Beyond the Circadian Rhythm: Variable Cycles of Regularity Found in Long-Term Sleep Tracking

Ji Won Chung
Brown University
Providence, RI, USA
ji_won_chung@brown.edu

Robin Yuan
Brigham and Women's Hospital & Harvard Medical School
Boston, MA, USA
ryuan@bwh.harvard.edu

Kirsi-Marja Zitting
Brigham and Women's Hospital & Harvard Medical School
Boston, MA, USA
kzitting@bwh.harvard.edu

Jiahua Chen
Brown University
Providence, RI, USA
jiahua_chen2@brown.edu

Neil Xu
Brown University
Providence, RI, USA
neil_xu@brown.edu

Nediyana Daskalova
Brown University
Providence, RI, USA
nediyana888@gmail.com

Jeff Huang
Brown University
Providence, RI, USA
jeff_huang@brown.edu

Abstract

Sleep is more than resting eight hours a day—it contextualizes and shapes the routines during the day. Using a large-scale naturalistic dataset of 180,083 people from a popular sleep app, made possible by the widespread adoption of passive tracking, we find that people's lives have distinct natural rhythms that can be automatically inferred from sleep routines. We discover heterogeneous behaviors: the rhythm of sleep is different for each person, as there is a different cadence for each person to achieve consistency. Some are most consistent week-to-week, while others weeks-to-weeks. We investigate changes in overall daily routines and find the interval for each person at which they show the most consistency. Through a series of comparative case analyses, we investigate the implications of designing for the weekly 'norm'. Our tripartite analyses triangulate to one conclusion: we should design for people's natural routines to account for variable cycles of regularity.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; *Empirical studies in ubiquitous and mobile computing*; • **Applied computing** → *Consumer health.*

Keywords

routine management, longitudinal study, behavioral regularity, sleep regularity, naturalistic study, personal informatics, design

ACM Reference Format:
Ji Won Chung, Robin Yuan, Kirsi-Marja Zitting, Jiahua Chen, Neil Xu, Nediyana Daskalova, and Jeff Huang. 2025. Beyond the Circadian Rhythm: Variable Cycles of Regularity Found in Long-Term Sleep Tracking. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3713868

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713868

1 Introduction

Routines are simple, but maintaining them is not. Consequently, people seek help from health applications and fitness trackers [3, 5, 12, 17, 18, 36, 75, 91, 93, 94] to provide them with accountability and insights on how to achieve a routine lifestyle. Yet, technologies for behavioral regularity have known systemic flaws—their efficacy and design are limited by our knowledge of naturalistic, long-term human behavior and the methods used to study it [17, 48, 56, 57].

Much of the research and design on behavioral regularity technologies assume users follow a standard, weekly routine [1–3, 45, 64, 71, 93, 94, 99, 104]. However, routine lengths remain an understudied property of routines, and this assumption is based on user experiences from short-lived, controlled studies, instead of people's true, lived experiences [17, 47, 48, 56].

Without large-scale, longitudinal observations, we cannot observe long-term behavioral patterns, nor answer basic questions—what do people's routines look like outside the lab? What does an average person's routine look like? How prevalent are weekly and non-weekly routines? What kind of different routines exist and how many people do they affect? Simply put, we are unable to evaluate for whom, when, and how our technologies support or fail [56, 57, 73].

To improve the design of health technologies, we conduct an *exploratory* analysis of real, longitudinal sleep routine lengths on 180,083 Sleep as Android app users. Sleep routines offer insights into how people structure their overall day—they are a proxy signal for people's overall routines [6]. If you sleep at the same times, you are awake at the other times and have access to a structured lifestyle. Routine lengths contextualize other behaviors during the day and highlight the diversity of user behaviors. They indicate how long it takes to repeat a behavioral pattern and imply how one's vocation [51, 64] or social obligations [1, 2, 71, 78] dictate their waking hours.

While there are many approaches in investigating routines, we look at sleep because it is a universal behavior experienced by everyone, allowing us to analyze patterns beyond just specific individuals. Thus, we examine longitudinal, sleep routine lengths to understand the following research questions:

RQ1: How do people's routines vary?
- **(a)** Do daily routines change over time? In what ways?
- **(b)** What is the most common routine length?
- **(c)** Are there alternative routine lengths?

RQ2: What lessons can we learn from longitudinal sleep records, and how can these insights help us reflect on our current technology designs and identify any necessary changes?

To answer these questions, we collaborate with sleep medicine experts and design a process to find a set of good long-term recorders from a naturalistic dataset. By being transparent in our system design, which is often obscured in literature [4, 85], we provide insights into the practical limitations and implications of the system, offering guidance for future system designers **(RQ2)**. We select users with continuous streaks of records to refrain from interpolating missing records. We then conduct stationarity tests on each user's daily sleep regularity scores to detect overall changes in their routine **(RQ1a)**; this process also ensures that subsequent analyses are not based on random behavioral fluctuations. Next, instead of assuming that a user follows a traditional, weekly routine, we uncover each user's natural routine length by applying the Fast Fourier Transform on a series of their daily sleep patterns **(RQ1b)**. Lastly, our comparative, case analyses of non-weekly routines demonstrate the limitations of weekly designs **(RQ1c, RQ2)**.

Our investigation of routine lengths suggests that alternative, non-weekly cycles are more common than we think **(RQ2)**. We find that more than a half of user routines are better described by alternative, non-weekly rhythms, which, at times, span more than 8 weeks long. Furthermore, some individuals cannot be described by a singular type of rhythm—rhythms change over time, some are polyrhythmic, or best described by a mix of rhythms, and some have no rhythms at all.

As we will show by example and through our discussions with sleep medicine experts, considering a user's natural routine length leads to tailored recommendations that are otherwise masked by standard, weekly designs. Thus, this paper's main contribution is the discussion on how incorporating this simple design change—accounting for natural routine lengths—can create more transparent and equitable behavioral regularity and scheduling technologies.

2 Related Work

2.1 Routines in Technological Design

Routines capture the structure of an individual's day [6, 25, 28, 29] and are associated with different life styles, vocations, and stages of life [14, 16, 17, 102]. Polyphasic users have distinct patterns in the day because they 'split all sleep into a series of naps' [49]. Shift workers show 'routine variations' [6], or periodically different schedules for their shifts [49, 64]. Some menstruators [33, 65, 95] experience periodic behavioral changes with their cycle. The elderly commonly experience gradual disruptions in their day-to-day physical mobility [89].

By understanding characterizations of a user's 'normal' routine, we can identify *deviations* [6, 25], which help us design user-centric scheduling systems and timely digital interventions. In 2003, Begole et al. prototyped a new communication application to facilitate work coordination by investigating people's messaging rhythms and deviations, such as when they were away from work [9]. MAHI found technologies can assist newly diagnosed diabetic users by focusing on 'breakdowns', or moments when users realize stark changes in their daily routines [70]. Similarly, changes in routine internet usage were associated with depression in users who recently moved [92], and changes in sleep routines and online activities indicated the need for postpartum care [26].

An overlooked characteristic of routines is their lengths. Weekly and monthly designs are the common choice to visualize user insights in many consumer technologies [3, 36, 45, 93, 94, 99, 104]. These top-down designs have been criticized for reinforcing 'normative' behaviors that overlook individual nuances [8, 21, 25, 50, 55, 86] and providing limited support to users with 'abnormal' or irregular sleep patterns [49–51, 64, 85]. Many menstruators find it 'difficult to understand changes in the body' [95] if their biological rhythms are outside the normative 'monthly' cycle because technologies aren't designed to support such 'non-normative', irregular rhythms [33, 82, 95]. Users whose routines are interdependent on those of others often face challenges because 'normative' schedules do not align with their natural rhythms [25]. In other words, weekly designs lack equity [101] and prevent users from receiving 'actionable feedback' [85].

Previous works partially address this gap by designing technologies that cater to diverse users and their unique, 'natural' routines. Daskalova et al.'s SleepBandits [24] and SleepCoacher [23] reduce reliance on 'normative' designs by allowing users to gain self-directed insights via self-experimentation. Karlgren et al. [50, 51] expand the design space to include natural, sleep and body rhythms. SleepGuru proposes a scheduling algorithm that takes into consideration the real-life constraints of irregular sleepers like researchers and airline employees [64]. Abdullah et al. [1, 2] and Murnane et al. [78] propose new computational models that account for people's biological daily, circadian rhythms. Davidoff et al. suggest alternative models for group-based routine coordination to accommodate families with intertwined routines [13, 25, 91].

However, there is a limit to understanding and designing for natural and non-normative routines with controlled, short-term studies. Many natural behavioral changes and routines only become apparent with long-term observations [37, 40, 57, 88, 96, 107]. Thanks to advancements in unobtrusive sleep tracking, we now have a new method to investigate longitudinal, natural routines: long-term sleep datasets.

2.2 Development of Sleep Tracking

Traditional methods of sleep tracking have not been able to capture a person's natural, long-term behaviors, as they rely on expert analyses and uncomfortable sensors in an unnatural sleeping environment. Polysomnographies, the gold-standard for clinical sleep studies, require the individual to wear multiple sensors at a laboratory under the guidance of a clinician [4, 15, 85]. While actigraphies require a single wristwatch wearable, they are unsuited for long-term tracking because they are uncomfortable to wear throughout the night [17, 46]. As a result, large-scale datasets such as the Multi-Ethnic Study of Atherosclerosis (MESA) [10] and the UK Biobank [97] resort to ecological momentary assessments that capture snippets of behavior up to a week long [10, 52, 69, 105, 106, 111].

Alternative methods, such as case studies, allow in-depth investigations [25, 37, 98], but demand high labor costs [15, 17, 47, 56, 57].

Thus, commercial sleep tracking is a promising alternative. Sleep tracking is a core feature integrated by many commercial applications with a large user base [3, 36, 93, 94, 99, 104]—Sleep As Android and Sleep Cycle each have 10 million downloads on Google Play [93, 94]. Users are not restricted to wearables as they can longitudinally record their sleep times with passive sensors [1, 3, 12, 36, 38, 38, 41, 53, 68, 75, 84, 99, 110], semi-automatically track [18, 64, 93, 94], or manually input their sleep records [16, 93].

Despite their benefits, data from commercial technologies are criticized across disciplines for their data quality, credibility, and inability to be validated in a clinical setting [4, 7, 15, 17, 58, 67, 85]. However, sensing devices have improved their accuracy [38, 68, 75, 84, 110] and recent validity tests corroborate advancement in accurate sensing of commercial devices [4, 34, 59]. Chinoy et al. show that seven commercial devices, tested against polysomnographies, perform comparably to actigraphies [15]. In other words, commercial devices are becoming a viable and cost-effective option for large-scale, longitudinal research, even for clinical standards [15].

Furthermore, with widespread, commercial integration of sleep tracking, we can now observe natural behavioral changes and routines that only become apparent with long-term observations [37, 40, 46, 57, 88, 96, 103, 107, 107, 109]. Jeong et al. was able to investigate smartwatch activity patterns of 50 users for 203 days [46] and Xu et al. introduced a 2-year dataset to model associations between sleep and associated health factors [107].

Within the sleep medicine community, Yuan et al. examined 2 years of data and found changes in sleep duration and sleep timing from Covid-19 [109]. Additionally, Viswanath et al. [103] extended Katori et al.'s work to discover 13 different sleep phenotypes [52], or behavioral traits, using data from 33,000 naturalistic Oura Ring users. As Arnardottir et al. state, 'the widespread use of wearables... enables research on sleeping patterns and behaviors on larger and more heterogeneous cohorts of people than ever before' [4].

Our investigation is distinct from previous works [46, 103, 107] in both our method and the characteristic of routines that we choose to focus on: routine lengths. Prior works on sleep routines focus primarily on daily sleep regularity [35, 69, 83, 105, 108] and investigate its association with some other factors such as mortality [105], dementia [108], and cardiometabolic risk [69]. However, we focus on deepening our knowledge of routines and expand our investigation beyond the day-to-day and week-by-week contexts by examining a *series* of daily sleep regularity scores. While our work also investigates long-term behaviors with a secondary dataset, Viswanath et al.'s work investigates routine transitions [103].

Furthermore, unlike Viswanath et al.'s work, we do not make assumptions on how long a sleep period or cycle length is; instead, we find the cycle length that best describes the user from their sleep records. In addition, our aim is not to categorize but to discuss how considering natural, longitudinal routine lengths introduces more equitable designs and actionable recommendations to users of behavioral regularity and scheduling technologies.

3 Finding Good Long-term Sleep Recorders

We analyze 180,083 users with up to a decade of records from the Sleep as Android App during June 1st, 2014 to April 23rd, 2024. All anonymous users consented to using their data for research purposes. Users self-report sleep start and end times with a method of choice. Users can also integrate wearables to enhance accuracy and utilize a sleep estimation feature, which uses activity logs to autofill probable sleep periods and prevent missing recordings. Thus, the app's core functionality is that of a semi-automatic sleep diary [93]. Multiple sleep events within a 24 hour period are recorded, including naps and interrupted sleep. Information on phone device, time zone, gender, and age are also available.

In the following sections, we outline the process and design considerations for developing the system that identifies 'good long-term sleep recorders'. While the system is applied to a single collection of sleep records, we document our design considerations for reproducibility and the number of users affected at each step to provide a practical evaluation of the system. This section also serves to provide transparency in the system design such that future designers can gain insights into the practical limitations and implications of the system **(RQ2)**.

First, we collaborate with domain experts specializing in research on sleep and circadian disorders to define 'good sleep records'. Then we analyze the number of sleep logs for each user to define an 'engaged' user. Next, we look at each user's ability to record uninterruptedly to find 'good long-term sleep recorders'. Finally, we discuss the trade-offs of using this dataset and evaluate for dataset validity and the biases of this system.

3.1 Methods

3.1.1 Good Sleep Records. Selecting which sleep records to include in analysis is a challenge in naturalistic data, due to the difficulty in corroborating heuristics of validity. 'Good sleep records' reflect expected sleep behavior. Each record should have logical start and end times and fall within the range of normal sleep duration. The former issue is negligible. Less than one percent of users mistakenly input their sleep end time before their sleep start times and such records are removed. The latter requires domain expertise, because uninformed thresholds can introduce biases in favor of users with particular behaviors.

In consultation with domain experts, we set the threshold of a 'good sleep record' as a range between 5 minutes and 16 hours. According to domain experts, it is *"extremely unlikely for a healthy, non-sleep deprived individual to have sleep episodes to exceed 16 hours… it is physiologically hard for someone to sleep over a day. They may be in bed for that long, but are probably not asleep."* We set the minimum threshold to 5 minutes to include naps.

Any sleep records outside the range are removed. The practical impact of each removal process is as follows. Excessively long sleep sessions accounted for only 0.24% of the total records and 251 users only recorded excessively, long sleep sessions. About a fifth of users recorded excessive sleep at least once, indicating that this recording mistake is relatively common among people. Over half the users (102,842) recorded extremely short sleep sessions at least once and approximately one percent of users (2,489) recorded only sleep sessions less than 5 minutes.

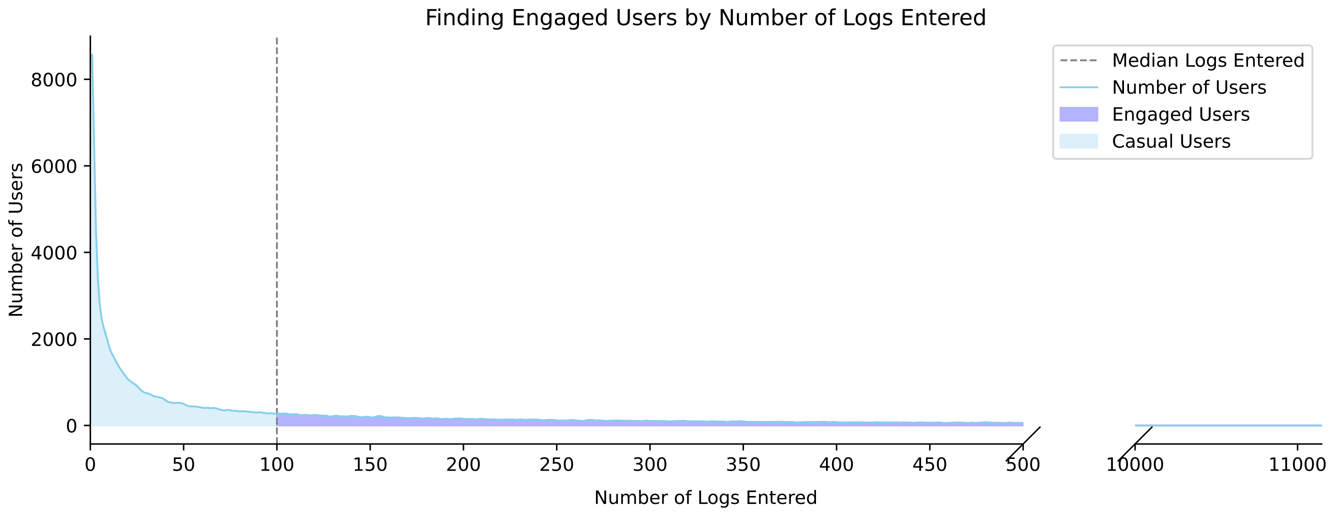


Figure 1: Half the users recorded at least 100 sleep sessions and more than a quarter of users recorded 445 sleep sessions or more. We use the median number of logs entered to define the 88,765 'engaged' users.

Thus, short sleep records, potentially caused by noise from passive tracking devices, are the most common reason for messy records. This process leaves 177,342 users with 'good sleep records'.

3.1.2 Engaged Users. From these 'good sleep records', we identify 'engaged users' with the number of sleep sessions recorded, a common proxy for user engagement [87]. We base the analysis using users' non-nap times to further minimize the influence of potential erroneous self-reports. Because the upper threshold of naps varies, even within the sleep community, we consider records that are at least an hour long, derived from the median nap time (Figure 2). As depicted in Figure 1, the sharp incline in the beginning signifies the casual recorders—a quarter of users record 15 sleep sessions or less. On the other hand, more than 50% of users log at least 100 sessions of sleep and more than a quarter of users record 445 sleep sessions or more. It is unlikely for such high numbers of entries to be random, given that users voluntarily self-report their sleep sessions for their own benefit. We take the top half of users that log at least 100 sleep sessions to ensure we analyze records from committed users. This process leaves us with 88,765 'engaged users'.

3.1.3 Long-term, Continuous Recording. While the number of logs is indicative of engagement, not all users are equally 'good trackers' who can record their sleep uninterrupted for extended time periods. Users may forget to track, choose to stop tracking, or are unable to record their sleep due to external factors (e.g., traveling with no Wi-Fi, broken phone, phones out of batteries, etc.) [31, 62, 63, 73, 80]. Typically, understanding how many times a user fails to record will tell us about how good of a tracker they are.

However, we find that defining 'good trackers' based on the number of missing records may be skewed because longitudinal users can have extended periods of inactivity [46, 74]. We extracted the median and maximum duration of gaps in recording. The median characterizes the common mistakes the user makes, while the maximum describes the long breaks the user takes.

In general, the median number of days users skip recording is 2 days in a row. Half the users *lapse* or pause recording [32] for at least 50 days in a row before returning to the application. Thus, looking at the percentage of missing records is misleading—a user who took a break is not necessarily a bad tracker.

Instead, we look at the longest *streaks* [46], or continuous records of their main sleep times, excluding naps. Two sleep records are considered continuous if the user wakes up on the same date or the next date after the previous wake-up date. While more than 50% of users record continuously for at least 56 days, we need long-term observations to observe routines, or repeat behavioral patterns. Thus, we focus on the top 30% of users who can record for 120 days at a time at least once. This allows us to observe at least 3 cycles, each spanning at least a month each. These 25,578 engaged users are what we define as 'good long-term sleep recorders' and are the basis of subsequent analyses.

3.2 Striving for Validity in Naturalistic Self-Reports

We choose a deliberate trade-off using this secondary dataset. Because the data is from an external platform where the collection began over a decade ago, there are inherent constraints to its use. Unlike lab settings, self-reports are messy with missing information. As we show in previous sections, users can forget and pause tracking for long periods of time. Many also omit their demographic data—80% of users' age and gender are unknown. While 95% of users report what smartphone device they used, the dataset does not provide information on what tracking method they used.

On the other hand, self-reports include users from all over the world and are recorded in-situ. Users are not subject to the reminder follow-ups or artificial sleep conditions of controlled studies. As a result, these are a diverse group of long-term users who track in their natural sleep environment.

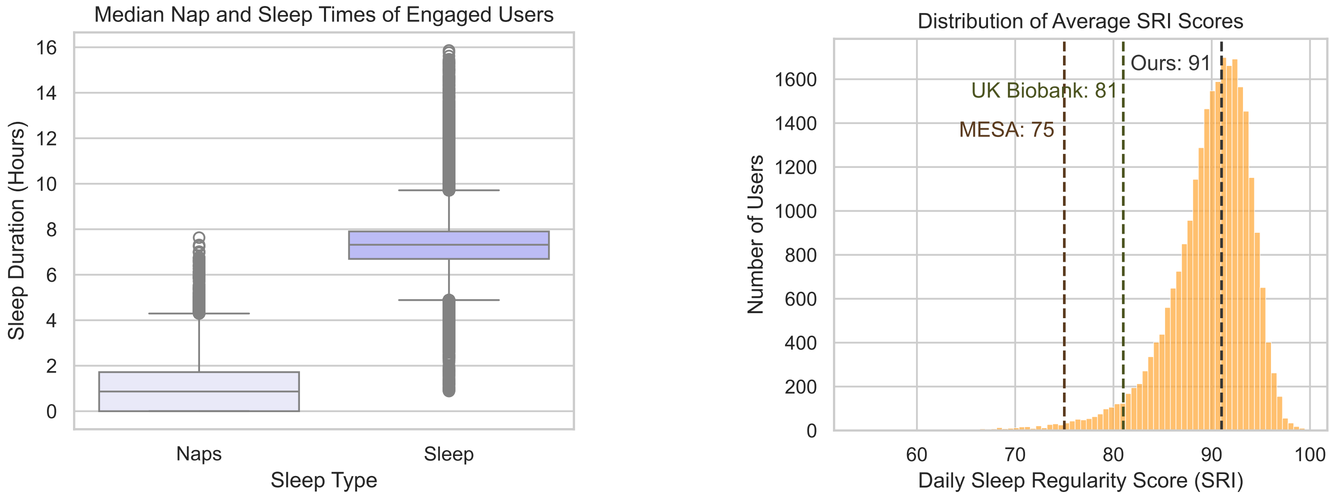


Figure 2: Dataset reflects expected sleep and nap lengths. Average naps are 0–1.9 hours with a median of 1 hour. Sleep sessions are 6.9–8 hours long with a median of 7.5 hours. Daily sleep routine scores are within expected ranges and exhibit similar left skew distribution shapes, as does previous works on daily sleep regularity [69, 83, 105]. High median shows our system has a bias towards users with more regular daily routines.

Analysis of each user's primary time zone, defined by the mode, indicates 46% of users are from Europe, 34% from the Americas, and 13% from Asia. The median indicates that more than 50% of users can log at least 100 sleep sessions and 25,578 users track without missing a record for at least 120 days. We believe a dataset that has these trade-offs provides a unique perspective, and complements other methodologies and datasets about human behavior and routines.

By corroborating the data using heuristics from the literature, we gain confidence that these self-reports have sufficient signal. While the records may have unintentional mistakes, self-reports have little incentive structure for users to intentionally obstruct the data for extended periods of time, as users were capturing the data for their own use, rather than in service for a research study or experimenter. We apply two heuristics to check the data: nap and sleep duration and average daily routine scores.

First, people's sleep and nap duration are within expected ranges. For each user, we use a Gaussian Mixture Model to form two clusters of each user's naps and main sleep duration. This unsupervised clustering method helps describe multimodal density plots. The model assumes that the data is composed by a mix of Gaussian distributions, a common assumption for nap and sleep duration in literature [35]. We apply this probabilistic clustering method because it is more robust to outliers and therefore naturalistic datasets, unlike other clustering methods such as k-means.

'Naps' are any sleep sessions that last less than half the mean of their primary sleep length (a definition provided by our domain experts). Our definition of nap times did not include the time of the nap to avoid systematically excluding users with non-normative schedules (e.g. night owls, shift workers). As a result, naps for users who have a main sleep duration of 5–6 hours would be up to 2.5–3 hours long. This could indicate fragmented sleep, but this would be indistinguishable from a nap that a user takes from their main sleep since we do not have ground truth from our users.

While we use naps as a behavioral check for both naps and sleep durations, given that naps are under-defined even within the sleep medicine community, the primary focus is on the main sleep duration. As summarized in Figure 2, the median 'nap' time is one hour, and naps range between 0–1.9 hours. Most users' sleep times range between 6.8–8 hours with a median of 7.4 hours. All of these indicate that the users in the dataset exhibit expected sleep behavior, and not random noise.

In addition, people's daily sleep routines also align with expectations from prior research, further supporting the validity of the dataset. We use the Sleep Regularity Index (SRI) because it offers a way to validate a series of daily natural routines [83], while accounting for behaviors we expect to occur in our naturalistic dataset such as naps and sleep fragmentation [35]. The metric is a similarity score that describes how similar two adjacent days are, on a scale of 0 to 100. A user who has perfect daily regularity, or is asleep and awake at the same times on day 1 and day 2, has an SRI of 100. A unit in SRI accounts for 14.4 minutes of discrepancies with the previous day. We use SRI scores to investigate daily routines because they include information on *both* the waking and sleeping hours of the day.

Previous studies of people's average SRI scores sampled one week periods of its users [69, 105]. Studies found the user SRI scores to have a left skew distribution and range between 60–100 [69] and 58–100 [105]. Our system also shows users with a similar left skew (Figure 2) and daily scores ranging from 53–100. However, the median is skewed towards users with higher daily regularity, likely due to the selection criteria favoring consistent long-term trackers with no missing records—there may be a confounding factor in which 'good recorders' are also the same people who maintain a daily routine. In addition, users who do not utilize semi-automatic methods may have more missing records, making them less likely to be included in the long-term continuous recorders group.

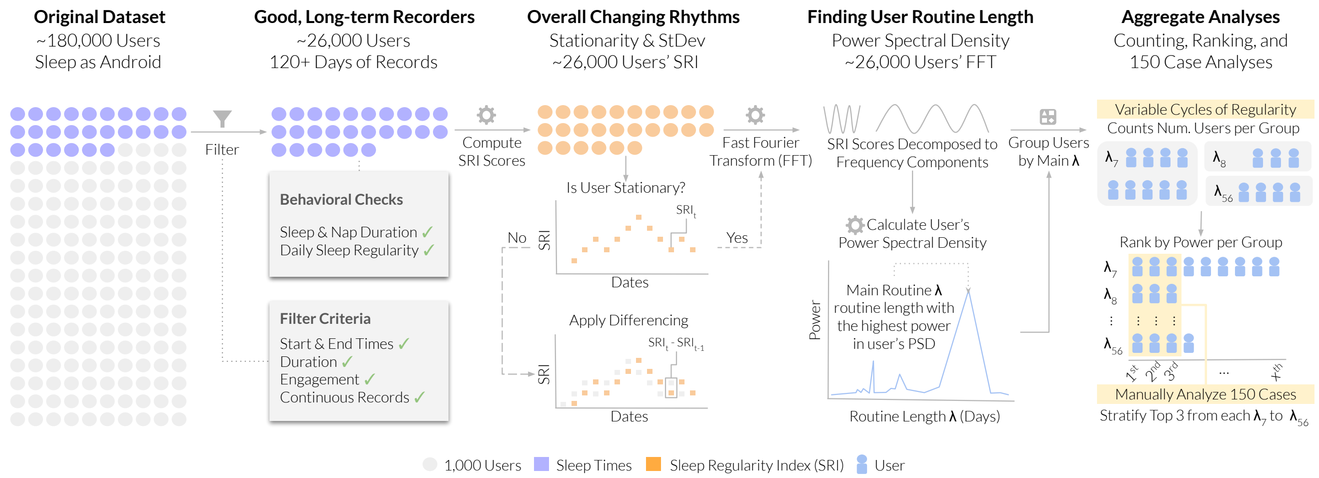


Figure 3: Flowchart of Analyses and Methods. Filter 'Original Dataset' to analyze a set of 'Good, Long-term Recorders'. Sleep and nap duration and daily sleep regularity are 'Behavioral Checks' to verify 'Good, Long-term Recorders' have reasonable sleep behaviors. Next, user's sleep times (purple) are used to compute their SRI or Sleep Regularity Index (orange). Each user's SRI is checked for stationarity; differencing is applied if not stationary. We compute the Fast Fourier Transform and Power Spectral Density on stationary users to find each user's main routine length, λ. Users are grouped by λ to analyze 'Variable Cycles of Regularity' (Figure 5). Users are ranked by their normalized power within each λ group. The top 3 ranked users from each λ_7 to λ_{56} groups make the '150 Cases'.

While we acknowledge these limitations, it is still the case that users' daily sleep routines are within expectations of prior research, and show something other than random behavior. Furthermore, because our study explores whether there are variations in user routines, if we observe variations even among a more regular group of users, this suggests that individuals with less consistent routines may also exhibit variability in their behaviors. More importantly, this dataset still provides us with an opportunity to investigate routines of 25,578 users in their natural settings, a mass investigation that has not been possible until now, thanks to the widespread adoption of sleep tracking.

4 Rhythms of Regularity

Average daily routines, while informative, do not capture *how* people's routines evolve over time. We need to look at a long *series* of daily routines to understand the different ways in which people behave, or their **rhythms of regularity**. We conduct a tripartite investigation to understand how people's routines vary **(RQ1)** and, in the process, we identify the lessons learned from the analyses **(RQ2)**. The overall process is summarized in Figure 3.

First, we look at overall changes in daily routines by analyzing standard deviations and conducting stationarity tests on each user's SRI scores **(RQ1a)**. Next, to find the most common cycle length **(RQ1b)**, we investigate routine lengths of $\lambda = n$ using a Fast Fourier Transform, a method to identify periodic components of a signal. Lastly, we conduct a series of comparative, case analyses and incorporate insights from our discussions with domain experts in sleep medicine to gain an in-depth understanding of how people's routines vary **(RQ1c)** and identify key lessons we learn from investigating longitudinal sleep records **(RQ2)**.

Throughout our analyses, we find that routines are heterogeneous and many are better described by changing or non-weekly routines and we further expand on the lessons learned in the discussion section (**RQ2**).

4.1 Changing Rhythms: How Daily Routines Change Over Time

We examine how people's daily routines change over time because it indicates how common behavioral changes are among users **(RQ1a)**. The question is, can we observe such changes using people's sleep records? And *how* do people's routines change overall? Do people have little to no changes in their day-to-day routines? Or does everyone have one consistent routine throughout?

4.1.1 Method. We answer these questions from two angles: users' standard deviations and stationarity tests of their daily SRI scores. Standard deviations help us understand *how much consistency* people have in their day-to-day routines. The higher the standard deviation, the more irregular the user's daily routines are on average—in other words, they struggle to maintain a consistent, daily routine. We convert the units into minutes for interpretability.

Stationarity tests are descriptive tests that help us examine changes in mean and variance over time and allow us to understand trends [44], or *how users evolve* over time, and whether users *adopt new routines*. Stationarity is also a precursor condition to examine cyclic patterns in subsection 4.2, because it allows us to observe real, periodic patterns, without the interference of trends or seasonal effects. Because stationarity tests require continuous, univariate time series data with no missing values, we apply them to each of the SRI scores of our 'good, long-term recorders'.

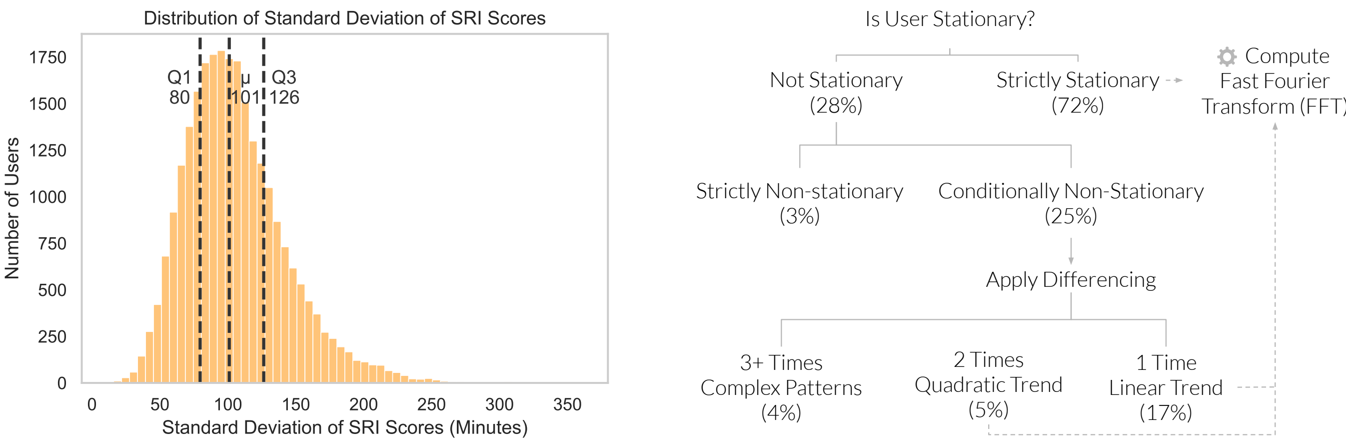


Figure 4: Half the users have 1.3—2.1 hours of variation in their average day-to-day routines (left). Only one percent of users have constant daily routines that vary less than 40 minutes day-to-day and these are the users with $\lambda = 1$ (left). 28% of users change their routines significantly over time (right). Users with 'strictly non-stationary' daily regularity scores and users who require more than two differencing have complex patterns which imply major behavioral changes over time. This means such users have no dominant routine, and are considered to have a $\lambda = 0$. We can remove trends from 'Quadratic Trend' and 'Linear Trend' users with differencing. For these users and the 'strictly stationary' users, we can compute their Fast Fourier Transform and examine cyclic patterns, or routine lengths.

We use two complementary stationarity tests to determine whether the user's daily routines are *strictly stationary*, *strictly non-stationary*, or *conditionally non-stationary*. The Augmented Dickey-Fuller (ADF) [30] tests for non-stationarity and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) [61] for whether there is a constant deterministic trend. For implementation, we used the constant regression model with automatically selected lags (ADF used the Akaike Information Criteria).

If both ADF and KPSS reject the null at $\alpha = 0.05$ [90], they are considered *strictly stationary*: they have no major changes in their daily routine. If they both agree to accept the null, they are *strictly non-stationary*, which means the user has major changes in their daily routine and has adopted a new routine over time. If the two tests disagree, this means the user's daily routines are *conditionally non-stationary* and may have underlying trends or seasonal effects in their daily routines.

One way to achieve stationarity is by removing these effects through differencing [44], the successive subtracting of the previous SRI score from the next. The number of times we difference also reveals underlying trends, or *how* people change. Achieving stationarity after one differencing indicates a linear trend, while two suggest a quadratic one. Users who required higher order differencing have complex trends and are not made stationary to avoid oversmoothing [44].

4.1.2 Findings. Our results show that most users don't follow the same exact routine every day—**it's rare to be regular, not the norm**. Half the users have 1.3–2.1 hours of inconsistencies in their day-to-day routines, 28% of users' daily routines are not stationary and adopt a new daily routine over time, and less than one percent of users follow the same, stationary routine with less than 40 minutes of variation day-to-day.

Furthermore, according to the number of times we difference to achieve stationarity, we find that some have linear while others have quadratic changes in their daily routines (Figure 4). All of these results show that the *ways* people change also vary.

The results suggest we need to design technologies that are forgiving of change. Technologies that model real-life patterns should expect users to naturally deviate from their average daily routines. Expecting perfect day-to-day consistency is unrealistic for most people. People adopt new routines and change their daily routines over time. Change is prevalent, but perfect, daily regularity is not, and we continue to see such behavioral variations in subsequent sections when examining routine lengths.

4.2 Routine Lengths: A Method to Investigate Variable Cycles of Regularity

Learning what natural routine lengths are most common among users **(RQ1b)** inform us for whom and how we should design behavioral, health technologies. Routine lengths are behavioral signatures, or distinctive patterns unique to an individual. Different routine lengths have different meanings. For example, the users with less than 40 minutes of day-to-day variations have a constant daily routine, while 9–5 workers follow a weekly one. Others, like the non-stationary users and users with complex trends, do not have a single, dominant routine because they change their routines significantly over time. Thus, investigating routine lengths informs us of the most common routine and of alternative routines (**RQ1c**).

4.2.1 Method. We can examine the user's routine length with the Fast Fourier Transform (FFT). This method deconstructs the user's stationary daily SRI scores into a combination of cosine and sine functions, or basic frequency components. The amplitude and phase information of each frequency component is stored in an FFT

Figure 5: Variable Cycles of Regularity. Weekly routines $\lambda = 7$ **are the most prevalent among 25,578, long-term users, but more than half of them have alternative routine lengths. Some are regular weeks-to-weeks, and others are months long. Others** ($\lambda = 0$) **don't have a singular, dominant routine because their daily routines change significantly over time.**

coefficient. The normalized squared magnitude of the coefficient represents the power of that frequency.

The Power Spectral Density (PSD) is the distribution of power across different frequencies. Peaks in the distribution indicate which routine lengths λ have more power in describing the overall structure of the signal. We search for the highest power λ routine length from $\lambda = 2$ up to $\lambda = n$, where n represents one-third of the total number of the user's recording days. This ensures that we observe at least three repeating cycles. To focus on the underlying routine patterns, we smooth the SRI scores with a Gaussian Filter ($\sigma = 1$), which dampens noise by placing less weight on scores further from the user's SRI mean, or daily sleep routines that are atypical for that user. In other words, by searching for the λ with the highest PSD value, we can find the user's most dominant routine length (i.e., their main routine).

We then count how many users have a weekly routine ($\lambda = 7$) and non-weekly ones. To understand what types of non-weekly routine lengths exist, we bin users by doubling the routine lengths: 1 to 2 weeks or $\lambda = 8$–14 days, 2 to 4 weeks, 4 to 8 weeks, and 8 weeks or longer. We also incorporate findings from subsection 4.1. Users with $\lambda = 0$ are those who do not have a singular, dominant routine as their daily routines change significantly over time—these are the strictly non-stationary and conditionally non-stationary users with complex trends from Figure 4. Users with $\lambda = 1$ are the one percent of users whose daily routines are constant as they vary less than 40 minutes day-to-day.

4.2.2 Findings. We find **variable cycles of regularity**—some users are consistent week-to-week while others are consistent weeks-to-weeks (Figure 5). While the most common routine length is a weekly one ($\lambda = 7$), over half the users follow a non-weekly routine, some of which span weeks-to-weeks ($\lambda \geq 14$). 6% of users have routine lengths less than a week, 10% of users have routine lengths $\lambda = 8$ to $\lambda = 14$ days, 15% of routines last between 2–4 weeks, 12% have routines that last between one and two months, and 6% of them have routines that last longer than 2 months. 7% of users don't have a singular, dominant routine because they are either non-stationary or have complex trends, meaning their routines change over time.

Given the prevalence of non-weekly routines among users, these findings suggest a need to redesign behavioral health technologies such that they account for people's natural routines. The current practice of designing technologies for the 'normative' $\lambda = 7$ systematically disregards more than half of users. While previous works have shown that many marginalized groups face mismatched expectations between their routines and product features [49, 51, 64, 95], our results suggest that the problem may affect a larger population and that we need to reconsider our assumptions on what is 'normal' or 'expected' behavior. To start addressing this problem, in the next section, we conduct a series of case analyses to understand in-depth the practical implications of current design practices and the issues that users with alternative routines may face.

4.3 Case Analyses: In-Depth Investigations of Alternative Routines

To gain insights on how to redesign for users' natural routines **(RQ2)**, we conduct a series of comparative case analyses. By 'situating' ourselves in the shoes of our users [43] and observing in detail how user behaviors unfold over a long period of time [42, 76, 113], we can understand the practical implications of current design practices, such as when and for whom current design practices fail **(RQ1c)**. Thus, the primary objective of this analysis is not to categorize users, but rather to provide illustrative examples that highlight the diversity of routines and emphasize the importance of observing natural routines longer than a week.

4.3.1 Methods. We focus on 150 cases, comprised of the top three users with the most distinct patterns, or highest PSD values, for each $\lambda = 7$–56. We focus on cases that span a week or longer, because a core objective of the paper is to fill the gap in understanding longitudinal behavioral patterns that short studies cannot capture.

To investigate unique behavioral patterns among users, we qualitatively analyze their PSD charts and sleep times with their natural λ. Each user's PSD is normalized by the user's total PSD value into a value between 0 and 1. The normalization allows us to visually compare the structure of users' PSD charts. Users with only one routine show one dominant peak in the PSD chart, while those with complex routines show multiple peaks (Figure 8).

Thus, PSD charts give us information on their routine lengths while the visualization of user sleep times helps us identify other distinct behavioral patterns such as abrupt shifts in sleep wake times. By combining our understanding of routine lengths from PSD charts and other behavioral patterns from sleep time charts, we can holistically identify user behavioral patterns.

4.3.2 Observations. In our case studies, we observe that how people change within their routine, or the **composition of rhythms**, must be a key consideration when designing for natural routines. We identify two unique composition types that can only be understood with long-term observations: *waves* and *alternations*. These are the two most common patterns in our visual comparisons of sleep times and PSD charts.

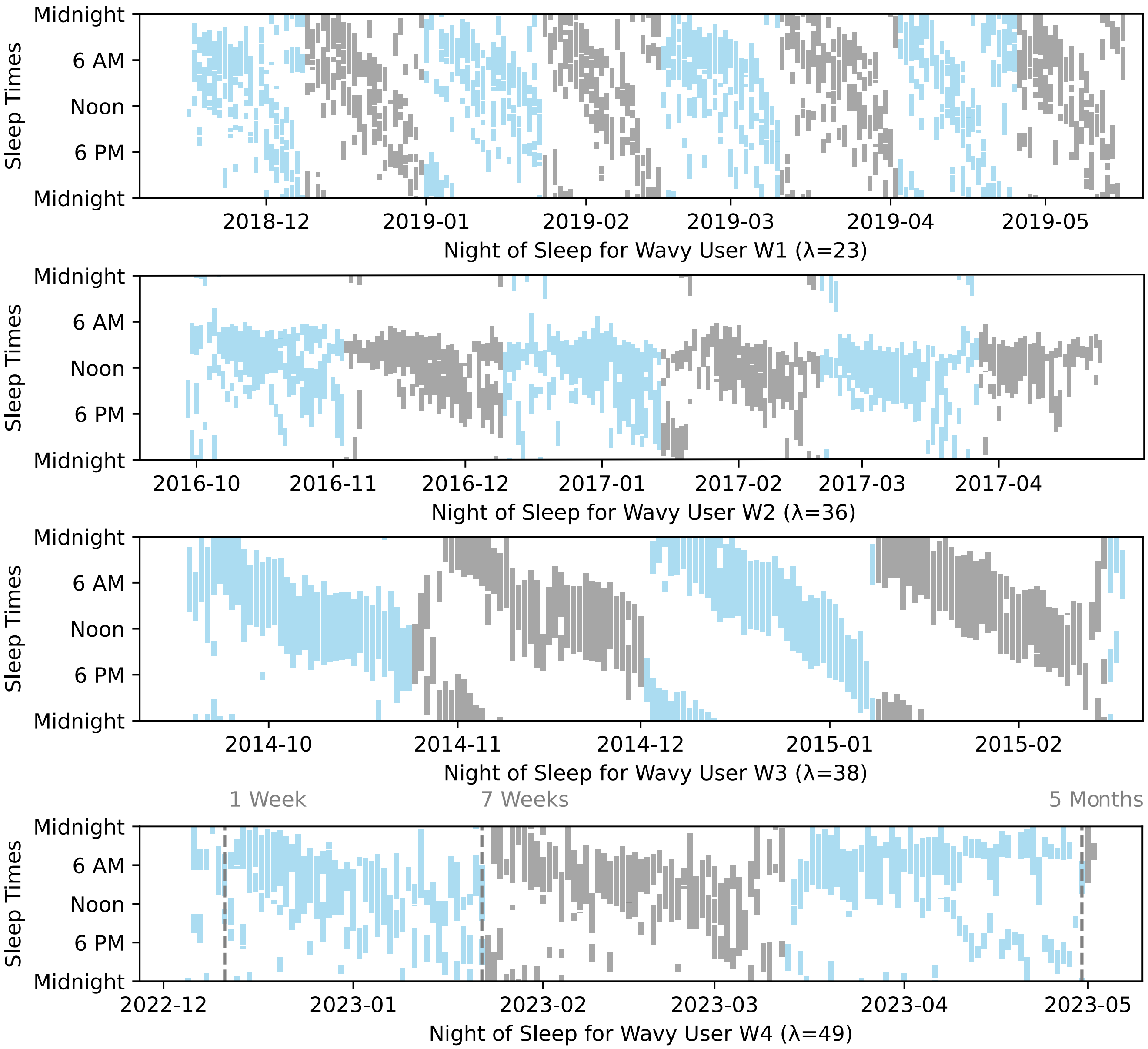


Figure 6: Wavy Users. Demonstrates routines can be long and requires long-term observations; some span even seven weeks (W4). Wavy users have distinct, slow changes in their sleep times that gradually return to their original sleep time every λ days. To highlight wave patterns, colors alternate every λ between sky blue and gray.

Other patterns described users with 'fragmented sleep' and 'hilly' sleep structures. However, these are not included in our analyses as their patterns were less visually apparent and identified in a small subset of users (less than 5). Users with *waves* gradually shift their sleep and awake times over their natural λ (Figure 6). On the other hand, users with *alternations* periodically *alternate* between their primary and secondary routine, typically through an abrupt change (Figure 7). We observe *waves* in 8% of the cases (12/150) and *alternations* in 11% of cases (17/150).

Beyond the Weekly Rhythm. *Wavy* and *alternating* users tell us that certain routines are describable only with long-term observations. Wavy user W4 in Figure 6 requires 5 months of records to observe that the user's sleep and awake times gradually shift every 7 weeks from midnight to midnight and exhibit *waves*. Similarly, we cannot identify alternating user A3's 3-week routine without long-term observations (Figure 7). If we only observe the first week of A3's routines, the user appears to be an irregular sleeper who has inconsistent sleep times between midnight to noon.

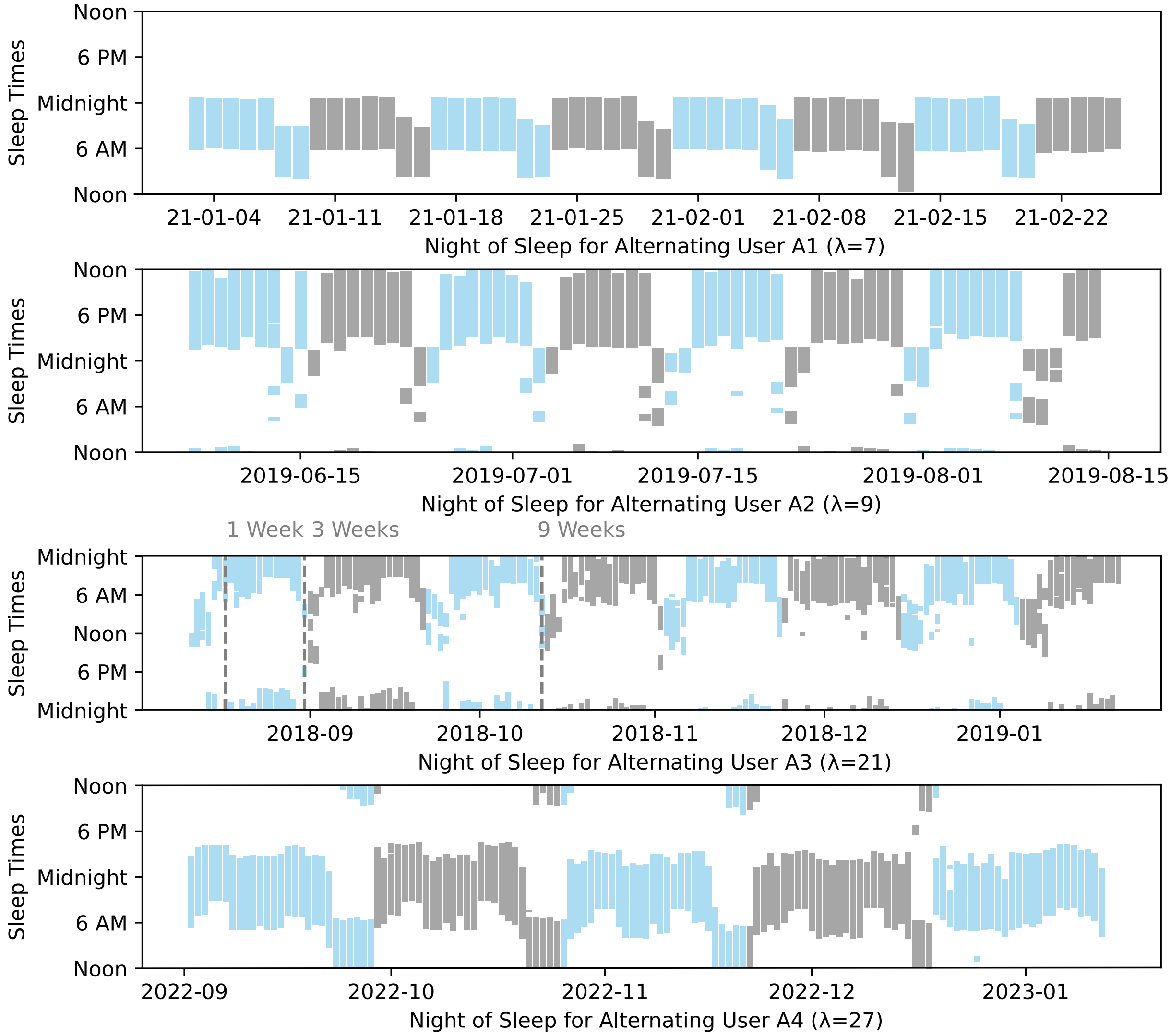


Figure 7: Alternating Users. Demonstrates that routines can have abrupt 'alternations', or changes in routines. Within a λ routine, users alternate between routines. The change is not gradual like wavy users, but abrupt.

With three weeks of observations, we might believe that the user had a singular, 'bad', first week, and that their real sleep times are between midnight and 6 am. It is only after observing multiple weeks of records that we can recognize that the first 'irregular' week is actually a routine variation [6], or an *alternation* in their routine, that is part of the user's larger 3 week routine.

Furthermore, comparing short-term routines with long-term routines reveals design opportunities for behavioral regularity tools. For example, when discussing cases A1 and A4, our domain experts raised the following questions: *"Somebody who alternates every 4 days (like a shift worker) will likely face health issues based on current research or conventional wisdom, but what about somebody who shifts on a monthly basis? Or when you shift rhythms, how long do you need to stay on the new rhythm to mitigate health risks?"*

While our domain experts provide perspectives from a healthcare standpoint, these insights suggest the need to design interfaces with more holistic views of routines. The length and transitions of routines are not only interconnected with other aspects of health, but also influence other external routines, such as productivity and work schedules. Incorporating such factors into personal informatics tools and scheduling systems could provide new ways for individuals to explore their data and manage their routines. While further research is needed to design systems that foster holistic self-exploration and self-experimentation without reinforcing unhealthy behaviors, considering such factors in the system design could enhance long-term user experiences with self-exploration tools. Thus, there is an opportunity to develop technologies that consider long-term interactions between different routines.

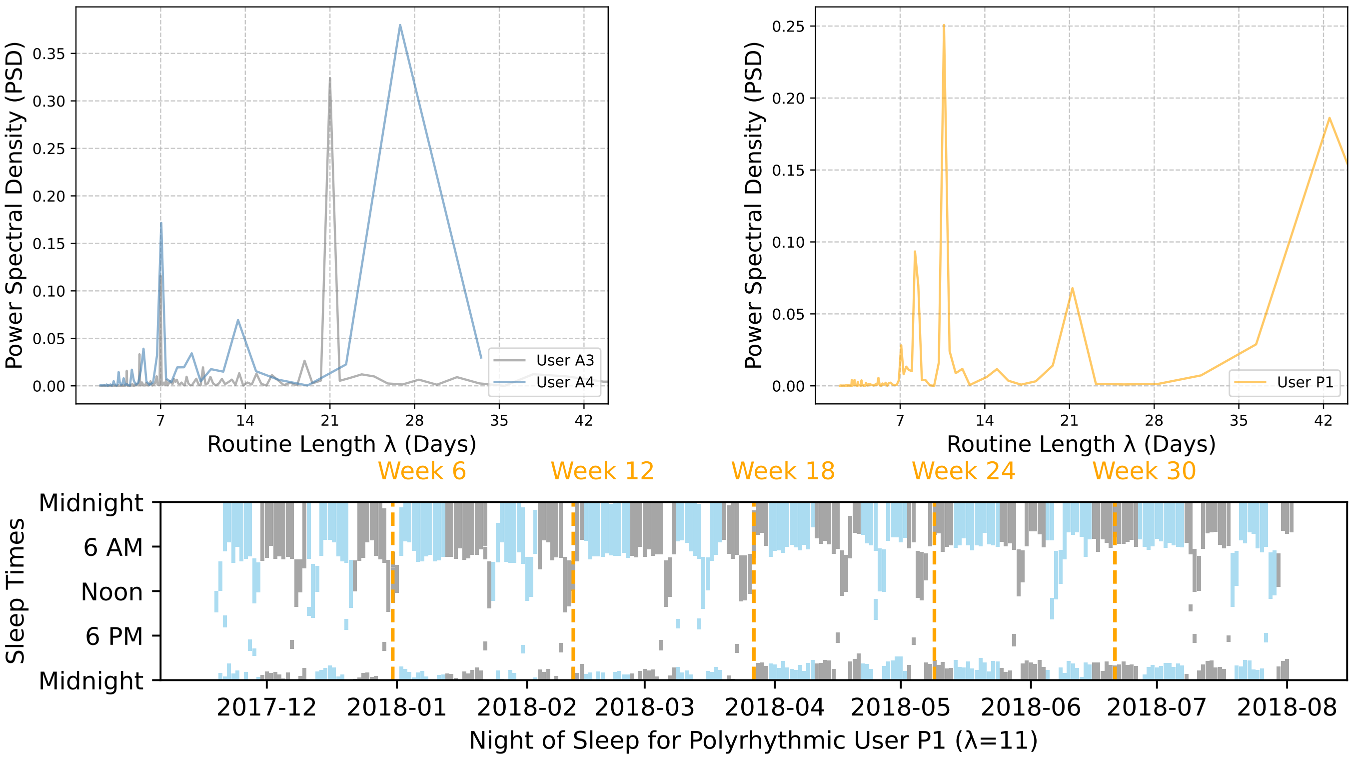


Figure 8: Polyrhythms. In contrast to users in Figure 6, polyrhythmic users show multiple high peaks in their PSD. They have a main routine and another subroutine. Users A3 and A4 show peaks at $\lambda = 7$ in addition to their main routines at $\lambda = 21$ and $\lambda = 27$, respectively. User P1 also has multiple subroutines (right). The top two peaks at $\lambda = 11, 42$ reflect shorter and longer term patterns (bottom). Examples demonstrate that routines are complex and are not always defined by a singular routine.

Polyrhythms. In addition, we learn from the long-term, alternating users that natural routines are complex, composed of subroutines—they are *polyrhythmic*. In our visual inspection of the alternating users' PSD charts, we observe a recurring pattern characterized by two or more prominent peaks in their PSD. This behavior contrasts with wavy users whose PSD typically exhibits a single dominant peak. We know that many people have 2 subroutines within a week because they suffer from social jet lag, or two different lifestyles during their workday and weekends [1, 64, 98], like alternating user A1.

However, subroutines aren't limited to a weekly framework. A2 has a 9-day main routine and a 3-day sub routine. As denoted by the two highest peaks in their PSD (Figure 8, left), A3 and A4 also alternate between a distinct, main and secondary routine. These subroutines are also non-weekly. Polyrhythmic user P1 has an 11-day and a 6 week long routine. In weeks 6–24 in Figure 8, the user shows a clear 6 week pattern with alternating sub-patterns. Within a 6 week cycle, the user sleeps between midnight and 6 a.m. in the first 11 days, and in the next three 11-day cycles, they have *alternations* towards the end of each 11-day cycle.

These observations show that some users maintain two distinct lifestyles. While we do not know whether these polyrhythms are due to personal choice or external factors like their vocation, they indicate that we shouldn't design technologies that impose a singular routine or lifestyle.

For example, if a system creates an 11-day routine for user P1 and continually alerts them that they are making mistakes during the next three 11-day cycles, despite this being part of their actual routine, those alerts could be perceived as discouraging. The user may feel that the system overrides their experience or ignores the context of their behavior [11, 96]. This misalignment in expectations can result in decreased engagement, or the user may give up on trying to improve their routines or using the system altogether. Thus, to minimize user frustration and maximize personalized recommendations, we need to design systems that accommodate for the user's subroutines.

Re-Designing with Natural Routines. Not only are we limited in our ability to recognize long-term patterns, but as we show by example, applying a normative week-to-week framework on such users will prevent us from providing tailored insights and sometimes even lead to incorrect recommendations. Let's take A3 and overlay each week of their routine, to observe weekly patterns (in essence, this is a visualization of a commonly implemented weekly average). We see in Figure 9 (left) that user A3 struggles every fifth and seventh day of their routine, as their SRI scores drops to 40.

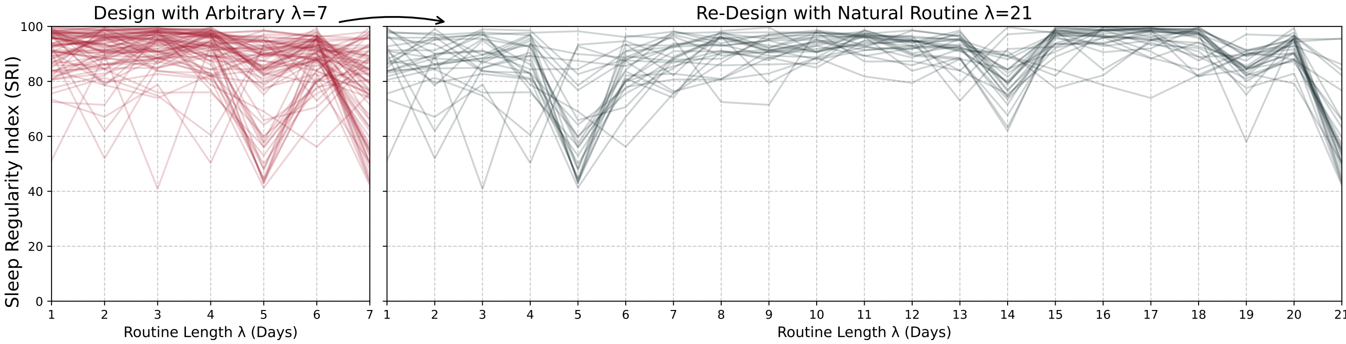


Figure 9: Re-designing Using Natural Routines. Insights using an arbitrary weekly $\lambda = 7$ routine is limiting and misleading for the alternating user A3 (left). The overlaid line plot on the left suggests the user struggles every fifth and seventh day of their routine. However, if we re-design the plot using their natural routine of $\lambda = 21$ (right), we find that the user should look at their routine on a 3-week basis, and focus on the 5th day, 14th, and last day of their routines to achieve regularity.

However, if we make a simple design change and display their daily routine scores using their natural $\lambda = 21$ routine in Figure 9 (right), we can understand user needs better and offer specific recommendations. We see that their behavior from 3 weeks prior best describe them, not the week prior. The individual struggles to maintain a consistent daily routine on the fifth, fourteenth, and last days of their cycle. In fact, their daily sleep routine score for the seventh day of their cycle does not drop to 40, as suggested by the $\lambda = 7$ design, but is instead around the 80s. Their score drops to 40 only on the fifth and the last day of their cycles. They also have irregularities to a lesser degree every two weeks, as it drops to 60 instead of 40. In other words, these are the nuances that become lost when we do not design with people's natural routines.

These examples illustrate that relying on a normative week-to-week framework prevents us from understanding the third of users with $\lambda \geq 15$ routines (Figure 5). By not considering their natural routines, these users lose the opportunity to find meaningful insights and create actionable items. Pre-existing dashboards and models that do not consider natural routines are misinformed and can lead users to misguided conclusions.

Re-designing with longer natural rhythms is also an opportunity to develop better recommendation systems and data exploration tools for both clinicians and individuals. According to our discussions with sleep medicine experts, such re-designs could help identify whether *"specific recurring behaviors (due to work, social, etc.) correspond [and] be helpful in the clinic when discussing recommendations with users to take into account behavioral factors."* In addition, such re-designs could help identify patients with seasonal affective disorder or patients with non-24-hour disorder. The latter are *"usually drifting a few minutes each day, [and] the larger rhythm occurs on the scale of months [... This] might be obfuscated if you were averaging data across 1 week, but would show up more clearly on a larger scale."* Integrating such re-designs into consumer technologies could also help individuals recognize long-term behavioral patterns that are difficult to identify in clinical settings and enhance their access to timely, personalized care and self-knowledge.

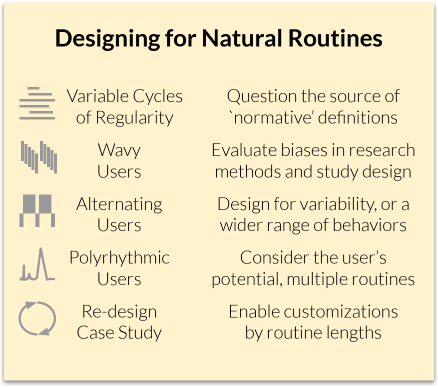
5 Discussion

5.1 Long-term Designs that Expect Change

We've investigated users' natural routines in three different ways. First, we examined overall changes in daily routines. Next, we explored the diversity of routine lengths. Finally, we analyzed sample cases of unique routine patterns. Across all three analyses, we found that changes in sleep and routines are not uncommon. 28% of user showed major changes in overall behavioral routines, more than half of users have non-weekly routines, and some long-term users have wavy, alternating, and polyrhythmic routines. As found in previous works [52, 64, 103], variations in routines are not limited to a select group of anomalous users or those with specific needs—they are experienced by a broader group of users in general. In other words, *changing routines is normal behavior.*

Variations in routines prompt us to reassess the assumptions underlying our current approaches in evaluating and designing systems. Many tracking technologies are evaluated on its predictive ability to detect or model sleep using short-term studies [5, 12, 38, 75, 77, 111]. While accurate sensing is important (as it has evidently made this study possible), the true value of consumer health technologies may lie in their ability to remain robust in the face of behavioral change.

Take for example, the polyrhythmic user in Figure 8 whose true routines only appear after long-term observations. Are current systems able to adapt to such users with different, non-weekly routines, gradual and abrupt changes in routines, and polyrhythmic ones? Should we instate different stabilization strategies and fewer reminders and nudges during periodic routine switches [37, 66, 112]? Can systems accommodate to *lapses* in recording that appear in half the users? Is it even possible to have a constantly high-accuracy system? Perhaps, given the prevalence of ever-changing users [79], we should instead design context-aware behavioral systems [27] that accommodate and accept a certain level of unpredictability [54] and adapt to behavioral transitions [79].



Problem	Description and Effect	Opportunities
Limited Insights	Insights and visualizations are not helpful and reminders are not timely, because user's true behaviors are not accounted for	Systems ask for user's schedule and enable views or filters by different routine lengths
Access Inequity	Users with `non-normative' schedules or long routines (e.g. shift workers) have limited access to personalized technologies	Systems, tracking devices, and interfaces account for a wider range of behaviors
Misinformation	Users who have different routines follow recommendations based on incorrect ones, preventing access to timely care	Systems provide more transparency on what routine characteristcs are accounted for
Bias Reinforcement	Users are asked to describe their experiences on a short-term or weekly basis, which bias the study results	Researchers first ask for user's routine schedule or expand scope of invesitgation

Figure 10: Design Guidelines. The yellow sticky note summarizes the five lessons to consider when designing for natural routines. The table on the right demonstrates problems that arise from not considering natural routines, describes the effects of such problems, and identifies potential design opportunities.

5.2 Re-evaluating Research Methods with Alternative Designs

Given that more than half of users showed non-weekly rhythms, designers and researchers should also re-evaluate whether their investigation methods are providing users with sufficient alternative designs [100]. Previous works often ask users to critique systems while providing only weekly or monthly views of their behaviors [18, 22–24, 26, 45, 64, 98]. However, as we learned from our polyrhythmic users and the redesign process, imposing a 'normative' framework is limiting and sometimes even misleading because 'one size does not fit all' [39]. This sentiment is further corroborated by previous works in which users with 'non-normative' behavioral patterns feel frustrated by pre-existing activity trackers because they do not fully explain their lived experiences [19, 25, 33, 49, 51, 55, 64, 95].

Thus, when evaluating cyclic system designs with users, we should ensure that the study does not bias users by providing them with 'normative' designs. User preferences to 'normative' weekly designs may be an artifact of liking familiar designs and may not be a reflection of what the user truly finds helpful. To draw better critiques and make sure we are "getting the right design and the design right", as Tohidi et al. suggests [100], we should re-evaluate our systems with users by providing them with alternative designs that accommodate their natural routines.

We may not have to look far for alternative designs—we can revisit and examine the systems created for the 'extreme' ends of the user spectrum [33, 50, 51, 64]. For example, previous research on individuals with irregular vocations [64] and family routines [25], have already suggested new calendar systems informed by people's natural schedules. Such designs for 'niche' users may already be more flexible and accommodating to changing routines because they are built to support a wide range of variation in behaviors and evolving user needs [32, 60, 79]. By integrating these approaches into more mainstream designs, we can benefit a broader range of users, including those with 'normative' needs.

5.3 Designing Transparent Tools with Long-term Natural Routines

As evident in our long-term recorders, self-tracking data is no longer simple, but complex. Users who wish to gain insights or conduct data-driven self-experiments need new tools that allow them to process and investigate long-term patterns without the effects of data overload. One solution is to develop sophisticated blackbox models that identify insights and patterns, but such systems often strip users' agency over their data and suffer from their lack of explainability and transparency [4, 85]. From a practical standpoint, it is also challenging to enforce companies to be transparent about proprietary models.

A concrete, but simple solution to achieve transparency is through re-designs that account for a user's natural routine and explicitly communicating what behavioral assumptions were considered in the design. The visual dashboards of six widely used commercial tracking applications (i.e. Whoop, SleepCycle, Sleep as Android, Apple Health, Oura, Fitbit) [3, 36, 93, 94, 99, 104] provide view filters only on a calendar basis (e.g. daily, weekly, monthly, and yearly). This implicit design choice can leave users with non-standard schedules unaware of the limitations of such designs.

Thus, systems should inform their users what types of behaviors they can support so that users can make informed decisions or provide adaptive, behavior-based interfaces to cater to different cohorts of users [22]. For example, systems could ask users for their scheduling routine before using the application or provide flexible visualization dashboards that customize view filters based on user input. As we show in our re-design case, such design changes can reduce misinformation without introducing much complexity and suggest more meaningful insights, concrete action items, and 'durable changes' on how a user should modify their behavior [37, 81, 85]. Ultimately, these changes could lead to more equitable systems by providing 'non-normative' users access to behavioral information that 'normative' users by default have access to due to the weekly design of existing systems (Figure 10).

5.4 Opportunities for Holistic and Equitable Scheduling Systems

Analyzing different proxy signals opens opportunities to design new holistic and equitable scheduling systems. Scheduling is a complex, social coordination problem that requires users to acknowledge each other's differences and accommodate individual needs for all parties to feel fairness [101]. Our case analyses suggest we can design more equitable scheduling systems by considering a user's routine length, how many routines they are managing, and how long it takes for an individual to recover from a change in their routine. A more holistic view of the individual could provide a more nuanced understanding of whether two people's routines are compatible with each other and help create more equitable group-based scheduling strategies [9, 25, 101].

Furthermore, understanding routines through such proxy signals could pave the way for new holistic scheduling systems that integrate both the individual's internal and external behavioral schedules. Similar to how SleepTight found that non-routineness in external behaviors affected people's sleep routines [18], we could provide users with more context-aware recommendations to users by analyzing both their external, messaging patterns and internal, sleep routines. Menstruators can find whether their personal struggles to maintain a routine stem from a misalignment between their longer, natural rhythms and their weekly schedules. Individuals who are aging or in postpartum could start quantitatively identifying whether the onset of declining health is due to the breakdown of their original routines. Thus, users could become more interoceptive, study the *interplay* between their internal and external rhythms, and design new personalized schedules that are compatible with their biological rhythms.

5.5 New Methods to Investigate Long-term Behaviors and Routines

In this paper, we show *by example* that proxy signals from a commercial sleep app have fine-grained detail about peoples' natural routines. *We can start seeing people's routines in their natural settings, without some of the biases from controlled lab settings.* As we discuss in subsection 3.2, secondary data from an external platform often cannot provide context and causes behind different behaviors like other short-term studies [6, 25, 37], as the data collection began a decade ago. However, as we have shown through our analyses, such datasets still complement pre-existing knowledge, deepen our understanding of current design flaws and unmet user needs, and inform the design of future behavioral systems.

While not everyone has access to longitudinal sleep records, this method of investigation, using a proxy signal to investigate long-term natural routines, is not limited to just sleep data. This method can be applied to other secondary datasets such as messaging data or internet usage. While such datasets may not inherently be purposed to track, they inherently contain other behavioral information and can be repurposed to investigate natural routines. The tools for it already exist—Sochiatrist can extract users' pre-existing natural messaging data in a privacy-sensitive way [72] and GLOBEM contains multidimensional, long-term tracking data from cross-institutions [107]. Cross-dataset investigations could also inform new data-driven, long-term personas [20] which designers can utilize to investigate long-term user needs. Thus, investigating long-term behavioral data is no longer limited by the lack of data or feasibility [17, 47, 56, 57, 73], but rather by the approaches we choose and the trade-offs we are willing to accept.

5.6 Limitations

One of the major limitations in this study is that we are unable to further contextualize the users. Unlike controlled studies which can interview and survey their users, analysis of a naturalistic, secondary dataset whose collection started many years ago is unable to cross-check the reasons behind behavioral changes. Thus, this study serves to complement pre-existing knowledge, not to replace it. In addition, because Sleep as Android allows users to record sleep data with the method and device of their choice, broader insights on the efficacy of specific sleep technologies is limited. Further investigations are needed to understand associations with user engagement and recording behaviors.

There could also exist an unmeasurable bias from investigating a subset of users with records at least 120 days long, especially if regularity in recording is associated with regularity in sleep. Thus using weekly $\lambda = 7$ routine as a baseline for future studies should be qualified. Even the subset of users who find, download, and regularly use the app are not a random sample of people who sleep, but rather people with some inherent interest in understanding or improving their sleep, limiting the ability for us to generalize any findings from the dataset.

6 Conclusion

By studying a large sample of sleepers for long periods of their lives, we have identified both common regularity patterns in some sleepers, as well as different variations of regularity in many other sleepers. The weekly pattern is common, where the day of week moderates when someone sleeps, but there are cycles less than a week and cycles more than week. Some individuals have shown to have asymmetrical cycles (wave-like patterns), shifting their sleep later over a couple of weeks before resetting back to the original sleep time. However, even for a single individual, the cycle of regularity can change over the period that they track, as we have seen clear transitions between one pattern of sleep to a new pattern, with different variance and sleep times, after a few months. As we have noted, changing sleep is normal behavior. So the characterization of regularity for someone who monitors their sleep for a while depends both on the time period in their life and the length of cycle, λ, in the analysis.

Using technology for long-term sleep tracking has allowed us to observe a broader existence of regularity. Apart from the clinical definition, differences between consecutive days or single days compared to an average, there is a concept of repeating cycles. And even those repeating cycles can change into a different set of repeating cycles, over a person's lifetime, making sleep tracking more interesting as a long-term affair. As we design technology for sleep, these lessons reveal that sleep is rarely a normative behavior, so taking a broader view of cyclical patterns can better adapt the technology for our complex lives.

Acknowledgments

This research is supported in part by the National Institutes of Health under R01MH124832 and R01MH135499. We thank the Sleep as Android team for providing us with the dataset and all the users who shared their data to SleepCloud for research, thereby making this study possible. We thank David H. Laidlaw and James Tompkin for their initial guidance, and Ishaan Agarwal and Katherine M. Kinnaird for earlier analyses that led to the research questions in this paper. We thank Anita de Mello Koch, Saket Tiwari, Rafael Rodriguez-Sanchez, Ashley Kwon, Zhicheng Huang, Justus Adam, Elijah Rivera, Zainab Iftikhar, Tongyu Zhou, Talie Massachi, and Nicole Nugent for their formative discussions and edits. We thank all members of the Brown HCI lab for their insightful feedback.

References

[1] Saeed Abdullah, Mark Matthews, Elizabeth L. Murnane, Geri Gay, and Tanzeem Choudhury. 2014. Towards circadian computing: "early to bed and early to rise" makes some of us unhealthy and sleep deprived. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Seattle, Washington) *(UbiComp '14)*. Association for Computing Machinery, New York, NY, USA, 673–684. doi:10.1145/2632048.2632100

[2] Saeed Abdullah, Elizabeth L. Murnane, Mark Matthews, Matthew Kay, Julie A. Kientz, Geri Gay, and Tanzeem Choudhury. 2016. Cognitive rhythms: unobtrusive and continuous sensing of alertness using a mobile phone. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 178–189. doi:10.1145/2971648.2971712

[3] Apple. 2024. *Track your sleep with Apple Watch*. Apple Watch User Guide. https://support.apple.com/guide/watch/sleep-apd830528336/watchos

[4] Erna Sif Arnardottir, Anna Sigridur Islind, and María Óskarsdóttir. 2021. The future of sleep measurements: a review and perspective. *Sleep medicine clinics* 16, 3 (2021), 447–464. doi:10.1016/j.jsmc.2021.05.004

[5] Yin Bai, Bin Xu, Yuanchao Ma, Guodong Sun, and Yu Zhao. 2012. Will you have a good sleep tonight? Sleep Quality Prediction with Mobile Phone. In *Proceedings of the 7th International Conference on Body Area Networks* (Oslo, Norway) *(BodyNets '12)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, BEL, 124–130. doi:10.4108/icst.bodynets.2012.250091

[6] Nikola Banovic, Tofi Buzali, Fanny Chevalier, Jennifer Mankoff, and Anind K. Dey. 2016. Modeling and Understanding Human Routine Behavior. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 248–260. doi:10.1145/2858036.2858557

[7] Argelinda Baroni, Jean-Marie Bruzzese, Christina A. Di Bartolo, and Jess P. Shatkin. 2016. Fitbit Flex: an unreliable device for longitudinal sleep measures in a non-clinical population. *Sleep and Breathing* 20 (2016), 853–854. doi:10.1007/s11325-015-1271-2

[8] Eric P.S. Baumer, Sherri Jean Katz, Jill E. Freeman, Phil Adams, Amy L. Gonzales, John Pollak, Daniela Retelny, Jeff Niederdeppe, Christine M. Olson, and Geri K. Gay. 2012. Prescriptive persuasion and open-ended social awareness: expanding the design space of mobile health. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. Association for Computing Machinery, New York, NY, USA, 475–484. doi:10.1145/2145204.2145279

[9] James "Bo" Begole, John C. Tang, and Rosco Hill. 2003. Rhythm modeling, visualizations and applications. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology* (Vancouver, Canada) *(UIST '03)*. Association for Computing Machinery, New York, NY, USA, 11–20. doi:10.1145/964696.964698

[10] Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L. Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L. Jackson, Michelle A. Williams, and Susan Redline. 2015. Racial/Ethnic Differences in Sleep Disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep* 38, 6 (06 2015), 877–888. doi:10.5665/sleep.4732 arXiv:https://academic.oup.com/sleep/article-pdf/38/6/877/26671856/aasm.38.6.877.pdf

[11] Yu-Peng Chen, Julia Woodward, Dinank Bista, Xuanpu Zhang, Ishvina Singh, Oluwatomisin Obajemu, Meena N. Shankar, Kathryn M. Ross, Jaime Ruiz, and Lisa Anthony. 2024. Investigating Contextual Notifications to Drive Self-Monitoring in mHealth Apps for Weight Maintenance. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 993, 21 pages. doi:10.1145/3613904.3641993

[12] Zhenyu Chen, Mu Lin, Fanglin Chen, Nicholas D. Lane, Giuseppe Cardone, Rui Wang, Tianxing Li, Yiqiang Chen, Tanzeem Choudhury, and Andrew T. Campbell. 2013. Unobtrusive sleep monitoring using smartphones. In *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare* (Venice, Italy) *(PervasiveHealth '13)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, BEL, 145–152. doi:10.4108/icst.pervasivehealth.2013.252148

[13] Anna Cherenshchykova and Andrew D. Miller. 2019. Family-Based Sleep Technologies: Opportunities and Challenges. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3290607.3312907

[14] Yi-Shyuan Chiang, Ruei-Che Chang, Yi-Lin Chuang, Shih-Ya Chou, Hao-Ping Lee, I-Ju Lin, Jian-Hua Jiang Chen, and Yung-Ju Chang. 2020. Exploring the Design Space of User-System Communication for Smart-home Routine Assistants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376501

[15] Evan D Chinoy, Joseph A Cuellar, Kirbie E Huwa, Jason T Jameson, Catherine H Watson, Sara C Bessman, Dale A Hirsch, Adam D Cooper, Sean PA Drummond, and Rachel R Markwald. 2021. Performance of seven consumer sleep-tracking devices compared with polysomnography. *Sleep* 44, 5 (2021), zsaa291. doi:10.1093/sleep/zsaa291

[16] Eun Kyoung Choe. 2011. Design of persuasive technologies for healthy sleep behavior. In *Proceedings of the 13th International Conference on Ubiquitous Computing* (Beijing, China) *(UbiComp '11)*. Association for Computing Machinery, New York, NY, USA, 507–510. doi:10.1145/2030112.2030193

[17] Eun Kyoung Choe, Sunny Consolvo, Nathaniel F. Watson, and Julie A. Kientz. 2011. Opportunities for computing technologies to support healthy sleep behaviors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 3053–3062. doi:10.1145/1978942.1979395

[18] Eun Kyoung Choe, Bongshin Lee, Matthew Kay, Wanda Pratt, and Julie A. Kientz. 2015. SleepTight: low-burden, self-monitoring technology for capturing and reflecting on sleep behaviors. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 121–132. doi:10.1145/2750858.2804266

[19] Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1143–1152. doi:10.1145/2556288.2557372

[20] Alan Cooper, Robert Reimann, David Cronin, and Christopher Noessel. 2014. *About Face: The Essentials of Interaction Design* (4th ed.). Wiley Publishing, Indianapolis, IN, USA.

[21] Nediyana Daskalova, Eindra Kyi, Kevin Ouyang, Arthur Borem, Sally Chen, Sung Hyun Park, Nicole Nugent, and Jeff Huang. 2021. Self-E: Smartphone-Supported Guidance for Customizable Self-Experimentation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 227, 13 pages. doi:10.1145/3411764.3445100

[22] Nediyana Daskalova, Bongshin Lee, Jeff Huang, Chester Ni, and Jessica Lundin. 2018. Investigating the Effectiveness of Cohort-Based Sleep Recommendations. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 101 (Sept. 2018), 19 pages. doi:10.1145/3264911

[23] Nediyana Daskalova, Danaë Metaxa-Kakavouli, Adrienne Tran, Nicole Nugent, Julie Boergers, John McGeary, and Jeff Huang. 2016. SleepCoacher: A Personalized Automated Self-Experimentation System for Sleep Recommendations. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 347–358. doi:10.1145/2984511.2984534

[24] Nediyana Daskalova, Jina Yoon, Yibing Wang, Cintia Araujo, Guillermo Beltran, Nicole Nugent, John McGeary, Joseph Jay Williams, and Jeff Huang. 2020. SleepBandits: Guided Flexible Self-Experiments for Sleep. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376584

[25] Scott Davidoff, John Zimmerman, and Anind K. Dey. 2010. How routine learners can support family coordination. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 2461–2470. doi:10.1145/1753326.1753699

[26] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 3267–3276. doi:10.1145/2470654.2466447

[27] Anind K. Dey. 2001. Understanding and Using Context. *Personal Ubiquitous Comput.* 5, 1 (Jan. 2001), 4–7. doi:10.1007/s007790170019

[28] Gemma Di Federico and Andrea Burattin. 2023. Do You Behave Always the Same?. In *Process Mining Workshops*, Marco Montali, Arik Senderovich, and Matthias Weidlich (Eds.). Springer Nature Switzerland, Cham, 5–17. doi:10.1007/978-3-031-27815-0_1

[29] Gemma Di Federico, Carlos Fernández-Llatas, Zahra Ahmadi, Mohsen Shirali, and Andrea Burattin. 2024. Identifying Variation in Personal Daily Routine Through Process Mining: A Case Study. In *Process Mining Workshops*, Johannes De Smedt and Pnina Soffer (Eds.). Springer Nature Switzerland, Cham, 223–234. doi:10.1007/978-3-031-56107-8_17

[30] David A Dickey and Wayne A Fuller. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association* 74, 366a (1979), 427–431. doi:10.2307/2286348

[31] Daniel A. Epstein, Monica Caraway, Chuck Johnston, An Ping, James Fogarty, and Sean A. Munson. 2016. Beyond Abandonment to Next Steps: Understanding and Designing for Life after Personal Informatics Tool Use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1109–1113. doi:10.1145/2858036.2858045

[32] Daniel A. Epstein, Jennifer H. Kang, Laura R. Pina, James Fogarty, and Sean A. Munson. 2016. Reconsidering the device in the drawer: lapses as a design opportunity in personal informatics. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 829–840. doi:10.1145/2971648.2971656

[33] Daniel A. Epstein, Nicole B. Lee, Jennifer H. Kang, Elena Agapie, Jessica Schroeder, Laura R. Pina, James Fogarty, Julie A. Kientz, and Sean Munson. 2017. Examining Menstrual Tracking to Inform the Design of Personal Informatics Tools. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6876–6888. doi:10.1145/3025453.3025635

[34] Kelly R Evenson, Michelle M Goto, and Robert D Furberg. 2015. Systematic review of the validity and reliability of consumer-wearable activity trackers. *International Journal of Behavioral Nutrition and Physical Activity* 12 (2015), 1–22. doi:10.1186/s12966-015-0314-1

[35] Dorothee Fischer, Elizabeth B. Klerman, and Andrew J.K. Phillips. 2021. Measuring sleep regularity: theoretical properties and practical usage of existing metrics. *Sleep* 44, 10 (Oct. 2021), zsab103. doi:10.1093/sleep/zsab103

[36] Fitbit. 2024. *Understand the Impact of Your Sleep.* Google. https://www.fitbit.com/global/us/technology/sleep

[37] Thomas Fritz, Elaine M. Huang, Gail C. Murphy, and Thomas Zimmermann. 2014. Persuasive technology in the real world: a study of long-term use of activity sensing devices for fitness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 487–496. doi:10.1145/2556288.2557383

[38] Tian Hao, Guoliang Xing, and Gang Zhou. 2013. iSleep: unobtrusive sleep quality monitoring using smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems* (Roma, Italy) *(SenSys '13)*. Association for Computing Machinery, New York, NY, USA, Article 4, 14 pages. doi:10.1145/2517351.2517359

[39] Helen Ai He, Saul Greenberg, and Elaine M. Huang. 2010. One size does not fit all: applying the transtheoretical model to energy feedback technology design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 927–936. doi:10.1145/1753326.1753464

[40] Eric B. Hekler, Predrag Klasnja, Jon E. Froehlich, and Matthew P. Buman. 2013. Mind the theoretical gap: interpreting, using, and developing behavioral theory in HCI research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 3307–3316. doi:10.1145/2470654.2466452

[41] Chen-Yu Hsu, Aayush Ahuja, Shichao Yue, Rumen Hristov, Zachary Kabelac, and Dina Katabi. 2017. Zero-Effort In-Home Sleep and Insomnia Monitoring using Radio Signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 59 (Sept. 2017), 18 pages. doi:10.1145/3130924

[42] Jeff Huang and Jing Qian. 2023. irchiver: A Full-Resolution Personal Web Archive for Users and Researchers. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (Austin, TX, USA) *(CHIIR '23)*. Association for Computing Machinery, New York, NY, USA, 449–453. doi:10.1145/3576840.3578286

[43] Jeff Huang, Thomas Zimmermann, Nachiappan Nagapan, Charles Harrison, and Bruce C. Phillips. 2013. Mastering the art of war: how patterns of gameplay influence skill in Halo. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 695–704. doi:10.1145/2470654.2470753

[44] Rob J. Hyndman and George Athanasopoulos. 2023. *Forecasting: principles and practice.* OTexts, Melbourne, Australia.

[45] Alaul Islam, Ranjini Aravind, Tanja Blascheck, Anastasia Bezerianos, and Petra Isenberg. 2022. Preferences and Effectiveness of Sleep Data Visualizations for Smartwatches and Fitness Bands. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 27, 17 pages. doi:10.1145/3491102.3501921

[46] Hayeon Jeong, Heepyung Kim, Rihun Kim, Uichin Lee, and Yong Jeong. 2017. Smartwatch Wearing Behavior Analysis: A Longitudinal Study. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 60 (sep 2017), 31 pages. doi:10.1145/3131892

[47] Evangelos Karapanos, Jhilmil Jain, and Marc Hassenzahl. 2012. Theories, methods and case studies of longitudinal HCI research. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI EA '12)*. Association for Computing Machinery, New York, NY, USA, 2727–2730. doi:10.1145/2212776.2212706

[48] Evangelos Karapanos, John Zimmerman, Jodi Forlizzi, and Jean-Bernard Martens. 2009. User experience over time: an initial framework. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 729–738. doi:10.1145/1518701.1518814

[49] Kasper Karlgren, Barry Brown, and Donald McMillan. 2022. From Self-Tracking to Sleep-Hacking: Online Collaboration on Changing Sleep. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 517 (Nov. 2022), 26 pages. doi:10.1145/3555630

[50] Kasper Karlgren and Donald McMillan. 2022. Designing for Extreme Sleepers: Rethinking the Rhythms of Sleep Technology. In *Nordic Human-Computer Interaction Conference* (Aarhus, Denmark) *(NordiCHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 62, 17 pages. doi:10.1145/3546155.3546685

[51] Kasper Karlgren and Donald Mcmillan. 2023. Sleep Planning with Awari: Uncovering the Materiality of Body Rhythms using Research through Design. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 686, 17 pages. doi:10.1145/3544548.3581502

[52] Machiko Katori, Shoi Shi, Koji L. Ode, Yasuhiro Tomita, and Hiroki R. Ueda. 2022. The 103,200-arm acceleration dataset in the UK Biobank revealed a landscape of human sleep phenotypes. *Proceedings of the National Academy of Sciences* 119, 12 (2022), e2116729119. doi:10.1073/pnas.2116729119

[53] Matthew Kay, Eun Kyoung Choe, Jesse Shepherd, Benjamin Greenstein, Nathaniel Watson, Sunny Consolvo, and Julie A. Kientz. 2012. Lullaby: a capture & access system for understanding the sleep environment. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (Pittsburgh, Pennsylvania) *(UbiComp '12)*. Association for Computing Machinery, New York, NY, USA, 226–234. doi:10.1145/2370216.2370253

[54] Matthew Kay, Shwetak N. Patel, and Julie A. Kientz. 2015. How Good is 85%? A Survey Tool to Connect Classifier Evaluation to Acceptability of Accuracy. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 347–356. doi:10.1145/2702123.2702603

[55] Bogoan Kim, Sung-In Kim, Sangwon Park, Hee Jeong Yoo, Hwajung Hong, and Kyungsik Han. 2023. RoutineAid: Externalizing Key Design Elements to Support Daily Routines of Individuals with Autism. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 333, 18 pages. doi:10.1145/3544548.3581048

[56] Maria Kjærup, Mikael B. Skov, Peter Axel Nielsen, Jesper Kjeldskov, Jens Gerken, and Harald Reiterer. 2021. Longitudinal Studies in HCI Research: A Review of CHI Publications From 1982–2019. In *Advances in Longitudinal HCI Research*, Evangelos Karapanos, Jens Gerken, Jesper Kjeldskov, and Mikael B. Skov (Eds.). Springer International Publishing, Cham, 11–39. doi:10.1007/978-3-030-67322-2_2

[57] Predrag Klasnja, Sunny Consolvo, and Wanda Pratt. 2011. How to evaluate technologies for health behavior change in HCI research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 3063–3072. doi:10.1145/1978942.1979396

[58] Ping-Ru T Ko, Julie A Kientz, Eun Kyoung Choe, Matthew Kay, Carol A Landis, and Nathaniel F Watson. 2015. Consumer sleep technologies: a review of the landscape. *Journal of clinical sleep medicine* 11, 12 (2015), 1455–1461. doi:10.5664/jcsm.5288

[59] Andrew G. Kubala, Bethany Barone Gibbs, Daniel J. Buysse, Sanjay R. Patel, Martica H. Hall, and Christopher E. Kline. 2020. Field-based measurement of sleep: agreement between six commercial activity monitors and a validated accelerometer. *Behavioral sleep medicine* 18, 5 (2020), 637–652. doi:10.1080/15402002.2019.1651316

[60] Sari Kujala, Virpi Roto, Kaisa Väänänen-Vainio-Mattila, Evangelos Karapanos, and Arto Sinnelä. 2011. UX Curve: A method for evaluating long-term user experience. *Interacting with computers* 23, 5 (2011), 473–483. doi:10.1016/j.

intcom.2011.06.005

[61] Denis Kwiatkowski, Peter C.B. Phillips, Peter Schmidt, and Yongcheol Shin. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics* 54, 1-3 (1992), 159–178. doi:10.1016/0304-4076(92)90104-Y

[62] Amanda Lazar, Christian Koehler, Theresa Jean Tanenbaum, and David H. Nguyen. 2015. Why we use and abandon smart devices. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 635–646. doi:10.1145/2750858.2804288

[63] Jonathan K. Lazar and Anthony F. Norcio. 2000. System and training design for end-user error. In *Human centered methods in information systems: Current research and practice*. IGI Global, Hershey, PA, USA, 76–90. doi:10.4018/9781878289643.ch005

[64] Jungeun Lee, Sungnam Kim, Minki Cheon, Hyojin Ju, JaeEun Lee, and Inseok Hwang. 2022. SleepGuru: Personalized Sleep Planning System for Real-life Actionability and Negotiability. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22, 52)*. Association for Computing Machinery, New York, NY, USA, Article 52, 16 pages. doi:10.1145/3526113.3545709

[65] Mirim Lee, Bon-chang Koo, Hee-seok Jeong, Joongsin Park, Juhee Cho, and Jundong Cho. 2015. Understanding Women's Needs in Menopause for Development of mHealth. In *Proceedings of the 2015 Workshop on Pervasive Wireless Healthcare* (Hangzhou, China) *(MobileHealth '15)*. Association for Computing Machinery, New York, NY, USA, 51–56. doi:10.1145/2757290.2757295

[66] Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2011. Mining behavioral economics to design persuasive technology for healthy choices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 325–334. doi:10.1145/1978942.1978989

[67] Zilu Liang and Bernd Ploderer. 2020. How Does Fitbit Measure Brainwaves: A Qualitative Study into the Credibility of Sleep-tracking Technologies. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 17 (mar 2020), 29 pages. doi:10.1145/3380994

[68] Chen Liu, Jie Xiong, Lin Cai, Lin Feng, Xiaojiang Chen, and Dingyi Fang. 2019. Beyond Respiration: Contactless Sleep Sound-Activity Recognition Using RF Signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 96 (sep 2019), 22 pages. doi:10.1145/3351254

[69] Jessica Lunsford-Avery, Matthew Engelhard, Ann Navar, and Scott Kollins. 2018. Validation of the Sleep Regularity Index in Older Adults and Associations with Cardiometabolic Risk. *Scientific Reports* 8 (09 2018). doi:10.1038/s41598-018-32402-5

[70] Lena Mamykina, Elizabeth Mynatt, Patricia Davidson, and Daniel Greenblatt. 2008. MAHI: investigation of social scaffolding for reflective thinking in diabetes management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 477–486. doi:10.1145/1357054.1357131

[71] Gloria Mark, Yiran Wang, Melissa Niiya, and Stephanie Reich. 2016. Sleep Debt in Student Life: Online Attention Focus, Facebook, and Mood. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5517–5528. doi:10.1145/2858036.2858437

[72] Talie Massachi, Grant Fong, Varun Mathur, Sachin R. Pendse, Gabriela Hoefer, Jessica J. Fu, Chong Wang, Nikita Ramoji, Nicole R. Nugent, Megan L. Ranney, Daniel P. Dickstein, Michael F. Armey, Ellie Pavlick, and Jeff Huang. 2020. Sochiatrist: Signals of Affect in Messaging Data. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 111 (Oct. 2020), 25 pages. doi:10.1145/3415182

[73] Valerie Mendoza and David G. Novick. 2005. Usability over time. In *Proceedings of the 23rd Annual International Conference on Design of Communication: Documenting & Designing for Pervasive Information* (Coventry, United Kingdom) *(SIGDOC '05)*. Association for Computing Machinery, New York, NY, USA, 151–158. doi:10.1145/1085313.1085348

[74] Jochen Meyer, Merlin Wasmann, Wilko Heuten, Abdallah El Ali, and Susanne C.J. Boll. 2017. Identification and Classification of Usage Patterns in Long-Term Activity Tracking. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 667–678. doi:10.1145/3025453.3025690

[75] Jun-Ki Min, Afsaneh Doryab, Jason Wiese, Shahriyar Amini, John Zimmerman, and Jason I. Hong. 2014. Toss 'n' turn: smartphone as sleep and sleep quality detector. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 477–486. doi:10.1145/2556288.2557220

[76] Thomas P. Moran, Patrick Chiu, Steve Harrison, Gordon Kurtenbach, Scott Minneman, and William van Melle. 1996. Evolutionary engagement in an ongoing collaborative work process: a case study. In *Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work* (Boston, Massachusetts, USA) *(CSCW '96)*. Association for Computing Machinery, New York, NY, USA, 150–159. doi:10.1145/240080.240243

[77] Dan Morris, T. Scott Saponas, Andrew Guillory, and Ilya Kelner. 2014. RecoFit: using a wearable sensor to find, recognize, and count repetitive exercises. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3225–3234. doi:10.1145/2556288.2557116

[78] Elizabeth L. Murnane, Saeed Abdullah, Mark Matthews, Tanzeem Choudhury, and Geri Gay. 2015. Social (media) jet lag: how usage of social technology can modulate and reflect circadian rhythms. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 843–854. doi:10.1145/2750858.2807522

[79] Jasmin Niess and Paweł W. Woźniak. 2018. Supporting Meaningful Personal Fitness: the Tracker Goal Evolution Model. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3173745

[80] Donald A. Norman. 1983. Design rules based on analyses of human error. *Commun. ACM* 26, 4 (April 1983), 254–258. doi:10.1145/2163.358092

[81] Oladapo Oyebode, Chinenye Ndulue, Dinesh Mulchandani, Ashfaq A. Zamil Adib, Mona Alhasani, and Rita Orji. 2021. Tailoring Persuasive and Behaviour Change Systems Based on Stages of Change and Motivation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 703, 19 pages. doi:10.1145/3411764.3445619

[82] Jessica R. Peters, Katja M. Schmalenberger, Ashley G. Eng, Allison Stumper, Michelle M. Martel, and Tory A. Eisenlohr-Moul. 2024. Dimensional Affective Sensitivity to Hormones across the Menstrual Cycle (DASH-MC): A transdiagnostic framework for ovarian steroid influences on psychopathology. *Molecular Psychiatry* 30 (2024), 1–12. doi:10.1038/s41380-024-02693-4

[83] Andrew J.K. Phillips, William M. Clerx, Conor S. O'Brien, Akane Sano, Laura K. Barger, Rosalind W. Picard, Steven W. Lockley, Elizabeth B. Klerman, and Charles A. Czeisler. 2017. Irregular sleep/wake patterns are associated with poorer academic performance and delayed circadian and sleep/wake timing. *Scientific reports* 7, 1 (2017), 3216. doi:10.1038/s41598-017-03171-4

[84] Tauhidur Rahman, Alexander T. Adams, Ruth Vinisha Ravichandran, Mi Zhang, Shwetak N. Patel, Julie A. Kientz, and Tanzeem Choudhury. 2015. DoppleSleep: a contactless unobtrusive sleep sensing system using short-range Doppler radar. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 39–50. doi:10.1145/2750858.2804280

[85] Ruth Ravichandran, Sang-Wha Sien, Shwetak N. Patel, Julie A. Kientz, and Laura R. Pina. 2017. Making Sense of Sleep Sensors: How Sleep Sensing Technologies Support and Undermine Sleep Health. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6864–6875. doi:10.1145/3025453.3025557

[86] Lara Reime, Vasiliki Tsaknaki, and Marisa Leavitt Cohn. 2023. Walking Through Normativities of Reproductive Bodies: A Method for Critical Analysis of Tracking Applications. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 658, 15 pages. doi:10.1145/3544548.3581450

[87] Kerry Rodden, Hilary Hutchinson, and Xin Fu. 2010. Measuring the user experience on a large scale: user-centered metrics for web applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 2395–2398. doi:10.1145/1753326.1753687

[88] John Rooksby, Mattias Rost, Alistair Morrison, and Matthew Chalmers. 2014. Personal tracking as lived informatics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1163–1172. doi:10.1145/2556288.2557039

[89] Jim Rowan and Elizabeth D. Mynatt. 2005. Digital Family Portrait Field Trial: Support for Aging in Place. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. Association for Computing Machinery, New York, NY, USA, 521–530. doi:10.1145/1054972.1055044

[90] Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*. SciPy, Austin, TX, 10–25080. doi:10.25080/Majora-92bf1922-011

[91] Ji Youn Shin, Wei Peng, and Hee Rin Lee. 2022. More than Bedtime and the Bedroom: Sleep Management as a Collaborative Work for the Family. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 234, 16 pages. doi:10.1145/3491102.3517535

[92] Irina Shklovski, Robert Kraut, and Jonathon Cummings. 2006. Routine patterns of internet use & psychological well-being: coping with a residential move. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*

(Montréal, Québec, Canada) *(CHI '06)*. Association for Computing Machinery, New York, NY, USA, 969–978. doi:10.1145/1124772.1124917

[93] Sleep as Android. 2024. *Sleep as Android*. Urbandroid. https://sleep.urbandroid.org/

[94] Sleep Cycle. 2024. *How Sleep Cycle Works*. Sleep Cycle. https://www.sleepcycle.com/how-sleep-cycle-works/

[95] Marie Louise Juul Søndergaard, Marianela Ciolfi Felice, and Madeline Balaam. 2021. Designing Menstrual Technologies with Adolescents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 260, 14 pages. doi:10.1145/3411764.3445471

[96] Katarzyna Stawarz, Anna L. Cox, and Ann Blandford. 2015. Beyond Self-Tracking and Reminders: Designing Smartphone Apps That Support Habit Formation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2653–2662. doi:10.1145/2702123.2702230

[97] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS medicine* 12 (03 2015), e1001779. doi:10.1371/journal.pmed.1001779

[98] Lie Ming Tang and Judy Kay. 2017. Harnessing Long Term Physical Activity Data—How Long-term Trackers Use Data and How an Adherence-based Interface Supports New Insights. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2, Article 26 (June 2017), 28 pages. doi:10.1145/3090091

[99] Oura Team. 2024. *Using Trends View on Oura*. Oura Health. https://ouraring.com/blog/trends

[100] Maryam Tohidi, William Buxton, Ronald Baecker, and Abigail Sellen. 2006. Getting the right design and the design right. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI '06)*. Association for Computing Machinery, New York, NY, USA, 1243–1252. doi:10.1145/1124772.1124960

[101] Alarith Uhde, Nadine Schlicker, Dieter P. Wallach, and Marc Hassenzahl. 2020. Fairness and Decision-making in Collaborative Shift Scheduling Systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376656

[102] Tjeu van Bussel, Roy van den Heuvel, and Carine Lallemand. 2022. Habilyzer: Empowering Office Workers to Investigate their Working Habits using an Open-Ended Sensor Kit. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 264, 8 pages. doi:10.1145/3491101.3519849

[103] Varun K. Viswanath, Wendy Hartogenesis, Stephan Dilchert, Leena Pandya, Frederick M. Hecht, Ashley E. Mason, Edward J. Wang, and Benjamin L. Smarr. 2024. Five million nights: temporal dynamics in human sleep phenotypes. *npj Digital Medicine* 7, 1 (2024), 150. doi:10.1038/s41746-024-01125-5

[104] WHOOP. 2024. *Metrics for Human Performance*. WHOOP. https://www.whoop.com/us/en/the-data/#sleep

[105] Daniel P. Windred, Angus C. Burns, Jacqueline M. Lane, Richa Saxena, Martin K. Rutter, Sean W. Cain, and Andrew J.K. Phillips. 2024. Sleep regularity is a stronger predictor of mortality risk than sleep duration: A prospective cohort study. *Sleep* 47, 1 (2024), zsad253. doi:10.1093/sleep/zsad253

[106] Daniel P. Windred, Samuel E. Jones, Alex Russell, Angus C. Burns, Philip Chan, Michael N. Weedon, Martin K. Rutter, Patrick Olivier, Céline Vetter, Richa Saxena, Jacqueline M. Lane, Sean W. Cain, and Andrew J.K. Phillips. 2021. Objective assessment of sleep regularity in 60 000 UK Biobank participants using an open-source package. *Sleep* 44, 12 (2021), zsab254. doi:10.1093/sleep/zsab254

[107] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. 2023. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 190 (jan 2023), 34 pages. doi:10.1145/3569485

[108] Stephanie R. Yiallourou, Lachlan Cribb, Marina G. Cavuoto, Ella Rowsthorn, Jessica Nicolazzo, Madeline Gibson, Andrée-Ann Baril, and Matthew P. Pase. 2024. Association of the sleep regularity index with incident dementia and brain volume. *Neurology* 102, 2 (2024), e208029. doi:10.1212/WNL.0000000000208029

[109] Robin K. Yuan, Kirsi-Marja Zitting, Liyaan Maskati, and Jeff Huang. 2022. Increased sleep duration and delayed sleep timing during the COVID-19 pandemic. *Scientific Reports* 12, 1 (2022), 1–9. doi:10.1038/s41598-022-14782-x

[110] Shichao Yue, Yuzhe Yang, Hao Wang, Hariharan Rahul, and Dina Katabi. 2020. BodyCompass: Monitoring Sleep Posture with Wireless Signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 66 (June 2020), 25 pages. doi:10.1145/3397311

[111] Bing Zhai, Ignacio Perez-Pozuelo, Emma A. D. Clifton, Joao Palotti, and Yu Guan. 2020. Making Sense of Sleep: Multimodal Sleep Stage Classification in a Large, Diverse Population Using Movement and Cardiac Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 67 (June 2020), 33 pages. doi:10.1145/3397325

[112] Renwen Zhang, Kathryn E. Ringland, Melina Paan, David C. Mohr, and Madhu Reddy. 2021. Designing for Emotional Well-being: Integrating Persuasion and Customization into Mental Health Technologies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 542, 13 pages. doi:10.1145/3411764.3445771

[113] Tongyu Zhou, Joshua Kong Yang, Vivian Hsinyueh Chan, Ji Won Chung, and Jeff Huang. 2024. PortalInk: 2.5D Visual Storytelling with SVG Parallax and Waypoint Transitions. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, 16 pages. doi:10.1145/3654777.3676376