of specific symptoms a user might report, to detecting underlying medical conditions. Equipped with the results of this work at the cycle level, future work will consist of identifying further differences at a finer grain, namely across the different menstrual phases.

Despite the strength of these results, there are several mitigating factors to bear in mind. We acknowledge that self-tracked data may be unreliable for several reasons, such as inconsistent user engagement or ambiguous symptomatic language. For example, there is potential overlap between similar-sounding symptoms, e.g., some users may track 'low energy,' whereas others may track 'exhausted.' Users can also engage inconsistently by tracking an unequal number of cycles or forgetting to track their period. We successfully ameliorate the latter issue by excluding unexpectedly long cycles utilizing our proposed procedure. For the former issue, we observe that the consistently highly variable group tracked a lower number of cycles on average (see Table 2). However, we note that the number of users who only tracked two cycles (after our preprocessing steps) is small across the entire cohort, representing 2.62% and 0.57% of the consistently highly and not highly variable groups, respectively.

In addition to the risk of inconsistent user engagement, inherent in the nature of self-tracked data is the challenge of disentangling user behavior from true physiological experiences. The design and selection of symptoms for the Clue app was based both on the scientific literature around the menstrual experience and research on which categories users deemed important. As such, in order to encompass a wide range of relevant menstrual and health experiences, the available tracking categories are broad and treated with equal importance. However, we acknowledge that since the symptoms in the app are not based on validated scales and are not designed for diagnosis of specific conditions, they are most likely not granular (or targeted) enough to make definite claims about specific conditions. Furthermore, while there are infotexts in the Clue app that explain each tracking category, self-reported data are influenced by individual user interpretation, and by how users use the app to meet their own needs; we cannot guarantee that each category holds the same meaning for each user.

In this paper, we take symptom tracking behavior to be a proxy for true physiological behavior. However, we are cognizant of the fact that these are not necessarily equivalent. Note that it is very difficult to know what the true physiological experience is in any circumstance: e.g., the experience of menopause varies greatly by culture[73]. With self-tracked data and without access to ground truth, it is complicated (if not impossible) to truthfully distinguish the experienced symptoms from the tracked ones, due to the presence of engagement artifacts and other unforeseen factors. As such, we have taken steps to reduce tracking artifacts with preprocessing techniques, but recognize that limitations remain. Nonetheless, it remains useful to examine these datasets to better understand not only women's menstrual experiences at scale, but also how to improve self-tracking technologies to enable clearer, more interpretable datasets in the future.

Overall, large-scale self-tracked mobile-health data allow us to quantitatively explore the question of characterizing menstrual behavior. Our findings reinforce the claim that menstruation is characterized by variability rather than by regularity[17,18,22,27,29]. We find variation in cycle length statistics as well as in self-reported symptoms, showcasing the spectrum of how women experience their menstruation. We reveal statistically significant relationships between the variability of cycle length and self-reported qualitative symptoms. The identified set of symptoms that show association with timing data (e.g., period flow and pain) are the diagnostic symptoms frequently leveraged for diagnosis of health-relevant conditions, such as endometriosis and PCOS, insights that are useful to both clinicians and users. More broadly, we also develop a methodology for identifying artifacts in self-tracked data, which can be extended to other self-reported menstrual tracking datasets. This work not only statistically verifies the variation of menstrual experience, but also presents promising opportunities for future statistical modeling, prediction, and the potential to inform diagnosis of menstrual-related disorders.

METHODS

Data overview

We leverage a deidentified version of the Clue data warehouse, a dataset of 117,014,597 self-tracking events for 378,694 users in our cohort of interest. Clue app users input overall personal information at sign up, such as age and hormonal birth control (HBC) type. The dataset contains information from 2015 to 2018 for users worldwide, covering countries within North and South America, Europe, Asia, and Africa (see Supplementary Information for a detailed count of cohort users per country). Users can self-track symptoms over time across the 20 available categories (see Table 3 for symptom list) and can preselect which

categories they wish to track when they sign up—all users do not track all categories.

Clue app users track an event by selecting a category and then choosing an associated symptom. Each row in the primitive dataset represents a tracked event e, with relevant fields being (i) the user u that tracked the event e_u, (ii) the reported symptom s in that event $e_u = s$, and (iii) the user-specific cycle c_e in which the event takes place. A menstrual cycle is defined as the span of days from the first day of a period through to and including the day before the first day of the next period[29]. A period consists of sequential days of bleeding (greater than spotting and within 10 days after the first greater-than-spotting bleeding event) unbroken by no more than one day on which only spotting or no bleeding occurred. Note that Clue considers a menses duration longer than 10 days as an outlier, as it would exceed mean period length plus 3 standard deviations for any studied population[29]. In addition, a user has the opportunity to specify whether a cycle should be excluded from their Clue history—for instance, if the user feels that the cycle is not representative of their typical menstrual behavior due to a medical procedure or changes in birth control.

Cohort definition

A cohort of users and cycles was selected for this analysis, based on factors, including age, HBC usage, cycle length, and engagement patterns. Recall that we restricted our data to users aged between 21 and 33 years, since menstrual cycles are relatively less variable in length and more likely to be ovulatory during this age range[18,22,29,65,66]. At younger ages, the reproductive axis (the hypothalamic–pituitary–ovarian axis) in some women, especially those who experienced a later-than-average age at menarche, may not be fully matured. At older ages, some women may be experiencing premature menopause. Restricting our sample to this age group substantially reduces the influence of confounders like undetected heterogeneity on our analyses. Per-age details like cycle and period length statistics are provided in the Supplementary Information.

Since HBC and copper IUDs have been shown to impact cycle length and other aspects of menstruation, we consider natural menstrual cycles only. Therefore, we ignore cycles from users who reported some form of HBC (patch, pill, injection, ring, and implant) or IUD (we excluded all cycles with evidence of IUD use, as there is no explicit distinction between hormonal and copper IUD usage in the dataset). This step removes about 45% of the cycle data, but is crucial to studying menstruation in a standardized way across users, else it would be unclear whether an exhibited menstrual behavior was due to physiology or the effect of birth control. We exclude cycles that a user deems to be anomalous to avoid potential artifacts in cycle patterns. In addition, we eliminate cycles greater than 90 days long, as well as users who have only tracked two cycles, to rule out cases that we argue indicate lack of engagement or noncontinuous use of the app. Finally, we exclude cycles where we believe the user forgot to track their period, hence resulting in an artificially long cycle length; we explain this procedure below. The effect of these filtering steps on the dataset is outlined in Fig. 6, with the final step indicating the removal of aforementioned suspected artificially long cycles. In total, the proposed data-filtering steps reduced the size of the cycle dataset by about 49%. However, the resulting age-specific, natural cycle-only user cohort and the corresponding dataset with potential artifacts removed enables us to study our research questions in a less noisy setting.

Ethics

The research presented here was exempt from Columbia University IRB approval, in accordance with 45CFR46.101(b), as all data are deidentified, and no participant risks are associated with taking part in the study. Participants do not receive direct benefit from this study, but their participation contributes to the general knowledge of menstrual cycles and their symptoms.

Characterizing longitudinal menstrual tracking via cycle length difference

There are many useful ways to characterize menstrual cycles, each of which offers its own advantages and disadvantages. For instance, cycle length provides insight into the length of time between periods and has been widely documented to vary across women[22–28], but is insufficient for understanding menstrual cycle length volatility, as it fails to characterize variability from one cycle to the next.

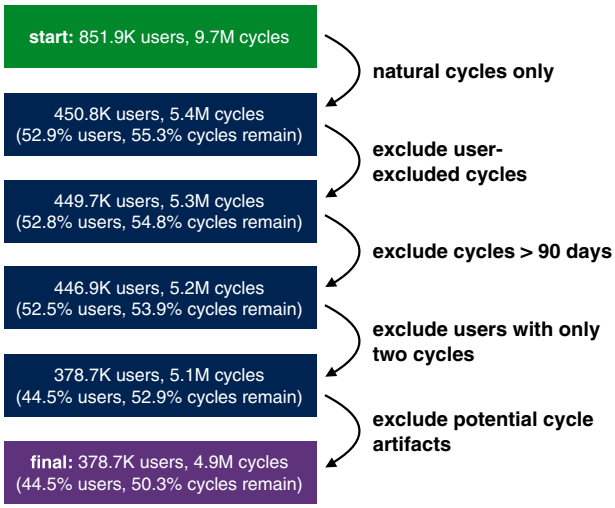

Fig. 6 Filtering process for computing user and cycle cohort. Step-by-step filtering process for computing the final user and cycle cohort. The percentage of users and cycles removed at each step is computed out of the initial numbers. Note that we only include users aged between 21 and 33 years, since women exhibit more stable menstrual behavior in their 'middle life' phase[18,22,29,65,66].

We propose computing cycle length differences (CLDs), which we define as the absolute differences in subsequent cycle lengths. CLDs represent a user's longitudinal cycle tracking history by quantifying their between-cycle volatility. This metric captures menstrual patterns, regardless of specific cycle lengths, allowing us to measure fluctuation over time and identify those who are consistently highly variable. This metric does not capture some other menstrual phenomena, such as cycle lengths growing at a constant rate—that is, if a cycle length grew consistently by two days with each cycle, the CLDs would all be equal to two, but there would be a large difference between the shortest and longest cycle lengths. However, CLDs and related metrics of median and maximum CLD do allow us to characterize users on the extreme ends of the between-cycle variability spectrum and identify potential cycle tracking artifacts, as described in the following sections. Figure 7 outlines the computation of CLDs and related statistics.

Quantifying engagement with cycle tracking

We propose a methodology for identifying cycles associated with lack of app engagement, specifically where users forgot to track their period, since this may inflate the corresponding computed cycle length. Our procedure allows us to distinguish physiological behavior (i.e., true 'long' cycle lengths) from tracking artifacts (i.e., artificially inflated cycle lengths), which allows us to more reliably utilize symptom tracking behavior as a proxy for true physiological behavior.

Figure 8 showcases how maximum CLD impacts our overall picture of user engagement. In particular, the multimodal nature of the histogram of maximum CLD (in blue) indicates that there may be cycles where users forgot to track their period, resulting in an overestimation of cycle length. Note the peaks around 30 and 60 days, which may correspond to users forgetting to track one or two periods, respectively. That is, consider a user who exhibits a perfectly uniform cycle length of 30 and hence always has a CLD of 0. If this user were to forget to track a period once in their history, then the app would record that they have a cycle length of 60 and a maximum CLD of 30—such a user would fall into the first peak of the histogram. The discrepancy between regular patterns (via the median CLD) and extreme events (maximum CLD) is further illustrated in Fig. 9.

We identify cycles that are 'atypically long' compared with the 'typical' cycle length for each user by examining the difference between each CLD and the median CLD of that user. An illustrative example is provided in the top panel of Fig. 7, where the third cycle appears to be 'atypically long.' Specifically, we flag cycles per user where the corresponding CLD exceeds the user's median CLD by at least 10 days as 'atypically long' (the longer of the two cycles corresponding to the CLD is flagged). This cutoff is based on an attempt to find in the data, rather than posit a priori, a feature that would distinguish 'typical' from 'extreme' (i.e., abnormally long) reported

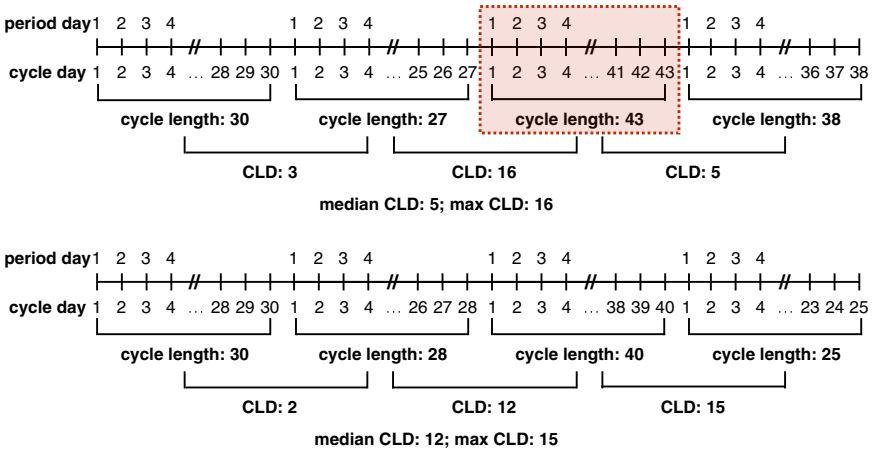

Fig. 7 Identifying cycle tracking artifacts and characterizing user regularity. We provide illustrative examples of identifying a cycle tracking artifact (top) and characterizing a user's regularity (bottom) based on CLD statistics. In each example, we display a user's cycle history with a total of four cycles. Cycle length is computed as the length of time between the first day of a period and the first day of the next period, and CLD is computed as the absolute difference between subsequent cycle lengths (i.e., if a user has n cycles tracked, they will have $n-1$ CLD values). Period length is computed by counting the number of sequential days on which there is menstrual bleeding greater than spotting ('light,' 'medium,' or 'heavy'). Two such sequences are considered one period if separated by no more than one day of non-bleeding/spotting. In the top example, the user's second CLD exceeds their median by at least 10, and thus we identify the corresponding 'artifically long' cycle in red—this cycle will be excluded from our analysis. In the bottom example, the user's median CLD is at least 9, and thus they will be classified as a consistently highly variable user.

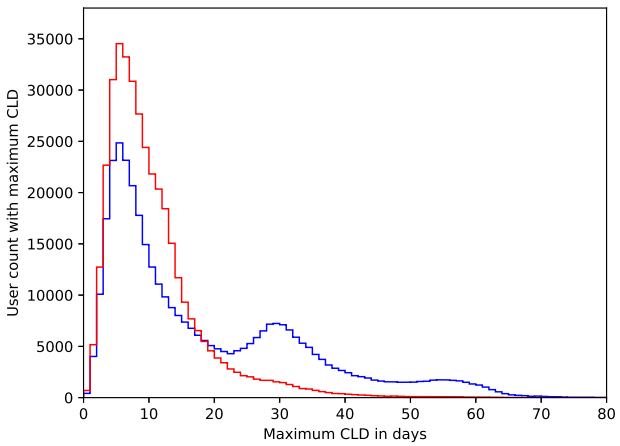

Fig. 8 Histogram of maximum CLD before and after excluding cycle artifacts. For each user, we compute the maximum CLD and plot a histogram before (blue) and after (red) excluding cycles without user engagement (i.e., cycles that are potential artifacts). We see that the multimodal behavior (peaks at around 30 and 60 days) is largely dampened upon removing these cycles. In addition, the fat right-hand tail in the red curve implies that we preserve the natural variation in cycle length—we are not simply removing long cycles.

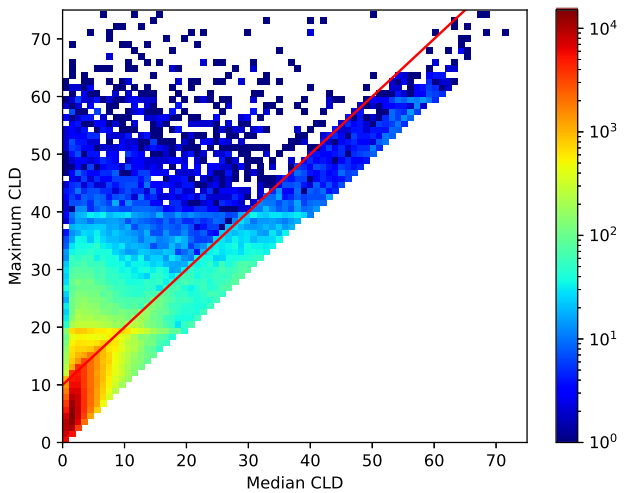

Fig. 9 Median CLD versus maximum CLD two-dimensional histogram. We plot a two-dimensional histogram of users' median CLD versus maximum CLD in logarithmic space, as well as the line where maximum CLD is equal to median CLD plus 10 in red. We can see that the line separates out a highly concentrated region of users, as well as a more scattered region of users. Specifically, the majority of the mass falls under this line, as showcased by the concentrated red color in the lower left-hand corner of the plot, and a diagonal band extending upward, while the region above the line is more spread out. Thus, we examine the cycles that fall above the line as possible cycle tracking artifacts.

cycles. To do so, we plot the two-dimensional histogram (see Fig. 9) of maximum CLD against median CLD; this is a histogram in which every example is a user. We observe a clear visual feature: a band of users we consider 'typical,' for whom maximum CLD was within 10 days of the median CLD, and a scatter of other users for whom their maximum CLD could be far larger. To capture this visually striking feature (see the diagonal red line along maximum CLD equal to 10 more than median CLD in Fig. 9), we defined extreme events as those at least 10 days above the median CLD. We consider the cycles flagged as 'atypically long' to be the result of cycle engagement artifacts and exclude them from our analysis.

As shown in Fig. 8 (red line), the multimodal shape is largely removed after eliminating the 'atypically long' cycles. We find that 42% of the cohort has at least one such cycle, and for these users, we exclude a small number (1.59) of cycles per user on average. This indicates that our method is stringent enough to identify artificially long cycles, but conservative enough to preserve the heterogeneity of the data.

We further validate our method by examining tracking activity during the interval where a user is expected to track their period for each of these excluded cycles and find that in 89.18% of such cases, there is no evidence of bleeding-related events during this interval, i.e., the user likely did not engage in period tracking. Note that we define this interval as the user's last reported period day plus their median cycle length, plus or minus their median period length. In the remaining 10.82% of excluded cycles, it is unclear whether the bleeding-related events tracked during this interval represent a period or some other non-period bleeding. Note that by our definition of a period, a single bleeding event is not synonymous with

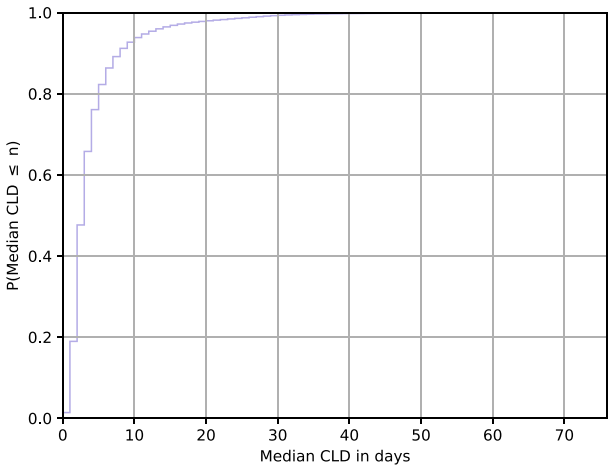

Fig. 10 Cumulative distribution of median CLD. Looking at the cumulative distribution of median CLD, we see that the curve flattens out significantly around the 'elbow' at 9 days; thus, we choose greater than 9 days as our cutoff for our definition of consistently highly variable.

period. As a conservative measure and to maintain consistency of our definitions for period and artificially long cycles, we exclude those cycles from our analysis. This ensures a coherent data preprocessing pipeline and impacts the results minimally (these excluded cycles with some bleeding-related events amount to only 0.56% of all cycles). Quantifying inconsistent tracking engagement allows us to ameliorate its impact on subsequent analyses.

Characterizing users according to cycle length variability

We acknowledge that there is a wide spectrum of variability in women's menstrual health experiences, and we wish to examine those who fall at opposite ends of the variability spectrum on the basis of their cycle pattern consistency. We choose the median CLD metric for our analysis, as it is robust to outliers (the mean would be more susceptible to being skewed by rare events). Upon examining the cumulative distribution of this metric across users in Fig. 10, we consider a median CLD of greater than 9 days to be an appropriate stringent cutoff for identifying consistently highly variable menstrual patterns. This choice aligns with previous work on menstrual pattern analysis: cycle length variability studies in Guatemalan, Bolivian, Indian, US, and European women noted differences in the maximum and minimum cycle length ranging from 6 to 14 days[23–28]. Our proposed cutoff separates users into two distinct groups of menstrual patterns: the vast majority (92.32%) of the population falls to the left of this threshold, and thus the consistently highly variable group (the remaining 7.68%) represent those whose variability is extreme. As discussed in the Results section, we observe that the highly variable group experiences more drastic fluctuations in cycle length. We confirm that the cycle length distributions differ significantly between the two groups using a two-sample Kolmogorov–Smirnov[74] test.

Quantifying symptom tracking behavior across user groups

We focus on symptom tracking behavior at the cycle level, evaluating how often throughout their longitudinal tracking users track each symptom, regardless of when within the cycle the tracking occurred (i.e., ignoring at which phase or day of the cycle the symptom occurred). Note that because cycle length varies both within each user's longitudinal tracking and across women, the number of tracking events per cycle would be skewed by cycle length. To combat this issue, we measure the per-user proportion of cycles where a symptom has been tracked.

We also want our metric to capture symptom tracking behavior for cycles where users were interested in tracking the associated category (recall how users do not have to track all categories). Specifically, our analysis focuses on how often a user u has a symptom s tracking event $e_u = s$ per cycle n, given that they have tracked symptoms within the associated category C at least once across all their cycles N_u. We refer to this metric as the 'proportion of cycles with symptom out of cycles with category,' mathematically denoted as

$$\lambda_{us} = \frac{\sum_{n=1}^{N_u} \mathbb{1}[\exists e_u = s]}{\sum_{n=1}^{N_u} \mathbb{1}[\exists e_u \in C]}. \quad (1)$$

That is, to account for whether a user is actually interested in the symptom at hand, we compute the proportion of cycles with a symptom being tracked out of the number of cycles where the user has tracked the category related to that symptom. For example, consider a user who tracked 8 cycles; out of these, they tracked any of the symptoms within the pain category for 5 cycles. Specifically, they tracked the symptom 'headache' for a single cycle and 'tender breasts' for 4 cycles. Our metric λ_s captures the tracking regularity of a given symptom across a user's cycles. When applied to the example user, 20% of cycles with pain have 'headache' tracked, while 80% of the same cycles have reports of 'tender breasts.' In essence, λ_{us} is the conditional probability that user u tracks the specific symptom s, given that they have tracked any symptom from the symptom's corresponding category. Our metric measures per-cycle symptom tracking frequency and is robust to (i) different cycle lengths and number of cycles (as it is normalized with respect to each user's number of cycles), (ii) different user app interests (as it is contingent on whether the user has shown interest in tracking a given category at least once), and (iii) different app usage behaviors (as it does not depend of how many times within a cycle the symptom is tracked).

We study the cumulative distributions of λ_s per group (i.e., λ_{us} for all users u within each variability group), as well as how such densities are different on their support boundaries across groups. Since we lack a mechanistic model of what distribution the data might follow and wish to use a test meaningful for any distribution, we utilize a nonparametric test suitable for any ordinal (as opposed to, e.g., binary or categorical data): the Kolmogorov–Smirnov (KS) test[74]. This test of the comparative equality of one-dimensional probability distributions arising from two samples allows us to quantify statistical differences in symptom tracking behavior between groups. Specifically, we compare the distributions of the proportion of cycles with a symptom tracked (out of cycles with its corresponding category) for the consistently not highly variable and consistently highly variable user groups. The KS statistic quantifies the distance between the empirical cumulative distributions of two samples, and the associated test is sensitive to differences in both location and shape of said distributions, allowing us to characterize *where* and *how much* the symptom tracking patterns (as measured by the proposed λ_s metric) differ between groups.

In the two-sample case, the null distribution of the KS statistic is calculated under the null hypothesis that the samples are drawn from the same distribution, where the distribution considered under the null hypothesis is an unrestricted continuous distribution (i.e., no distributional assumption is made on the symptom tracking patterns). The KS statistic depends on the number of data points within each of the populations (i.e., the number of observations that we have for each group when computing their per-symptom empirical cumulative density function). The null hypothesis is rejected at level α if

$$D_{n,m} > \sqrt{-\frac{1}{2}\ln\alpha \cdot \frac{n+m}{nm}}, \quad (2)$$

where n and m are the sizes of the first and second data samples, respectively, and $D_{n,m}$, the computed two-sample KS statistic. The p-values reported by the KS test consider observed sample sizes, accounting for the impact of whether certain symptoms are more or less frequently logged in each group.

In order to explore *how* the empirical distributions differ, we study their support boundaries, i.e., $p(\lambda_s > 0.95)$ and $p(\lambda_s < 0.05)$. These represent how likely users in each group are to either consistently track a symptom throughout their cycle history (i.e., in almost every cycle where they track the category), or to not track it at all (i.e., in very few of their cycles where they track the category). We compute the odds ratio of these values (on either the high or low extreme end of the proportion range) for the consistently highly variable group to the consistently not highly variable group. If we have an odds ratio greater than 1 for the high extreme end of the metric range for a symptom, this would indicate that the consistently highly variable group is more likely to report a very high proportion of cycles with that symptom. On the other hand, an odds ratio greater than 1 for the low extreme end of the proportion range (i.e., the proportion of cycles with a symptom tracked is close to zero) indicates that the

consistently highly variable group is more likely *not* to report such a symptom.

When possible, 95% confidence intervals have been added to reported KS values using bootstrap analysis. To do so, we draw 100,000 random samples—resampled with replacement—from each variability group and report the estimated mean KS statistic values and their 2.5 and 97.5 percentiles.

Reporting summary

Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The database that supports the findings of this study was made available by Clue by BioWink GmbH. While it is deidentified, it cannot be made directly available to the reader. Researchers interested in gaining access to the data can contact Clue by BioWink GmbH, and establish a data use agreement with them.

CODE AVAILABILITY

Our code has been developed using open-source tools in Python with common statistical libraries (e.g., Pandas and SciPy). The code required for data preprocessing and producing results is available in the public GitHub repository https://github.com/iurteaga/menstrual_cycle_analysis.

Received: 11 October 2019; Accepted: 23 March 2020;
Published online: 26 May 2020

REFERENCES

1. Popat, V. B., Prodanov, T., Calis, K. A. & Nelson, L. M. The menstrual cycle: a biological marker of general health in adolescents. *Ann. N. Y. Acad. Sci.* **1135**, 43–51 (2008).
2. Bedford, J. L., Prior, J. C. & Barr, S. I. A prospective exploration of cognitive dietary restraint, subclinical ovulatory disturbances, cortisol, and change in bone density over two years in healthy young women. *J. Clin. Endocrinol. Metab.* **95**, 3291–3299 (2010).
3. Zittermann, A. et al. Physiologic fluctuations of serum estradiol levels influence biochemical markers of bone resorption in young women. *J. Clin. Endocrinol. Metab.* **85**, 95–101 (2000).
4. Solomon, C. G. et al. Menstrual cycle irregularity and risk for future cardiovascular disease. *J. Clin. Endocrinol. Metab.* **87**, 2013–2017 (2002).
5. Carmina, E. & Lobo, R. A. Polycystic ovary syndrome (PCOS): arguably the most common endocrinopathy is associated with significant morbidity in women. *J. Clin. Endocrinol. Metab.* **84**, 1897–1899 (1999).
6. Giudice, L. C. Endometriosis. *N. Engl. J. Med.* **362**, 2389–2398 (2010).
7. Barbosa, C. P., Souza, A. B. D., Bianco, B. & Christofolini, D. The effect of hormones on endometriosis development. *Minerva Gynecol.* **63**, 375–386 (2011).
8. Shuster, L. T., Rhodes, D. J., Gostout, B. S., Grossardt, B. R. & Rocca, W. A. Premature menopause or early menopause: long-term health consequences. *Maturitas* **65**, 161–166 (2010).
9. Mahoney, M. M. Shift work, jet lag, and female reproduction. *Int. J. Endocrinol.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2834958/ (2010).
10. Jordan, J., Craig, K., Clifton, D. K. & Soules, M. R. Luteal phase defect: the sensitivity and specificity of diagnostic methods in common clinical use. *Fertil. Steril.* **62**, 54–62 (1994).
11. Crawford, N. M., Pritchard, D. A., Herring, A. H. & Steiner, A. Z. Prospective evaluation of luteal phase length and natural fertility. *Fertil. Steril.* **107**, 749–755 (2017).
12. Prior, J. C. Perimenopause The complex endocrinology of the menopausal transition. *Endocr. Rev.* **19**, 397–428 (1998).
13. Landgren, B.-M. et al. Menopause transition: annual changes in serum hormonal patterns over the menstrual cycle in women during a nine-year period prior to menopause. *J. Clin. Endocrinol. Metab.* **89**, 2763–2769 (2004).
14. Prior, J. C. & Hitchcock, C. L. The endocrinology of perimenopause: need for a paradigm shift. *Front. Biosci.* **3**, 474–486 (2011).
15. Prior, J. C., Vigna, Y., Sciarretta, D., Alojado, N. & Schulzer, M. Conditioning exercise decreases premenstrual symptoms: a prospective, controlled 6-month trial. *Fertil. Steril.* **47**, 402–408 (1987).
16. Barr, S., Janelle, K. & Prior, J. C. Energy intakes are higher during the luteal phase of ovulatory menstrual cycles. *Am. J. Clin. Nutr.* **61**, 39–43 (1995).
17. Arey, L. B. The degree of normal menstrual irregularity. *Am. J. Obstet. Gynecol.* **37**, 12–29 (1939).
18. Treloar, A. E., Boynton, R. E., Behn, B. G. & Brown, B. W. Variation of the human menstrual cycle through reproductive life. *Int. J. Fertil.* **12**, 77–126 (1967).
19. Snowden, R., Christian, B. & World Health Organization. *Patterns and perceptions of menstruation: a World Health Organization international collaborative study in Egypt, India, Indonesia, Jamaica, Mexico, Pakistan, Philippines, Republic of Korea, United Kingdom and Yugoslavia* (World Health Organization by Croom Helm, 1983).
20. Ferrell, R. J. et al. The length of perimenopausal menstrual cycles increases later and to a greater degree than previously reported. *Fertil. Steril.* **86**, 619–624 (2006).
21. Gorrindo, T. et al. Lifelong menstrual histories are typically erratic and trending: a taxonomy. *Menopause* **14**, 74–88 (2007).
22. Chiazze, L., Brayer, F. T., John, J., Macisco, J., Parker, M. P. & Duffy, B. J. The length and variability of the human menstrual cycle. *J. Am. Med. Assoc.* **203**, 377–380 (1968).
23. Mn¨ster, K., Schmidt, L. & Helm, P. Length and variation in the menstrual cycle—a cross-sectional study from a Danish county. *Br. J. Obstet. Gynaecol.* **99**, 422–429 (1992).
24. Belsey, E. M. & Pinol, A. P. Menstrual bleeding patterns in untreated women. Task force on long-acting systemic agents for fertility regulation. *Contraception* **55**, 57–65 (1997).
25. Burkhart, M. C., de Mazariegos, L., Salazar, S. & Hess, T. Incidence of irregular cycles among Mayan women who reported having regular cycles: implications for fertility awareness methods. *Contraception* **59**, 271–275 (1999).
26. Vitzthum, V. J., Spielvogel, H., Caceres, E. & Gaines, J. Menstrual patterns and fecundity among non-lactating and lactating cycling women in rural highland Bolivia: implications for contraceptive choice. *Contraception* **62**, 181–187 (2000).
27. Creinin, M. D., Keverline, S. & Meyn, L. A. How regular is regular? An analysis of menstrual cycle regularity. *Contraception* **70**, 289–292 (2004).
28. Williams, S. R. Menstrual cycle characteristics and predictability of ovulation of Bhutia women in Sikkim. *India J. physiological Anthropol.* **25**, 85–90 (2006).
29. Vitzthum, V. J. The ecology and evolutionary endocrinology of reproduction in the human female. *Am. J. Phys. Anthropol.* **140**, 95–136 (2009).
30. American College of Obstetricians and Gynecologists. Menstruation in girls and adolescents: using the menstrual cycle as a vital sign. *Obstet. Gynecol.* **126**, 143–6 (2015).
31. Bobel, C. in *The Managed Body: Developing Girls and Menstrual Health in the Global South* (ed. Bobel, C.) 281–321 (Springer International Publishing, Cham, 2019).
32. Lippe Taylor, Inc. Scientific forum addresses menstrual cycle as vital sign. https://www.eurekalert.org/pub_releases/2004-09/lti-sfa092004.php (2004).
33. American Academy of Pediatrics, and American College of Obstetricians and Gynecologists. Menstruation in girls and adolescents: using the menstrual cycle as a vital sign. *Pediatrics* **118**, 2245–2250 (2006).
34. Solomon, C. G. et al. Long or highly irregular menstrual cycles as a marker for risk of type 2 diabetes mellitus. *Jama* **286**, 2421–2426 (2001).
35. Solomon, C. G. et al. Menstrual cycle irregularity and risk for future cardiovascular disease. *J. Clin. Endocrinol. Metab.* **87**, 2013–2017 (2002).
36. Hripcsak, G. et al. Characterizing treatment pathways at scale using the OHDSI network. *Proc. Natl Acad. Sci. USA* **113**, 7329–7336 (2016).
37. Li, I., Dey, A. & Forlizzi, J. A stage-based model of personal informatics systems. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 557–566 (ACM, 2010).
38. Kohane, I. S. Ten things we have to do to achieve precision medicine. *Science* **349**, 37–38 (2015).
39. Fox, S. & Duggan, M. *Tracking for Health*. Tech. Rep. (Pew Research Center, 2013).
40. Krebs, P. & Duncan, D. T. Health app use among US mobile phone owners: a national survey. *JMIR mHealth and uHealth* **3**, e101 (2015).
41. Althoff, T. Population-scale pervasive health. *IEEE Pervasive Comput.* **16**, 75–79 (2017).
42. Althoff, T. et al. Large-scale physical activity data reveal worldwide activity inequality. *Nature* **547**, 336–339 (2017).
43. Chan, Y.-F. Y. et al. The Asthma Mobile Health Study, a large-scale clinical observational study using ResearchKit. *Nat. Biotechnol.* **35**, 354 (2017).
44. Webster, D. E. et al. The Mole Mapper Study, mobile phone skin imaging and melanoma risk data collected using ResearchKit. *Sci. Data* **4**, 170005 (2018).
45. Egger, H. L. et al. Automatic emotion and attention analysis of young children at home: a ResearchKit autism feasibility study. *npj Digital Med.* **1**, 20 (2018).
46. Bot, B. M. et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. *Sci. Data* **3**, 160011 (2016).
47. Dagum, P. Digital biomarkers of cognitive function. *npj Digital Med.* **1**, 10 (2018).
48. Smets, E. et al. Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *npj Digital Med.* **1**, 67 (2018).
49. Byambasuren, O., Sanders, S., Beller, E. & Glasziou, P. Prescribable mHealth apps identified from an overview of systematic reviews. *npj Digital Med.* **1**, 12 (2018).

50. Ata, R. et al. Clinical validation of smartphone-based activity tracking in peripheral artery disease patients. *npj Digital Med.* **1**, 66 (2018).
51. Torous, J. et al. Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia. *npj Digital Med.* **1**, 15 (2018).
52. Urteaga, I., McKillop, M., Lipsky-Gorman, S. & Elhadad, N. Phenotyping endometriosis through mixed membership models of self-tracking data. in *2018 Machine Learning for Healthcare (MLHC)* (2018).
53. Wartella, E., Rideout, V., Montague, H., Beaudoin-Ryan, L. & Lauricella, A. Teens, health and technology: a national survey. *Media Commun.* **4**, 13–23 (2016).
54. Fox, S. & Duggan, M. *Mobile Health 2012*. Tech. Rep. (Pew Research Center, 2012).
55. Clue by BioWink GmbH, Adalbertstraße 7-8, 10999 Berlin, Germany. https://helloclue.com/ (2019).
56. *Dot: A Fertility Tracker App*. https://www.dottheapp.com/ (2019).
57. *Glow: An App for Fertility & Beyond*. https://glowing.com/glow (2019).
58. *Spot On: A Birth Control and Period Tracker App powered by Planned Parenthood*. https://shortyawards.com/9th/spot-on (2019).
59. *Natural Cycles: Digital Birth Control*. https://www.naturalcycles.com (2019).
60. Pierson, E., Althoff, T., Thomas, D., Hillard, P. & Leskovec, J. The menstrual cycle is a primary contributor to cyclic variation in women's mood, behavior, and vital signs. *bioRxiv*. https://doi.org/10.1101/583153 (2019).
61. Bull, J. R. et al. Real-world menstrual cycle characteristics of more than 600,000 menstrual cycles. *npj Digital Med.* **2**, 83 (2019).
62. Symul, L., Wac, K., Hillard, P. & Salathé, M. Assessment of menstrual health status and evolution through mobile apps for fertility awareness. *npj Digital Med.* **2**, 64 (2019).
63. Epstein, D. A. et al. Examining menstrual tracking to inform the design of personal informatics tools. in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 6876–6888 (2017).
64. Moglia, M. L., Nguyen, H. V., Chyjek, K., Chen, K. T. & Castaño, P. M. Evaluation of smartphone menstrual cycle tracking applications using an adapted applications scoring system. *Obstet. Gynecol.* **127**, 1153–1160 (2016).
65. Ferrell, R. J. et al. Monitoring reproductive aging in a 5-year prospective study: aggregate and individual changes in steroid hormones and menstrual cycle lengths with age. *Menopause* **12**, 567–757 (2005).
66. Harlow, S. D. et al. Executive summary of the stages of reproductive aging workshop+ 10: addressing the unfinished agenda of staging reproductive aging. *J. Clin. Endocrinol. Metab.* **97**, 1159–1168 (2012).
67. Druet, A. What is an "irregular" menstrual cycle? https://helloclue.com/articles/cycle-a-z/whatis-an-irregular-menstrual-cycle (2018).
68. Johannisson, E., Landgren, B.-M., Rohr, H. P. & Diczfalusy, E. Endometrial morphology and peripheral hormone levels in women with regular menstrual cycles. *Fertil. Steril.* **48**, 401–408 (1987).
69. Fehring, R. J., Schneider, M. & Raviele, K. M. Variability in the phases of the menstrual cycle. *J. Obstet., Gynecologic, Neonatal Nurs.* **35**, 376–84 (2006).
70. Lenton, E. A., Landgren, B.-M. & Sexton, L. Normal variation in the length of the luteal phase of the menstrual cycle: identification of the short luteal phase. *BJOG: Int. J. Obstet. Gynaecol.* **91**, 685–689 (1984).
71. Lenton, E. A., Landgren, B.-M., Sexton, L. & Harper, R. Normal variation in the length of the follicular phase of the menstrual cycle: effect of chronological age. *BJOG: Int. J. Obstet. Gynaecol.* **91**, 681–684 (1984).
72. Pierson, E., Althoff, T. & Leskovec, J. Modeling individual cyclic variation in human behavior. in *Proceedings of the 2018 World Wide Web Conference, WWW '18*, 107–116 (International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2018).
73. Jones, E. K., Jurgenson, J. R., Katzenellenbogen, J. M. & Thompson, S. C. Menopause and the influence of culture: another gap for Indigenous Australian women? *BMC Womenas Health.* **12**, 43 (2012).
74. Kolmogorov, A. N. Sulla determinazione empirica di una legge di distribuzione. *G. dellalstituto Italianodegli Attuari* **4**, 83–91 (1933).

ACKNOWLEDGEMENTS

The authors are deeply grateful to all Clue users whose deidentified data have been used for this study.

AUTHOR CONTRIBUTIONS

K.L. and I.U. contributed equally to this work. K.L., I.U., C.H.W., and N.E. conceived the proposed research and designed the experiments. K.L. and I.U. processed the dataset, conducted the experiments, and wrote the first draft of the paper. C.H.W., A.D., A.S., V.J.V., and N.E. reviewed and edited it. All authors read and approved the paper.

COMPETING INTERESTS

K.L. is supported by NSF's Graduate Research Fellowship Program Award #1644869. I.U., C.H.W., and N.E. are supported by NSF Award #1344668. K.L., I.U., C.W., and N.E. declare that they have no competing interests. A.D., A.S., and V.J.V. were employed by Clue by BioWink GmbH at the time of this research project.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41746-020-0269-8.

Correspondence and requests for materials should be addressed to N.E.