Learning predictive models from menstrual cycle data

Kathy Li

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022

Abstract

Learning predictive models from menstrual cycle data

Kathy Li

Despite being a physiological phenomenon that impacts billions of womxn worldwide, menstruation has long been understudied. In this dissertation, we first explore the menstrual characteristics of nearly 380,000 womxn, as collected via a self-tracking mobile health (mHealth) app, Clue. We examine how variation in menstrual cycle length is related to volatility in other experienced symptoms, helping to debunk the idea that menstrual cycles should be 'regular.' We then develop predictive models for menstruation utilizing this dataset, demonstrating first how a fully generative model that explicitly accounts for the possibility that self-tracked data may be flawed in terms of reliability can both outperform baselines and aid in the detection of self-tracking artifacts (i.e., instances where a user supposedly did not experience a period event, but in reality forgot or otherwise neglected to track it). Finally, we explore a hierarchical, deep generative model for symptom tracking, where we utilize a deep neural network to learn per-user parameters for tracking and retain a mechanism for modeling per-user likelihood of adherence. We find that leveraging symptom data at the time series level allows us to predict occurrence of next bleeding and non-bleeding tracking events with high accuracy. This work demonstrates the great potential that large-scale mHealth data holds to better understanding menstruation as a whole, as well as the importance of treating such data carefully.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am very fortunate to have a strong support system that has assisted me throughout this journey.

I would first like to thank my collaborators and advisors, whose knowledge, insight, and generosity made this dissertation possible. Thank you first and foremost to my advisor, Prof. Chris Wiggins, for always offering an honest opinion, helping me organize my (often) scattered thoughts, and encouraging me to take breaks. I am lucky to have had an advisor whom I also consider a friend. Next, thank you to Dr. Inigo Urteaga for his patience, willingness and ability to explain things in a way that just seems to 'click' in my brain, and his commiseration during times of stress. Finally, thank you to my dissertation committee, Prof. Kyle Mandli, Prof. Marc Spiegelman, and Prof. Noemie Elhadad for their words of encouragement, useful insights, and flexibility.

Secondly, I would like to express my immense gratitude to my parents, whose fierce belief in my ability to do anything has guided me throughout my life, and especially when I faced setbacks and difficulties in my research process. Thank you mom and dad for always supporting me, encouraging me, and checking in (and for offering to proofread).

Next, I would like to thank the friends, old and new, whose paths crossed with mine in the last five years. The friends who listened to my concerns and complaints, boosted me up in

times of doubt, and celebrated my wins, no matter how small. I love you all, and I couldn't have done it without you — in particular, thank you Linda and Lena, for being there always.

To cap it off, since I'm a little cheesy, I would like to thank the city of New York for being a constant source of inspiration, a literal and metaphorical breath of fresh air, and one of my favorite places in the world.

To Mom and Dad

Chapter 1

Introduction

Menstruation serves as an important health indicator for womxn [1], and is an experience that womxn are accustomed to anticipating and monitoring from a young age, learning about terms like PMS (premenstrual syndrome) and TSS (toxic shock syndrome) [6] and being prepared for when their period arrives each month. From adolescence, many womxn are often introduced to menstruation as a taboo and embarrassing topic to discuss and find their experiences of pain to be invalidated or ignored [7; 8]. At best, menstruation can be a confusing phenomenon to navigate; at worst, it can be disruptive, distressing, and a source of great shame [9; 10]. It affects physical, mental, and social health; it can indicate the presence of myriad conditions ranging from fertility issues [11; 12] and menopause [13; 14; 15] to cardiovascular disease [16]. In fact, menstruation is such a compelling source of insight for womxn that it has been hypothesized as "the fifth vital sign" [17; 18; 19; 20]. Beyond individual health insights, it raises broader concerns about contraception, education around how to use period products like pads and tampons (as well as access to such products), and fertility.

[1] In this dissertation, we refer to Clue users or menstruators with the term 'womxn,' which is often considered to be more gender-inclusive, acknowledging that not all menstruators are women and vice versa.

However, while menstruation is a large part of many peoples' lives and plays a key role in understanding womxn's health, it has continued to mystify researchers across different contexts and remains largely misunderstood and understudied. Historically, this neglect has occurred for a variety of reasons — among them, societal stigma associated with discussing menstruation candidly, the normalization of womxn's pain, insufficient knowledge related to menstrual physiology, and lack of access to large-scale, reliable datasets [21; 22] have limited advancements. Nearly 15,000 publications in the past decade related to seminal fluid can be found in PubMed; by comparison, only about 400 publications exist that mention menstrual blood [21]. The company Pantone released a new "period red" shade in 2020 meant to de-stigmatize menstruation; however, the effort fell short, receiving criticism for the bright red color failing to represent menstrual blood and the marketing, which inaccurately depicted a menstrual cup inside a uterus [23]. This stark difference in research interest for and broader understanding of male and female health conditions, even though menstruation affects half of the world's population, clearly demonstrates the importance of this field.

In particular, open questions relating to menstruation include how to improve inclusivity around menstruation, how to reliably characterize the length and nature of menstrual cycles, how menstruation relates to other aspects of mental and social wellness, and how to predict the occurrence of menstrual cycles. Each of these questions requires collaboration across fields — experts on the physiological nature of menstruation can inform assumptions around typical menstrual behavior, those who investigate the social impact of menstruation can provide context for the needs of menstruators, and quantitative researchers can develop models to assist in predicting various aspects of menstruation.

While this field has grown in popularity in recent years, it is still relatively new and offers

limitless opportunities for advancement. In particular, with the rise in usage of mobile tracking apps such as menstrual trackers that allow users to input information about their menstrual cycles, we now have access to data at a size and scale that was previously unavailable. Although this data offers great potential for in-depth, quantitative investigation of menstruation, it also holds its own reliability risks due to its self-tracked nature (for instance, users may not always track exactly what they experience, or may forget to track altogether). In this dissertation, I will explore how to develop accurate, interpretable, and flexible predictive models for menstruation, with a particular focus on how to consider the inherently unreliable nature of mobile health data.

Specifically, in Chapter 2 I will provide background on mobile health data and its potential to contribute to our knowledge of health behavior at-large; menstrual cycle definitions as they pertain to our self-tracked mobile health dataset from Clue, a popular menstrual health tracker; an overview of the Clue dataset; and machine learning, including graphical modeling as it relates to depicting generative statistical processes for observed data and a brief overview of deep learning. In Chapter 3, I will characterize our dataset, which spans millions of cycles and hundreds of thousands of users, paying special attention to developing a quantitative definition of menstrual cycle variability and helping to debunk the idea that menstruation should be 'regular.' In addition, I will demonstrate how variation in menstrual cycle length relates to variation in menstrual cycle symptoms.

In Chapters 4 and 5, I will move to proposing two different predictive models for menstruation. Starting with a fully generative model in Chapter 4, I will describe in more detail how inconsistent user adherence to self-tracking apps impacts data reliability and motivate the need to consider such behavior. I will then introduce a hierarchical, generative model for menstrual

cycle lengths that parameterizes separately per-user typical cycle length behavior and per-user self-tracking adherence, showcasing how this model outperforms baselines, particularly as the cycle proceeds, as well as how it can be practically applied to improve mobile health apps. In Chapter 5, I will introduce a second model that utilizes time series representations of the data (rather than menstrual cycle lengths) as input, spanning both period flow and other related qualitative symptoms. In contrast to the model in Chapter 4, this model will be a deep generative model, leveraging the power of deep learning to learn parameters related to symptom tracking in order to predict the occurrence of the next period (i.e., next cycle length) as well as the occurrence of future symptoms. As in the model introduced in Chapter 4, this model will also incorporate a hierarchical component for user adherence.

By providing an in-depth exploration of the dataset, as well as careful considerations of different types of predictive models, I will demonstrate how machine learning models serve as powerful tools to not only predict menstruation, but also to understand it. Such insights can benefit users, clinicians, app designers, and researchers across specialties in the field, and more broadly can inform those who work with self-tracked mobile health data.

Chapter 2

Background

In this chapter, we introduce relevant issues and considerations for utilizing mobile health data. In addition, we provide context for menstrual cycle definitions, which will be useful to understanding the dataset that we use, and present a summary of the dataset. Finally, we provide an overview of graphical modeling notation and definitions, as well as describe the structure and utility of deep learning models.

2.1 Mobile health data

The rise of data-powered health has enabled more nuanced, quantitative understanding of various health conditions and user behaviors. For instance, observational health data sources have shed light on individual clinical trajectories [24], increased self-awareness about individual health [25], and helped deliver on the promise of precision medicine [26]. Meanwhile, mobile health solutions have also enabled a high-resolution view of a large, highly diverse range of individuals over time [27; 28; 29; 30] and can provide insights into chronic diseases and behaviors [31; 32; 33; 34; 35; 36; 37; 38; 39; 40]. Menstrual trackers in particular have become increasingly common — they are the second most popular app for adolescent girls and the fourth most popular for adult womxn [41; 42] — meaning that millions of womxn around the

world now routinely track their menstrual cycles and a variety of contextual factors and symptoms, accumulating high volumes of temporal, heterogeneous data via several different apps [43; 44; 45; 46; 47]. This growth in access to menstrual health data has enabled researchers to identify menstrual patterns at scale and explore their relationships with a broad set of symptoms. Such research is exemplified by studies connecting the menstrual cycle to variations in womxn's mood, behavior, and vital signs [48], which showcase the insights that self-tracked data can provide into cycle characteristics [49], ovulation timing, and the evolution of reproductive health for large populations [50]. Furthermore, these insights can empower informed decision-making through increased self-awareness [51].

2.2 Menstrual cycle definitions

We define a **self-tracking event**, in the context of mobile health data, as an instance when a user logs a symptom in the app. Relatedly, period self-tracking events refer to instances when a user self-reports days where they have experienced period flow. We use such period tracking events to determine length of the menstrual **cycle**, which we define as the span of days from the first day of a period through to and including the day before the first day of the next period [4]. A **period** consists of sequential days of bleeding (greater than spotting and within ten days after the first greater-than-spotting bleeding event) unbroken by no more than one day on which only spotting or no bleeding occurred.

2.3 Data overview

We utilize a de-identified user-tracked dataset from Clue by BioWink [43], one of the most popular and accurate menstrual trackers worldwide [52]. Clue users can track period data and

symptom information in categories like exercise, pain, and sexual activity (see Figure 2.1). Note that Clue users input personal information at sign up, such as birth control usage and age, but are not required to specify gender; information on race or ethnicity is also not collected. This large-scale dataset provides a high resolution, long-term view of variation in both physiology (e.g., period and cycle duration) and symptoms (e.g., pain and mood) across menstrual cycles, enabling us to study the shared information between quantitative, temporal attributes and qualitative, symptomatic attributes of menstrual experiences.

Figure 2.1: Sample screenshots of the Clue app. Users can track daily symptoms across 20 categories. On the left, for example, the app displays what day the user is currently on in their cycle. On the right, a user can choose from 'cramps,' 'headache,' 'ovulation,' or 'tender breasts' symptoms for the category 'pain.'

Users are able to self-track their symptom experiences across 20 different categories, both directly related to period physiology like 'period flow' and not, like 'social activity'. These

categories are selected at sign up, and not all users track all categories. As described above, a 'self-tracking event' refers to an instance when a user logs an event by selecting a category, such as 'period flow,' and then choosing an associated symptom out of the available options ('light,' 'medium,' 'heavy,' or 'spotting,' in this example). Each row in the primitive dataset represents a tracked event e, with the relevant information being (i) the user u that tracked the event e_u, (ii) the reported symptom s associated with that event $e_u = s$, and (iii) the user-specific cycle c_e in which the event takes place.

2.4 Ethics

Since we utilize mobile health data in this work, it is imperative that we take ethical concerns into consideration. In particular, mobile health data contains individual-level health information that can be sensitive to the user. We have taken care to approach this data with respect for ethics, as defined by the Belmont Report principles for conducting research involving human subjects (namely, respect for persons, beneficence, and justice) [53]. Our data are de-identified, ensuring that sensitive information is not personally identifiable. Furthermore, our work seeks to benefit future users by providing insights that can have positive impact on their understanding of their menstrual cycles. We have worked closely with experts at Clue to ensure the data being used and the areas being investigated both serve this purpose without jeopardizing ethics for the user. Finally, the research presented here was exempt from Columbia University IRB approval, in accordance with 45CFR46.101(b), as all data are de-identified, and no participant risks are associated with taking part in the study. Although participants do not receive direct benefit from this study, their participation contributes to the general knowledge of menstrual cycles and their symptoms.

2.5 Machine learning

2.5.1 Graphical modeling

A graphical model [54] is a visual representation for explaining and reasoning about a probabilistic model, which outlines the relationships between latent variables, observed data, parameters, and hyperparameters. It provides a diagram for a generative model, which is a hypothesis for how data are generated (and can be used to generate synthetic data to, for instance, check the consistency of an inference or prediction method). We provide a simple example of a graphical model in Figure 2.2. In this type of representation, shaded circles represent observed data, open circles represent latent (unobserved) variables, and dots represent hyperparameters. Latent variables are the quantities we wish to infer from the data. The lines and arrows drawn between shapes represent conditional dependencies between variables. 'Plates' (the rectangular boxes) represent groups of variables which share the same repeated conditional dependence relations; for instance, a plate could represent many users or many instances of time for representing a temporal process. By following the arrows in a graphical model, one can see the hypothetical process by which the data are generated.

For instance, in Figure 2.2, θ represents the hyperparameter; λ_t represents the latent variable; and x_t represents the observed data. The plate around λ_t and x_t with the label T indicates that λ and x are indexed per-time, whereas one θ exists for the whole dataset. We can consider, for instance, that λ_t are drawn from a probability distribution hyperparameterized by θ. The lines and arrows indicate that x_t is dependent on λ_t, and λ_t is dependent on θ — the generative process begins with θ, from which λ_t is determined, and then x_t.

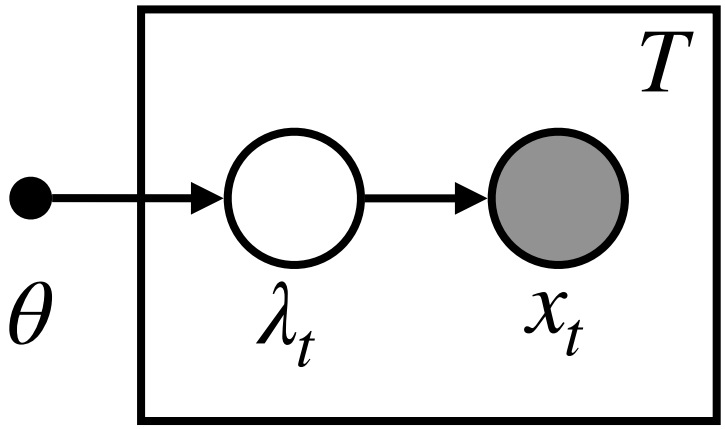


Figure 2.2: Simplified example of a graphical model, indicating the relationship between observed data x_t, latent variable λ_t, and hyperparameter θ. x_t and λ_t are replicated per t for $t = 1, \ldots T$.

2.5.2 Deep learning

Deep learning is a subset of machine learning that focuses specifically on progressively learning from input data to output by utilizing multilayer methods to process information [55]. In the context of this work, we focus on deep (artificial) neural networks, which are meant to simulate the way biological neural networks process information. Deep learning can be thought of in terms of arbitrary function approximation, or as a method for probabilistic inference [56]; in general, it refers to a set of methods for learning data by learning a set of weights for successive layers of non-linear transformations that lead from input to output.

There are many different architectures used in deep learning; we outline the key ones briefly below:

- **Multilayer perceptrons** (also called deep feedforward networks) [56], which are defined

as mappings from input to output where information is only passed forward from input, to intermediate functions, to the output. These networks typically utilize composition of multiple functions, where the first function is considered the first layer, the second is considered the second layer, and so on, with the final layer referred to as the output layer. The length of the chain of functions that are composed together is referred to as the **depth** of the network. Since the training data does not have the output for each of the individual functions (layers), these are referred to as **hidden layers**.

- **Convolutional neural networks** (CNNs) [57], which are defined as a set of networks typically used for grid-like data (for instance, images, which represent a 2-D grid of pixels). CNNs differ from other neural networks in that a convolution is used instead of a general matrix multiplication in at least one of their layers. A **convolution** is a mathematical operation defined as the integral of the product of two functions, where one is reversed and shifted. This integral is evaluated for all values of the shift, which produces the convolution function. For CNNs, the function that is shifted is referred to as the kernel and the output is referred to as the feature map. CNNs can utilize multi-dimensional kernels to process multi-dimensional data (like images). A layer of a CNN also often contains a pooling layer, which replaces the output of the layer with a summary statistic of nearby outputs (such as the maximum or average of a specific rectangular neighborhood).

- **Recurrent neural networks** [58; 59] (RNNs), which are networks developed for sequential data and differ from feedforward networks in that they utilize iterative function loops to process information. The key idea is that at each hidden layer, not only is

the current external input utilized, but also the activations from the hidden layer of the previous timestep. That is, whereas feedforward networks can only map from input to output, RNNs can map from the entire history of the input to the output, allowing for the network to retain '**memory**' of previous inputs (since prior elements influence the output). Gated RNNs, like long short-term memory networks [60] (LSTMs), allow for accumulation of information over a longer duration.

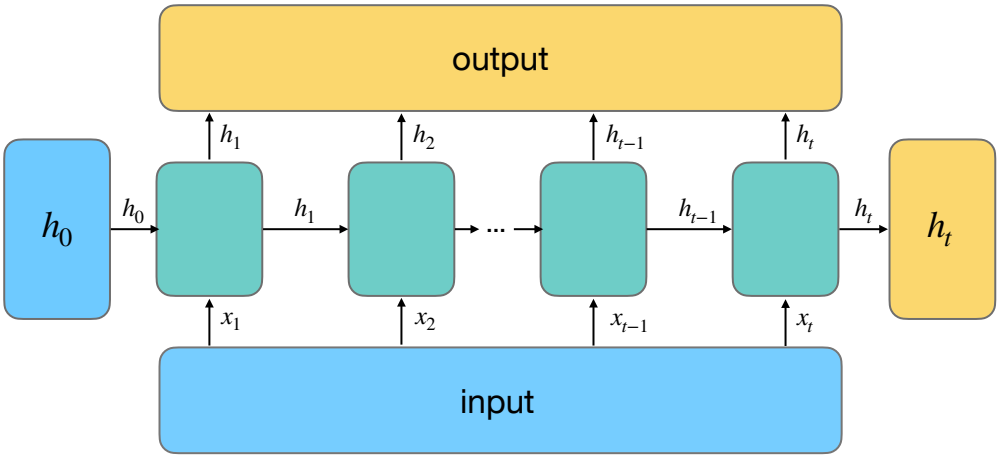


Figure 2.3: Sample graphic for a one-layer RNN, showcasing how input x maps to output, with hidden states h that are dependent on previous hidden state and current input.

In Chapter 5, we will utilize an RNN-based deep generative model for symptom tracking information, which is sequential. We provide a depiction of an RNN in Figure 2.3, which shows how prior hidden state h_{t-1} and current input x_t influence the output at each timestep.

Since deep learning involves using multiple hidden layers with non-linear functions, it is possible to learn more abstract patterns by expressing complex representations in terms of simpler ones. However, this can come at the sacrifice of interpretability. In Chapter 5, we

utilize a deep generative model, which hypothesizes a generative model for the data, and learns the proposed parameters via deep learning in order to represent the complex symptomatic behavior. This is inspired by a hierarchical, deep generative modeling approach for a different biological application of modeling how different cancer cell lines respond to experimental drugs, which utilizes a generative model for the experimental setup and a deep learning model for biological complexities [61]. This method allows for balance between predictive power offered by deep learning and interpretability offered by a generative approach.

Chapter 3

Characterization of self-tracked menstrual cycle data

In this chapter, we characterize a large-scale dataset of self-tracked menstrual cycle information, including menstrual cycle lengths and a variety of qualitative symptoms. By exploring this dataset, we are able to develop a quantitative definition for menstrual cycle 'regularity,' a concept that remains open to exploration among researchers in the field. Additionally, we showcase that users with different menstrual cycle patterns also exhibit different symptomatic experiences, which sets the stage for the importance of individual-level modeling.

3.1 Introduction

3.1.1 Menstruation as an understudied topic

As discussed in Chapter 1, menstruation continues to be an understudied area of research, despite its key role in understanding womxn's health. Due to such neglect, womxn are often left with unaddressed pain and confusing or inaccurate diagnoses [62]. The existence of pain associated with menstruation is very common — dysmenorrhea, or painful menstruation associated with symptoms like abdominal cramps and headaches — is estimated to affect up to 91% of womxn of reproductive age [63]. Dysmenorrhea has also been shown to be associated with quality of life conditions like depression, anxiety, decreased productivity, and fatigue [64;

65], as well as menstruation-related disorders like polycystic ovary syndrome (PCOS) and endometriosis, which can cause infertility, intense pelvic pain, and limited mobility [66; 67]. By better understanding the day-to-day patterns of menstruation, researchers can work to close the systemic gaps that exist in addressing conditions that disproportionately affect womxn [68], providing healthcare professionals with the tools and vocabulary to better identify such conditions and empowering womxn with the knowledge to manage them.

3.1.2 Variation in menstrual cycles

Although the phenonemon of menstruation has been long been a subject of curiosity for researchers in a diverse set of fields, spanning medicine and healthcare to sociology, some misconceptions about the menstrual experience still remain. In particular, the notion of a 'regular 28-day long cycle' continues to exist, despite empirical evidence that variation is both a natural and likely part of the menstrual experience. Clinical studies and recent analyses of menstrual self-tracking app data [49; 50] have supported the claim that "complete regularity in menstruation through extended time is a myth," [69; 1] and have shown that variation in period and cycle length (the number of days between subsequent periods) between cycles, between and within womxn, and among populations is the norm [2; 70; 71; 72; 73; 74; 75; 4; 76]. In order to quantitatively solidify these findings and assist in dismantling the idea of menstruation as a homogenous experience, we aim to develop a definition for menstrual variability that is grounded in a large-scale dataset. In doing so, we can make progress in answering the open question: what exactly does it mean to be 'regular' [74]?

In addition to variation in the length and consistency of menstrual cycles, each womxn's qualitative menstrual symptoms are unique. As previously mentioned, menstruation can influ-

ence not only physiological symptoms like abdominal pain and headaches, but can also impact quality of life, affecting mood and energy levels. With access to a large-scale, longitudinal dataset, we also seek to characterize how womxn who track their cycle lengths differently may also track their symptom experiences distinctly. Finally, we also acknowledge that users may not always track their cycles perfectly and therefore develop a methodology for identifying and removing such cycles from our dataset (a concept which we will further motivate in Chapter 4).

3.2 Methods

As mentioned in the Background section, we leverage a dataset from Clue, where users are able to track symptoms across 20 categories. Table 3.1 provides a description of the available Clue categories, their corresponding symptoms, and the frequency with which they are tracked in the dataset (note that the definition of a user as 'consistently highly variable' or 'consistently not highly variable' will be described later in this chapter).

Table 3.1: Description of the Clue app tracking categories and corresponding symptoms, along with the per-symptom number of tracking observations (and their corresponding proportion with respect to the total number of observations) for the 'consistently not highly variable' and 'consistently highly variable' user groups.

Category	**Description**	**Symptoms**	**No. of tracking events (%) for consistently not highly variable group**	**No. of tracking events (%) for consistently highly variable group**
period	Period flow	spotting, light, medium, heavy	22,096,884 (19.71)	913,403 (18.56)
emotion	Emotional state	happy, sensitive, sad, PMS	11,377,997 (10.15)	501,610 (10.19)
pain	Type of pain experienced	cramps, tender breasts, headache, ovulation pain	9,730,958 (8.68)	406,710 (8.26)
energy	Energy level	low, high, exhausted, energized	8,710,403 (7.77)	410,216 (8.34)
sleep	Hours of sleep	0-3, 3-6, 6-9, > 9	8,597,769 (7.67)	405,726 (8.24)
skin	Skin health	acne, good, oily, dry	5,896,540 (5.26)	263,258 (5.35)
mental	Mental state	calm, distracted, focused, stressed	5,871,137 (5.24)	252,621 (5.13)
sex	Sexual health	unprotected sex, high sex drive, protected sex, withdrawal sex	5,813,292 (5.19)	271,540 (5.52)

motivation	Motivation level	motivated, unmotivated, productive, unproductive	5,467,728 (4.88)	236,052 (4.80)
craving	Food cravings	sweet, salty, carbs, chocolate	4,867,777 (4.34)	224,751 (4.57)
digestion	Digestive health	great, bloated, gassy, nauseated	4,825,627 (4.30)	209,651 (4.26)
social	Social behavior	sociable, withdrawn, supportive, conflict	4,178,744 (3.73)	186,110 (3.78)
poop	Stool health	normal, constipated, great, diarrhea	3,889,471 (3.47)	172,716 (3.51)
hair	Hair health	good, bad, oily, dry	3,128,384 (2.79)	147,844 (3.00)
fluid	Vaginal discharge type	creamy, egg white, sticky, atypical	2,378,211 (2.12)	106,782 (2.17)
collection method	Method for period collection	pad, tampon, panty liner, menstrual cup	2,027,258 (1.81)	84,270 (1.71)
exercise	Physical exercise	running, yoga, biking, swimming	1,222,568 (1.09)	44,946 (0.91)
party	Party-related experiences	drinks, cigarettes, big night, hangover	900,444 (0.8)	40,779 (0.83)
medication	Type of medication taken	pain, cold / flu, antihistamine, antibiotic	561,540 (0.5)	21,030 (0.43)

ailment	Physical maladies	cold / flu, allergy, injury, fever	550,951 (0.49)	20,899 (0.42)

Cohort definition

The cohort for this chapter's study comprises $117,014,597$ self-tracking events for $378,694$ users located on all continents from 2015-2018, aged 21–33 years old (see Appendix A for tables of detailed summary statistics, detailed count of cohort users per country, and age-specific statistics). At sign up, users can input overall personal information like age and hormonal birth control (HBC) type. We select users from the Clue dataset in this age range to ensure the consistency of our dataset (because menstrual cycle lengths are relatively less variable and cycles are more likely to be ovulatory during this age interval [1; 2; 3; 4; 5]). Specifically, the reproductive axis (the hypothalamic-pituitary-ovarian axis) may not be fully matured for younger womxn, especially those who experienced a later than average age at menarche. On the other end, older womxn may be experiencing premature menopause. By restricting our cohort to this age range, we substantially reduce the influence of confounders like undetected heterogeneity on our results.

In addition, we select users with natural menstrual cycles only (i.e., no HBC or intrauterine device (IUD)) to control for the impact of hormonal contraception, which has been shown to impact cycle length and other aspects of menstruation. Specifically, we remove cycles from users who reported some form of HBC (patch, pill, injection, ring, implant) or IUD (there is no explicit distinction between hormonal and copper IUD usage in the dataset). Although this step reduces the dataset size by about 45%, it ensures that the exhibited menstrual behavior is due to physiology and not the effect of birth control.

Users are also able to specify whether a cycle should be excluded from their tracking history — for instance, if they feel that the cycle is not representative of their typical menstrual behavior due to a medical procedure or changes in birth control, they can indicate this in the app; we exclude such cycles. We also eliminate cycles longer than 90 days and users who have only tracked two cycles, to rule out cases where there may be lack of engagement or non-continuous app usage.

Finally, we exclude cycles where it's possible that the user forgot to track their period, hence resulting in an artificially long cycle length. We refer to such artificially long lengths as 'self-tracking artifacts.' The effect of these filtering steps on the dataset is outlined in Figure 3.1; the final step indicates the removal of self-tracking artifacts. In total, these data filtering steps reduced the size of the cycle dataset by about 49%, but the resulting age-specific, natural cycle-only user cohort and corresponding dataset with potential artifacts removed enables us to study our research questions in a less noisy setting.

For this resulting cohort, the average user is 25.49 (median of 25) years old (per-country and per-age detailed statistics are provided in the Appendix A). As reported in Table 3.2, the average number of cycles tracked per user is 12.89 (median of 11), with an average cycle length of 29.73 (median of 29) days and mean period length of 4.08 (median of 4) days. Note that a menses duration longer than 10 days is considered an outlier by Clue, since it exceeds mean period length plus 3 standard deviations for any studied population [4].

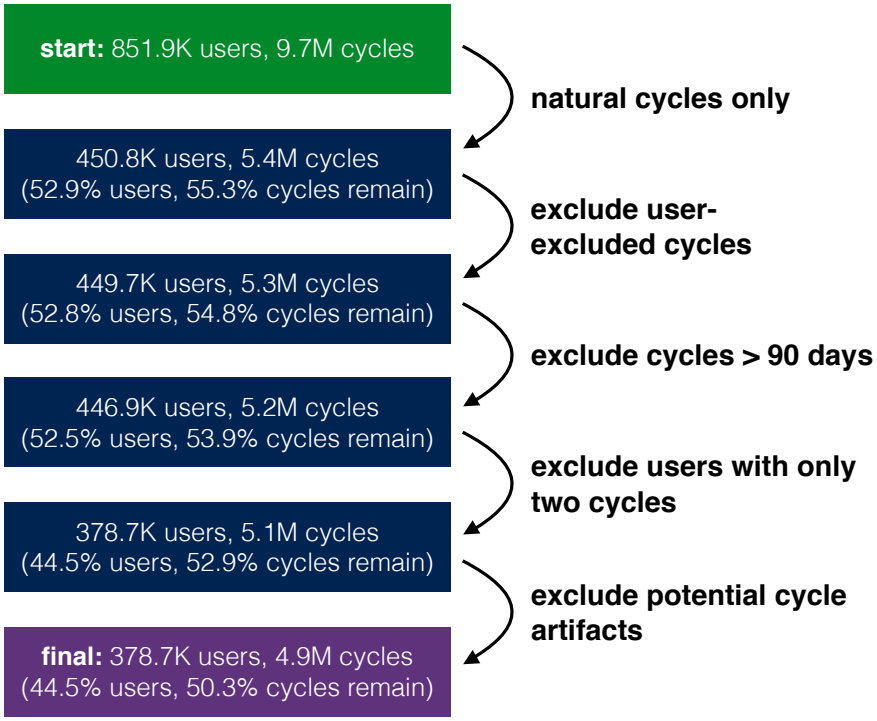


Figure 3.1: Step-by-step filtering process for computing the final user and cycle cohort. The percentage of users and cycles removed at each step is computed out of the initial numbers. Note that we only include users aged between 21-33 years, since womxn exhibit more stable menstrual behavior in their 'middle life' phase [1; 2; 3; 4; 5].

3.2.1 Defining variability groups based on cycle tracking history

The first question we seek to answer is whether we can quantitatively characterize cycle 'regularity.' To that end, we develop a definition for cycle variability based on the characteristics of the studied users.

We first propose the computation of cycle length differences, or CLDs, which we define as the absolute differences between consecutive cycle lengths. These are computed per-user — if

Table 3.2: Per-user cycle characteristics

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

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

we define a user's C cycle lengths as $d = [d_0, d_1, d_2, \ldots, d_C]$, then the CLDs are computed as

$$[|d_1 - d_0|, |d_2 - d_1|, \ldots, |d_C - d_{C-1}|]. \tag{3.1}$$

For instance, a user with cycle lengths $d = [30, 40, 25, 30]$ has corresponding CLDs of $[10, 15, 5]$. CLDs allow us to understand volatility from one cycle to the next. Although cycle lengths have been shown to vary widely among womxn [2; 70; 71; 72; 73; 74; 75], they fail to capture between-cycle dynamics. In contrast, regardless of specific cycle lengths, CLDs capture menstrual patterns in a user's longitudinal tracking history. This allows us to measure fluctuation over time and identify users who are more or less consistent in their cycle volatility.

CLDs do not capture certain menstrual phenomena, such as a cycle length that grows at a constant rate; for instance, if a user's cycle length increases consistently by two days every cycle, the CLDs would all equal two, but there would indeed be a large differential between

the shortest and longest cycle length (i.e., there would be volatility that isn't necessarily captured). However, CLDs and related metrics of median and maximum CLD do allow us to characterize those who fall on the extreme ends of the between-cycle variability spectrum and identify potential self-tracking artifacts. Figure 3.2 outlines how CLDs and related statistics are computed, as well as describes how potential artifacts are detected.

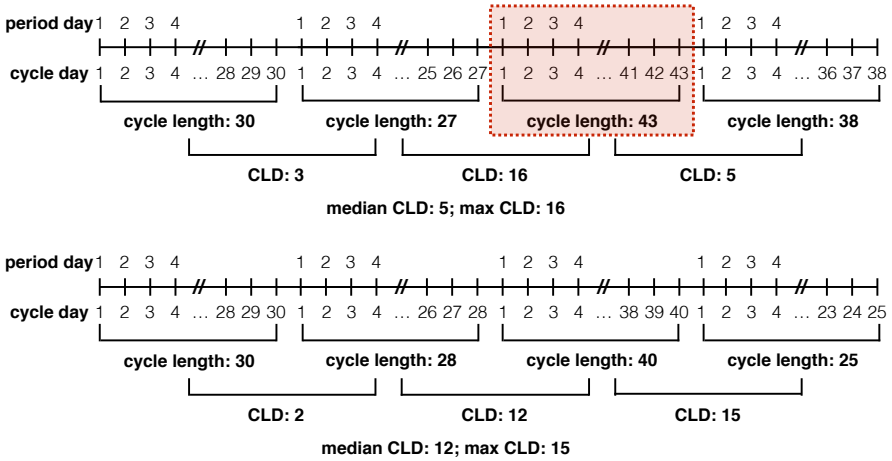


Figure 3.2: We provide illustrative examples of identifying a cycle tracking artifact (top) and characterizing a user's regularity (bottom) based on CLD statistics. In each example, we display a user's cycle history with a total of 4 cycles. Cycle length is computed as the length of time between the first day of a period and the first day of the next period, and CLD is computed as the absolute difference between subsequent cycle lengths (i.e., if a user has n cycles tracked, they will have $n-1$ CLD values). Period length is computed by counting the number of sequential days on which there is menstrual bleeding greater than spotting ('light,' 'medium,' or 'heavy'). Two such sequences are considered one period if separated by no more than one day of non-bleeding/spotting. In the top example, the user's second CLD exceeds their median by at least 10, and thus we identify the corresponding 'artifically long' cycle in red — this cycle will be excluded from our analysis. In the bottom example, the user's median CLD is at least 9, and thus they will be classified as a consistently highly variable user.

Variability of womxn's menstrual experiences exists on a broad spectrum. In order to

define and examine groups of users who fall on different ends of the variability spectrum, we utilize the per-user metric of median CLD. We choose the median because it is able to characterize the overall consistency of users' cycles while remaining robust to outliers (versus the mean, which would be more susceptible to being skewed by rare events). We choose a cutoff of greater than 9 days for identifying users with consistently highly variable menstrual patterns, based on examining the cumulative distribution function for median CLD across users as seen in Figure 3.3. We believe this cutoff is an appropriately stringent choice because it aligns with existing work on analyzing menstrual patterns — cycle length variability studies conducted for womxn in Guatemala, Bolivia, India, Europe, and the US noted differences in the maximum and minimum cycle length ranging from 6 to 14 days [70; 71; 72; 73; 74; 75].

Our proposed cutoff for median CLD separates users into two distinct groups of menstrual patterns: the vast majority (92.32%) of the population falls to the left of this threshold in the **not consistently highly variable** group. The remaining 7.68% of the population, who we consider to be the **consistently highly variable** group, represent those whose variability is extreme — these users experience more drastic fluctuations in cycle length, as seen in the Results section below. We use a two-sample Kolmogorov–Smirnov (KS) [77] test to confirm that the cycle length distributions differ significantly between the two groups, which we describe in further detail below.

3.2.2 Excluding cycles lacking user engagement

Since our data are self-tracked, they pose the possibility of users not engaging reliably with the app; namely, users may not track a physiological event, even if it happened. In

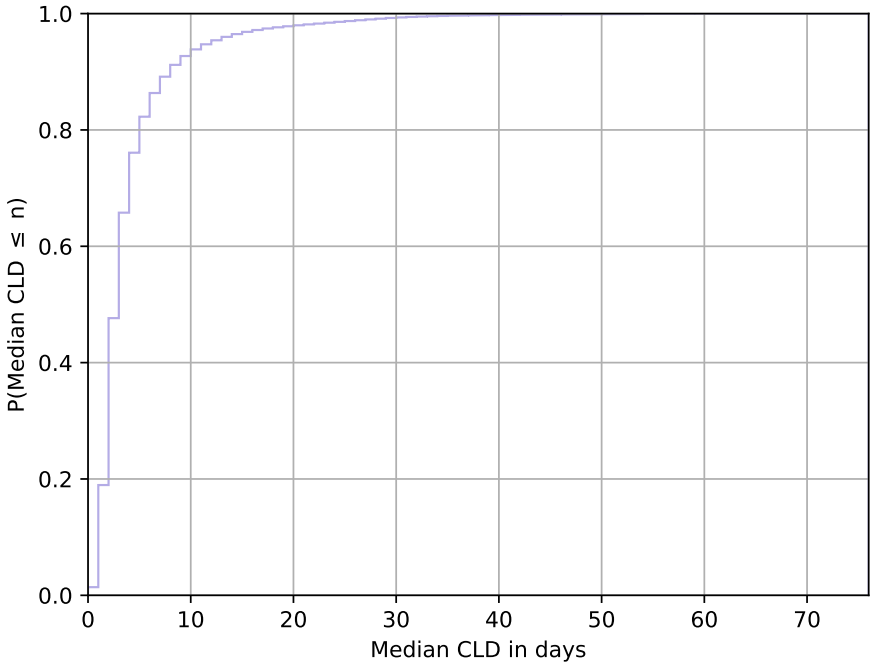


Figure 3.3: Looking at the cumulative distribution of median CLD, we see that the curve flattens out significantly around the 'elbow' at 9 days; thus, we choose greater than 9 days as our cutoff for our definition of consistently highly variable.

particular, if a user forgets or otherwise skips tracking of their period, the corresponding computed cycle length will appear artificially inflated (since cycle length is computed as the number of days between subsequently-tracked periods) — we refer to such an instance as a cycle engagement or self-tracking artifact. In order to combat this, we develop a methodology for identifying and removing cycles we believe lack user engagement, which allows us to distingush physiological behavior (i.e., true 'long' cycle lengths) from self-tracking artifacts (i.e., artificially inflated cycle lengths) to more reliably consider symptom tracking behavior as a proxy for true physiological behavior.

Median CLD and maximum CLD provide a view into how each users' cycle lengths fluctuate, and to what degree of extremity. In particular, examining median CLD allows us to

characterize typical patterns, while maximum CLD allows us to identify outliers in cycle length variation (and therefore potential cases of inconsistent user engagement). In Figure 3.4, we showcase a histogram of maximum CLD across users — the multimodality of the curve (in blue) indicates that there may be instances where users skipped tracking of their period, thus resulting in an overestimation of cycle length. Specifically, peaks at 30 and 60 days suggest instances where a user may have self-tracking artifacts corresponding to an inflation of one or two cycle lengths (i.e., skipped tracking of one or two periods), respectively. For instance, consider a user who exhibits perfectly uniform cycle lengths of 30 days each and corresponding CLDs of 0. If this user were to skip tracking of one period in their history, then their maximum CLD would be 30 (with an artificially inflated cycle length of 60 days) — such a user would fall in the first peak of the maximum CLD histogram.

In order to identify where self-tracking artifacts occur, we compare the median and maximum CLD of each user and flag cycles where the corresponding CLD exceeds the user's median CLD (which represents their 'typical' cycle variability) by at least 10 days as a possible 'atypically long' cycle. Specifically, the longer of the two cycles corresponding to the CLD is flagged. We provide an illustrative example of this procedure in the top panel of Figure 3.2, where the third cycle has been identified as a potential instance of skipped tracking. The cutoff of 10 days is based on an attempt to locate a feature in the data (rather than posit a priori) that distinguishes 'typical' from 'extreme' reported cycles. In particular, we plot a two-dimensional histogram of median versus maximum CLD, where each example is one user in Figure 3.5, illustrating how median and maximum CLD demonstrate the discrepancy between per-user typical cycle patterns and extreme events. In particular, we see a clear visual feature — a band of users where maximum CLD is within 10 days of median CLD, and a scatter of other

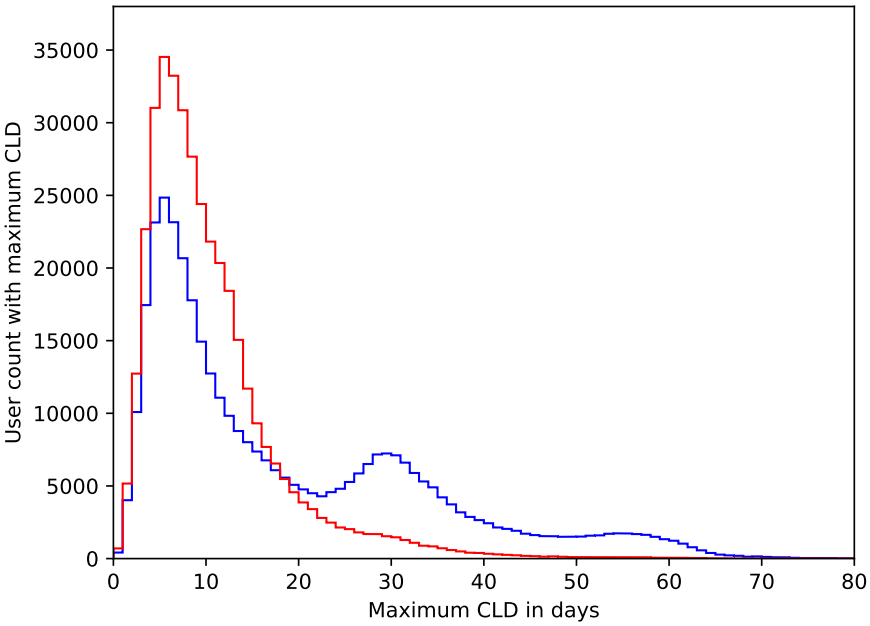


Figure 3.4: For each user, we compute the maximum CLD and plot a histogram before (blue) and after (red) excluding cycles without user engagement (i.e., cycles that are potential artifacts). We see that the multi-modal behavior (peaks at around 30 and 60 days) is largely dampened upon removing these cycles. In addition, the fat right-hand tail in the red curve implies that we preserve the natural variation in cycle length — we are not simply removing long cycles.

users for whom the maximum CLD far exceeds the median; this is displayed as a diagonal red line along where maximum CLD is equal to 10 more than the median CLD. To capture this, we define extreme events as those corresponding to cycle lengths where the CLD is at least 10 days more than the median, and consider other events to fall within the spectrum of normalcy for that particular user. These flagged cycles are considered 'atypically long' as the result of self-tracking artifacts and are excluded from our analysis.

After we exclude such cycles from our analysis, we see that the multimodality of the maximum CLD histogram is largely removed (see red line in Figure 3.4). However, note that while

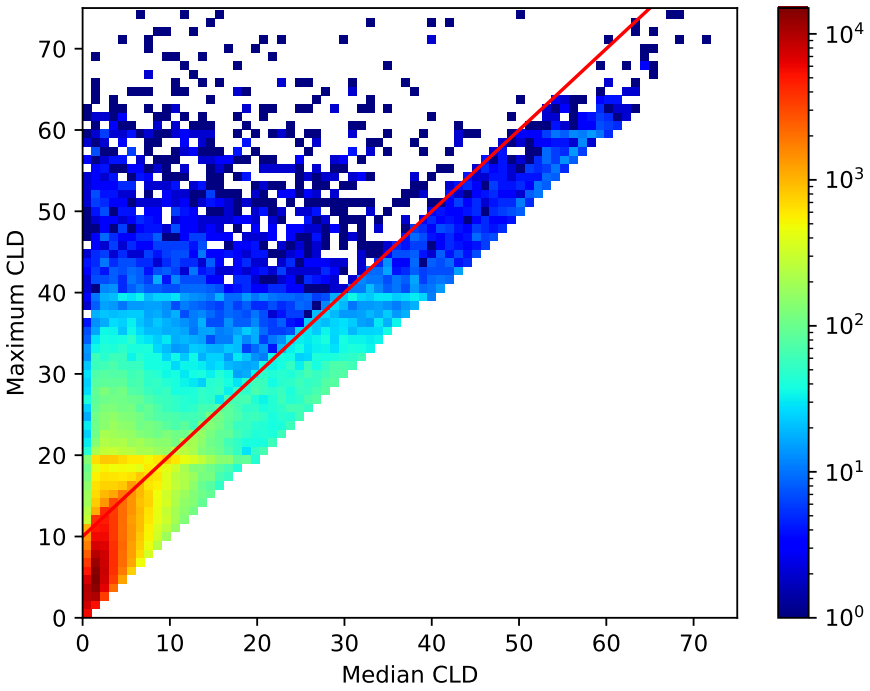


Figure 3.5: We plot a two-dimensional histogram of users' median CLD versus maximum CLD in logarithmic space, as well as the line where maximum CLD is equal to median CLD plus 10 in red. We can see that the line separates out a highly concentrated region of users, as well as a more scattered region of users. Specifically, the majority of the mass falls under this line, as showcased by the concentrated red color in the lower lefthand corner of the plot and a diagonal band extending upwards, while the concentration in the region above the line is more dispersed. Thus, we examine the cycles that fall above the line as possible cycle tracking artifacts.

our method is stringent enough to identify self-tracking artifacts, it is also conservative enough to preserve the heterogeneity of the data (and the multitude of menstrual experiences that it represents), as seen in the long righthand tail of the red line. In particular, we find that for the 42% of users who have at least one 'atypically long' cycle, we exclude a small number of cycles (1.59) per user on average.

In order to further validate that the cycles we exclude are likely due to skipped tracking, we examine user tracking activity during the interval where they are expected to track their period for each of these excluded cycles to see if any bleeding-related events were tracked. We define this interval as the user's last reported period day plus their median cycle length, plus or minus their median period length. In 89.18% of these cycles, no bleeding-related events were tracked, indicating that the user likely did not engage in period tracking; in the remaining 10.82% of the excluded cycles, it is unclear whether the bleeding-related events tracked by the user during this interval represent period or non-period bleeding. However, by our definition, a single bleeding event is not sufficient to be considered a period. To be conservative and maintain consistency of our definitions for period and artificially inflated cycle lengths, we exclude these cycles from our analysis, ensuring a coherent data preprocessing pipeline. Note that this impacts results minimally, since excluded cycles with some bleeding-related events account for only 0.56% of all cycles. By quantifying inconsistent engagement with tracking, we can ameliorate its impact on subsequent analyses.

3.2.3 Characterizing symptom tracking variability

To quantify symptom tracking behavior, we consider how often throughout each user's longitudinal tracking history they track each symptom, regardless of when (i.e., which phase or day) within the cycle the tracking occurred. In other words, we focus on symptom tracking at the cycle level. Since cycle length varies both within and between users' tracking histories, the number of tracking events per cycle would be skewed by cycle length; to combat this, we measure the per-user proportion of cycles where a symptom has been tracked.

Since not all users track all categories, we wish to capture symptom tracking behavior for

cycles where users were interested in tracking the associated category. We consider a user to be interested in tracking the associated category if they have tracked any symptom in that category at least once across all of their cycles, and we compute a metric we refer to as 'proportion of cycles with symptom out of cycles with category,' which is how often a user u has a symptom s tracking event $e_u = s$ per cycle n, given that they have tracked symptoms within the associated category C at least once across all their cycles N_u. This is mathematically denoted as

$$\lambda_{us} = \frac{\sum_{n=1}^{N_u} \mathbb{1}[\exists e_u = s]}{\sum_{n=1}^{N_u} \mathbb{1}[\exists e_u \in C]} \,. \tag{3.2}$$

That is, to account for user interest in tracking the symptom at hand, we compute the proportion of cycles with a symptom being tracked out of the number of cycles where the user has tracked the category related to that symptom. For instance, consider a user who tracked 9 cycles; out of these, they tracked any of the symptoms within the 'mental state' category for 4 cycles. For only 1 of these cycles, they tracked the symptom 'distracted,' while for 3 of these cycles, they tracked the symptom 'stressed.' For this example user, 25% of the cycles with 'mental state' have 'distracted' tracked, while 75% of the same cycles have reports of 'stressed.' Our metric λ_{us} captures the tracking regularity of each symptom across a user's cycles; it essentially represents the conditional probability that user u tracks the specific symptom s given that they have tracked any symptom from the symptom's corresponding category. This metric is robust to (*i*) different cycle lengths and number of cycles (it is normalized with respect to each user's number of cycles), (*ii*) different user app interests (it is contingent on whether the user has shown interest in tracking a given category), and (*iii*) different app usage behaviors (it is independent of how many times within a cycle a given symptom is tracked).

3.2.4 Kolmogorov–Smirnov test

In order to understand if and how symptom tracking differs between variability groups, we utilize a two-sample Kolmogorov–Smirnov [77] (KS) test. This test is nonparametric and suitable for any ordinal (as opposed to, e.g., binary or categorical data), which is useful since we lack a mechanistic model of what distribution the data may be drawn from. It compares the equality of one-dimensional probability distributions arising from two samples, and can be used to assess statistical differences in symptom tracking behavior between variability groups. Using this test, we can examine how the cumulative distributions of λ_s per group (i.e., λ_{us} for all users u within each variability group) differ, and in particular, how these densities are distinct on their support boundaries across groups (i.e., the consistently not highly variable and consistently highly variable user groups). The KS statistic quantifies the distance between the empirical cumulative distributions of two samples (i.e., between the two groups); the associated KS test is sensitive to differences in both location and shape of the distributions, allowing us to characterize *where* and *how much* the symptom tracking patterns (as measured by the proposed λ_s metric) differ between groups.

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

The null hypothesis is rejected at level α if

$$D_{n,m} > \sqrt{-\frac{1}{2}\ln\alpha \cdot \frac{n+m}{nm}}\,, \tag{3.3}$$

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

In addition to determining whether the symptom tracking distributions differ between groups, we also seek to explore *how* they differ. To do so, we study the support boundaries of the distributions for each group, i.e., where $p(\lambda_s > 0.95)$ and $p(\lambda_s < 0.05)$. These probability intervals represent how likely users in each variability group are to either consistently track a symptom throughout their cycles (i.e., in almost all of the cycles where they track the category, they track the specific symptom), or to not track it at all (i.e., in almost all of the cycles where they track the category, they do not track the specific symptom). We then compute the odds ratio of these values for the consistently highly variable group to the consistently not highly variable group, for both the high extreme and low extreme end of the proportion range. If the odds ratio is greater than 1 for the high extreme end of the range for a symptom, this indicates that the consistently highly variable group is more likely than the consistently not highly variable group to report a very high proportion of cycles with that symptom. On the contrary, if the odds ratio is greater than 1 for the low extreme end, this indicates that the consistently highly variable group is more likely to report a very low proportion of cycles with that symptom (i.e., the consistently highly variable group is more likely *not* to report such a symptom)

Note that when possible, 95% confidence intervals have been added to reported KS values

using bootstrap analysis. To do so, we draw 100,000 random samples — resampled with replacement — from each variability group and report the estimated mean KS statistic values and their 2.5 and 97.5 percentiles.

3.3 Results

3.3.1 Cycle length characteristics

We examine the cycle length characteristics of users in each of the variability groups. First, we visualize tracking histories on an individual level by plotting a time series embedding of consecutive cycle lengths for one randomly sampled user from each variability group in Figure 3.6. Specifically, we sample one user each from the consistently highly variable and consistently not highly variable groups who have the median number of cycles tracked (11 cycles) and plot their consecutive cycle lengths on the x, y, and z axes, respectively. This allows us to visualize how much each user's cycle lengths change over the tracked cycles. We find that the consistently highly variable (orange) user's cycles wander through the space, indicating that they experience consistently volatile cycle lengths throughout their history. On the other hand, the consistently not highly variable (teal) user's cycles occupy a much tighter area, indicating that their cycle lengths are consistently stable throughout their history.

Next, we extend our visualization of this phenomenon to a population level. To do so, we randomly sample three consecutive cycles from each user's tracking history for the entire cohort and once again plot these lengths on the x, y, and z axes, respectively. This time series embedding of cycle length for the entire population is seen in Figure 3.7a, where each point represents one user's three consecutive cycles (in contrast to Figure 3.6, where we plotted the entire cycle histories of each user). If a user is perfectly consistently not highly variable

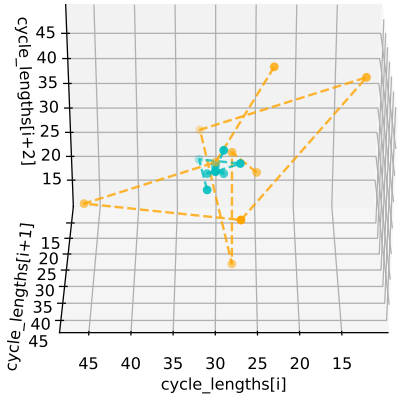


Figure 3.6: We sample one consistently highly variable and one consistently not highly variable user, each with the median number of cycles (11), from the user cohort and plot each set of three consecutive cycles on the x, y and z axes, respectively. This allows us to visualize how much a user's cycle lengths change throughout their entire cycle tracking history — we would expect that a not consistently highly variable user would have points that cluster closer together in space. We see that the consistently not highly variable (teal) user occupies a small region, while the consistently highly variable (orange) user's points move through the space. This indicates that the teal user's cycle lengths are consistently very similar to one another, whereas the orange user experiences more consistent fluctuation in cycle lengths. Thus, we see that separating users into groups on the basis of median CLD identifies those who are more and less consistently highly variable.

(i.e., they always track the exact same cycle length from one cycle to the next), then their representative point would fall on the $x = y = z$ line. If a user's cycle lengths fluctuate at all, then they would fall somewhere outside of this line; the degree of their fluctuation would determine their position. We observe a phenomenon in Figure 3.7a that is consistent with the one seen in Figure 3.6 — the consistently not highly variable group (teal) occupies a tighter region of space than the consistently highly variable one (orange). Specifically, this region is clustered around the $x = y = z$ line. This indicates that a user in the consistently highly

variable group is more likely to experience volatile menstrual patterns (i.e., highly varying cycle lengths), and that therefore our median CLD metric reasonably separates out groups of users based on their cycle length fluctuations.

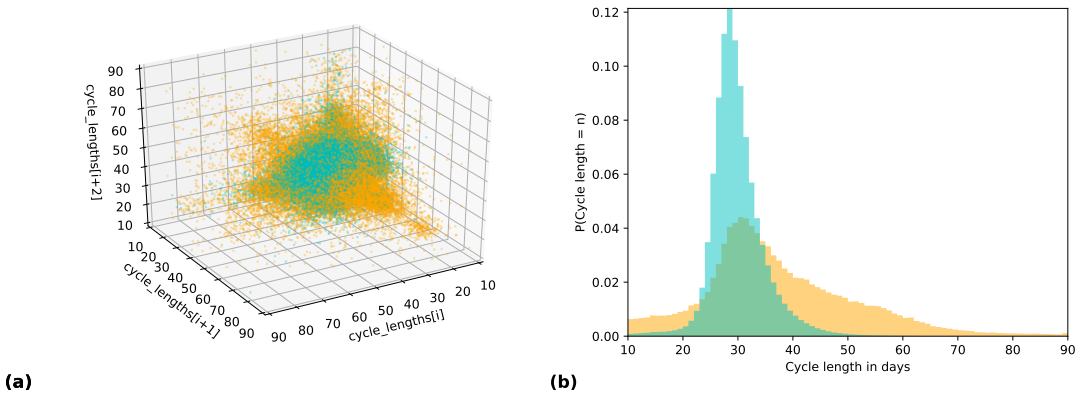


Figure 3.7: Time series embedding (**a**) and probability distributions (**b**) of cycle length for the consistently not highly variable (teal) and consistently highly variable (orange) groups. **(a)** The cycle lengths of three consecutive randomly sampled cycles from each user in the cohort are plotted on the x, y, and z axes. Each consistently not highly variable user is represented by a teal point, and each consistently highly variable user by an orange point. It is visually evident that the teal cluster of users occupies a tighter region of the space around the $x = y = z$ line, with the orange cluster fanning outward. **(b)** The cycle length probability distributions of the cohort, where we note that the orange group's distribution has a much wider spread and is less peaked than the teal group. Cycle lengths are more heterogeneous, or widely distributed, for the orange group, confirming that the consistently highly variable group represents those with more fluctuation in cycle length. The cumulative distributions per-group differ significantly (as per a two-sample KS test).

We also study the empirical cycle length distributions per group in order to assess how their location and shape may differ. In Figure 3.7b, we see that not only do cycle length statistics such as mean and median cycle length differ between the groups, but the shapes of each group's distribution are also distinct — in addition to being centered at longer cycle lengths (median of

34 days versus 29 days), the cycle length distribution for the consistently highly variable group is less peaked with a wider spread. In other words, this group's distribution of cycle lengths encompasses a more volatile range, has much heavier tails, and is skewed towards longer cycle lengths. As expected, we also find with a two-sample KS test that these distributions differ significantly — the KS statistic is 0.377 with a 95% confidence interval of (0.375, 0.378).

3.3.2 Period length characteristics

In addition to cycle length, we also examine how period length characteristics differ between the two variability groups and find that our metric (median CLD) identifies two distinct groups of users based on their cycle (not period) length variability. This is because while womxn in the two variability groups differ significantly in their cycle lengths, as seen above, their period length distributions are much less variable — period lengths fluctuate similarly between the groups. Specifically, Figure 3.8 showcases the distributions of period length for each group — period length is centered around the same median of 4 days for both groups, and the shapes of the distributions are similar. Therefore, we see that variability in cycle length (as separated out by median CLD) is not due to period length differences between groups, as period length varies the same amount across all womxn. Note that although period length distributions do differ significantly under the two-sample KS test, the KS statistic for the period length distributions is 0.066 with a 95% confidence interval of (0.064, 0.068), which is nearly an order of magnitude smaller than the associated values for the cycle length distributions, showcasing that the cycle length distributions differ more drastically and with much higher probability than the corresponding period length distributions.

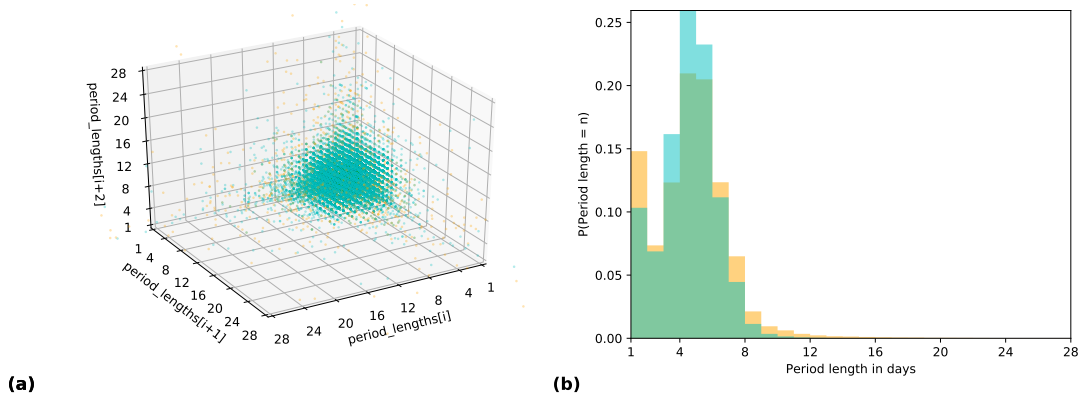


Figure 3.8: Time series embedding (**a**) and probability distributions (**b**) of period length for the consistently not highly variable (teal) and consistently highly variable (orange) groups. **(a)** The period lengths of three consecutive randomly sampled cycles from each user in the cohort are plotted on the x, y, and z axes. Visually, we observe that both groups occupy a very similar region of the period length space (few orange points are placed outside the region occupied by the teal cluster). **(b)** The period length probability distributions of the cohort, where we observe that the orange and teal distributions are largely overlapping, with the same median of 4 days and a similar shape, indicating that period lengths are distributed very similarly for the two groups. We notice a slight peak in single day period reports in both groups, which we argue is reminiscent of app usage behavior: some users are interested in knowing (approximately) when they had their period, not in tracking how long it was, so they may only track the day it occurred and not continue tracking after that.

3.3.3 Length statistics over the app usage timeline

In order to determine how cycle and period variability may change over time, we examine per-group cycle statistics over the app usage timeline. In particular, we align users by their subsequently-tracked cycles (not absolute time) using cycle ID, i.e., a cycle ID of 1 corresponds to the first cycle of a user, 2 to their second cycle, and so on. Figure 3.9 demonstrates that cycle and period length statistics are stationary over time at the group level, and that the

consistently highly variable group consistently displays a higher average cycle length over time (as compared to the consistently not highly variable group). This is verified in Table 3.2 — the mean cycle length for the consistently not highly variable group is 29.45 days (median of 29), and the mean is 37.04 days (median of 34) for the consistently highly variable group. In addition, we find that although average cycle and period length are stable over time for all examined cohorts (the consistently highly variable group, consistently not highly variable group, and the entire user cohort), the consistently highly variable users exhibit a wider spread (i.e., higher volatility) across cycles. Consequently, we see that the consistently highly variable group accounts for a large degree of the volatility in the data, a detail that would likely be 'smoothed out' and lost if we considered the population as a whole, rather than separating the users into two groups. Since cycle and period length statistics are constant within groups across app usage, we are confident that median CLD is not merely capturing spurious correlations related to the length of time the user stays with the app.

3.3.4 Symptom tracking differences

We find that there exists a relationship between median CLD and symptom tracking behavior — despite CLD only ostensibly being a measure of cycle length variability, it can also provide information about symptom experiences. Firstly, we find that womxn located at different ends of the menstrual variability spectrum exhibit different symptom patterns. Specifically, we observe that while users in the different variability groups exhibit similar tracking frequencies (i.e., the total number of times they track certain symptoms over their history) per category (as in Table 3.1), their symptom tracking patterns (i.e., how they track throughout history) are distinct. We assess the degree to which these symptom tracking patterns differ by using

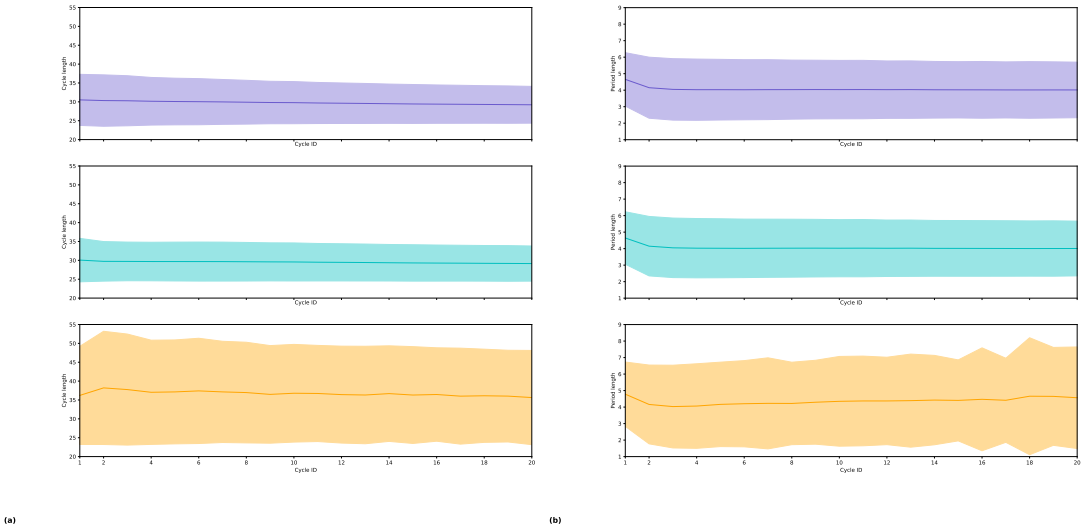


Figure 3.9: For each user's cycles (indexed by cycle ID), we average cycle (**a**) and period length (**b**) across three different groups: the entire user cohort (top, purple), the consistently not highly variable user cohort (middle, teal), and the consistently highly variable user cohort (bottom, orange). This allows us to visualize how cycle and period length vary over time for each group on average and in terms of standard deviation (for illustrative purposes, we restrict the cycle ID to 20). Cycle and period length statistics are stationary over the app usage timeline within each plot. We note that the top and middle plots look similar in each figure (i.e., the consistently not highly variable group looks similar to the overall population in terms of both cycle and period length), but the wider shaded orange spread of the bottom plot demonstrates the higher degree of variability in the consistently highly variable group. In addition, this spread is consistently wider for the orange plot over time. This showcases that the consistently highly variable group represents a large degree of the variability that we see in the data overall.

the KS test to evaluate how the population-level distributions of 'proportion of cycles with symptom out of cycles with category' (as in Equation 3.2) differ for each symptom. We find that these distributions differ between the user groups across most categories, and that these differences are significant for all symptoms within the period, pain, and emotion categories.

This result may prove clinically useful for assessing menstrual conditions and overall wellness, and the KS test results for symptoms within these categories are presented in Table A.10.

We also find that womxn in the consistently highly variable group display more heterogeneous behavior. In other words, their behavior is more unpredictable. For instance, we consider the symptoms encompassing period flow — womxn in the consistently highly variable group are significantly more likely not to report heavy periods throughout their cycle history (odds ratio of 1.734 on the low extreme end of the proportion range in Table A.11). For the symptom of 'spotting,' the tracking pattern is more heterogeneous for the consistently highly variable group, as shown by the higher odds ratios on both extremes of the proportion range, (i.e., either in all or none of their cycle history) shown in Tables A.11 and A.12.

In addition, we find that consistently highly variable users have generally more heterogeneous experiences for non-bleeding related symptoms like pain. One particularly interesting finding is that those who are consistently highly variable are much more likely to track headaches and tender breasts in at least 95% of their cycles, with odds ratios of 1.663 and 1.715, respectively (see Table A.12).

3.4 Significance

By exploring the large-scale Clue dataset, we are able to quantitatively interrogate antiquated notions of menstrual regularity, gain insight into how cycle length experiences are related to symptomatic ones, and showcase the utility of self-tracked mobile health datasets (when handled with care). We characterize the menstrual experience as a broad spectrum and validate that variability is the norm. In addition, we identify statistically significant relationships between cycle timing and symptoms like period flow and pain, which can prove useful to

clinicians and users (such symptoms are frequently leveraged for diagnosis of health-relevant conditions like endometriosis and polycystic ovary syndrome (PCOS)). On a broader scale, the methodology we've developed for identifying potential self-tracking artifacts can be applied to other menstrual self-tracking datasets (or similar datasets where cyclic behavior is expected). This research sets the groundwork for modeling and prediction of menstrual phenomena, which we will explore in Chapters 4 and 5.

Our work quantitatively solidifies prior claims of menstrual variability. In Appendix A we provide a comparison of our summary statistics against those of related studies that use self-tracked menstrual cycle data [49; 50]; although our studies differ slightly in terms of population demographics, we believe they provide a reasonable basis for comparison. Overall, our period and cycle length statistics are similar, and we draw similar conclusions about cycle lengths having slightly higher values (median of 29 in our dataset) and wider ranges than previously commonly believed (see Appendix A for full details). Our high-level cycle statistics also align well with previous clinical studies [1; 2; 74].

By proposing a definition of variability based on fluctuations between cycle lengths rather than on cycle lengths themselves (e.g., mean cycle length), we are able to depict a more nuanced view of the range of menstrual experiences. In particular, studies on menstrual variability have shown that even when cycle lengths appear consistent based on mean cycle length, this is a misconception, as womxn with such cycles still experience significant cycle variability [74]. Utilizing median CLD to define variability allows us to meaningfully separate users into two groups that differ significantly in their cycle and symptom tracking behaviors; we are unaware of any single figure of merit that can likewise helpfully separate users into distinct segments. Although cycle length has been proposed as a biomarker of menstrual health (e.g., very long

and very short cycles are associated with a higher risk of infertility), our work suggests that cycle variability may also be a useful biomarker. Clue uses the International Federation of Gynecology and Obstetrics (FIGO) definitions for clinically irregular cycles in the app [78], but has not found connections with differences in tracking.

The discovery of associations between cycle timing and symptoms, and more specifically, the distinct expressions of symptom experiences between user groups, enables further investigation of clinical associations and can potentially aid in diagnosing menstrual health conditions. This is a step forward in studying the relationship between menstrual patterns and symptomatic variables, which has been limited — there is ample work on how hormone levels change during the cycle [79; 80; 81; 82], but relatively much less on how the broad array of symptoms are related. Recent work using self-tracked data has explored this concept, but over a limited set of symptoms [83] and without discriminating by age or birth control usage (which can bias results) [48]. A method for estimating ovulation timing based on Fertility Awareness Method observations (i.e., basal body temperature (BBT), cervical mucus, cervix position, and vaginal sensation) has been presented [50], but such data are inaccessible to us due to data privacy concerns (the sharing of sensitive fields such as pregnancy tests and BBT) and the European Union's General Data Protection Regulation. In contrast to existing work, we explore symptoms of interest explicitly and comment quantitatively and qualitatively on a broad set of symptoms. Our study provides insight into which high-signal self-tracked symptom patterns can be potentially useful either for predicting each other (e.g., predicting cycle variability from symptoms) or health consequences (e.g., PCOS).

Despite the strength of our results, we must bear in mind several mitigating factors. Most prominently, in this work we take tracking behavior as a proxy for true, physiological experi-

ence. However, there are multiple reasons why self-tracked data can be unreliable or inaccurate, such as ambiguous symptomatic language or inconsistent user engagement. With respect to the former, there is the possibility of overlap in symptom or category names, such as the symptoms 'low energy' or 'exhausted' — a user could reasonably elect to track one or the other for arbitrary reasons. While the Clue app provides explanatory infotexts for each tracking category, users are ultimately influenced by their own interpretations and how they use the app to meet their own needs; each category may not necessarily mean the same thing to each user. Relatedly, the Clue app was designed based on scientific literature and research on which categories users deemed important, and therefore to cater to a broad array of experiences and needs, the tracking categories are treated as equally important. While this allows for users to track a variety of individual needs, the symptoms in the app are not based on validated scales or designed with specific diagnoses in mind, meaning the symptom names may not be granular or targeted enough to make definite, condition-specific claims.

In addition to the risk of imperfect category and symptom descriptions, users may also engage inconsistently with the app; we consider two forms of inconsistent user adherence: tracking an unequal number of cycles or forgetting to track period. For the former matter, we observe that the consistently highly variable group tracked a lower number of cycles on average (see Table 3.2), but that the number of users who only tracked two cycles (after our preprocessing steps) is small across all users; such instances represent 2.62% and 0.57% in the consistently highly and not highly variable groups, respectively. Therefore, although the number of cycles tracked may differ among users, we argue that the extreme cases of only two cycles tracked is low enough to be negligible. For the latter issue, we utilize our procedure of excluding unexpectedly long cycles to ameliorate the impact of inaccurate cycle lengths

due to skipped period tracking. However, we acknowledge that even with this effort in place, it is complicated, if not impossible, to know what the true physiological experience is based on self-tracked data, since there is no access to ground truth. In addition to engagement artifacts, there could be unforeseen factors like cultural differences [84] impacting individual experiences and how they are tracked. While we have utilized preprocessing techniques to reduce the likelihood of self-tracking artifacts, we recognize that limitations nonetheless remain. Regardless, examining such datasets remains useful to better understanding both womxn's menstrual experiences at scale and how to improve self-tracking technologies to enable clearer, more interpretable datasets in the future.

Significance to users: This work provides concrete evidence from a large-scale, self-tracked database that variability in menstrual cycles and symptoms not only exists, but is the norm. In this sense, it is a step forward in validating each womxn's own unique and diverse set of experiences. The fact that we utilize self-tracked data is important, not only from a research method standpoint, but because it also demonstrates to the user that such self-tracking can be useful in a broader sense to understanding menstruation, and that such research can be done in an ethical and nuanced manner.

Chapter 4

A hierarchical, generative model for menstrual cycle lengths that models skipped period tracking

In this chapter, we propose a hierarchical, generative model for menstrual cycle lengths that characterizes each user by per-individual parameters for typical cycle length and propensity to skip tracking. We showcase this model's ability to aid in detection of self-tracking artifacts (i.e., instances where users skip tracking of their cycles) and the utility of accounting for such artifacts in our model by showcasing its performance relative to baselines over time.

4.1 Introduction

4.1.1 User adherence to mobile health apps

The rise of data-powered health has enabled high-resolution views into large, highly diverse populations over time [27; 28; 29; 30]. In particular, mobile health (mHealth) tracking apps enable users to self-manage their personal health by giving them the ability to instantaneously and flexibly track information anytime, anywhere [25; 27]. Such mHealth solutions provide insights into a broad range of conditions and behaviors, from compliance and chronic diseases [85; 86] to endometriosis [87] and fertility care [88] to asthma [31], giving users increased awareness

of and autonomy over various facets of their individual health. In addition, the existence of such apps provide researchers with new, large-scale datasets to obtain detailed observations of chronic conditions [32; 33; 34; 35; 36; 37; 38; 39; 40] that were previously unavailable at such size and scope.

However, while such apps have expanded the opportunities for self-management of health and related research objectives, they also present risks. In particular, self-tracking apps are dependent on user adherence (i.e., insights from an app can only be based on what the user actually tracks); therefore, if a user engages inconsistently with the app, the representation of their health may be skewed or inaccurate. Studies have shown that the design of an app is crucial to user engagement [89; 90], and that engagement can vary widely between users. Factors like the app's user interface and notification system, as well as device fatigue [91; 92; 93], can influence how often and how consistently users interact with the app. Apps that provide predictions and analytics to the user can only derive such learnings from what they track — the existence of imperfect tracking therefore raises the question of how to distinguish true health phenomena from tracking behavior in order to provide the most accurate picture of an individual's health.

4.1.2 Menstrual trackers as use case

We ground our exploration of this issue in the context of menstrual trackers, a category of mobile tracking apps that have become increasingly common: they are the second most popular app for adolescent girls and the fourth most popular for adult womxn [41; 42]. Menstrual trackers [43; 44; 45; 46; 47] serve as a rich source of temporal, heterogeneous data from millions of womxn worldwide. Access to such large-scale, longitudinal datasets has

enabled quantitative investigation into menstrual cycle characteristics and dynamics, including studies describing menstrual cycles and related symptoms [94; 49; 50; 95] and efforts to better understand ovulation [96]. The user populations for these datasets vary in their intention for utilizing such apps (for instance, some may be looking to simply track their menstrual cycle, while others may be interested in fertility awareness and family planning); however, most users are interested in knowing when their next period will occur and what symptoms to expect. To meet this need, many apps provide users with insight into their individual menstrual behavior, fertility, and more, giving them a deeper understanding of their menstrual experience [97]. However, while such predictions are available, they may fall short in accuracy [51]. In particular, aforementioned adherence artifacts resulting from inconsistent tracking may obfuscate health-related conclusions. For menstrual trackers, this manifests as inflated cycle length computations if a user forgets to track their period.

As mentioned in Chapter 3, the menstrual experience is inherently variable, rendering prediction of a user's next cycle start difficult (even assuming tracked information is exactly representative of each user experience). Self-tracking data introduces an additional source of variability due to differing user tracking behavior (e.g., some users may track their information consistently, while others may skip tracking, whether intentionally or by accident). Since multiple sources of uncertainty must be taken into account, modeling menstruation is especially difficult when utilizing such data. Therefore, in order to properly harness the power of mobile app data sources, researchers must be able to develop accurate predictive models that can address the specific nature of such data. While we utilize menstrual trackers as a use case, this issue will be prevalent in any kind of self-tracked data.

4.2 Methods

4.2.1 Data cohort

We leverage the same de-identified self-tracked dataset from Clue as in Chapter 3, focusing on period self-tracking events only. Recall that a 'self-tracking event' refers to when a user logs a symptom in the app, and period self-tracking events refer to instances when a user self-reports days where they have experienced period flow; we use these events to compute cycle lengths (i.e., the number of days between subsequent periods).

We utilize the same data exclusion criteria as in Chapter 3, focusing our analysis on users aged 21-33 with natural cycles only, excluding user-identified anomalous cycles, and removing cycles longer than 90 days. We utilize cycle length information only as the input to our proposed model, taking the first 11 cycles for all 186,106 menstruators with more than 11 cycles tracked (since 11 is the median number of cycles tracked in the full dataset).

See Table 4.1 for comparison of summary statistics for all cycles of users in the selected cohort versus the first 11 cycles only for the same users. We see that cycle length and period length statistics differ very minimally between these two sets of cycles, indicating that using the first 11 cycles is a reasonable representation of user history.

4.2.2 Definition of adherence artifact

Since period tracking exactly determines cycle length, it is crucial to consider the possibility that users may not always track their period accurately, and therefore their observed cycle length may not reflect their true experience. Here, we refer to a 'self-tracking (or adherence) artifact' as a mismatch between the self-tracked event and the true, experienced physiological

Table 4.1: Summary statistics for selected self-tracked menstruator dataset

Summary statistic	Selected cohort	Selected cohort (first 11 cycles only)
Total number of users	186,106	186,106
Total number of cycles	3,857,535	2,047,166
Number of cycles (mean±sd, median)	20.73±8.35, 18.00	11.00±0.00, 11.00
Cycle length (mean±sd, median)	30.45±7.73, 29.00	30.71±7.90, 29.00
Period length (mean±sd, median)	4.07±1.76, 4.00	4.13±1.80, 4.00
Age (mean±sd, median)	26.07±3.56, 26.00	25.59±3.61, 25.00

Summary statistics for selected self-tracked menstruator dataset for the whole dataset, as well as the selected first 11 cycles only. Total number of users and age are the same for the selected cohort and selected cohort's first 11 cycles only, since they represent the same set of users. We see that cycle length and period length statistics differ very minimally between the selected cohort and the selected cohort's first 11 cycles only, indicating that using the first 11 cycles is a reasonable representation of each user's history.

phenomenon (focusing on period self-tracking events) — for instance, a user may experience period flow on one day, but not report it in the app. We investigate the impact that such self-tracking artifacts can have on modeling and prediction of next menstrual cycle start date. In particular, we examine how they inflate cycle length computation and result in the appearance of apparent cycle 'skips.' We provide an illustrative example of this phenomenon in Figure 4.1.

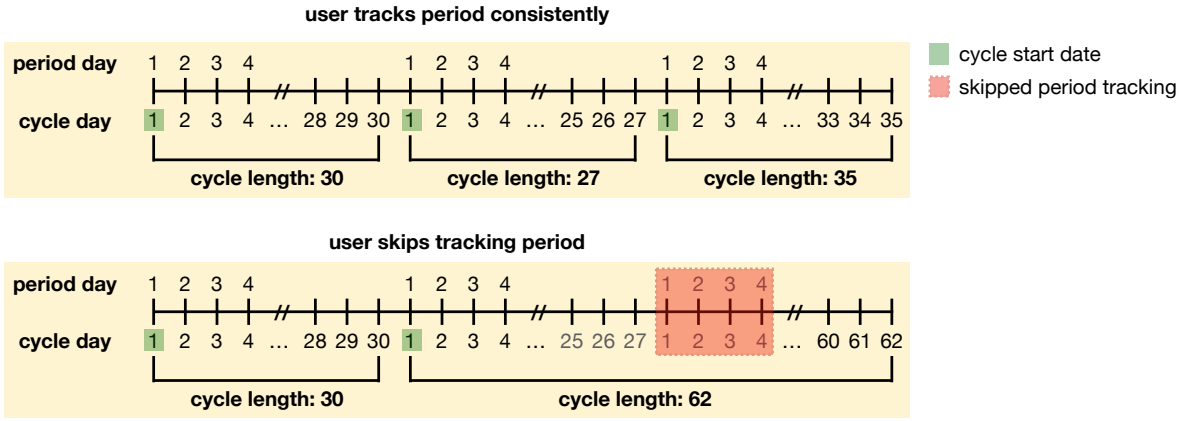


Figure 4.1: Example cycle tracking history for the same user, demonstrating two scenarios: where they track all of their periods (top) and where they skip tracking of one of their periods (bottom). Cycle start dates are highlighted in green and skipped period tracking is highlighted in red. The bottom panel showcases how skipping tracking of one period can result in inflated observed cycle lengths — instead of two subsequent cycles of length 27 and 35, respectively, because the user skips tracking of a period, it appears that they have one cycle of length 62. This is because cycle length is determined by the number of days between tracked periods. This phenomenon holds analogously if a user skipped more than one period (in which case three subsequent cycle lengths would appear as if it were a single, inflated cycle length).

4.2.3 Proposed generative model

We propose a probabilistic machine learning model with three main features: 1) it accounts explicitly for self-tracking artifacts by probabilistically factoring in the possibility that users

may have skipped period tracking and have inflated observed cycle lengths; 2) it dynamically updates predictions of cycle length and skipping probability as the cycle proceeds, providing insight into how these predictions evolve over time; and 3) it prioritizes the unique nature of each user's menstrual experience by modeling individual cycle length histories and providing individual predictions, while also incorporating population-wide knowledge.

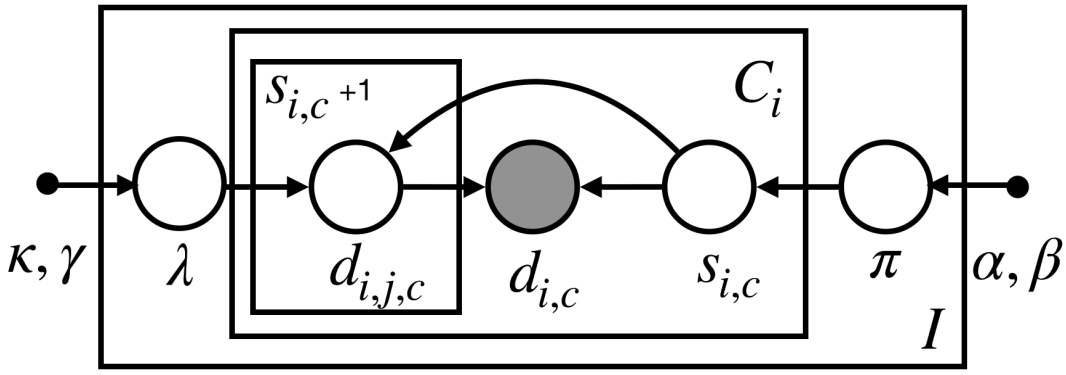


Figure 4.2: Hierarchical graphical model for proposed generative process. In our graphical model, variables within the outer plate are replicated for users $i = 1, \cdots, I$, variables within the inner plate are replicated for each per-user cycle $c = 1, \cdots, C_i$, and variables within the innermost plate are replicated for each skipped cycle $j = 0, \cdots, s_{i,c}$. Individual-level parameters λ_i (average cycle length without skipping) and π_i (probability of skipping a cycle) are drawn from population-level distributions characterized by hyperparameters $u = [\kappa, \gamma, \alpha, \beta]$. $s_{i,c}$ represents number of skipped cycles for user i and cycle number c; $d_{i,c}$ represents observed cycle length. We model observed data (cycle lengths $d_{i,c}$) as the sum of true (unobserved) cycle lengths $d_{i,j,c}$ skipped $s_{i,c}$ times (so that an observed cycle length $d_{i,c}$ contains $1 + s_{i,c}$ unobserved cycle lengths $d_{i,j,c}$).

Since our model is generative, we hypothesize the distributions from which each of our proposed variables is drawn and describe their relationships to one another [98]. We showcase the proposed generative process (i.e., candidate probabilistic model for generating the observed

data) as a probabilistic graphical model in Figure 4.2.

In particular, our model posits that each user can be characterized by two latent quantities that govern the observed data: λ_i, their typical cycle length patterns; and π_i, their propensity to skip tracking. The observed cycle lengths $d_{i,c}$ for user i and cycle c are modeled as the sum of the latent, true (unobserved) cycle lengths $d_{i,j,c}$, which are skipped $s_{i,c}$ times (j indexes the skipped cycles).

Specifically, we provide details on the generative process for cycle lengths $d_{i,j,c}$, which draws per-user specific parameters from population level shared priors:

- **Observed variables:** Observed cycle length $d_{i,c}$, with $c = \{1, \cdots, C_i\}$ cycle lengths for each individual $i = \{1, \cdots, I\}$. Each true cycle length (for user i, cycle c, out of the number of skipped cycles j) is drawn from a Poisson distribution, $d_{i,j,c} \sim p(d_{i,j,c}|\lambda_i) = Pois(d_{i,j,c}|\lambda_i)$. The sum of independent Poissons is a different Poisson distribution, so the observed cycle length ($d_{i,c} = \sum_{j=0}^{s_{i,c}+1} d_{i,j,c}$) is also drawn from a Poisson, conditioned on the number of skipped cycles,

$$d_{i,c} \sim Pois(\lambda_i(s_{i,c} + 1)) \; . \tag{4.1}$$

- **Latent variables:** $s_{i,c}$ denotes the number of skipped (not reported) cycles, with $c = \{1, \cdots, C_i\}$ cycle lengths for each individual $i = \{1, \cdots, I\}$. The number of skipped cycles is drawn from a truncated Geometric distribution with a maximum number of skipped cycles S,

$$s_{i,c} \sim p(s|\pi_i) = \frac{\pi_i^s(1-\pi_i)}{\sum_{s=0}^{S} \pi_i^s(1-\pi_i)} = \frac{\pi_i^s}{\sum_{s=0}^{S} \pi_i^s} = \frac{\pi_i^s(1-\pi_i)}{(1-\pi_i^{(S+1)})} \quad \text{for } s \in \mathbb{N} \; . \tag{4.2}$$

- **Parameters λ_i:** the Poisson rate parameters for each individual $i = \{1, \cdots, I\}$. Per-user

Poisson rate parameters λ_i are drawn from a population-level Gamma distribution

$$\lambda_i \sim p(\lambda|\kappa,\gamma) = \frac{\gamma^\kappa}{\Gamma(\kappa)}\lambda^{\kappa-1}e^{-\gamma\lambda} \quad \text{for } \lambda > 0 \text{ and } \kappa,\gamma > 0. \tag{4.3}$$

- **Hyperparameters of the Poisson rate parameter:** κ, γ of a Gamma distribution prior for the Poisson rate at the population level.

- **Parameters π_i:** the probability of skipping a cycle for each individual $i = \{1, \cdots, I\}$. The probability of an individual skipping a cycle is drawn from a population-level Beta distribution

$$\pi_i \sim p(\pi|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}\,, \quad \text{for } \pi \in [0,1] \text{ and } \alpha,\beta > 0. \tag{4.4}$$

- **Hyperparameters of the geometric distribution parameters:** α, β of the population level Beta distribution prior on skipping probabilities.

Proposing separate probability distributions from which each of the per-user parameters (λ_i and π_i) are drawn enables us to disentangle true, per-user cycle behavior from self-tracking adherence. Therefore, we can gain intepretable insight into not only per-user cycle length behavior, but also per-user cycle skipping behavior. To do so, we learn our per-user parameters on the basis of observed self-tracked cycle lengths, accommodating the latent (unobserved) variables via marginalization of their uncertainties.

In addition to being generative, our model is also hierarchical — we are able to borrow information between users, taking advantage of population-wide knowledge, while also computing individual-level parameters and predictions. This is ideal for a model of menstruation, since it maintains the integrity of each individual's unique experience. As seen above, we represent

individual-level information with the aforementioned individual-level variables for typical cycle patterns and self-tracking adherence and incorporate population-wide characteristics in the form of hyperparameters that are learned at the population level. These hyperparameters then influence the populations from which the individual-level quantities are drawn. For instance, if the most likely cycle length is around 30 days for the whole user base, the population-wide distribution will represent this. However, when individual-level typical cycle length is drawn from population-wide distribution, it will be influenced by each user's own cycle tracking history. In this way, we are able to consider common patterns that exist among users and individual differences.

4.2.4 Parameter inference

Specifically, given a dataset of C_i cycle lengths for I users, we perform hyperparameter inference via type-II maximum likelihood estimation. We compute a Monte Carlo (MC) approximation to the negative log-likelihood: $-\ln(p(d|u)) = -\ln(\sum_i p(d_i|u))$. Due to the impossibility of integrating out the number of skipped cycles $s_{i,c}$ analytically, we compute a MC approximation to each cycle length likelihood $p(d_i|u)$ with M samples,

$$p(d_i|u) = \frac{1}{M}\sum_m p(d_i|\theta_m)\ , \theta_m \sim p(u) \tag{4.5}$$

where u represents the hyperparameters $[\alpha, \beta, \kappa, \gamma]$ of the distributions from where samples θ_m, representing the parameters $[\lambda_m, \pi_m]$, are drawn. We compute the probability $p(d_i|\theta_m)$ by integrating out the probability of skipping $s_{i,c}$, which is drawn from a truncated geometric

distribution as in Eqn. (4.2):

$$p(d_i|\theta_m) = \prod_{c=1}^{C_i} p(d_{i,c}|\theta_m) = \prod_{c=1}^{C_i} \sum_{s=0}^{S} p(d_{i,c}|\lambda_m, s)p(s|\pi_m) \tag{4.6}$$

$$= \prod_{c=1}^{C_i} \sum_{s=0}^{S} ((\lambda_m(s+1))^{d_{i,c}} e^{-\lambda_m(s+1)}/d_{i,c}!) \left(\frac{\pi_m^s(1-\pi_m)}{\sum_{s=0}^{S} \pi_m^s(1-\pi_m)} \right) \tag{4.7}$$

$$= \prod_{c=1}^{C_i} \frac{\lambda_m^{d_{i,c}} e^{-\lambda_m}}{d_{i,c}!} \sum_{s=0}^{S} ((s+1)^{d_{i,c}} e^{-\lambda_m s}) \left(\frac{\pi_m^s}{\sum_{s=0}^{S} \pi_m^s} \right) \tag{4.8}$$

$$= \prod_{c=1}^{C_i} \phi(\lambda_m) \frac{\sum_{s=0}^{S} (s+1)^{d_{i,c}} (\pi_m e^{-\lambda_m})^s}{\sum_{s=0}^{S} \pi_m^s} \tag{4.9}$$

$$= \prod_{c=1}^{C_i} \phi(\lambda_m) \frac{\sum_{s=0}^{S} (s+1)^{d_{i,c}} (\pi_m e^{-\lambda_m})^s}{\frac{1-\pi_m^{S+1}}{1-\pi_m}} \tag{4.10}$$

$$= \prod_{c=1}^{C_i} \frac{1-\pi_m}{1-\pi_m^{S+1}} \phi(\lambda_m) \sum_{s=0}^{S} (s+1)^{d_{i,c}} (\pi_m e^{-\lambda_m})^s \tag{4.11}$$

where $d_{i,c}$ represents one cycle length c for a given user i, C_i is the number of cycles for user i, S is the maximum value of s, and ϕ is the Poisson density.

4.2.5 Computing predictions

The generative nature of our model enables us to update predictions as the cycle proceeds; we refer to each day of the next cycle as 'current day.' For instance, we can predict on current day 0 (when the next cycle first starts), day 1, and so on. We can update our predictions of both when the next period will occur (i.e., cycle length) and how likely the user is to have skipped tracking of a period. Furthermore, since our model is generative and we have specified the number of skipped cycles s as a latent quantity, we can consider two possibilities when computing predictions: we can assume the next reported cycle length will be truth (i.e., the next observed cycle will not be skipped), setting $s = 0$; or we can assume the next reported cycle may not be truth (i.e., accounting for the user possibly skipping their next cycle

tracking), setting $s \geq 0$. Assuming $s \geq 0$ allows us to account for as many skipped cycles as desired, allowing us to assess the impact of accounting for self-tracking artifacts on predictive performance.

That is, in order to update our predictions of per-user cycle length as each subsequent day passes, we are interested in the posterior of the next reported cycle length d^*, conditioned on previous cycle lengths d_i for a user i and the day of the current cycle $d_{current}$,

$$p(d^*|d^* > d_{current}, d_i, \hat{u}) = \frac{p(d^*, d^* > d_{current}|d_i, \hat{u})}{p(d^* > d_{current}|d_i, \hat{u})} = \frac{p(d^*|d_i, \hat{u})[d^* > d_{current}]}{p(d^* > d_{current}|d_i, \hat{u})} \tag{4.12}$$

where we explicitly indicate that $p(d^*, d^* > d_{current}|d_i, \hat{u}) = 0$ if $d^* \leq d_{current}$.

In addition to characterizing the full distribution, we are interested in computing the expectation of the conditional predictive posterior as a point estimate for the next cycle length,

$$\begin{aligned}
E[p(d^*|d^* > d_{current}, d_i, \hat{u})] &= \sum_{d^*} d^* p(d^*|d^* > d_{current}, d_i, \hat{u}) && (4.13)\\
&= \sum_{d^*} d^* \frac{p(d^*|d_i, \hat{u})[d^* > d_{current}]}{p(d^* > d_{current}|d_i, \hat{u})} && (4.14)\\
&= \frac{\sum_{d^*} d^* p(d^*|d_i, \hat{u})[d^* > d_{current}]}{p(d^* > d_{current}|d_i, \hat{u})} && (4.15)\\
&= \frac{\sum_{d^* > d_{current}} d^* p(d^*|d_i, \hat{u})}{p(d^* > d_{current}|d_i, \hat{u})} && (4.16)\\
&= \frac{\sum_{d^*=d_{current}+1}^{D} d^* p(d^*|d_i, \hat{u})}{\sum_{d^*=d_{current}+1}^{D} p(d^*|d_i, \hat{u})}. && (4.17)
\end{aligned}$$

The key term above is $p(d^*|d_i, \hat{u})$:

$$p(d^*|d_i, \hat{u}) = \frac{\int d\lambda d\pi q(\lambda) b(\pi) \sum_{s*} p(s^*|\pi) p(d^*|s^*, \lambda) p(d_i|\lambda, \pi)}{\int d\lambda d\pi q(\lambda) b(\pi) p(d_i|\lambda, \pi)}\ , \tag{4.18}$$

where d_i are the cycle lengths for a user i and s_i are the number of skipped cycles for a user, and d^*, s^* are the next reported cycle length and next number of skipped cycles, respectively.

For the truncated geometric distribution on skipping probabilities, we compute the above as

$$p(d^*|d_i, \hat{u}) = \frac{\sum_{m=1}^{M} \frac{1-\pi_m}{1-\pi_m^{S+1}} \sum_{s^*=0}^{S} \pi_m^{s^*} p(d^*|s^*, \lambda_m) p(d_i|\lambda_m, \pi_m)}{\sum_{m=1}^{M} p(d_i|\lambda_m, \pi_m)} \,. \tag{4.19}$$

We compute $p(d^*|d_i, \hat{u})$ for a range of cycle length days $d^* = \{0, \cdots, D\}$, normalizing appropriately over d^* for each value of $d_{current}$, using $p(d_i|\lambda_m, \pi_m)$ (as specified in Eqn. (4.6) of the description of inference) and $p(d^*|s^*, \lambda_m) = Pois(\lambda(s^* + 1))$ (i.e., the Poisson PMF), where we must also normalize $p(d^*|s^*, \lambda)$ over $d^* = \{0, \cdots, D\}$.

4.2.6 Model training, prediction task, and evaluation

We evaluate the average prediction accuracy of our model across all users with root mean square error (RMSE). This RMSE between true cycle lengths d_i and predicted cycle lengths $\hat{d}_i$ is computed for a given model and N users at each current day of the next cycle, where each of the N users as their own prediction, as $RMSE = \sqrt{\frac{\sum_{i=1}^{N}(d_i - \hat{d}_i)^2}{N}}$.

To evaluate model accuracy on a per-user basis, we use absolute error and median absolute error (MAE), where absolute error between an actual data point d_i and a prediction $\hat{d}_i$ is computed as $|d_i - \hat{d}_i|$.

To evaluate menstrual regularity, we use the metric median cycle length difference (CLD), based on previous work on characterizing menstruation [95], as seen in Chapter 3. Recall that CLDs are computed per-user as the absolute differences between consecutive cycle lengths — if we define a user's C cycle lengths as $d = [d_0, d_1, d_2, \ldots, d_C]$, then the CLDs are computed as $[|d_1 - d_0|, |d_2 - d_1|, \ldots, |d_C - d_{C-1}|]$. In this way, CLDs measure variability from one cycle to the next, and a higher median CLD indicates users with generally more volatile cycle tracking histories (and vice versa) [95].

Note that in discussing prediction below, we denote a user's cycle history as d_i, the predicted

next cycle length as d^*, the predicted number of skips in the next cycle as s^*, the learned hyperparameters as $\hat{u}$, and the current day of the next cycle (on which we are computing the predictions) as $d_{current}$.

4.2.7 Alternative baselines

To evaluate the predictive performance of our proposed model, we consider summary statistic-based and neural network-based baselines:

- `Mean and median baselines:` the predicted next cycle for each user is the average (or median) of their previously observed cycle lengths.
- `CNN:` a 1-layer convolutional neural network with a 3-dimensional kernel.
- `RNN:` a 1-layer bidirectional recurrent neural network with a 3-dimensional hidden state.
- `LSTM:` a 1-layer Long Short-Term Memory neural network with a 3-dimensional hidden state.

As with the proposed model, we train these baselines on the first 10 cycle lengths and predict next cycle start for the 11th cycle. Since these are not generative models and their output is only next cycle start date, we cannot predict the likelihood of skipped tracking or update predictions dynamically. We also test other neural network architectures (changing kernel or hidden state dimensionality and number of layers) and find no meaningful difference in performance; see Appendix B for details.

Although menstrual trackers utilize proprietary solutions for cycle prediction (and therefore we cannot exactly compare our predictions to theirs), we believe our baselines provide a reasonable and fair picture of alternative approaches for our predictive task. We choose summary

statistic-based baselines because they represent the common conception that menstruation is 'regular,' and that consequently the mean or median of several cycle lengths could reasonably estimate the next cycle length. In addition to this simplified predictive approach, we choose neural network-based baselines since they have been shown to be powerful predictive models in many healthcare applications.

4.3 Results

In this section, we demonstrate the key results of our work. We start by showcasing our model's successful detection of self-tracking artifacts, which can be utilized in mHealth apps to alert users of potential missed tracking. Next, we present our model's posterior predictive distribution for cycle length, which is interpretable and representative of the data.

We then discuss our model's predictive performance. In particular, we highlight our model's optimal performance relative to alternative baselines in predicting next cycle start, especially on later days of the cycle and most prominently when typical cycle length has passed, which demonstrates the benefit of dynamically updating beliefs about both cycle length and cycle skips. These insights can help users better understand cycle behavior as the cycle evolves. Finally, we showcase the effect of individual variability on cycle length predictions, which highlights the importance of modeling unique experiences.

4.3.1 Detecting self-tracking artifacts

We showcase our model's ability to detect when a user has skipped period tracking on simulated data (i.e., where we know the ground truth of when a user has skipped tracking in their history). We train on the first 10 cycles and predict the likelihood of skipping during the

11th cycle (as in our real menstruator data experiments) and provide details on the simulated data in Appendix B. As demonstrated in Figure 4.1, identifying when a user has skipped tracking of their period is vital to modeling self-tracked cycle lengths accurately; otherwise, there is a risk of mistaking observed, artificially inflated cycle lengths for true ones.

We start with a simple example in Figure 4.3, which showcases how our individual posterior predicted probability of skipping the upcoming cycle (i.e., $p_i\left(s^* = 1 | d_{current}\right)$, where s^* is the predicted number of skipped cycles in the upcoming cycle) evolves over the current day of the 11th cycle for a selected simulated user. We draw vertical lines at days 30 and 40 of the next cycle, and the markers represent the predicted probability of having skipped one cycle on those days. We choose these days because 30 days is around the average cycle length for this user, and day 40 represents when the user has surpassed their typical cycle length.

We see that for this user who has skipped a cycle before (in their set of 10 training cycle lengths), their probability of skipping a cycle in the 11th (unseen) cycle is low up until around current day 30 of that cycle, but increases substantially afterward (e.g., on day 40, the probability of skipping rises to around 0.8). Therefore, our model predicts that before the average cycle length of this user (30 days) has passed, the user is unlikely to have skipped tracking their period (i.e., the probability is low); however, when the typical cycle length is exceeded, the likelihood of skipping spikes. This increased probability as time proceeds demonstrates how our model accurately detects when a user is likely to have skipped upcoming period tracking based on their individual cycle length history and updates these beliefs over time.

In Figure 4.3, we only showcase the probability of the user skipping tracking of one cycle (i.e., skipping one period tracking, or $s^* = 1$) in the upcoming cycle. However, we can compute probabilities of skipping any number of cycles — we can model the likelihood that a user has

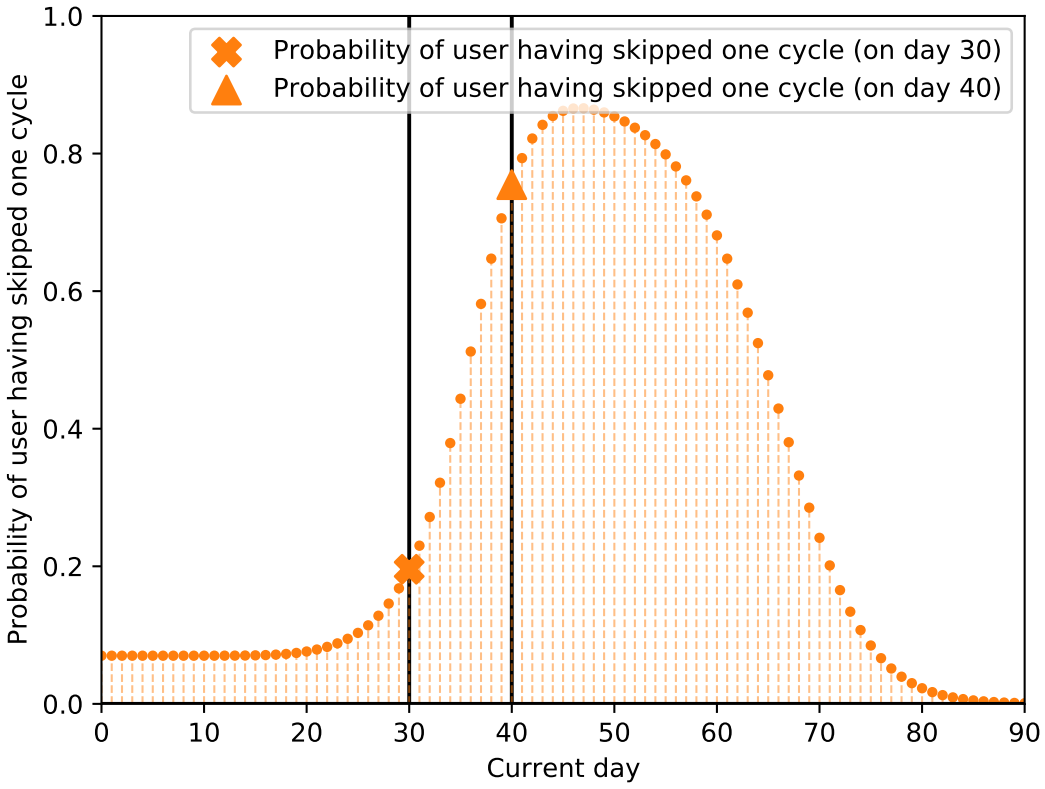


Figure 4.3: Predicted probability of skipping one cycle over time for a simulated user. Orange curve represents probability of user having skipped one cycle; markers indicate probability of having skipped one cycle on day 30 or 40 of the upcoming cycle. We see that the probability of having skipped one cycle in the upcoming cycle is low until day 30. However, past day 30, we see that this probability increases; on day 40, it is around 0.8 (versus 0.2 on day 30). Thus, the model detects that the user is likely to have skipped a cycle on day 40, when their typical cycle length has been passed. Because data in this experiment are simulated, we know that this user has skipped a cycle before in their history and does actually skip the next cycle. Our inferred probabilities recover this, showing that our model can accurately detect when a user is likely to have skipped an upcoming cycle based on their individual cycle length histories and update these beliefs over time.

skipped zero cycles, one cycle, two cycles, and so on in the upcoming reported cycle. To demonstrate this more deeply, in Figure 4.4 we showcase the probabilities of possible cycle

skips, shorthanded as $p\left(s^{*}|d_{current}\right)$, for $s^{*}=0$ and $s^{*}=1$ on simulated data for two different users: (a) showcases a simulated user who has skipped in their history, and (b) showcases a simulated user who has never skipped in their history. We again draw vertical lines at days 30 and 40 of the next cycle (to represent before and after the typical cycle length has passed) and use markers to represent the predicted probability of skipping zero or one cycle on those days.

By comparing these users, we see how our model is able to incorporate historical information to detect differences in skipping behavior as the cycle proceeds. For the user who has skipped in their history (a), as the typical cycle length is passed without tracking (e.g., on day 40), their probability of skipping one cycle is around 0.8, and their probability of skipping zero cycles on that day is around 0.2, a significant drop from a near 0.8 probability on day 30. Thus, as we saw in Figure 4.3, the model incorporates knowledge that the user has skipped before in their tracking history into its prediction of how likely they are to have skipped in this unseen cycle. In contrast, for the user who has never skipped in their tracking history (b), their probability of skipping zero or one cycle on day 40 hovers around 0.5 — in other words, it is less clear whether this user may have skipped a cycle, because they have never skipped before. For instance, the model recognizes that for this user (b), it's more likely that this is a true long cycle (which may occur across menstruators in response to internal or external stimuli) than for user (a).

While Figures 4.3 and 4.4 focus on $s^{*}\in\{0,1\}$, note that this behavior holds analogously for $s^{*}=2$ and beyond. For instance, $p\left(s^{*}=2\right)$ is low early in the next cycle and peaks past day 60, just as $p\left(s^{*}=1\right)$ starts low and peaks past day 30. This is because 60 represents two typical cycle lengths. Our model's ability to detect and alert users of potential tracking artifacts is important not only to accurately predicting when the next cycle will occur, but also

to improving the design of mHealth apps and the quality of their data for menstrual health research.

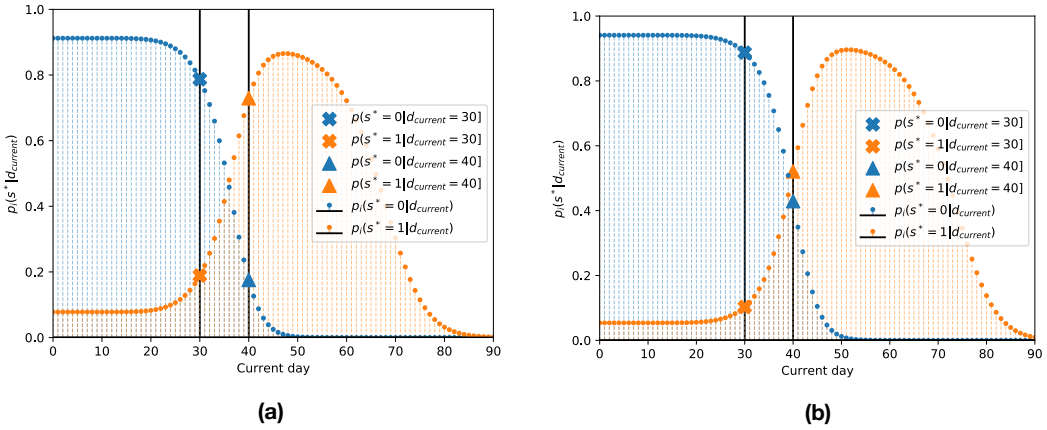


Figure 4.4: Individual posterior predictive probability of skipping upcoming cycle, $p_i\left(s^*|d_{current}\right)$, over current day of next cycle $d_{current}$ for two users from simulated data: one who has skipped a cycle in their history (**a**) and one who has never skipped a cycle (**b**). Our personalized model detects differences in predicted skipping behavior for the two users. Blue and orange curves represent probabilities of skipping zero or one cycle, respectively; markers indicate probability of skipping zero or one cycle on day 30 or 40 of the upcoming cycle. Note that users can also skip more than one cycle. For both example users, we see that the probability of having skipped zero cycles in the upcoming cycle ($p_i\left(s^*=0|d_{current}\right)$) is high until day 30. However, past day 30, the model detects that the user (a) who has skipped in their history is more likely to have skipped the upcoming cycle than for the user (b) who has never skipped. This demonstrates how the model takes into account the previous non-skipping behavior of this user. Because data in this experiment are simulated, we know that the user in (a) does actually skip the next cycle, while the user in (b) does not. Our inferred probabilities recover this, showing that our model can accurately detect when a user is likely to have skipped an upcoming cycle based on their individual cycle length histories and update these beliefs over time.

4.3.2 Representing multimodality in cycle length distribution

One key advantage of our model is that we are able to explicitly represent the skipping behavior of users, which is reflected in the posterior predictive distribution for cycle length. In particular, this distribution is multimodal, which aligns with the idea that users who skip tracking of cycles appear to have artificially inflated cycle lengths (in the case of skipping one cycle, effectively doubled), and that therefore the probability of certain cycle lengths peak as time passes without tracking. We showcase this in Figure 4.5, where we plot our model's posterior predictive distribution for cycle length $p(d^*|\hat{u}, d_i, d^* > d_{current})$. This represents the probabilistic next cycle predictions for a specific user based on their previous cycle length history as the next cycle proceeds (as denoted by $d_{current}$). Specifically, we show the probability (z-axis) of a user's next cycle being a specific length (x-axis) for the current day of the next cycle (y-axis), assuming that their next observed cycle **(a)** is truth (no skipped cycles, $s = 0$) or **(b)** may contain skipped cycles (possible skipped cycles, $s \geq 0$).

We notice how in scenario **(a)** (assuming the next cycle is truth), the posterior predictive distribution is unimodal, reflecting how the probability of the next cycle length being an increased length is consistently increasing as time passes. In contrast, in scenario **(b)** (assuming the next cycle may not be truth), the posterior predictive distribution is multimodal, with peaks around $d^* = 30, 60, 90$. This demonstrates our model's ability to update its beliefs about likelihood of skipping over time in order to provide more accurate cycle length predictions. In particular, such multimodality is the result of (i) conditioning on the day of the next cycle $d_{current}$ and (ii) the explicit modeling of cycle skips, s. By doing so, our posterior predictive distribution (when $s \geq 0$) mirrors the skipping phenomena seen in the dataset — when a

user passes their 'typical' cycle length (i.e., around 30 days in this instance), their likelihood of having skipped tracking increases. This multimodal distribution is not only easily interpretable, but is also crucial to representing self-tracking artifacts in mHealth data and providing accurate cycle length predictions.

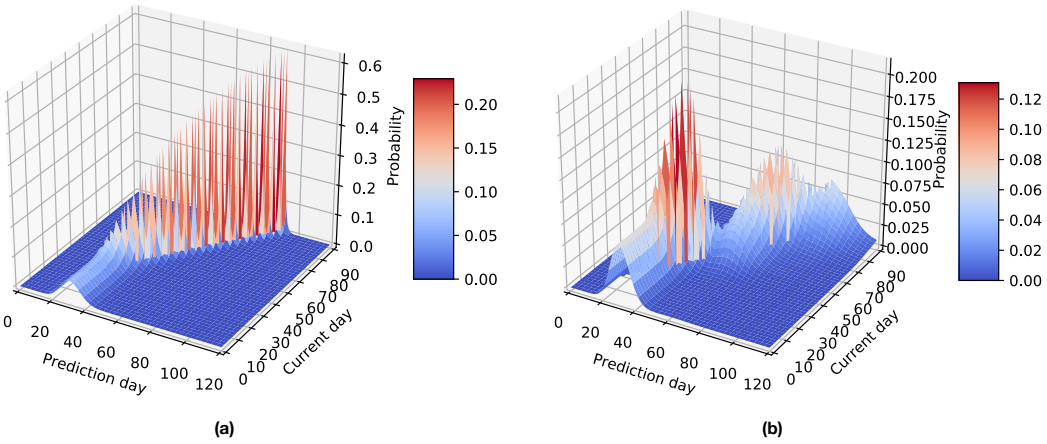


Figure 4.5: Posterior predictive distribution for cycle length over prediction day d^* (i.e., what the next reported cycle is predicted to be) and current day $d_{current}$ (i.e., day in next cycle) for the same user from menstruator data, assuming either that next observed cycle is truth (**a**) or that next observed cycle may contain skipped cycles (**b**). **(a)** When we assume the next observed cycle is true as reported ($s = 0$), our posterior predictive distribution is unimodal. The probability of the next cycle length is peaked around 30 until around day 30 of the next cycle, after which the peak moves consistently to the right, indicating that our cycle length predictions are consistently increasing past day 30 and not adjusting for the likelihood of skipped cycles. **(b)** When we account for the possibility of skipped cycles with $s \geq 0$, our posterior predictive distribution is multimodal. Prior to day 30 of the next cycle, the distribution is similarly peaked around 30 days, as with the $s = 0$ case. However, when the cycle passes day 30, the distribution shows a peak around day 60, indicating the possibility that a user may have skipped a cycle. This behavior holds analogously past day 60. Our explicit modeling of cycle skips allows us to identify when a user may have missed tracking a cycle.

4.3.3 Model performance as cycle proceeds

Our model outperforms the studied baselines in prediction accuracy, particularly as the cycle proceeds; we showcase performance over 'current day' in Figure 4.6. We also showcase specific RMSE values on particular days of the next cycle in Table 4.2.

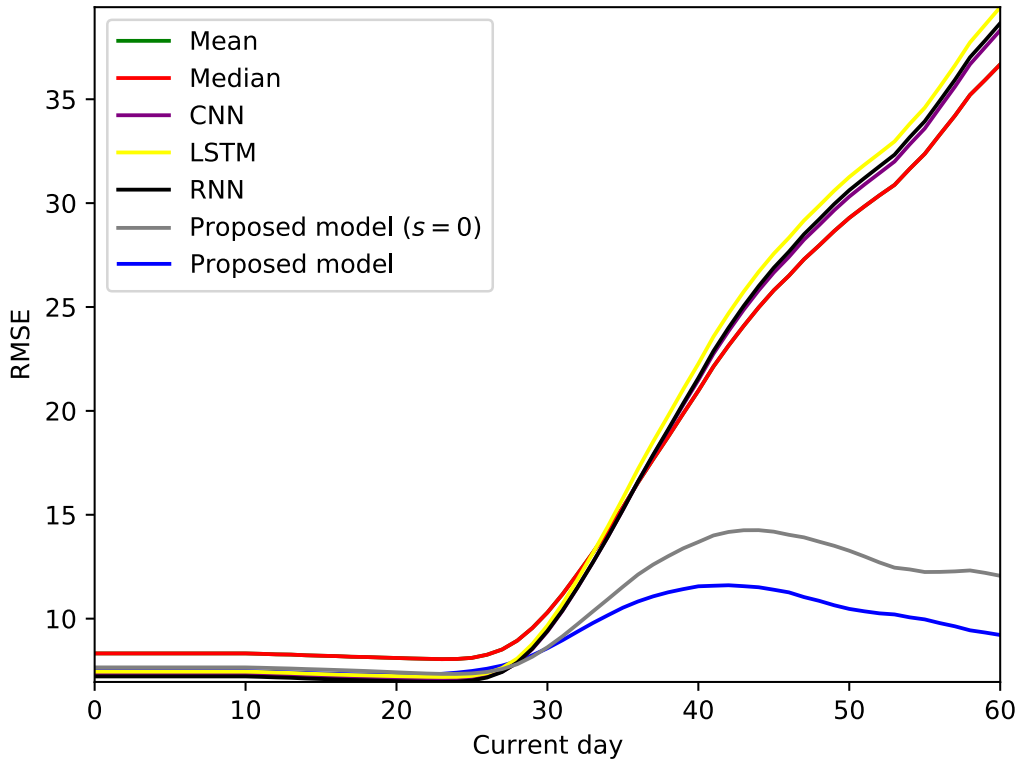


Figure 4.6: Prediction RMSE for proposed model and baselines over current day of the next cycle on the menstruator data, averaged over all users. Both models' superior performance is magnified past around day 30 of the next cycle; they are able to update predictions dynamically, as compared to static baselines. In particular, accounting for skipped cycles ('full' version of our proposed model, blue line) proves especially beneficial to prediction accuracy versus assuming the next reported cycle is truth ('alternative' version of our proposed model, gray line) — by anticipating the possible presence of skipped cycles, we are able to make more accurate predictions and avoid the bump in RMSE seen in the gray line.

As seen in Table 4.2, our model outperforms all alternative baselines on day 0 (the first day of the next cycle), and in particular as the cycle evolves past day 29 — this superior performance is also demonstrated in Figure 4.6, where our models (gray line, $s = 0$ and blue line, $s \geq 0$) display much lower RMSE than baselines. Specifically, accounting for potential skipped cycles becomes increasingly important as the cycle proceeds; this model (blue line) is advantageous to the one assuming the next observed cycle contains no self-tracking artifacts (gray line).

Accounting for skipped cycles is increasingly crucial to predictive accuracy as the cycle proceeds past day 29 because the likelihood of skipped cycles increases as the typical cycle length passes with no tracking activity. Our model is able to incorporate this scenario into its predictions, whereas baselines cannot — although the likelihood of a cycle skip increases over time, not all models can account for this when computing cycle length predictions. Consequently, a benefit offered by our proposed generative model is that it can both account for this phenomenon and dynamically update predictions; this value is reflected in its superior performance relative to baselines.

To evaluate the robustness of our training and predictive performance with respect to different modeling choices, we tested different dataset sizes and ordering of cycle lengths. We find that our model's performance is generally stable across different training set sizes and reordering of cycle lengths. In particular, we shuffled the order of each user's cycles to account for possible time dependency of tracked cycles. See Appendix B for details.

Table 4.2: Prediction RMSE results by model on day 0 and day 40

Model	Day 0		Day 40	
	37K	186K	37K	186K
Mean	7.602	7.497	22.276	21.915
Median	7.586	7.489	23.675	23.394
CNN	8.102	8.027	24.741	24.506
LSTM	7.548	7.402	23.025	22.681
RNN	7.597	7.763	23.474	22.954
Proposed model (predict with $s = 0$)	7.712	7.562	15.114	14.778
Proposed model	7.483	7.382	11.840	11.774

Prediction RMSE for proposed model and baselines on day 0 and day 40 for a subset of the menstruator data ($I = 37,222$) and the full menstruator data ($I = 186,106$). Note that here we train on $C = 10$ cycles and predict the next one. 'Proposed model ($s = 0$)' indicates an alternative version of our proposed model, assuming the next observed cycle contains no self-tracking artifacts; 'Proposed model' indicates the full version of our proposed model, accounting for the presence of potential self-tracking artifacts. Our model outperforms summary statistic-based and neural network-based baselines on day 0 when we account for skipped cycles and does so on only a subset of the data.

4.3.4 Impact of cycle variability

It is imperative for models relating to menstruation to consider the inherent variability of the menstrual experience between and within users. In order to account for the role that menstrual variability may play in producing accurate predictions, we examine our predictive results on an individual level (in addition to averaging them over the whole population). The ability to learn population-wide information and make individualized predictions is a direct benefit of our hierarchical modeling approach.

In particular, we showcase a violin plot of per-user median cycle length difference (CLD) versus absolute error in predicted cycle length in Figure 4.7. For each variability group (as defined by the median CLD value on the x-axis), the middle white point represents its corresponding median absolute error, and the thick gray bar represents its interquartile range. This plot demonstrates how variability impacts prediction accuracy — more variable users are generally more difficult to predict, underscoring the importance of taking into account each individual's experience.

We also note that outliers within a user's cycle length history (e.g., instances where users may have never skipped in their history, but skip the last cycle), which represent a small proportion of the user base, can greatly skew RMSE computations. For instance, users with very consistent cycle lengths (i.e., a median CLD of 0) have a median absolute error (or MAE) as low as 1.5 days, even though the RMSE for this group is 6.15.

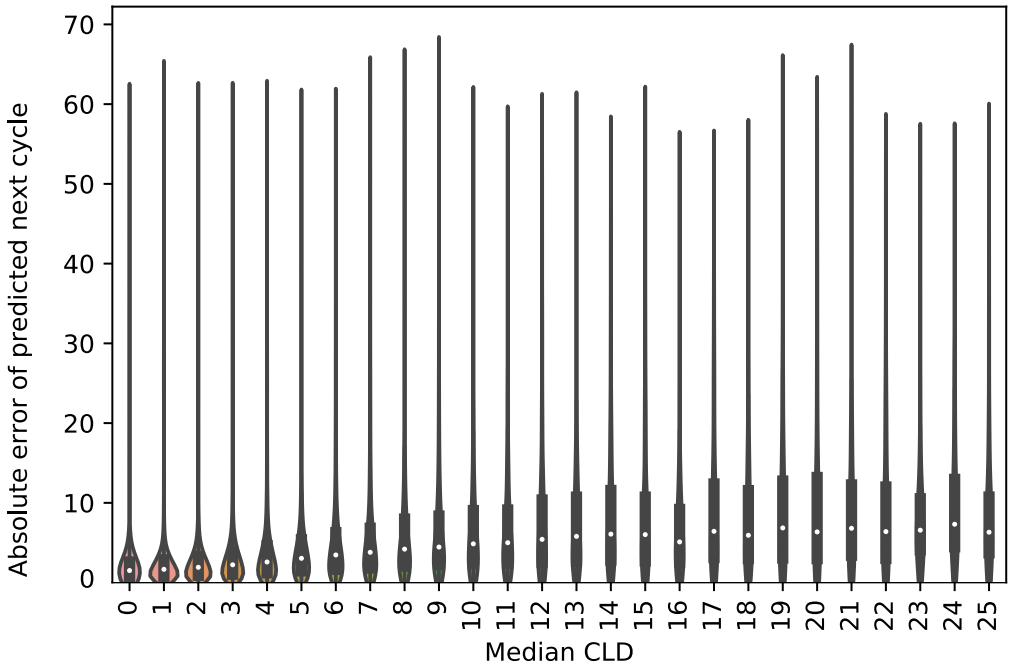


Figure 4.7: Violin plot of per-user absolute error of predicted next cycle length, stratified by user median cycle length difference (CLD) on the menstruator data. We see from the increasing trend in absolute error with median CLD that more variable users are typically more difficult to predict, showcasing that consideration of per-individual behavior is vital to the integrity of our model.

4.4 Significance

By proposing a generative, probabilistic model for menstrual cycle lengths, we are able to characterize the underlying mechanisms of menstruation as collected via mobile tracking apps, a step to better understanding menstruation as a whole. In particular, our model offers the advancement of flexibly accounting for adherence artifacts by explicitly considering the possibilities that users track their information inconsistently and separating this cycle-skipping behavior from typical cycle length patterns. Although heuristics for identifying such

self-tracking artifacts have been proposed (for instance, locating cycle lengths that are anomalous on a per-user basis [95]), these definitions can be limiting. In contrast, we are able to examine the likelihood of skipped cycles in a specific, probabilistic manner, which enables us to distinguish true cycle lengths from self-tracking adherence. Consequently, we can gain insight into both menstrual tracking behavior, as well as learn practical implications for mHealth users and designers.

By computing dual productions (i.e., predictions of both cycle length and possible cycle skips), we are able to provide users with a more accurate, detailed picture of when their next cycle will occur, even when they may not be consistent with their tracking. In addition, rather than providing an option for users to exclude self-identified faulty cycles after the fact, our methodology provides the possibility of proactively alerting users when they may have skipped tracking, allowing users to better self-manage their menstruation. Specifically, users could be alerted when their cycle skipping probability is high, such as at the peak of the skipping probability distribution shown in Figure 4.3. Furthermore, it is important to note that since cycle variability is common, longer cycle lengths can also be due to true physiological phenomena and not just skipped tracking; our model captures this context, which can also be provided to users in these alerts. Implementing this type of informed alerting increases efficacy and accuracy of self-reporting and helps reduce user notification fatigue (which can occur if, for example, everyday alerts are sent out instead of targeted ones); both of these factors are crucial to creating more reliable datasets for the future.

Our model demonstrates the importance of considering the specific nature of mHealth data, which can subsequently aid researchers and users in better understanding menstruation and the underlying structure behind observed cycle lengths. Furthermore, the insights it provides

to mHealth app developers can be used to add nuance to user alerting systems. As self-tracking apps become increasingly popular among users and important among researchers as a source of information for healthcare interventions, these insights can aid in improving the quality of mHealth data and ensuring it is being treated responsibly.

Other efforts to model user-reported menstrual cycle lengths focus on issues like how to represent between-womxn and within-womxn variability, including utilizing hierarchical models [99], linear random effects models [100] that account for how menstrual cycle behavior evolves with age [101], and mixture models of standard cycles (cycles 43 days and shorter) and nonstandard cycles (cycles longer than 43 days) [102]. These studies capture many important aspects of menstruation, like the vitality of considering each user's individual cycle behavior, and include exclusion criteria for users who may not have reported their cycles accurately. However, they fail to explicitly address the user adherence issues inherent in self-reported mHealth data, rendering it difficult to determine whether the observed nonstandard cycles actually resulted from skipped tracking. Furthermore, these studies may be limited in their definition of a standard or nonstandard cycle, the size of the dataset, and the scope of the information available — one advantage of our analysis is that we are able to utilize a large dataset of natural menstrual cycles only, alleviating issues regarding confounding factors like hormonal birth control.

Additionally, since sparsity is a prevalent issue with self-tracked data, a performant model with the minimal type of information needed is beneficial. In this case, cycle length information is both the minimal type of information and the data most commonly recorded by users who use menstrual tracking apps. With observed cycle lengths as our only model input, we are able to achieve error comparable to prior studies. For instance, an RMSE of 1.6 was achieved in a

related study [103]. This error, however, is based on standard cycles only and uses self-tracking data from a mHealth app designed for female athletes (a specific subset of individuals that does not necessarily represent the diversity of womxn). If we similarly consider non-variable cycles only (based on the definition of menstrual regularity as represented in Figure 4.7), our model achieves a comparable median absolute error of 1.5 days. Therefore, we are able to achieve reasonable performance with the whole dataset of cycle lengths only, and even more so when we restrict it to more standard cycles. As compared to other menstrual tracking apps, Clue has a broader target audience, and therefore may be prone to the presence of outliers (due to unexpected cycle skips) that increase the RMSE.

Our study has limitations. Firstly, a risk inherent to our work (and any study that utilizes self-tracking data) is the lack of access to ground truth: knowledge of what the true, experienced cycle lengths are. Relatedly, we do not have explicit user information about events that may disrupt menstruation, like pregnancy or miscarriage. To account for this, we conservatively remove cycles longer than 90 days, as well as those self-identified by the user as being anomalous. Another limitation of this work is that it does not leverage any menstrual symptom information. However, such observations offer great potential to extend this model — our previous work [95] demonstrated how cycle timing and symptom experiences are related, and other studies have included symptom covariates, like cramps and period flow in their models, to examine how these impact reported menstrual cycle length [103]. Including symptomatic information in addition to cycle lengths is crucial to understanding menstrual variability more holistically [68] and may impact cycle prediction accuracy.

By demonstrating our model's ability to successfully detect self-tracking artifacts and outperform alternative baselines in predicting next cycle start, we have showcased the potential

that self-tracking data holds to advance understanding of previously enigmatic physiological processes. Our fully generative model allows for interpretable insight into the mechanisms behind self-tracking behavior, and specifically, skipping behavior. In Chapter 5, we explore extending this fully generative model to a deep generative model that incorporates symptomatic information.

Significance to users: Our results showcase the impact that skipped period tracking can have on accurately anticipating next cycle length (assuming that menstrual cycle tracking apps utilize past cycle lengths to predict the next one, as we have done here). If these results are leveraged by menstrual cycle tracking apps, they can be used to build features that would assist users in tracking their data more consistently and accurately (for instance, with an intelligent tracking alert feature based on the probability of skipped tracking, as discussed in Figure 4.4). Moreover, these insights can be used to not only help users track more effectively, but can also be used to help users understand the mechanisms behind the predictions they may see in apps — for example, a probability of next cycle length could be shown alongside the predicted length itself.

Chapter 5

A hierarchical, deep generative model for menstrual symptoms that accounts for skipped tracking

5.1 Introduction

In Chapter 3, we explored the complex experience of menstruation, which spans symptoms beyond bleeding to impact social, emotional, and physical well-being. In Chapter 4, we hypothesized a fully generative model for reported menstrual cycle length using cycle lengths only as input and predicting next reported cycle length. In this chapter, we will leverage deep learning, which holds potential for greater predictive power than simpler, more interpretable statistical models. In particular, we utilize symptomatic information from the Clue dataset to develop a hierarchical, deep generative model that takes as input per-user time series representations of symptom tracking (including, but not limited to, bleeding tracking) and predicts next occurrence of the tracking event.

In addition to utilizing time series information (which allows us to leverage real-time tracking information and use incomplete cycles, as opposed to only complete cycle lengths), we also learn a population-wide distribution distribution for likelihood of adherence, from which per-user likelihoods are drawn, allowing us to retain the modeled mechanism that tracked in-

formation may not always match with actual experienced behavior. Incorporating available symptom information showcases the flexibility of our model to not only predict bleeding events, but also related symptoms. Modeling these phenomena with a deep generative model, which separates a complex model of symptom dependence from an interpretable model of user adherence, allows for interpretability while also harnessing the power of deep learning. We train our model on bleeding information only, as well as bleeding and other symptom information, utilizing an RNN that learns across symptoms.

5.2 Methods

5.2.1 Data cohort

We utilize the same Clue dataset cohort as described in Chapter 3, leveraging the first $I = 20,000$ users with $T = 180$ days of symptom tracking information per user. We focus on the four most commonly tracked symptom categories — 'bleeding,' 'pain,' 'emotion,' and 'energy.' In Table 5.1, we provide a summary of how many users have tracked these symptoms in the cohort, as well as the proportion of tracking events to total events.

5.2.2 Data preprocessing

Whereas in the previous chapter we utilized cycle lengths (i.e., the number of days between subsequent periods) as data input to our model, here we utilize a time series representation of our tracking data, where a user tracking an event on a given day results in a '1' and absence of tracking results in a '0.' This allows us to more flexibly and accurately represent the symptom tracking experience. For each category, we combine symptoms at the category level, i.e., 'bleeding' represents whether a user has tracked any of 'light,' 'medium,' or 'heavy' on a given

Table 5.1: Overview of dataset

Symptom	No. of users tracked	Prop. of users tracked	Prop. of tracking events to total
Bleeding	19,942	0.997	0.145
Pain	15,342	0.767	0.072
Emotion	12,977	0.649	0.100
Energy	11,681	0.584	0.105

Summary of number of users who tracked a given symptom category in the training set, the proportion of users who tracked out of the total number of users, and the proportion of tracking events to total (i.e., $(1/IT)\sum_{I,T} x_{i,t}$).

day (see Table 3.1 for symptoms associated with each category). Our model acknowledges the possibility that a '0' can represent either true lack of physiological event, or lack of user tracking.

5.2.3 Proposed hierarchical, deep generative model

Our proposed hierarchical, deep generative model learns per-user and per-symptom transition probabilities as the output of a deep RNN and learns population-wide hyperparameters for a population-wide distribution from which per-user likelihoods of adherence are drawn (similarly to the model in Chapter 4). We also learn per-user parameters for likelihood of tracking on the first day of the dataset. The graphical model for this proposed model is shown in Figure 5.1.

Specifically, we propose that $x_{i,t} = z_{i,t}g_{i,t}$, where

- $x_{i,t}$ is the observed (binary) data for user i, day t

- $z_{i,t}$ is the true (binary) data for the same day
- $g_{i,t}$ is an indicator that represents whether or not the user skipped tracking on that day

Intuitively, if $g_{i,t} = 1$, then the user has not skipped tracking on that day (and hence $x_{i,t} = z_{i,t}$).

We propose that $g_{i,t} \sim Bern(b_i)$, where b_i represents the probability of adherence for user i. Note that for this model, we assume the probability of adherence is per user, and this probability is utilized across time to generate $g_{i,t}$ (i.e., if $b_i = 1$, then $g_{i,t} = 1$ for all t, and therefore the user has not skipped tracking). In this model, we also propose that $b_i \sim Beta(\alpha, \beta)$, a population-wide prior distribution for user-level adherence.

Finally, we also define the marginal distribution for the initial $z_{i,0}$ by $\theta_{0,i}$, where $p(z_{i,0} = 1) = 1 - p(z_{i,0} = 0) = \theta_{0,i}$. That is, whereas $z_{i,t}$ for $t \geq 1$ is dependent on $z_{i,t-1}$, initial $z_{i,0}$ is dependent only on $\theta_{0,i}$. We denote θ_0 as a vector where each entry is $\theta_{0,i}$ per user i. We wish to infer α, β, θ_0, and $p(z_{i,t}|z_{i,t-1})$.

The proposed observation likelihood model (based on a Bernoulli indicator $g_{i,t}$) results in the following probabilities:

$$p(g_{i,t}|b_i) = b_i^{g_{i,t}}(1 - b_i)^{1-g_{i,t}} \tag{5.1}$$

and $x_{i,t}$:

$$p(x_{i,t}|z_{i,t}, g_{i,t}) = \begin{cases} x_{i,t} = z_{i,t}, & g_{i,t} = 1 \\ x_{i,t} = 0, & g_{i,t} = 0 \end{cases} \tag{5.2}$$

$$= [x_{i,t} = z_{i,t}]^{g_{i,t}}(1 - x_{i,t})^{(1-g_{i,t})}. \tag{5.3}$$

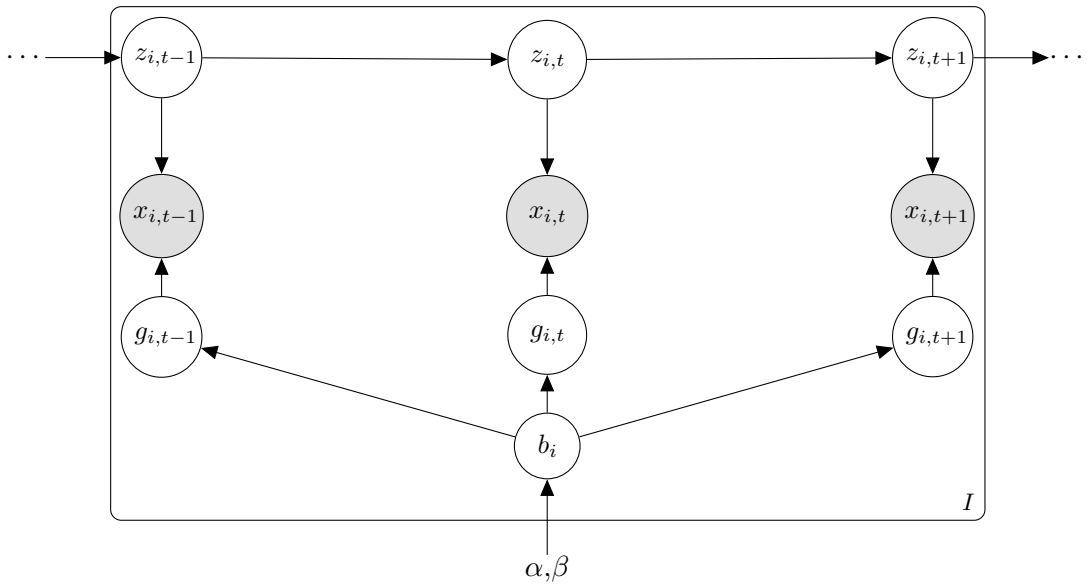


Figure 5.1: Graphical model for deep generative model. $x_{i,t}$ represents observed binary data for user i at time t (0 if tracking is not observed, 1 if tracked is observed), $g_{i,t}$ is an indicator of whether tracking was skipped, b_i represents the probability that a user adhered to tracking, and $z_{i,t}$ represents the true binary data. b_i are drawn from a population-wide Beta distribution, $Beta(\alpha, \beta)$. True data $z_{i,t}$ ranges from $t = 0, \cdots, T$, observed data $x_{i,t}$ ranges from $t = 1, \cdots, T$, and user index i ranges from $1, \cdots, I$. Note: initial emission probability $p(z_{i,0} = 1) = \theta_{0,i}$ is not pictured, but is learned per-user.

5.2.4 Description of RNN

In order to learn the desired latent transition probabilities $p(z_{i,t}|z_{i,t-1})$, we leverage a deep recurrent neural network (RNN). Note that in the following descriptions, we drop per-user dependence i for conciseness and refer to these transition probabilities as $p(z_t|z_{t-1})$ (i.e., when we refer to z_t, this includes a dimension of size I, the number of users). Each layer of an RNN

computes the hidden state

$$h_t = \tanh(\theta_{w,kh} z_{t-1} + \theta_{b,kh} + \theta_{w,hh} h_{t-1} + \theta_{b,hh}) \tag{5.4}$$

where h_t is the hidden state at time t, z_{t-1} is the input at time t, and h_{t-1} is the hidden state of the RNN at time $t-1$ (or, at time 0, the initial hidden state). k indexes the input size (i.e., the number of features in the input z_{t-1}, which is the number of symptoms S). We refer to the weights and biases of the RNN collectively as $\theta_h = [\theta_{w,kh}, \theta_{w,hh}, \theta_{b,kh}, \theta_{b,hh}]$, where

- $\theta_{w,kh}$ represents the learnable input-hidden weights, applied to the input, z_{t-1} and is of shape (`hidden size`, S)
- $\theta_{b,kh}$ represents the learnable input-hidden biases and is of shape (`hidden size`)
- $\theta_{w,hh}$ represents the learnable hidden-hidden weights, applied to the previous hidden state, h_{t-1} and is of shape (`hidden size`, `hidden size`)
- $\theta_{b,hh}$ represents the learnable hidden-hidden biases and is of shape (`hidden size`)

We can then refer to the computation of the hidden state as

$$h_t = \tanh(\theta_{w,kh} z_{t-1} + \theta_{b,kh} + \theta_{w,hh} h_{t-1} + \theta_{b,hh}) \tag{5.5}$$

$$= f_h(z_{t-1}; h_{t-1}, \theta_h) \tag{5.6}$$

The final (output) layer of the model applies a linear transformation to the final hidden state of the RNN (i.e., $\text{Linear}(x, A, b) = xA^T + b$) and utilizes a log sigmoid function (i.e., $\text{LogSigmoid}(x) = \log\left(\frac{1}{1+\exp(-x)}\right)$) to provide the the output of interest at time t, i.e., the transition probabilities. A small detail here is that the output is computed in log space, which we then exponentiate.

To summarize, the desired latent quantities to be learned are $\theta = [\theta_0, \theta_h, \theta_t, \alpha, \beta]$, where

- θ_0 represents the initial emission probabilities, i.e., $p(z_{i,0} = 1)$, which are per-user.
- θ_h represents the weights and biases of the neural network.
- α and β represent the population-wide hyperparameters for the population-wide Beta distribution for adherence probabilities, i.e., $b_i \sim Beta(\alpha, \beta)$. These b_i can be referred to as θ_e, the emission probabilities.

The transition probabilities $p(z_{i,t} = 1|z_{i,t-1})$ (represented as θ_t) are the output of the RNN:

$$p(z_{i,t} = 1|z_{i,t-1}) = f_p(z_{i,t-1}; h_{t-1}, \theta_h) \tag{5.7}$$

where f_p is a deep RNN, as described above; h_{t-1} is the last hidden state of the RNN; and θ_h are the weights and biases of the neural network. That is, our transition probabilities $p(z_{i,t}|z_{i,t-1})$ are the output of an RNN that takes as input $z_{i,t-1}$, the previous hidden state h_{t-1}, and the previous learned parameters θ_h. The RNN learns across symptoms and returns per-symptom and per-user transition probabilities. The remaining quantities θ_0, α, and β (which are used to draw θ_e, i.e., b_i) are parameters of the hierarchical, deep generative model that are learned via backpropagation. We utilize PyTorch for our RNN implementation [104].

5.2.5 Inference using the approximate expected log likelihood

It may be useful to consider our model from the paradigm of a Hidden Markov Model (HMM) [105], where the latent state space is $z \in \{0, 1\}$ (for each symptom), the observed data are $x \in \{0, 1\}$, the emission probabilities are θ_e (drawn from a $Beta(\alpha, \beta)$, where α and β are learned population-wide hyperparameters), and the transition probabilities are θ_t.

To infer the desired parameters of an HMM, we can use expectation-maximization, or EM [105]. In EM, we first compute the expected log likelihood of the model parameters θ

with respect to the current posterior distribution of the latent states $z_{0:T}$ given some observed sequence $x_{1:T}$ (i.e., $p(z_{0:T}|x_{1:T})$) and the current parameter estimates (this is the expectation, or E step), and then find the parameters θ that maximize this quantity (this is the maximization, or M step). In order to compute the expected log likelihood as our loss function, which we refer to as Q, we begin by writing the joint likelihood as:

$$p(x_{1:T}, z_{0:T}|\theta) = \sum_{g_{1:T}} p(x_{1:T}, g_{1:T}, z_{0:T}|\theta) \tag{5.8}$$

$$= \sum_{g_{1:T}} \left(p(z_0|\theta_0) \prod_{t=1}^{T} p(x_t, g_t, z_t|\theta, z_{t-1}, h_t) \right) \tag{5.9}$$

$$= \sum_{g_{1:T}} \left(p(z_0|\theta_0) \prod_{t=1}^{T} p(z_t|h_t;\theta_t) \prod_{t'=1}^{T} p(x_{t'}, g_{t'}|z_{t'};\theta_e) \right) \tag{5.10}$$

$$= p(z_0|\theta_0) \prod_{t=1}^{T} p(z_t|h_t;\theta_t) \sum_{g_{1:T}} \left(\prod_{t'=1}^{T} p(x_{t'}, g_{t'}|z_{t'};\theta_e) \right) \tag{5.11}$$

$$= p(z_0|\theta_0) \prod_{t=1}^{T} p(z_t|h_t;\theta_t) \left(\prod_{t'=1}^{T} p(x_{t'}|z_{t'},\theta_e) \right) \tag{5.12}$$

where the hidden state of the RNN at time t is represented as $h_t = f_h(h_{t-1}, z_{t-1};\theta_h)$ and f_h represents the hidden layer computation (we utilize f_p to refer to the full RNN, including the output layer, whereas f_h refers to the hidden layer computation only).

We then write the expected log likelihood Q as

$$Q(\theta) = E_{z_{0:T}|x_{1:T},\theta} \log p(x_{1:T}, z_{0:T}|\theta) \tag{5.13}$$

$$= \sum_{z_{0:T}} p(z_{0:T}|x_{1:T},\theta) \log \left(p(z_0|\theta_0) \prod_{t=1}^{T} p(z_t|h_t;\theta_t) \left(\prod_{t'=1}^{T} p(x_{t'}|z_{t'},\theta_e) \right) \right) \tag{5.14}$$

$$= \sum_{z_{0:T}} p(z_{0:T}|x_{1:T},\theta) \left(\log p(z_0|\theta_0) + \sum_{t=1}^{T} \log p(z_t|h_t;\theta_t) + \sum_{t'=1}^{T} \log p(x_{t'}|z_{t'},\theta_e) \right) \tag{5.15}$$

where again, $h_t = f_h(h_{t-1}, z_{t-1};\theta_h)$.

Note that our model differs from an HMM in that our transition from one state to the next

at time t is implicitly dependent on all time points prior to t through the hidden state of the RNN, h_{t-1} (i.e., h_t depends on h_{t-1}, which subsequently depends on the previous one and so forth). That is, from a generative standpoint, in order to generate z_t, we would need to not only know z_{t-1}, but also $[z_0, \ldots, z_{t-2}]$, and we cannot sum over all of the possible paths of $z_{0:T}$, since there are 2^S (where S is the number of symptoms) possible states at each timestep t, and this sum would explode exponentially (for instance, if there are two symptoms, the possible states at a timestep are $[0, 0]$, $[0, 1]$, $[1, 0]$, and $[1, 1]$, since z_t can be 0 or 1 for each symptom).

We can approach this by instead approximating the sum using the Viterbi path, i.e., the most probable path of latent $\tilde{z}_{1:T}$ based on $p(z_{1:T}|x_{1:T})$, as computed by iterating forward in time through the data $x_{1:T}$, and then updating estimates backwards. In addition to computing the most likely latent states, we also need to keep track of the hidden states h_t associated with these latent states.

We can then write our Viterbi path-based approximation to Q as

$$\hat{Q}(\theta) = \sum_{z_{0:T}} [z_{0:T} = \tilde{z}_{0:T}] \left(\log p(z_0|\theta_0) + \sum_{t=1}^{T} \log p(z_t|h_t; \theta_t) + \sum_{t'=1}^{T} \log p(x_{t'}|z_{t'}, \theta_e) \right) \tag{5.16}$$

$$= \log p(\tilde{z_0}|\theta_0) + \sum_{t=1}^{T} \log p(\tilde{z_t}|h_t; \theta_t) + \sum_{t'=1}^{T} \log p(x_{t'}|\tilde{z}_{t'}, \theta_e) \tag{5.17}$$

where $\tilde{z}$ represents the Viterbi path; we provide details for computing $\tilde{z}$ below. In order to learn our parameters θ (via the M step), we optimize $\hat{Q}$ numerically, updating $\hat{\theta} = \arg\max_\theta \hat{Q}(\theta)$ at each iteration. We provide optimization details later on in subsection 5.2.8.

5.2.6 Computing the Viterbi path, the most probable path iterating forward and backward through x

For the RNN-based model, we have

- Hidden states as deterministic functions of previous hidden state h_{t-1} and previous latent state $z_{t-1} = s'$, where $s' \in \{0, \ldots, 2^S - 1\}$ (S is the number of symptoms, and there are 2^S possible states; for the reduced case of one symptom, $s' \in \{0, 1\}$):

$$h_t = f_h(z_{t-1} = s', h_{t-1}; \theta_h) \tag{5.18}$$

- The time-varying transition probabilities as computed using the RNN:

$$p(z_t = s | z_{t-1} = s', h_{t-1}) = p\left(z_t = s | h_t = f_h(z_{t-1} = s', h_{t-1}; \theta_h)\right) \tag{5.19}$$

$$= f_p(z_{t-1}; h_{t-1}, \theta_h) \tag{5.20}$$

For the Viterbi algorithm, the key quantities to compute are

- $T_{prob}[s, t]$, which keeps track of the maximum probability of being at state s at time t

$$T_{prob}[s, t] = \max_{z_{t-1}=s'} \left(T_{prob}[z_{t-1} = s', t-1] p(z_t = s | z_{t-1} = s') p(x_t | z_t = s)\right) \tag{5.21}$$

- $T_{state}[s, t]$, which keeps track of the state $z_{t=1} = s'$ at $t - 1$ that leads to the most likely probability $T_{prob}[s, t]$

$$T_{state}[s, t] = \arg\max_{z_{t-1}=s'} \left(T_{prob}[z_{t-1} = s', t-1] p(z_t = s | z_{t-1} = s') p(x_t | z_t = s)\right) \tag{5.22}$$

- $T_{hidden}[s, t]$, which keeps track of the hidden state $h_t(s')$ at t that leads to the most likely probability $T_{prob}[s, t]$

$$T_{hidden}[s, t] = h_t(T_{state}[s, t]) = h_t(z_{t-1} = T_{state}[s, t], T_{hidden}[T_{state}[s, t], t-1])). \tag{5.23}$$

As such, we can compute the Viterbi path $\tilde{z}$ by iterating forward and backward through time, as shown in Algorithm 1. Note that since some of our models are trained on multiple symptoms, the Viterbi path is over states of size 2^S, where S represents the number of symptoms.

Algorithm 1 Algorithm for computing Viterbi path $\tilde{z}$ by first iterating forward through time, for $t = 1, \cdots, T$ to compute most likely z_t and associated hidden states, then iterating backwards through time $T, \cdots, 1$ to update predictions of z_t as final path $\tilde{z}$. States are represented by s', where $s' \in \{0, \ldots, 2^S - 1\}$ (S is the number of symptoms, and there are 2^S possible states).

Input: θ_0, h_0, $p(x_t|z_t)$

Input: Sequence of observations $x_{0:T}$

Initialization for each possible latent state s'

- Initialize $T_{prob}[s, 0]$ for observation x_0

$$T_{prob}[s = s', 0] = p(z_0 = s') \cdot p(x_0|z_0 = s') \tag{5.24}$$

- Initialize $T_{state}[s, 0]$

$$T_{state}[s = s', 0] = 0 \tag{5.25}$$

- Initialize $T_{hidden}[s, 0]$

$$T_{hidden}[s = s', 0] = h_0 \tag{5.26}$$

for $t = 1, \cdots, T$ **do**

Compute hidden state for each possible latent state $z_{t-1} = s'$

$$h_t(s') = h_t(z_{t-1} = s', h_{t-1}(s') = T_{hidden}[s', t-1]) \tag{5.27}$$

Compute $T_{prob}[s, t]$ for observation x_t over possible latent states at prior timestep $z_{t-1} = s'$ for each latent state s^*

$$T_{prob}[s = s^*, t] = \max_{z_{t-1}=s'} \left(T_{prob}[z_{t-1} = s', t-1] p(z_t = s^*|z_{t-1} = s', h_{t-1}(s')) p(x_t|z_t = s^*) \right) \tag{5.28}$$

$$= \max_{z_{t-1}=s'} \left(T_{prob}[z_{t-1} = s', t-1] p(z_t = s^*|h_t(s') p(x_t|z_t = s^*) \right) \tag{5.29}$$

7: Keep track of $T_{state}[s,t]$ over possible latent states at prior timestep $z_{t-1}=s'$ for each latent state s^*

$$T_{state}[s=s^*,t] = \underset{z_{t-1}=s'}{\arg\max} \left(T_{prob}[z_{t-1}=s',t-1]p(z_t=s^*|z_{t-1}=s',h_{t-1}(s'))p(x_t|z_t=s^*)\right) \tag{5.30}$$

$$= \underset{z_{t-1}=s'}{\arg\max} \left(T_{prob}[z_{t-1}=s',t-1]p(z_t=s^*|h_t(s'))p(x_t|z_t=s^*)\right) \tag{5.31}$$

8: Keep track of $T_{hidden}[s,t]$ for each latent state s^*

$$T_{hidden}[s=s^*,t] = h_t(T_{state}[s=s^*,t]) = h_t(z_{t-1}=T_{state}[s=s^*,t], h_{t-1}(T_{state}[s=s^*,t])) \tag{5.32}$$

9: **end for**

10: Select final most likely state $\tilde{z}_T = \arg\max_s T_{prob}[s,T]$

11: **for** $t=T,\cdots,1$ **do**

12: Recover state sequence

$$\tilde{z}_{t-1} = T_{state}[\tilde{z}_t, t] \tag{5.33}$$

13: **end for**

5.2.7 Simulated periodic data

In order to evaluate the ability of our model to recover true tracking adherence, we generate simulated periodic data, in which each user follows roughly the same regular tracking pattern with some variance, reporting tracking events for a given reported duration (i.e., a given number of days in a row) followed by a length of days with no tracking. In basic terms, the data for each symptom will be a string of 0s and 1s, repeated with roughly the same periodicity. For each generated symptom, we define the length of days with no tracking as a set periodicity (i.e., 28) divided by the symptom number. For instance, if we generated two symptoms, the first symptom would have a periodicity of about 28 (about 28 days in a row with no tracking), whereas the second would have a periodicity of about 14. We then generate the indicator for whether a user tracked the true event ($g_{i,t}$) according to a true tracking adherence probability b_i and compute $x_{i,t} = z_{i,t}g_{i,t}$ to generate the observed data x.

Specifically, each user's first spike is random within the first 5 days (drawn from a uniform) and their per-symptom length of no tracking (i.e., number of 0s in a row) varies from a set periodicity (for instance, 28 days for symptom 1, or 14 days for symptom 2) plus a random variance of up to 2 days. The reported duration for each symptom (i.e., the number of 1s that are tracked in a row) is the set periodicity length (for instance, 2) plus a random variance of up to 2 days. This allows us to generate a simulated dataset that is fairly regular, but with some variation between users, symptoms, and cycles.

5.2.8 Data selection, training, and optimization

We split our data x into training and testing based on time; if T_{train} represents the number of training days and T_{test} represents the number of testing days (where $T_{train} + T_{test} = T$, the

total number of days per user in the dataset), then

$$x_{train} = x_{i=1:I,t=1:T_{train}} \tag{5.34}$$

$$x_{test} = x_{i=1:I,t=T_{train}+1:T} \tag{5.35}$$

In addition, we implement an option to check whether users have tracked a given symptom in their training set, i.e., in x_{train}, otherwise we utilize the full $I = 20,000$. If we are utilizing more than one symptom, we check that the user has tracked all symptoms at least once in their training set.

We utilize the Adam optimizer [106] with $-\hat{Q}(\theta)$ as our loss function, running the optimization procedure to 2000 epochs with a loss epsilon convergence criteria of $1e-6$, a learning rate of 0.0001, and a batch size of 1000. We utilize a 3-layer RNN with a hidden size of 30 and utilize two initializations for α and β, and since weights and biases for the neural network are initialized randomly, we run each experiment for 3 distinct, random seeds. For all experiments, we initialize the initial emission parameter $\theta_0 = 0.5$.

5.2.9 Prediction by day

We compute our predictions per-user and per-symptom by day, i.e., we update our predictions at each timestep. We initialize our predictions $\hat{z}_{i,0}$ at time $t = 0$ with our learned θ_0, i.e., $\hat{z}_{i,0} = [\theta_0 \geq 0.5]$. In other words, if $\theta_0 \geq 0.5$, then $\hat{z}_{i,0} = 1$.

For subsequent predictions, we first update our prediction $\hat{z}_{i,t}$ by computing the most likely Viterbi path including the observed data point $x_{i,t}$ (and utilizing the last timestep in the updated path as $\hat{z}_{i,t}$). We then utilize this updated estimate of the most likely last latent state as the input to our RNN (along with the updated most likely hidden state) and utilize the

output $p(z_t|z_{t-1})$ to compute the most likely next z_t. As with $\hat{z}_{i,0}$, $\hat{z}_{i,t} = [p(z_{i,t} = 1|z_{i,t-1}) \geq 0.5]$.

On each day of the dataset, we predict 40 days into the future for each user (based on our transition probabilities and the Viterbi path updated with the data for that day). That is, for each day of prediction, we compute the Viterbi path including the new observed day x_t, and then compute predictions from day $t : t + 40$. This results in a prediction matrix of size $(I, T, S, 40)$, where I is the number of users, T is the number of days per-user in the dataset, S is the number of symptoms, and 40 is the prediction window (the number of days we predict into the future on each day). Our prediction approach is outlined in Algorithm 2. This extends the predictions from Chapter 4, which represented updating cycle length predictions on each day (i.e predicting one day out into the future).

Algorithm 2 Algorithm for computing predictions on each day of the dataset for a prediction window of size 40. We omit the subscript over i for clarity, but predictions are made per-user.

Initialize predictions $\hat{z}_{i,0}$ at time $t = 0$ with $\hat{z}_{i,0} = [\theta_0 \geq 0.5]$
for $t = 1, \cdots, T$ **do**
    Compute an updated Viterbi path $\tilde{z}_{1:t}$ and associated hidden states based on including new observation x_t
    Set $\hat{z}_t$ as the last latent state in the updated Viterbi path, i.e., $\hat{z}_t = \tilde{z}_t$
    for $w = 1, \cdots, 40$ **do**
        Pass $\hat{z}_{t+w-1}$ and last hidden state as input to trained RNN; output transition probabilities $p(z_{t+w}|z_{t+w-1})$ and last hidden state
        Update $\hat{z}_{t+w}$ based on current transition probabilities, $\hat{z}_{t+w} = [p(z_{t+w} = 1|z_{t+w-1}) \geq 0.5]$ and keep track of transition probabilities to later predict x
    end for
end for
Predict $\hat{x}_t = [p(x_t = 1|z_{t-1}) \geq 0.5]$

We then compute our predictions for $\hat{x}_{i,t}$ by first computing $p(x_{i,t} = 1|z_{i,t-1})$, marginalizing over $z_{i,t}$, i.e.,

$$p(\hat{x}_{i,t} = 1|z_{i,t-1}) = \sum_{z_{i,t}} p(x_{i,t} = 1|z_{i,t})p(z_{i,t}|z_{i,t-1}) \tag{5.36}$$

$$= p(x_{i,t} = 1|z_{i,t} = 0)p(z_{i,t} = 0|z_{i,t-1}) + p(x_{i,t} = 1|z_{i,t} = 1)p(z_{i,t} = 1|z_{i,t-1}) \tag{5.37}$$

$$= p(x_{i,t} = 1|z_{i,t} = 1)p(z_{i,t} = 1|z_{i,t-1}) \tag{5.38}$$

$$= b_i p(z_{i,t} = 1|z_{i,t-1}). \tag{5.39}$$

Then, as with our predictions on $z_{i,t}$, we choose to threshold the probability in order to compute our prediction, i.e., $\hat{x}_{i,t} = [p(x_{i,t} = 1|z_{i,t-1}) \geq 0.5]$.

Note that we have two options for the predicted outcome — we can predict $\hat{z}_{i,t}$, i.e., the

true experienced event, or $\hat{x}_{i,t}$, the tracked event. We focus on comparing $\hat{x}_{i,t}$ to observed $x_{i,t}$, since this comparison is one-to-one when we only have access to data $x_{i,t}$ (on simulated data, for instance, we can compare $\hat{z}_{i,t}$ to true $z_{i,t}$).

Since we learn a prior on b_i, when we compute predictions, we would like to use the posterior b_i (after observing x_{train}). To do so, we compute the approximate expected posterior b_i given observed x_{train} and $\tilde{z}_{train}$ (based on the Viterbi path having observed x_{train}), $\hat{E}(b|x_{train}, \tilde{z}_{train})$ with a Monte Carlo estimate. While we omit the subscripts for conciseness, this is computed per user and symptom.

Specifically, we start by writing out the desired expectation:

$$E(b|x_{train}, \tilde{z}_{train}) = \int_b b \cdot p(b|x_{train}, \tilde{z}_{train}) db. \tag{5.40}$$

Since we cannot compute this analytically, we approximate it with a sum, starting first by writing the posterior for b (based on observing the training set, x_{train}, and computing the most likely latent state path $\tilde{z}_{train}$) as

$$p(b|x_{train}, \tilde{z}_{train}) = \frac{p(x_{train}|b, \tilde{z}_{train})p(b)}{p(x_{train}|\tilde{z}_{train})} \tag{5.41}$$

$$\propto p(x_{train}|b, \tilde{z}_{train})p(b). \tag{5.42}$$

We can then approximate this with a Monte Carlo estimate, where

$$\hat{E}(b|x_{train}, \tilde{z}) = \sum_{m=1}^{M} \bar{w}^{(m)} b^{(m)} \tag{5.43}$$

In order to compute the weights, $\bar{w}^{(m)}$, we first draw $M = 100$ samples of $b^{(m)}$ from our learned prior, i.e., $b^{(m)} \sim Beta(\hat{\alpha}, \hat{\beta})$, and compute $w^{(m)} = p(x_{train}|b^{(m)}, \tilde{z}_{train})$. We then normalize our weights, computing $\bar{w}^{(m)} = \frac{w^{(m)}}{\sum_m w^{(m)}}$. This normalization ensures that the weights sum to 1, and that therefore the approximate posterior probability distribution for

b also sums to 1 over the drawn samples. Again, note that we compute these posteriors per individual and symptom.

5.2.10 Evaluation

In order to evaluate predictions, we align user data and predictions on 'day 0' of the cycle in their test set, i.e., the first day of bleeding in the test set that is at least 7 days away from the last day of bleeding in the train set. Then, to compute the predicted next cycle start for each user, we look at the first day of predicted bleeding that is at least 7 days away from 'day 0.'

We evaluate our models by computing the AUC (area under the receiving operator characteristic, or ROC, curve) [107], which evaluates our ability to distinguish true / false negatives and positives at different false positive rate thresholds. We refer to this metric as $\text{AUC}(x_{i,t}, p(\hat{x}_{i,t} = 1|z_{i,t-1}))$ (computed across time t and users i). We evaluate this AUC overall in order to compare it with a baseline, as well as on specific days (for instance, we can compute the AUC of predicting day 29 of the cycle). Since we have a window of 40 predictions for each day of the cycle, we can showcase this AUC as we approach the desired day. For instance, to evaluate predictions of day 26, we can look at prediction on day 1 for a window of 25, day 2 for a window of 24, and so on. AUC allows us to evaluate our model across prediction thresholds, and our prediction cutoff of 0.5 corresponds to a particular true positive and false positive rate (therefore corresponding to a specific point on the ROC curve).

In addition to AUC, we compute RMSE of our predicted next cycle start versus true (observed) next cycle start. Again, we can showcase this RMSE as a function of prediction day (day of the test data, aligned at day 0 per user).

5.2.11 Alternative baseline

We utilize a one-day lag baseline, which predicts $\hat{x}_{i,t} = x_{i,t-1}$. Note that for this baseline, we cannot provide a window of predictions on a given day (since this model relies on seeing each day in order to predict the next).

5.3 Results

In this section, we demonstrate the key results of our work. Firstly, we provide loss plots over epochs for different optimization methods, showcasing Adam's superior performance relative to alternatives. We then showcase our model's ability to correctly learn the empirical adherence b (where lack of adherence indicates that $x \neq z$, i.e., not considering cases of $x = z = 0$) on simulated data, which demonstrates proper parameter inference. Next, we outline our model's performance on the real data (based on results averaged across 3 seeds), focusing on two main prediction tasks: predicting events in the future and predicting the next cycle start. For the former task, we showcase AUC of predicting a day where the event is most likely to have happened for that symptom. We also showcase the AUC of predicting day 29 of bleeding, which is the most likely day of the next cycle start in our dataset. Note that in plots of AUC, we label the y-axis as 'AUC_future' to represent the AUC of predicting into the future. For the latter task, we showcase RMSE (and predictions) of next cycle start over prediction day for a model that uses only bleeding, bleeding and emotion, bleeding and pain, and bleeding and energy, assessing how additional symptoms affect performance.

Note that we see minimal differences between initializing prior $\alpha = 2, \beta = 2$ (uninformative) and $\alpha = 5, \beta = 1$ (informative), so for conciseness we focus here on the results for initializing

$\alpha = 5, \beta = 1$. In all experiments, we fit and evaluate on users who have tracked the given symptom(s) in their training data (i.e., ensuring that there are no users for whom their training data are all 0s).

5.3.1 Evaluation of optimization

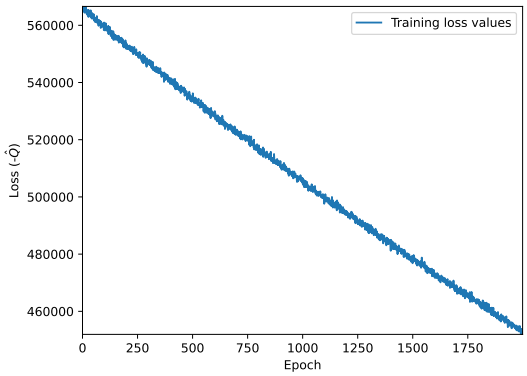
In order to assess the viability of different optimization methods, we test our model's learning using three popular methods for training deep models [56]: stochastic gradient descent (SGD) and two adaptive gradient methods, Adadelta [108] and Adam [106] on a simulated dataset of $I = 5,000$ and $T = 180$. We find that Adam learns the quickest and is able to reach an optimal minimum compared to the other two methods, as seen in Figure 5.2.

5.3.2 Inference of b on simulated periodic data

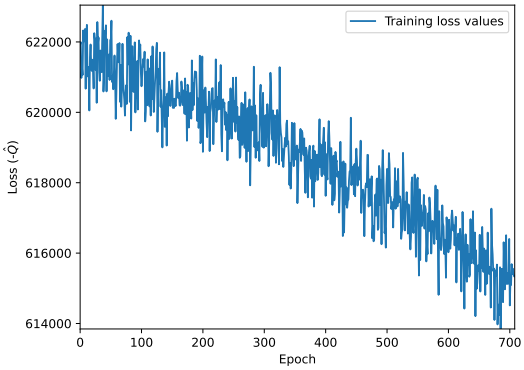
In order to assess our model's ability to correctly learn b, we generate simulated data for $I = 5,000$ and $T = 180$, as described in the Methods section above. Utilizing a generative $b = 0.8$, we evaluate our model's learning of b in Figure 5.3 across two initializations ($\alpha = 2$, $\beta = 2$ and $\alpha = 5$, $\beta = 1$) by examining the learned prior and posterior values. We choose these initializations since $(2, 2)$ represents an uninformative prior (with an expected value of $E(b) = 0.5$) and $(5, 1)$ represents an informative prior (with an expected value of $E(b) = 0.83$). Based on our results in Chapter 3 with excluding suspected cycle tracking artifacts, our belief is that user adherence will be above 0.5.

Additionally, note that we compare these learned values to 'adherence b,' which we define as

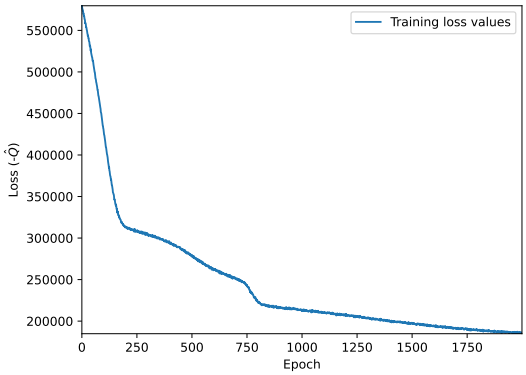
$$\frac{n_{[x_t=1,z_t=1]}}{n_{[x_t=0,z_t=1]} + n_{[x_t=1,z_t=1]}} \tag{5.44}$$



(a) Loss over epochs using SGD optimization method.



(b) Loss over epochs using Adadelta optimization method.



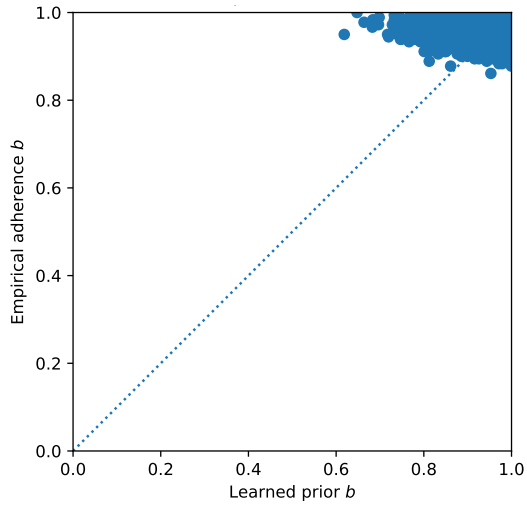
(c) Loss over epochs using Adam optimization method.

Figure 5.2: Loss ($-\hat{Q}(\theta)$) over epochs for different optimization methods.

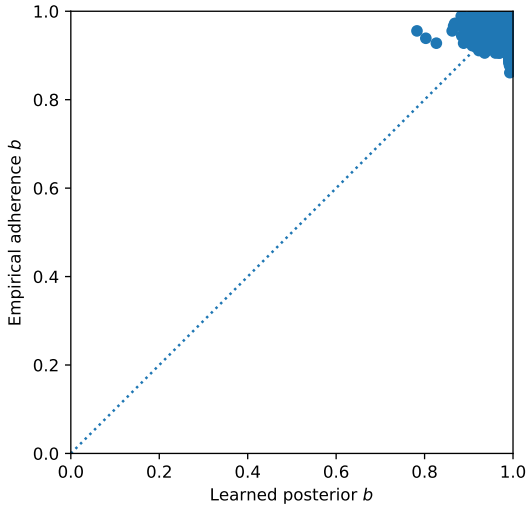
where $n_{[x_t=i,z_t=j]}$ represent the number of instances in the observed data where $x_t = i$ and $z_t = j$. This effectively means that instances of $x \neq z$ indicate lack of adherence and differs from typical empirical b, since it does not consider instances where $x = z = 0$ to be skipped tracking (even if the generated $g = 0$, since we do not know what the true z is in this case). The lower bound of adherence b is determined by the proportion of 0s to 1s in the observed data. Our simulated dataset has a high proportion of 0s to 1s, since this mirrors the true dataset most closely, which means that the adherence b will generally be high. For an explanation of

computing the MLE of b, i.e., maximizing $p(x|z, b)$, and why this MLE is the adherence b (not the generative b), see Appendix C.

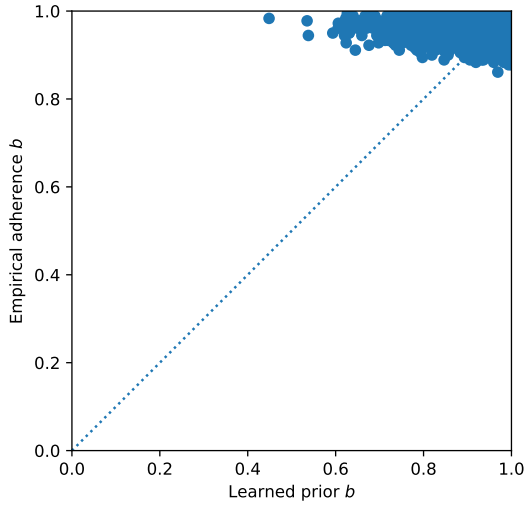
Figure 5.3 showcases that across the two initializations, our learned posterior b matches the true adherence b (and that we are able to learn this regardless of where the prior is initialized). Therefore, our model is able to successfully recover the truth on simulated data. See Figure C.1 of Appendix C for example plots of learned α and β values for bleeding only model with initializations of $(2, 2)$ and $(5, 1)$ on real data.



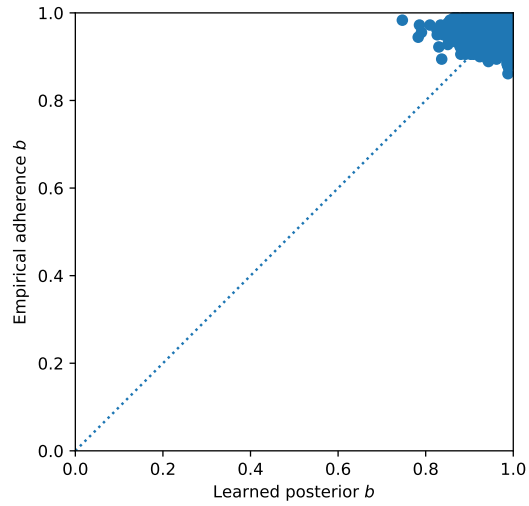
(a) Prior initialized to $(2, 2)$; adherence b versus learned prior. Scatter plot of learned prior versus empirical adherence b.



(b) Prior initialized to $(2, 2)$; adherence b versus learned posterior. Scatter plot of learned posterior versus empirical adherence b.



(c) Prior initialized to $(5, 1)$; adherence b versus learned prior. Scatter plot of learned prior versus empirical adherence b.



(d) Prior initialized to $(5, 1)$; adherence b versus learned posterior. Scatter plot of learned posterior versus empirical adherence b.

Figure 5.3: Learned prior and posterior b_i vs. true adherence b_i on simulated data with different initializations of (α, β). We see that the learned prior b_i values have more spread across users, whereas the learned posterior b_i values cluster around the adherence b, showcasing our model's ability to successfully recover the truth (i.e., the value on the y-axis).

5.3.3 Predicting future event

In this section, we examine our model's ability to predict the occurrence of a future event. First, we provide overall per-symptom test AUC for our model (based on predicting the next day out) versus the one-day lag baseline in Table 5.2. We see that across symptom models, our model outperforms the baseline.

Table 5.2: Overall test AUC vs. baseline, evaluated per symptom. Models are either trained on bleeding only or bleeding and another symptom.

Input symptom(s)	Model AUC per symptom	Baseline AUC per symptom
Bleeding	0.95	0.85
Bleeding and pain	0.95, 0.89	0.84, 0.73
Bleeding and emotion	0.95, 0.92	0.84, 0.79
Bleeding and energy	0.95, 0.92	0.84, 0.81

Note that as mentioned previously, we choose a prediction threshold of 0.5. For a visual of how this threshold can impact predictions, see Figure 5.4, where we showcase an example of computed per-user $p(\hat{x}_{i,t} = 1|z_{i,t-1})$ (labeled as $p(\hat{x} = 1)$ in the plot for brevity) over day of the dataset, based on the learned transition probabilities $p(z_{i,t}|z_{i,t-1})$ and adherence b_i. These predictions are based on updating predictions by day (and predicting one day out). The vertical blue lines indicate where the observed data $x_{i,t}$ is 1, the dotted red line indicates the prediction threshold of 0.5, and the blue triangles indicate the computed $p(\hat{x}_{i,t} = 1|z_{i,t-1})$ based on learned parameters at the last epoch (in this case, epoch 249). We see that while in

general our prediction threshold allows us to capture tracking successfully, in this instance, a slightly lower prediction threshold would've allowed for the identification of a few more true positives, such as the 1s before day 140. This is not the case across all users, but can be a useful consideration for interpreting our reported metrics.

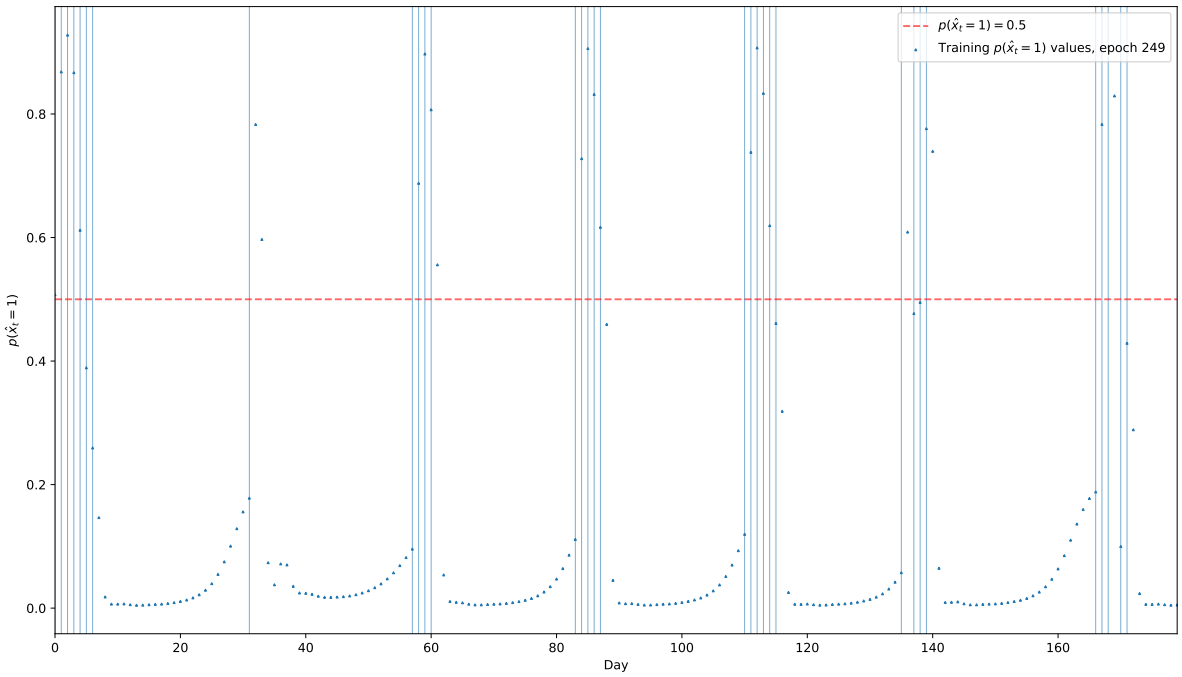
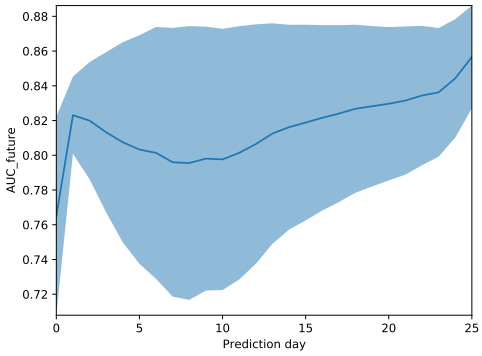


Figure 5.4: Computed prediction probabilities $p(\hat{x}_t = 1)$ (more precisely, $p(\hat{x}_{i,t} = 1|z_{i,t-1})$) over time for a particular user and particular seed, utilizing the bleeding only model and predicting one day out. Vertical lines represent where the observed data contains a 1, i.e., where the user tracked an event, and the dotted red line indicates our prediction threshold of 0.5. We see how our model captures prediction probabilities over time in a multimodal manner — probabilities generally increase when a tracking event is coming up and decrease after the tracking period has finished. In this instance, a lower prediction threshold may have allowed for us to identify more true positives.

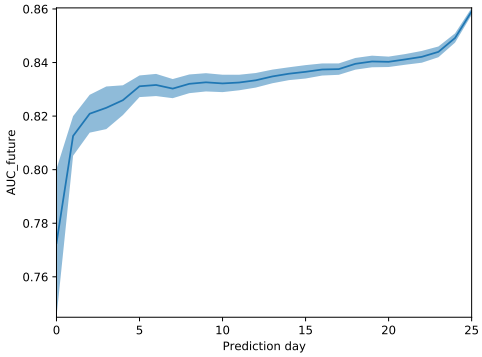
Next, we showcase our model's ability to predict a future symptom event (other than bleeding) in Figure 5.5 — for each symptom, we choose a day in the test set (aligned at day 0)

for which there is a high proportion of tracking and plot the AUC of predicting that day as a function of prediction day (i.e., as the test set approaches the day we are trying to predict). We see that across all symptoms, our model achieves an AUC of between 0.72 and 0.88, and that this AUC increases as a function of prediction day, indicating that our performance improves as we approach the event day. Consequently, for models where we include a symptom in addition to bleeding, we can provide the user with an accurate sense of when that symptom will occur in the next cycle.

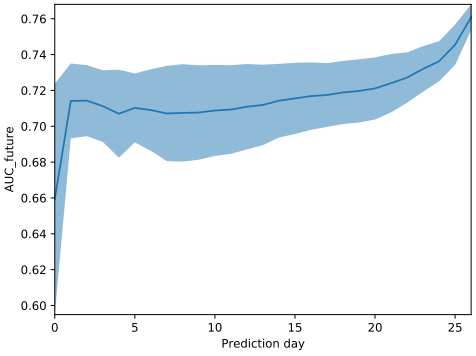
We also examine the AUC of predicting day 29 of bleeding for all models in Figure 5.6, since this is the most common day of the next cycle start. We see that for all models, AUC increases over prediction day, indicating that as we approach the day of the bleeding event, our models are able to more accurately predict its occurrence. Between models, we see minor differences between AUC on this particular day (this effect changes slightly depending on the day in question). However, bear in mind that the user sets differ between models (and that therefore the proportion and distribution of bleeding events also differs). For instance, the bleeding only model has $16,672$ eligible users, i.e., users with day 0 at least 29 days before the end of the test set; the model with energy has $9,776$; the model with emotion has $10,841$; and the model with pain has $12,845$. We see that for models with a symptom included, the variance of the AUC between seeds is lower than that of the bleeding only model.



(a) AUC of predicting day 26 of energy with bleeding and energy model.

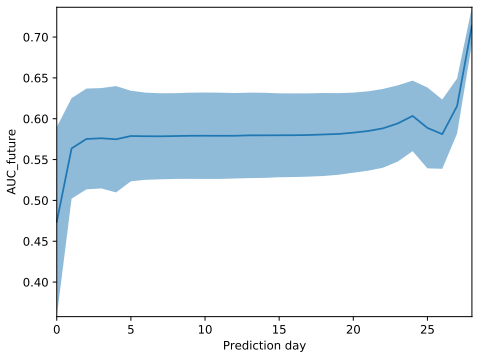


(b) AUC of predicting day 26 of emotion with bleeding and emotion model.

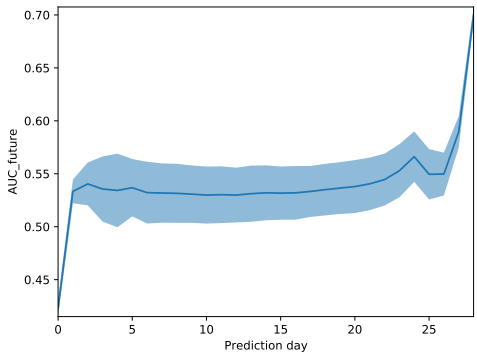


(c) AUC of predicting day 27 of pain with bleeding and pain model.

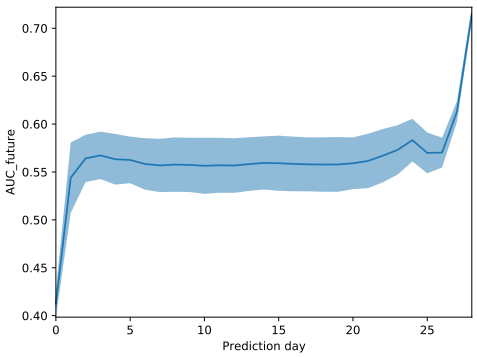
Figure 5.5: AUC of predicting future symptom events for each model with a symptom in addition to bleeding as input. We see that across models, we are able to predict future symptom events well, and that this performance improves as the prediction day approaches the day of the event we are trying to predict.



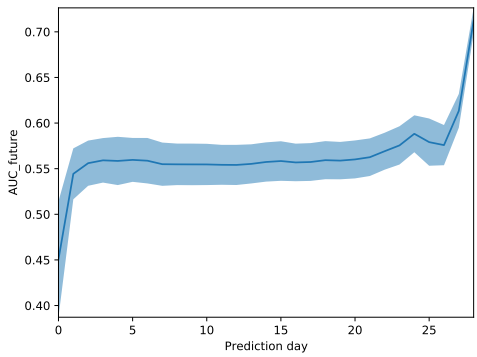
(a) AUC of predicting day 29 of bleeding with bleeding only model.



(b) AUC of predicting day 29 of bleeding with bleeding and energy model.



(c) AUC of predicting day 29 of bleeding with bleeding and emotion model.



(d) AUC of predicting day 29 of bleeding with bleeding and pain model.

Figure 5.6: AUC of predicting day 29 of bleeding across models. We see that across models, we are able to predict day 29 of bleeding (the most common cycle length in the dataset) well, with an AUC of about 0.7 as we approach day 29.

5.3.4 Predicting next cycle start

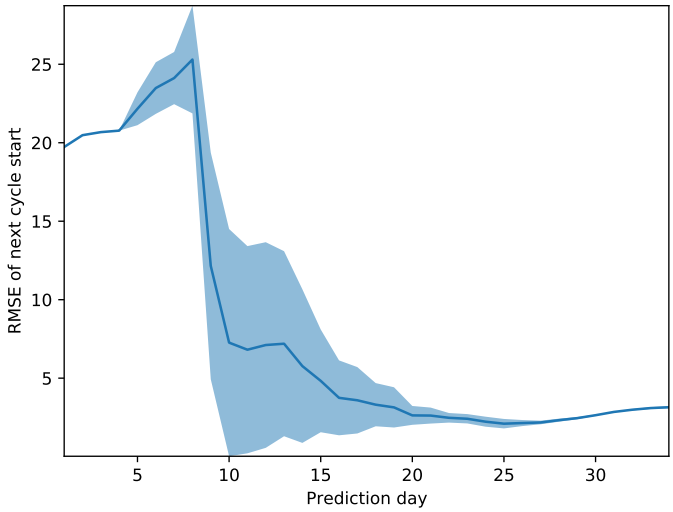
5.3.4.1 Bleeding only model

We start by utilizing bleeding as the only input symptom, checking for users who have tracked bleeding in the training set. In Figure 5.7, we showcase the RMSE of predicted next cycle start over prediction day (aligned with day 0 per user), with specific values presented in

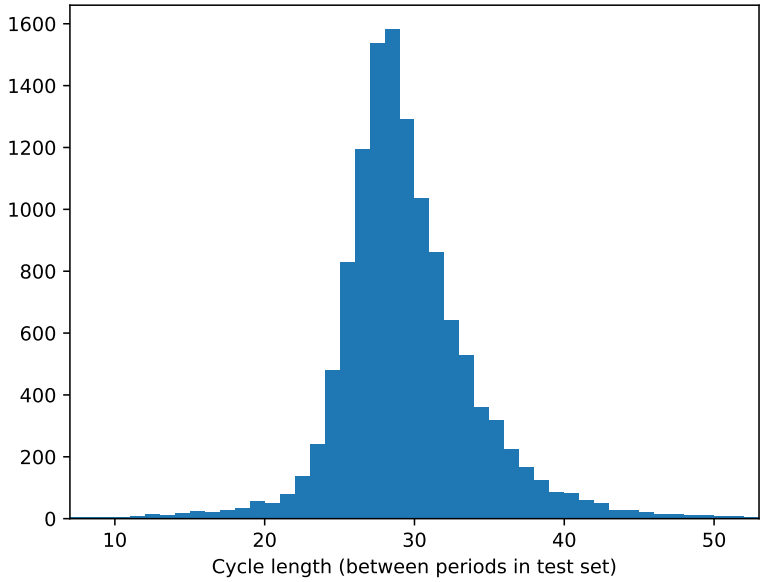
Table C.1 of Appendix C. We see that the RMSE starts high early in the cycle, but declines rapidly past around day 10 of the test set (on day 10, the RMSE is 7.26). This performance continues to improve as we proceed in the test set, dropping to an RMSE of 2.36 on day 20. See Figure C.2 of Appendix C for RMSE of predicting next cycle start with initialization of $(2, 2)$ for α and β.

In comparison to our prior generative model for cycle lengths in Chapter 4, this model's predictive performance is a significant improvement — when we are within about 10 days of the typical next cycle start, we are able to predict its occurrence within 2 or 3 days. We provide a histogram of observed cycle lengths in Figure 5.7b, which shows that cycle lengths are peaked around day 29, demonstrating why we see a drop in RMSE as we approach this typical next cycle start day.

Although our model performs well as the cycle proceeds, the RMSE in the beginning of the test set is high — this is due to the fact that in the beginning of the cycle, the model is not yet confident about predicting next cycle start (in Appendix C, we provide the number of predictions available on each day of the prediction window — in the beginning of the cycle, next cycle start predictions are relatively few). Instead, the model is predicting 'current' bleeding (not next bleeding); that is, the first day of predicted bleeding may not fall within the criteria of being considered 'next' cycle start (i.e., within 7 days of day 0). This is reasonable, because the cycle has not proceeded far enough at this point where the model is predicting next cycle (instead, it is predicting the bleeding of the current cycle).



(a) RMSE of predicted next cycle start over prediction day for model with bleeding only.



(b) Histogram of observed cycle length on full dataset.

Figure 5.7: RMSE of predicted next cycle start, using model with bleeding only over prediction day (**a**), and histogram of observed cycle length for the full dataset (**b**). We see that cycle lengths are peaked around day 29, and that prediction RMSE drops past around day 10 of prediction. This RMSE decreases as we approach the typical cycle length.

We showcase this effect by presenting tables of AUC for predicting early days of bleeding in the test set (days 2 and 3) in Tables 5.3 and 5.4. We can see that our model is making accurate predictions of bleeding early in the test set; however, early in the cycle, these are predictions of bleeding for the current cycle, not the next one. For AUC of predicting bleeding on days 4, 5, 6, and 7, see Appendix C.

The histogram of number of events per day in the test set in Figure 5.8b further showcases how early in the cycle, we have bleeding events for the current cycle, while from days 7 to about day 20, there are not many bleeding events observed. Therefore, we can consider that in the beginning of the cycle, our model is not confident yet in predicting the next cycle start (and that consequently, there are few users for whom we have a predicted next cycle start early on). However, our predictions for next cycle start become increasingly common and accurate as we proceed and observe more events closer to when the next cycle is expected to begin.

Table 5.3: AUC of predicting bleeding on day 2 of the test set over prediction day (aligned at day 0) for bleeding only model. The average and SD are computed across 3 seeds.

Prediction day	Mean AUC	SD AUC
0	0.67	0.04
1	0.73	0.03

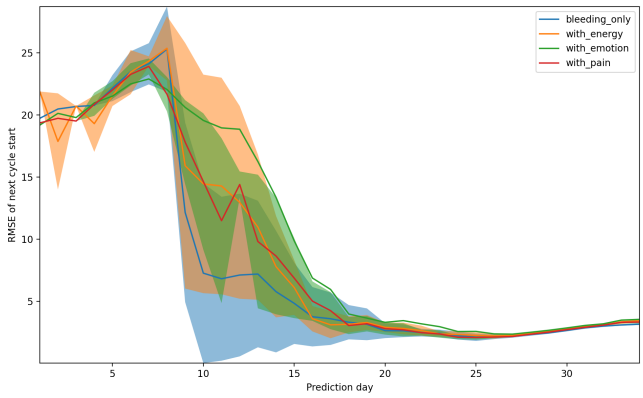
Table 5.4: AUC of predicting bleeding on day 3 of the test set over prediction day (aligned at day 0) for bleeding only model. The average and SD are computed across 3 seeds.

Prediction day	Mean AUC	SD AUC
0	0.64	0.04
1	0.7	0.05
2	0.76	0.04

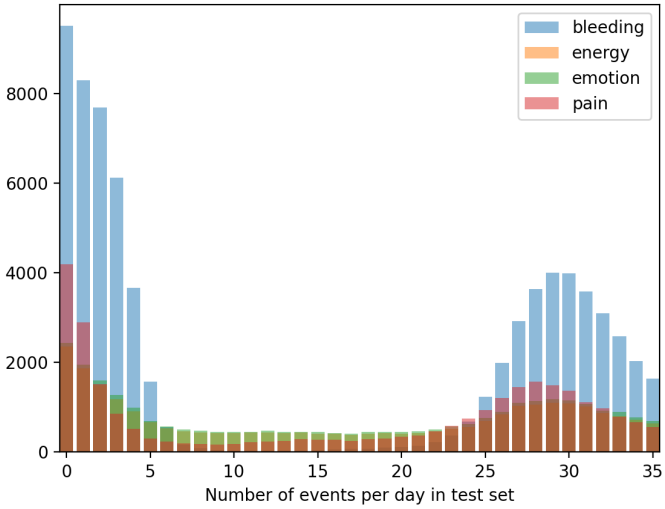
5.3.4.2 Bleeding and additional symptom model

In this section, we explore the impact of adding a symptom in addition to bleeding. Specifically, we test adding energy, emotion, and pain symptoms.

In general, we find that including a symptom in addition to bleeding does not have a significant impact on predicting next cycle start. Looking at prediction RMSE over prediction day as in Figure 5.8, we see that past day 15, the models converge to about the same point, reaching an RMSE of 2-3 days as the cycle proceeds (see Tables C.3, C.2, and C.4 in Appendix C for specific values).



(a) RMSE of predicted next cycle start, using model with bleeding only and bleeding with another symptom over prediction day.



(b) Histogram of observed number of events per symptom on each day of the test set.

Figure 5.8: RMSE of predicted next cycle start, using model with bleeding only and bleeding with another symptom over prediction day (**a**), and histogram of observed number of events per symptom on each day of the test set (**b**). We see that predictive performance is similar among models (i.e., whether we include another symptom or not), due to the fact that symptom events are aligned with when bleeding events occur, as seen in the histogram of tracking events.

There are a few considerations to bear in mind with this result; the first is that we are learning a more complex model with multi-symptom input. Secondly, the predictions made for different symptom models represent different user sets, in the sense that we can only compute the RMSE based on users who are eligible (i.e., for users who have a predicted next cycle start on a given day), which varies depending on the symptom we are considering. For instance, for the dataset where we check for tracking of pain in the training set, there are $9,501$ eligible users, i.e., users with an observed next cycle start; for emotion, there are $8,007$; and for energy, there are $7,225$. Therefore, there may be variation in the user sets we are comparing between results. Finally, as observed in Figure 5.8b, the other symptom events are generally aligned with bleeding, i.e., inclusion of these symptoms may not be adding additional information beyond what is offered when considering bleeding alone. We also provide a normalized version of Figure 5.8b in Figure C.3 of Appendix C.

In summary, when we train the model on bleeding and another symptom, predictive performance mirrors that of the model trained on bleeding only. That is, we are able to learn a more complex model without sacrificing performance in predicting next cycle start (and without needing to train separate models per symptom, which would be computationally less efficient). Earlier we also presented how including additional symptoms allows us to predict future symptom events (for the input symptom that we included) with high accuracy. This is beneficial because in a real-world setting, bleeding is the most commonly tracked information, and we have seen how using symptom-level bleeding information (instead of cycle lengths) allows us to predict next cycle start with high accuracy. Furthermore, if there is also access to other symptom information, including that in the model as well both allows us to still predict bleeding reliably, but also predict the other symptom.

5.4 Significance

The results of our proposed hierarchical, deep generative model for symptom tracking events showcase the potential that self-tracked data holds to inform users of future events, even with the presence of self-tracking artifacts (i.e., instances where a user does not track an event that they actually experienced). This natural extension of our fully generative model in Chapter 4 allows us to leverage more nuanced information from the dataset, while still providing estimates of next symptom occurrence and likelihood of adherence. Specifically, we capture the menstrual experience with 'bleeding' symptoms, rather than cycle lengths, and add nuance by utilizing symptoms beyond 'bleeding.' Furthermore, by using day-by-day information, we can use incomplete cycles, rather than being restricted to complete, observed cycle lengths.

Specifically, we are able to provide more accurate predictions (in comparison to the cycle length model in Chapter 4) and nuanced predictions as a function of prediction day for bleeding and other symptoms. We find that our predictions of next cycle start and next symptom event become increasingly accurate as we proceed in the test set, and notably, our prediction RMSE for next cycle start drops to around 2 about 10 days before the anticipated cycle start across all models. In addition, we find that we can predict next symptom occurrence with an AUC of around $0.72 - 0.88$ for symptoms and around $0.55 - 0.7$ for bleeding, with this AUC increasing as we approach the day we are trying to predict. Although we have focused on predictions of $\hat{x}$ in this chapter (since we can compare this to the observed data x), we can also use our model to generate predictions of $\hat{g}$ (the indicator of tracking adherence) to provide the user with predictions of skipped tracking. Furthermore, we can also provide the user with their

learned posterior b to give an overall sense of their tracking adherence, or use this to develop informed alerting on a per-user basis, as in Chapter 4. Finally, we find that when we learn a more complex model with multi-symptom input, we maintain our predictive performance for next cycle start while also generating these accurate predictions for next symptom occurrence.

In considering further exploration of a multi-symptom model for menstruation, a few key paths come to mind: the first is consider symptoms on a more granular level. That is, when we consider symptoms at the category level, we are benefited by less data sparsity, but may be losing more specific information about the cycle. For instance, it may be the case that certain symptoms can provide more anticipatory information for next cycle start (as compared to being synced with the next cycle start, which is what we have seen in our experiments). In addition, as we have discussed above, when we currently compare the bleeding only model against others, we are comparing different user sets (and for instance, since the bleeding only dataset is more inclusive, the population-wide hyperparameters have more users to learn from). In order to evaluate the models in a more one-to-one fashion, we could consider training the model for bleeding only on the users for whom there is both bleeding and the symptom in question (therefore, when comparing bleeding only to bleeding and symptom, we are considering fitting and evaluating on the same user set). Additionally, we could consider evaluating only those who have tracked the symptom in the test set as well, to further isolate the impact of including additional symptoms. Finally, we currently make predictions based on a probability threshold of 0.5 — changing this threshold to a lower one would allow for more predictions of 1s, which could impact our next cycle start prediction and favor more 'early' cycle start predictions; this threshold could also be based on user behavior or preference. From a metric perspective, this would increase our false positive rate (which is currently low for our model), but from a user

point of view, this may be desirable. Relatedly, while AUC is useful from a model validation perspective, we can also consider metrics that may be useful to the user (other than RMSE of next cycle start). For instance, we can consider the accuracy of predicting next cycle start or symptom occurrence within a window. The question of which metrics most benefit the user merits further investigation.

Significance to users: While other models exist for predicting the onset of menstruation (and related information, such as menstrual phases or period length), our model also focuses on predicting symptoms outside of bleeding. That is, the user can get an accurate sense not only of when their next period bleeding event will occur, but also what other symptoms to expect. Since menstruation is a multi-faceted experience, providing this nuance to the user can help them feel informed and prepared for upcoming symptom events. Additionally, we have developed a model that is flexible to symptom input. That is, if bleeding is the only symptom that is available, we can provide users with a dependable idea of when their next bleeding event (or next cycle start) will occur. However, if other symptoms are also available, we can include them in the model and predict them well, too, without sacrificing bleeding predictive performance. Since we provide a window of predictions on each day, we can continually update users with when we think the next cycle will occur. We have seen that these predictions improve over time and that at least 7 days before the expected cycle start, the prediction RMSE is between $2-3$ days. Finally, in comparison to other deep learning approaches, the generative component of our model facilitates interpretability by modeling user adherence and per-user temporal dynamics explicitly.

Chapter 6

Conclusions and future work

Understanding the physiological mechanisms behind menstruation and how they manifest in womxn will continue to be an important task for researchers, the broad population of menstruators this research impacts, and clinicians responsible for diagnosing and advising patients on issues related to menstruation. While there have been advancements in recent years (particularly with the rise of access to data and the increased interest in understanding womxn's health), there is still more progress to be made. In this dissertation, we've not only showcased how to understand and handle the nuances of self-tracked mobile health data, but we've also proposed two models for predicting menstrual cycle lengths and related symptoms that show promising results. Importantly, we've taken a holistic approach to investigating menstruation — we've integrated concerns related to data reliability, effective mobile app design, and the needs of menstruators with powerful statistics to produce models that are realistic, effective, and interpretable.

The results from our investigation of the data and our design of these flexible models provide quantitative insight into understanding how menstruation is not a 'regular' process for most womxn (nor should it be expected to be one): we saw that for our dataset spanning millions of womxn, users occupied different ends of the variability spectrum for cycle lengths. In addition,

we found that users who experience more variable cycles also generally experience more variable symptoms — i.e., knowledge about cycle length volatility can inform our understanding of symptom experiences. Furthermore, since we explicitly considered the possibility of skipped cycles with a parameter representing the likelihood of skipping tracking in our models, we were able to accurately detect when a cycle length may not have been accurately tracked by the user (i.e., when period tracking may have been missed), which can be useful to app designers to help proactively alert users who may be prone to missed tracking. Finally, we proposed a fully generative model that uses cycle lengths as the only input, which outperforms alternative baselines using only the most commonly tracked type of information, as well as a hierarchical, deep generative model that utilizes time series representations of symptom information, providing a more nuanced and flexible approach to tracking event prediction. In doing so, we achieved high predictive ability for next symptom tracking event without sacrificing interpretability — we proposed a generative process for the data and can provide learned parameter values in addition to predictions. That is, while the mechanism for learning these parameters is a deep one, their meaning is interpretable.

This work is significant to users because it quantitatively shows that menstrual variability is more common than not, revamping prior definitions of what is normal. Furthermore, it provides practical considerations to users: for users who have a typical (median) between-cycle length variability that is greater than 9 days, they may expect greater variance in their symptom experiences as well. For healthcare professionals, this 9 day variability threshold can also serve as an updated, more nuanced guideline for discussing menstrual variability with patients. In addition to these considerations for variability of the menstrual experience, this work also provides practical predictive models for next cycle length and next symptom occurrence,

which can be applied in many different data availability scenarios: cycle length information only, bleeding symptom information only, or bleeding and other symptom information. Moreover, since these models have a generative component, they can provide more interpretable insight behind the predictions; for instance, since we account for the possibility of self-tracking artifacts, we can provide the user with their learned self-tracking adherence, in addition to their cycle length or symptom predictions. This flexibility and interpretability means that the user can gain personalized insight into their cycle across a variety of settings. In addition, since we learn population-wide distributions for self-tracking adherence, this can provide a view into how a large population of mobile health users adheres at-large.

We've laid the groundwork for future research into prediction of menstrual cycle lengths and related symptoms. While we have shown the great potential of learning from individual symptoms (whether that be cycle lengths, bleeding events, or other symptoms), we would like to extend this work to develop models that leverage insight across symptoms by exploring a larger set of symptoms in addition to bleeding that may offer anticipatory signal for next cycle start. We also aim to further examine evaluation metrics that may be useful to the user (and how modeling choices, such as prediction thresholds, can impact these metrics). In doing so, we hope to further showcase the utility that self-tracked mobile health data holds to improve individual-level and broader understanding of menstruation and deliver these insights to the user in an impactful way.

Appendix A

Supplementary information for Chapter 3

A.1 Supplementary Information: Cohort and dataset

A.1.1 Study dataset

Table A.1: Summary statistics of this study's cohort dataset, compared with state of the art references on menstrual health studies through mobile apps.

Variable	This cohort	Cohort in [49]	Cohort in [50]
Number of users	378,694 (100.00%)	124,646 (32.92%)	212,967 (56.24%)
Number of observations	117,014,597 (100.00%)	NA	7,496,316 (6.41%)
Number of days of observation	34,056,343 (100.00%)	NA	33,675,453 (98.88%)
Number of cycles	4,881,697 (100.00%)	612,613 (12.55%)	2,732,424 (55.97%)

A.1.2 User demographics

Table A.2: High-level characteristics for this study's cohort dataset, compared with state of the art references on menstrual health studies through mobile apps.

	Full cohort		Cohort in [49]		Cohort in [50]	
Variable	Mean ± sd	Median	Mean ± sd	Median	Mean ± sd	Median
Age	25.49 ± 3.66	25	30.3	NA	30 ± 6	NA
Number of cycles	12.89 ± 9.11	11.00	8.6	NA	12.83 (NA)	NA
Cycle length	29.73 ± 5.73	29.00	29.3 ± 5.2	NA	NA	28
Period length	4.08 ± 1.76	4.00	4.0 ± 1.5	NA	NA	NA

Table A.3: Per-age number of users and cycles for the full cohort, as well as for the consistently not highly variable and consistently highly variable user groups.

	Full cohort		Consistently not highly variable		Consistently highly variable	
Age	Number users	Number cycles	Number users	Number cycles	Number users	Number cycles
21	71,511	557,083	65,520	526,413	5,991	30,670
22	36,723	500,736	33,338	478,394	3,385	22,342
23	33,943	466,999	30,984	447,498	2,959	19,501
24	32,225	442,053	29,529	424,706	2,696	17,347
25	30,651	422,465	28,191	406,519	2,460	15,946
26	29,377	402,905	27,066	388,306	2,311	14,599
27	27,757	380,662	25,802	368,043	1,955	12,619
28	25,257	353,535	23,518	342,245	1,739	11,290
29	22,991	325,875	21,535	316,637	1,456	9,238
30	20,744	297,814	19,462	289,725	1,282	8,089
31	18,424	269,125	17,358	262,045	1,066	7,080
32	16,444	244,483	15,521	238,957	923	5,526
33	12,647	217,962	11,782	212,206	865	5,756

Table A.4: Per-country user count in the full cohort, as well as for the consistently not highly variable and consistently highly variable user groups.

Country	Full cohort	Consistently not highly variable	Consistently highly variable
United States	97955	6911	91044
United Kingdom	32676	2486	30190
Mexico	32155	3102	29053
Brazil	27275	2535	24740
Germany	21538	1360	20178
France	19106	1371	17735
China	16529	1435	15094
Canada	15507	963	14544
Australia	14211	1103	13108
Spain	13574	804	12770
Italy	12775	685	12090
Japan	8716	692	8024
Denmark	7520	580	6940
Russia	7203	396	6807
Taiwan	5192	538	4654
Colombia	5024	475	4549
India	3976	424	3552
Switzerland	3380	216	3164
Sweden	3190	167	3023
Philippines	2876	346	2530
Argentina	2783	211	2572
Hong Kong	2706	266	2440
Singapore	2635	220	2415

Country	Full cohort	Consistently not highly variable	Consistently highly variable
South Korea	1910	205	1705
New Zealand	1902	171	1731
Peru	1897	205	1692
Netherlands	1832	135	1697
Austria	1512	117	1395
Portugal	1257	110	1147
Indonesia	1187	96	1091
Malaysia	1127	104	1023
Ireland	1115	84	1031
Chile	1080	100	980
Ecuador	1041	105	936
Turkey	835	78	757
Poland	710	43	667
Venezuela	690	51	639
Finland	482	44	438
Belgium	389	38	351
Saudi Arabia	387	27	360
Ukraine	382	29	353
Vietnam	299	42	257
Guatemala	82	12	70
South Africa	76	6	70

A.1.3 Cycle statistics per user age

Table A.5: Per-age average number of cycles per user for the full cohort, as well as for the consistently not highly variable and consistently highly variable user groups.

	Full cohort		Consistently not highly variable		Consistently highly variable	
Age	Mean ± sd (95% CI)	Median	Mean ± sd (95% CI)	Median	Mean ± sd (95% CI)	Median
21	7.79 ± 3.88 (1.00,14.00)	8.00	8.03 ± 3.88 (1.00,14.00)	8.00	5.12 ± 2.63 (1.00,11.00)	5.00
22	7.74 ± 3.92 (1.00,14.00)	8.00	7.97 ± 3.92 (1.00,14.00)	8.00	4.77 ± 2.52 (1.00,11.00)	4.00
23	7.77 ± 3.94 (1.00,14.00)	8.00	8.00 ± 3.93 (1.00,14.00)	8.00	4.73 ± 2.48 (1.00,10.00)	4.00
24	7.78 ± 3.96 (1.00,14.00)	8.00	7.99 ± 3.96 (1.00,14.00)	8.00	4.74 ± 2.46 (1.00,10.00)	4.00
25	7.82 ± 3.97 (1.00,14.00)	8.00	8.03 ± 3.96 (1.00,14.00)	8.00	4.71 ± 2.47 (1.00,10.00)	4.00
26	7.85 ± 3.99 (1.00,14.00)	8.00	8.05 ± 3.98 (1.00,14.00)	8.00	4.68 ± 2.40 (1.00,10.00)	4.00
27	7.86 ± 4.02 (1.00,14.00)	8.00	8.05 ± 4.02 (1.00,14.00)	8.00	4.68 ± 2.48 (1.00,10.00)	4.00
28	7.93 ± 4.03 (1.00,14.00)	8.00	8.11 ± 4.03 (1.00,14.00)	8.00	4.70 ± 2.43 (1.00,10.00)	4.00
29	8.00 ± 4.06 (1.00,14.00)	8.00	8.18 ± 4.05 (1.00,14.00)	8.00	4.61 ± 2.41 (1.00,10.00)	4.00
30	8.08 ± 4.09 (1.00,15.00)	8.00	8.26 ± 4.08 (1.00,15.00)	9.00	4.60 ± 2.36 (1.00,10.00)	4.00
31	8.13 ± 4.11 (1.00,15.00)	8.00	8.28 ± 4.11 (1.00,15.00)	9.00	4.81 ± 2.42 (1.00,10.00)	4.00
32	8.23 ± 4.15 (1.00,15.00)	8.00	8.39 ± 4.13 (1.00,15.00)	9.00	4.56 ± 2.46 (1.00,10.00)	4.00
33	8.85 ± 3.85 (3.00,15.00)	9.00	9.05 ± 3.80 (3.00,15.00)	9.00	4.88 ± 2.25 (2.00,10.00)	4.00

Table A.6: Per-age average cycle length per user for the full cohort, as well as for the consistently not highly variable and consistently highly variable user groups.

	Full cohort		Consistently not highly variable		Consistently highly variable	
Age	Mean ± sd (95% CI)	Median	Mean ± sd (95% CI)	Median	Mean ± sd (95% CI)	Median
21	30.24 ± 6.23 (20.00,45.00)	29.00	29.86 ± 5.25 (21.00,42.00)	29.00	36.83 ± 13.66 (13.00,69.00)	34.00
22	30.16 ± 6.02 (20.00,44.00)	29.00	29.85 ± 5.20 (21.00,42.00)	29.00	36.82 ± 13.67 (13.00,69.00)	34.00
23	30.10 ± 5.95 (21.00,44.00)	29.00	29.81 ± 5.17 (21.00,42.00)	29.00	36.88 ± 13.67 (13.00,69.00)	34.00
24	30.03 ± 5.84 (21.00,44.00)	29.00	29.74 ± 5.09 (21.00,42.00)	29.00	36.96 ± 13.62 (13.00,68.00)	34.00
25	29.95 ± 5.81 (21.00,44.00)	29.00	29.66 ± 5.06 (21.00,42.00)	29.00	37.14 ± 13.76 (13.00,69.00)	34.00
26	29.85 ± 5.74 (21.00,44.00)	29.00	29.58 ± 5.00 (22.00,41.00)	29.00	37.25 ± 13.76 (13.00,69.00)	35.00
27	29.71 ± 5.65 (21.00,43.00)	29.00	29.44 ± 4.92 (22.00,41.00)	29.00	37.38 ± 13.92 (13.00,71.00)	35.00
28	29.57 ± 5.56 (22.00,43.00)	29.00	29.32 ± 4.88 (22.00,41.00)	29.00	37.27 ± 13.60 (13.00,69.00)	35.00
29	29.42 ± 5.45 (22.00,42.00)	29.00	29.18 ± 4.80 (22.00,41.00)	28.00	37.34 ± 13.99 (13.00,71.00)	34.00
30	29.24 ± 5.35 (22.00,42.00)	28.00	29.01 ± 4.71 (22.00,40.00)	28.00	37.37 ± 13.81 (14.00,70.00)	35.00
31	29.06 ± 5.23 (22.00,42.00)	28.00	28.84 ± 4.62 (22.00,40.00)	28.00	37.21 ± 13.45 (13.00,67.00)	35.00
32	28.85 ± 5.08 (22.00,41.00)	28.00	28.66 ± 4.53 (22.00,39.00)	28.00	37.10 ± 13.71 (14.00,68.00)	34.00
33	28.66 ± 5.05 (22.00,40.00)	28.00	28.45 ± 4.39 (22.00,39.00)	28.00	36.57 ± 13.74 (13.00,70.00)	33.00

Table A.7: Per-age average period length per user for the full cohort, as well as for the consistently not highly variable and consistently highly variable user groups.

	Full cohort		Consistently not highly variable		Consistently highly variable	
Age	Mean ± sd (95% CI)	Median	Mean ± sd (95% CI)	Median	Mean ± sd (95% CI)	Median
21	4.18 ± 1.74 (1.00,7.00)	4.00	4.18 ± 1.70 (1.00,7.00)	4.00	4.23 ± 2.33 (1.00,8.00)	4.00
22	4.17 ± 1.76 (1.00,7.00)	4.00	4.16 ± 1.71 (1.00,7.00)	4.00	4.36 ± 2.59 (1.00,9.00)	4.00
23	4.14 ± 1.76 (1.00,7.00)	4.00	4.13 ± 1.72 (1.00,7.00)	4.00	4.29 ± 2.55 (1.00,9.00)	4.00
24	4.12 ± 1.75 (1.00,7.00)	4.00	4.12 ± 1.71 (1.00,7.00)	4.00	4.32 ± 2.55 (1.00,9.00)	4.00
25	4.11 ± 1.75 (1.00,7.00)	4.00	4.10 ± 1.71 (1.00,7.00)	4.00	4.32 ± 2.53 (1.00,9.00)	4.00
26	4.09 ± 1.77 (1.00,7.00)	4.00	4.08 ± 1.73 (1.00,7.00)	4.00	4.34 ± 2.62 (1.00,9.00)	4.00
27	4.06 ± 1.75 (1.00,7.00)	4.00	4.05 ± 1.73 (1.00,7.00)	4.00	4.34 ± 2.39 (1.00,9.00)	4.00
28	4.04 ± 1.75 (1.00,7.00)	4.00	4.03 ± 1.72 (1.00,7.00)	4.00	4.28 ± 2.57 (1.00,9.00)	4.00
29	4.01 ± 1.76 (1.00,7.00)	4.00	4.00 ± 1.73 (1.00,7.00)	4.00	4.22 ± 2.61 (1.00,9.00)	4.00
30	3.99 ± 1.77 (1.00,7.00)	4.00	3.98 ± 1.72 (1.00,7.00)	4.00	4.28 ± 2.88 (1.00,10.00)	4.00
31	3.97 ± 1.77 (1.00,7.00)	4.00	3.97 ± 1.74 (1.00,7.00)	4.00	4.19 ± 2.73 (1.00,9.02)	4.00
32	3.95 ± 1.78 (1.00,7.00)	4.00	3.95 ± 1.76 (1.00,7.00)	4.00	4.14 ± 2.47 (1.00,9.00)	4.00
33	3.91 ± 1.78 (1.00,7.00)	4.00	3.91 ± 1.76 (1.00,7.00)	4.00	4.01 ± 2.52 (1.00,9.00)	4.00

Table A.8: Per-age average median CLD per user for the full cohort, as well as for the consistently not highly variable and consistently highly variable user groups.

	Full cohort		Consistently not highly variable		Consistently highly variable	
Age	Mean ± sd (95% CI)	Median	Mean ± sd (95% CI)	Median	Mean ± sd (95% CI)	Median
21	4.49 ± 5.07 (1.00,19.00)	3.00	3.38 ± 2.39 (1.00,9.00)	3.00	16.82 ± 9.00 (5.19,40.00)	14.00
22	4.32 ± 4.83 (1.00,17.00)	3.00	3.40 ± 2.56 (1.00,9.50)	3.00	16.32 ± 9.36 (4.00,42.00)	13.50
23	4.23 ± 4.72 (1.00,17.00)	3.00	3.36 ± 2.52 (1.00,9.00)	3.00	16.42 ± 9.23 (4.00,41.00)	14.00
24	4.10 ± 4.53 (1.00,16.00)	3.00	3.30 ± 2.51 (1.00,9.00)	2.50	16.03 ± 8.98 (3.00,39.35)	13.50
25	4.07 ± 4.57 (1.00,16.00)	3.00	3.26 ± 2.44 (1.00,9.00)	2.50	16.50 ± 9.30 (4.00,41.29)	13.50
26	3.99 ± 4.61 (1.00,16.00)	3.00	3.19 ± 2.48 (1.00,9.00)	2.50	16.59 ± 9.67 (3.00,43.00)	13.50
27	3.86 ± 4.43 (0.50,15.50)	2.50	3.13 ± 2.37 (0.50,9.00)	2.50	16.59 ± 9.54 (3.34,42.66)	14.00
28	3.81 ± 4.38 (1.00,15.00)	2.50	3.10 ± 2.39 (0.50,9.00)	2.50	16.60 ± 9.49 (4.00,43.00)	13.50
29	3.70 ± 4.25 (1.00,14.50)	2.50	3.05 ± 2.38 (0.50,9.00)	2.50	16.60 ± 9.45 (4.00,42.00)	13.50
30	3.59 ± 4.16 (1.00,14.00)	2.50	2.95 ± 2.18 (0.50,8.50)	2.00	16.73 ± 9.65 (3.00,41.00)	13.50
31	3.52 ± 4.04 (0.50,14.00)	2.50	2.92 ± 2.28 (0.50,8.50)	2.00	16.42 ± 9.00 (4.00,37.95)	14.00
32	3.42 ± 4.01 (0.50,13.00)	2.00	2.87 ± 2.22 (0.50,8.50)	2.00	16.87 ± 10.05 (3.00,43.00)	13.50
33	3.44 ± 4.25 (1.00,14.00)	2.00	2.73 ± 1.99 (1.00,8.00)	2.00	17.58 ± 9.49 (7.00,45.00)	14.00

Table A.9: Per-age average maximum CLD per user for the full cohort, as well as for the consistently not highly variable and consistently highly variable user groups.

	Full cohort		Consistently not highly variable		Consistently highly variable	
Age	Mean ± sd (95% CI)	Median	Mean ± sd (95% CI)	Median	Mean ± sd (95% CI)	Median
21	9.48 ± 7.29 (1.00,30.00)	8.00	8.18 ± 5.34 (1.00,21.00)	7.00	23.81 ± 10.10 (9.00,51.00)	22.00
22	9.14 ± 6.91 (1.00,28.00)	7.00	8.08 ± 5.26 (1.00,21.00)	7.00	23.05 ± 10.14 (7.00,50.00)	21.00
23	8.97 ± 6.84 (1.00,28.00)	7.00	7.96 ± 5.23 (1.00,21.00)	7.00	23.10 ± 10.19 (7.00,50.00)	21.00
24	8.70 ± 6.65 (1.00,27.00)	7.00	7.76 ± 5.11 (1.00,20.00)	7.00	22.78 ± 10.16 (5.00,49.00)	21.00
25	8.67 ± 6.72 (1.00,28.00)	7.00	7.72 ± 5.12 (1.00,20.00)	7.00	23.39 ± 10.26 (7.00,51.00)	22.00
26	8.51 ± 6.64 (1.00,27.00)	7.00	7.59 ± 5.09 (1.00,20.00)	6.00	23.15 ± 10.29 (6.00,50.00)	21.00
27	8.26 ± 6.47 (1.00,26.00)	7.00	7.40 ± 4.93 (1.00,19.00)	6.00	23.22 ± 10.48 (6.00,50.32)	21.00
28	8.17 ± 6.40 (1.00,26.00)	6.00	7.35 ± 4.94 (1.00,19.00)	6.00	23.09 ± 10.17 (7.00,49.15)	21.00
29	8.01 ± 6.35 (1.00,26.00)	6.00	7.24 ± 4.94 (1.00,19.00)	6.00	23.42 ± 10.31 (6.00,50.35)	22.00
30	7.82 ± 6.13 (1.00,25.00)	6.00	7.07 ± 4.71 (1.00,18.00)	6.00	23.16 ± 10.29 (7.00,50.00)	21.00
31	7.71 ± 6.04 (1.00,25.00)	6.00	7.00 ± 4.74 (1.00,18.00)	6.00	23.00 ± 9.65 (8.00,48.00)	22.00
32	7.54 ± 5.88 (1.00,24.00)	6.00	6.91 ± 4.64 (1.00,18.00)	6.00	22.94 ± 10.35 (5.00,52.00)	21.00
33	7.72 ± 6.21 (2.00,26.00)	6.00	6.90 ± 4.63 (1.00,18.00)	6.00	24.01 ± 10.10 (11.00,51.92)	22.00

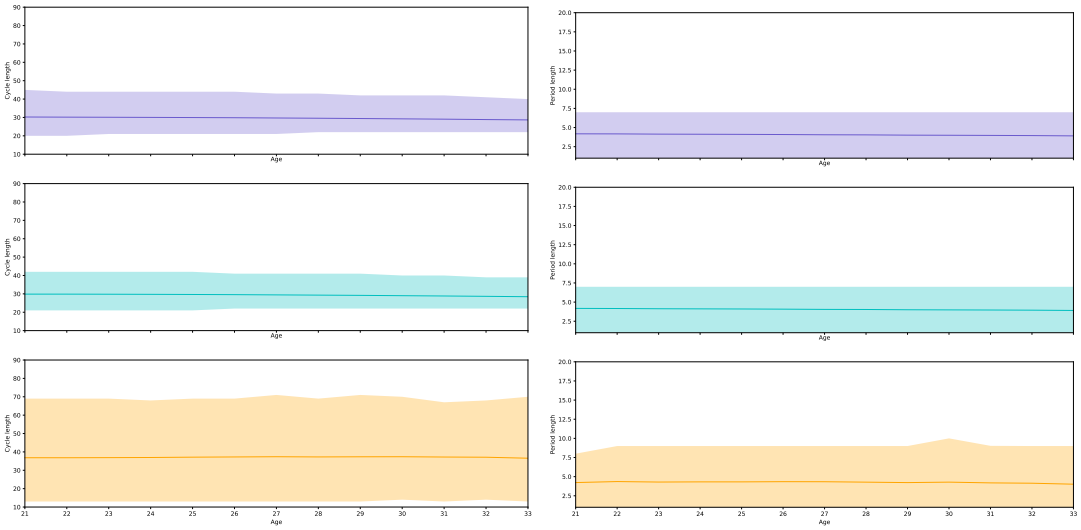


Figure A.1: For users with cycles at a specific age, we average cycle (left) and period length (right) across three different groups: the entire user cohort (top, purple), the consistently not highly variable user cohort (middle, teal), and the consistently highly variable user cohort (bottom, orange). This allows us to visualize how cycle and period length vary with age for each group, on average and in terms of standard deviation. We observe that cycle and period length statistics are stationary over the studied age range within each plot. We note that the the top and middle plots look similar in each figure (i.e., the consistently not highly variable group looks similar to the overall population in terms of both cycle and period length), but the wider shaded orange spread of the bottom plot demonstrates the higher degree of variability in the consistently highly variable group. In addition, this spread is consistently wider for all ages in the orange plot. This showcases that the consistently highly variable group represents a large degree of the variability that we see in the data overall.

A.2 Supplementary Information: Results

A.2.1 Assessing differences in reported symptoms across user groups

The following table provides the Kolmogorov-Smirnov statistic for the empirical cumulative distributions of the proportion of cycles with symptom out of cycles with category (λ_s) for the different user groups.

Table A.10: Kolmogorov-Smirnov test results for symptoms per-group

Category	Symptom	KS statistic (95% CI)	p-value
Period flow	heavy	0.181 (0.178,0.183)	< 0.000000
Stool health	normal	0.135 (0.130,0.140)	< 0.000000
Period flow	medium	0.134 (0.132,0.137)	< 0.000000
Social behavior	sociable	0.127 (0.121,0.132)	< 0.000000
Mental state	distracted	0.123 (0.118,0.127)	< 0.000000
Period flow	light	0.121 (0.118,0.124)	< 0.000000
Food cravings	sweet craving	0.120 (0.115,0.125)	< 0.000000
Energy level	low energy	0.118 (0.114,0.121)	< 0.000000
Motivation level	unproductive	0.117 (0.112,0.122)	< 0.000000
Digestive health	bloated	0.116 (0.111,0.122)	< 0.000000
Emotional state	sensitive	0.115 (0.112,0.118)	< 0.000000
Digestive health	gassy	0.114 (0.109,0.119)	< 0.000000
Emotional state	happy	0.108 (0.105,0.111)	< 0.000000
Mental state	calm	0.104 (0.099,0.108)	< 0.000000
Type of pain experienced	cramps	0.101 (0.097,0.104)	< 0.000000
Hours of sleep	3-6	0.100 (0.097,0.103)	< 0.000000

Category	Symptom	KS statistic (95% CI)	p-value
Food cravings	carbs craving	0.098 (0.094,0.103)	< 0.000000
Motivation level	motivated	0.098 (0.094,0.103)	< 0.000000
Motivation level	unmotivated	0.098 (0.092,0.103)	< 0.000000
Type of pain experienced	ovulation pain	0.096 (0.093,0.099)	< 0.000000
Skin health	acne skin	0.093 (0.088,0.098)	< 0.000000
Social behavior	withdrawn	0.093 (0.087,0.098)	< 0.000000
Skin health	oily skin	0.093 (0.089,0.096)	< 0.000000
Hair health	bad hair	0.092 (0.087,0.097)	< 0.000000
Vaginal discharge type	creamy	0.091 (0.086,0.095)	< 0.000000
Type of pain experienced	headache	0.089 (0.087,0.092)	< 0.000000
Hair health	good hair	0.089 (0.083,0.095)	< 0.000000
Period flow	spotting	0.089 (0.087,0.092)	< 0.000000
Emotional state	pms	0.086 (0.083,0.089)	< 0.000000
Digestive health	great digestion	0.085 (0.081,0.089)	< 0.000000
Skin health	good skin	0.085 (0.081,0.088)	< 0.000000
Food cravings	salty cravings	0.084 (0.080,0.089)	< 0.000000
Method for period collection	pad	0.083 (0.077,0.090)	< 0.000000
Type of pain experienced	tender breasts	0.082 (0.080,0.084)	< 0.000000
Hours of sleep	6-9	0.079 (0.076,0.083)	< 0.000000
Mental state	stressed	0.079 (0.074,0.083)	< 0.000000
Stool health	constipated	0.078 (0.074,0.083)	< 0.000000
Sexual health	unprotected sex	0.078 (0.074,0.081)	< 0.000000
Physical maladies	cold/flu	0.077 (0.067,0.087)	< 0.000000
Method for period collection	tampon	0.076 (0.070,0.083)	< 0.000000
Type of medication taken	cold/flu	0.076 (0.067,0.085)	< 0.000000
Emotional state	sad	0.076 (0.073,0.079)	< 0.000000

Category	Symptom	KS statistic (95% CI)	p-value
Social behavior	supportive	0.075 (0.071,0.079)	< 0.000000
Physical exercise	running	0.074 (0.067,0.081)	< 0.000000
Party-related experiences	cigarettes	0.074 (0.067,0.081)	< 0.000000
Stool health	diarrhea	0.071 (0.066,0.076)	< 0.000000
Motivation level	productive	0.071 (0.067,0.075)	< 0.000000
Food cravings	chocolate cravings	0.071 (0.066,0.075)	< 0.000000
Mental state	focused	0.069 (0.066,0.073)	< 0.000000
Vaginal discharge type	atypical	0.069 (0.065,0.074)	< 0.000000
Sexual health	protected sex	0.069 (0.065,0.073)	< 0.000000
Method for period collection	menstrual cup	0.067 (0.063,0.072)	< 0.000000
Skin health	dry skin	0.067 (0.063,0.072)	< 0.000000
Hair health	dry hair	0.067 (0.061,0.073)	< 0.000000
Hair health	oily hair	0.067 (0.062,0.072)	< 0.000000
Vaginal discharge type	sticky	0.066 (0.062,0.070)	< 0.000000
Energy level	exhausted	0.066 (0.063,0.069)	< 0.000000
Stool health	great	0.065 (0.060,0.071)	< 0.000000
Digestive health	nauseated	0.064 (0.059,0.069)	< 0.000000
Energy level	high energy	0.063 (0.061,0.066)	< 0.000000
Party-related experiences	big night party	0.063 (0.057,0.071)	< 0.000000
Social behavior	conflict	0.062 (0.059,0.068)	< 0.000000
Vaginal discharge type	egg white	0.062 (0.058,0.067)	< 0.000000
Physical exercise	yoga	0.062 (0.055,0.068)	< 0.000000
Physical maladies	allergy	0.061 (0.053,0.069)	0.000001
Hours of sleep	> 9	0.061 (0.057,0.064)	< 0.000000
Method for period collection	panty liner	0.057 (0.053,0.061)	< 0.000000
Physical exercise	biking	0.056 (0.049,0.062)	< 0.000000

Category	Symptom	KS statistic (95% CI)	p-value
Party-related experiences	hangover	0.055 (0.051,0.063)	< 0.000000
Energy level	energized	0.052 (0.049,0.055)	< 0.000000
Sexual health	high sex drive	0.052 (0.051,0.055)	< 0.000000
Type of medication taken	pain	0.046 (0.041,0.054)	0.000548
Sexual health	withdrawal sex	0.045 (0.044,0.048)	< 0.000000
Physical maladies	fever	0.044 (0.037,0.054)	0.001015
Type of medication taken	antibiotic	0.044 (0.036,0.053)	0.001040
Party-related experiences	drinks party	0.042 (0.037,0.050)	0.000028
Hours of sleep	0-3	0.041 (0.039,0.044)	< 0.000000
Physical maladies	injury	0.040 (0.034,0.049)	0.003686
Physical exercise	swimming	0.040 (0.034,0.045)	0.000003
Type of medication taken	antihistamine	0.032 (0.029,0.041)	0.032955

Table A.11: Likelihood of low proportion ($\lambda_s < 0.05$) of cycles with symptom out of cycles with category per group, with the associated odds ratio of how likely users in the consistently highly variable group to the consistently not highly variable group are not to track a symptom throughout their cycle history (i.e., in very few of their cycles). 95% confidence intervals attained via bootstrapping with 100,000 samples are shown in parentheses.

Category	Symptom	High variability group	Low variability group	Odds ratio
Period flow	medium	0.009 (0.009,0.009)	0.003 (0.003,0.003)	3.140 (2.826,3.522)
Period flow	light	0.036 (0.036,0.036)	0.014 (0.013,0.015)	2.568 (2.445,2.700)
Period flow	heavy	0.170 (0.169,0.170)	0.098 (0.096,0.100)	1.734 (1.703,1.766)
Type of pain experienced	cramps	0.105 (0.104,0.105)	0.073 (0.071,0.074)	1.436 (1.404,1.470)
Skin health	acne skin	0.174 (0.173,0.176)	0.132 (0.129,0.135)	1.319 (1.286,1.353)
Period flow	spotting	0.314 (0.313,0.315)	0.239 (0.237,0.241)	1.314 (1.300,1.328)
Mental state	stressed	0.243 (0.242,0.245)	0.186 (0.182,0.189)	1.312 (1.286,1.340)
Type of medication taken	pain	0.212 (0.209,0.215)	0.167 (0.160,0.174)	1.274 (1.220,1.334)
Emotional state	sad	0.348 (0.346,0.349)	0.273 (0.270,0.276)	1.273 (1.260,1.287)
Emotional state	pms	0.395 (0.394,0.396)	0.310 (0.307,0.313)	1.273 (1.261,1.286)
Motivation level	unmotivated	0.168 (0.167,0.170)	0.133 (0.129,0.136)	1.271 (1.237,1.307)

Category	Symptom	High variability group	Low variability group	Odds ratio
Party-related experiences	drinks party	0.166 (0.164,0.168)	0.131 (0.126,0.136)	1.270 (1.219,1.325)
Emotional state	sensitive	0.176 (0.175,0.177)	0.143 (0.140,0.145)	1.234 (1.214,1.254)
Stool health	diarrhea	0.369 (0.367,0.371)	0.299 (0.295,0.304)	1.234 (1.213,1.255)
Social behavior	withdrawn	0.215 (0.213,0.216)	0.176 (0.172,0.180)	1.218 (1.188,1.249)
Hours of sleep	6-9	0.161 (0.160,0.162)	0.133 (0.130,0.135)	1.218 (1.196,1.240)
Type of pain experienced	headache	0.326 (0.325,0.327)	0.269 (0.266,0.272)	1.212 (1.199,1.225)
Energy level	exhausted	0.312 (0.311,0.313)	0.258 (0.255,0.261)	1.208 (1.194,1.223)
Vaginal discharge type	egg white	0.359 (0.357,0.361)	0.298 (0.293,0.303)	1.206 (1.186,1.226)
Physical maladies	cold/flu	0.234 (0.231,0.238)	0.195 (0.187,0.202)	1.204 (1.158,1.254)
Social behavior	conflict	0.379 (0.377,0.381)	0.318 (0.313,0.323)	1.194 (1.174,1.215)
Digestive health	gassy	0.219 (0.217,0.221)	0.184 (0.180,0.188)	1.189 (1.162,1.217)
Motivation level	unproductive	0.207 (0.205,0.208)	0.175 (0.171,0.179)	1.179 (1.152,1.207)
Energy level	low energy	0.129 (0.128,0.130)	0.110 (0.108,0.112)	1.174 (1.151,1.198)
Digestive health	nauseated	0.427 (0.425,0.429)	0.365 (0.360,0.370)	1.170 (1.153,1.187)
Digestive health	bloated	0.151 (0.150,0.153)	0.130 (0.126,0.133)	1.165 (1.133,1.199)
Stool health	constipated	0.358 (0.356,0.360)	0.309 (0.304,0.314)	1.160 (1.141,1.180)
Food cravings	chocolate craving	0.350 (0.348,0.351)	0.302 (0.297,0.306)	1.159 (1.142,1.178)
Motivation level level	productive	0.354 (0.352,0.356)	0.308 (0.304,0.313)	1.148 (1.130,1.167)
Food cravings	salty craving	0.295 (0.293,0.296)	0.257 (0.253,0.261)	1.147 (1.127,1.168)
Food cravings	sweet craving	0.144 (0.143,0.146)	0.126 (0.123,0.129)	1.146 (1.116,1.178)
Type of pain experienced	tender breasts	0.366 (0.365,0.367)	0.320 (0.317,0.322)	1.145 (1.134,1.156)
Food cravings	carbs craving	0.310 (0.309,0.312)	0.271 (0.267,0.276)	1.144 (1.125,1.164)
Physical exercise	running	0.250 (0.248,0.253)	0.219 (0.214,0.224)	1.144 (1.116,1.174)
Sexual health	protected sex	0.533 (0.531,0.534)	0.466 (0.462,0.469)	1.143 (1.134,1.152)
Type of pain experienced	ovulation pain	0.721 (0.720,0.722)	0.633 (0.630,0.636)	1.139 (1.133,1.144)
Party-related experiences	big night party	0.522 (0.519,0.525)	0.460 (0.452,0.468)	1.136 (1.116,1.156)
Hair health	oily hair	0.363 (0.361,0.365)	0.320 (0.314,0.325)	1.135 (1.114,1.157)
Method for period collection	tampon	0.630 (0.628,0.633)	0.557 (0.551,0.563)	1.131 (1.119,1.144)
Physical exercise	yoga	0.551 (0.548,0.553)	0.489 (0.483,0.496)	1.125 (1.110,1.141)
Hair health	good hair	0.217 (0.215,0.219)	0.194 (0.189,0.199)	1.120 (1.091,1.150)
Party-related experiences	hangover	0.512 (0.509,0.515)	0.458 (0.450,0.465)	1.119 (1.100,1.139)
Stool health	great	0.595 (0.593,0.597)	0.533 (0.527,0.538)	1.118 (1.106,1.130)
Hours of sleep	3-6	0.259 (0.258,0.260)	0.232 (0.229,0.235)	1.117 (1.102,1.131)
Sexual health	high sex drive	0.469 (0.467,0.470)	0.420 (0.417,0.424)	1.115 (1.105,1.124)
Hours of sleep	> 9	0.587 (0.586,0.588)	0.530 (0.526,0.533)	1.108 (1.101,1.115)
Vaginal discharge type	sticky	0.439 (0.437,0.441)	0.399 (0.394,0.404)	1.101 (1.086,1.115)
Hair health	bad hair	0.324 (0.322,0.326)	0.295 (0.289,0.300)	1.099 (1.078,1.121)
Mental state	distracted	0.204 (0.202,0.205)	0.187 (0.183,0.190)	1.091 (1.069,1.115)

Category	Symptom	High variability group	Low variability group	Odds ratio
Skin health	good skin	0.384 (0.382,0.386)	0.352 (0.348,0.357)	1.091 (1.076,1.105)
Vaginal discharge type	creamy	0.342 (0.340,0.344)	0.315 (0.310,0.319)	1.087 (1.071,1.105)
Sexual health	unprotected sex	0.378 (0.376,0.379)	0.348 (0.344,0.351)	1.086 (1.075,1.097)
Energy level	high energy	0.394 (0.393,0.395)	0.363 (0.360,0.367)	1.085 (1.075,1.095)
Physical exercise	biking	0.715 (0.712,0.717)	0.660 (0.654,0.666)	1.083 (1.072,1.093)
Method for period collection	menstrual cup	0.880 (0.879,0.882)	0.814 (0.809,0.818)	1.082 (1.075,1.088)
Mental state	focused	0.407 (0.405,0.409)	0.377 (0.372,0.381)	1.081 (1.067,1.095)
Type of medication taken	cold/flu	0.569 (0.565,0.573)	0.527 (0.517,0.536)	1.080 (1.060,1.101)
Motivation level	motivated	0.299 (0.297,0.301)	0.278 (0.273,0.282)	1.075 (1.057,1.094)
Sexual health	withdrawal sex	0.596 (0.595,0.598)	0.556 (0.552,0.559)	1.073 (1.065,1.080)
Social behavior	supportive	0.412 (0.410,0.414)	0.386 (0.380,0.391)	1.069 (1.054,1.085)
Physical maladies	fever	0.704 (0.701,0.708)	0.661 (0.653,0.670)	1.065 (1.050,1.080)
Hair health	dry hair	0.441 (0.439,0.443)	0.415 (0.409,0.421)	1.063 (1.047,1.079)
Type of medication taken	antibiotic	0.712 (0.709,0.716)	0.671 (0.662,0.680)	1.061 (1.047,1.076)
Skin health	dry skin	0.493 (0.491,0.494)	0.464 (0.460,0.469)	1.060 (1.049,1.072)
Physical maladies	injury	0.732 (0.728,0.735)	0.692 (0.684,0.701)	1.057 (1.044,1.071)
Energy level	energized	0.625 (0.624,0.626)	0.593 (0.590,0.596)	1.054 (1.047,1.060)
Method for period collection	panty liner	0.553 (0.551,0.555)	0.525 (0.519,0.531)	1.053 (1.040,1.066)
Skin health	oily skin	0.372 (0.371,0.374)	0.355 (0.351,0.360)	1.048 (1.034,1.062)
Physical exercise	swimming	0.841 (0.840,0.843)	0.803 (0.798,0.808)	1.047 (1.040,1.054)
Hours of sleep	0-3	0.762 (0.761,0.763)	0.731 (0.728,0.734)	1.043 (1.038,1.047)
Type of medication taken	antihistamine	0.767 (0.763,0.770)	0.736 (0.727,0.744)	1.042 (1.030,1.055)
Social behavior	sociable	0.218 (0.217,0.220)	0.210 (0.206,0.215)	1.038 (1.015,1.062)
Physical maladies	allergy	0.581 (0.578,0.585)	0.560 (0.551,0.569)	1.037 (1.019,1.056)
Emotional state	happy	0.281 (0.280,0.282)	0.275 (0.272,0.278)	1.024 (1.013,1.035)
Mental state	calm	0.293 (0.292,0.295)	0.290 (0.286,0.295)	1.010 (0.995,1.027)
Digestive health	great digestion	0.388 (0.386,0.390)	0.388 (0.383,0.393)	1.002 (0.988,1.016)
Stool health	normal	0.181 (0.179,0.182)	0.181 (0.177,0.185)	0.998 (0.975,1.022)
Vaginal discharge type	atypical	0.664 (0.662,0.666)	0.673 (0.668,0.678)	0.986 (0.978,0.993)
Party-related experiences	cigarettes	0.581 (0.578,0.585)	0.608 (0.601,0.616)	0.956 (0.943,0.969)
Method for period collection	pad	0.214 (0.212,0.216)	0.236 (0.231,0.241)	0.907 (0.886,0.929)

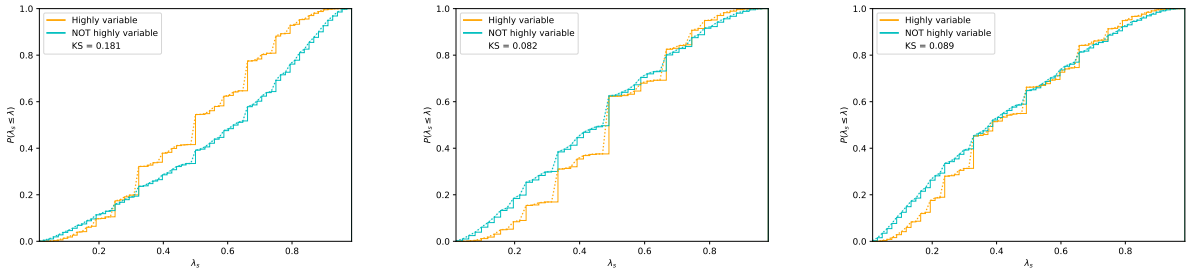
Table A.12: Likelihood of high proportion ($\lambda_s > 0.95$) of cycles with symptom out of cycles with category per group, with the associated odds ratio of how likely users in the consistently highly variable group to the consistently not highly variable group are to consistently track a symptom throughout their cycle history (i.e., in almost every cycle where they track the category). 95% confidence intervals attained via bootstrapping with 100,000 samples are shown in parentheses.

Category	Symptom	High variability group	Low variability group	Odds ratio
Hours of sleep	0-3	0.035 (0.034,0.035)	0.020 (0.019,0.021)	1.750 (1.667,1.839)
Period flow	spotting	0.067 (0.066,0.067)	0.039 (0.037,0.040)	1.729 (1.679,1.782)
Type of pain experienced	tender breasts	0.193 (0.192,0.194)	0.113 (0.111,0.115)	1.715 (1.684,1.746)
Vaginal discharge type	atypical	0.100 (0.099,0.101)	0.059 (0.056,0.061)	1.706 (1.636,1.780)
Energy level	energized	0.075 (0.074,0.075)	0.044 (0.043,0.046)	1.686 (1.633,1.741)
Type of pain experienced	headache	0.218 (0.217,0.219)	0.131 (0.129,0.133)	1.663 (1.636,1.691)
Skin health	dry skin	0.155 (0.154,0.157)	0.096 (0.093,0.098)	1.626 (1.579,1.676)
Type of medication taken	cold/flu	0.179 (0.176,0.182)	0.112 (0.107,0.118)	1.590 (1.506,1.681)
Skin health	oily skin	0.250 (0.248,0.251)	0.159 (0.155,0.162)	1.575 (1.540,1.611)
Hair health	dry hair	0.170 (0.169,0.172)	0.109 (0.105,0.113)	1.565 (1.510,1.624)
Digestive health	great digestion	0.241 (0.239,0.243)	0.158 (0.154,0.162)	1.528 (1.490,1.567)
Social behavior	supportive	0.215 (0.213,0.216)	0.141 (0.138,0.145)	1.519 (1.477,1.562)
Emotional state	happy	0.307 (0.306,0.308)	0.202 (0.200,0.205)	1.518 (1.498,1.538)
Skin health	good skin	0.242 (0.241,0.244)	0.160 (0.156,0.163)	1.518 (1.485,1.552)
Hair health	bad hair	0.266 (0.264,0.268)	0.175 (0.171,0.180)	1.514 (1.474,1.557)
Digestive health	nauseated	0.170 (0.168,0.171)	0.112 (0.109,0.116)	1.511 (1.466,1.558)
Stool health	great	0.101 (0.100,0.102)	0.068 (0.065,0.071)	1.487 (1.428,1.549)
Emotional state	sad	0.171 (0.170,0.172)	0.115 (0.113,0.117)	1.486 (1.459,1.513)
Method for period collection	panty liner	0.174 (0.172,0.175)	0.118 (0.114,0.122)	1.471 (1.422,1.523)
Stool health	constipated	0.246 (0.244,0.248)	0.169 (0.165,0.173)	1.454 (1.420,1.491)
Mental state	focused	0.218 (0.216,0.219)	0.150 (0.147,0.153)	1.451 (1.417,1.486)
Mental state	calm	0.327 (0.325,0.328)	0.225 (0.221,0.229)	1.450 (1.424,1.477)
Vaginal discharge type	sticky	0.214 (0.212,0.216)	0.148 (0.145,0.152)	1.442 (1.406,1.479)
Type of medication taken	antihistamine	0.099 (0.096,0.101)	0.069 (0.064,0.074)	1.437 (1.337,1.548)
Hours of sleep	3-6	0.322 (0.321,0.324)	0.225 (0.222,0.228)	1.431 (1.413,1.450)
Motivation level	motivated	0.321 (0.319,0.322)	0.225 (0.220,0.229)	1.428 (1.401,1.457)
Hours of sleep	> 9	0.093 (0.092,0.094)	0.065 (0.064,0.067)	1.425 (1.388,1.464)
Physical exercise	swimming	0.061 (0.060,0.062)	0.043 (0.040,0.045)	1.423 (1.339,1.516)
Motivation level	unproductive	0.387 (0.386,0.389)	0.272 (0.268,0.277)	1.422 (1.398,1.447)

Category	Symptom	High variability group	Low variability group	Odds ratio
Mental state	distracted	0.407 (0.405,0.409)	0.286 (0.282,0.290)	1.422 (1.400,1.444)
Type of pain experienced	ovulation pain	0.044 (0.043,0.044)	0.031 (0.030,0.032)	1.419 (1.369,1.473)
Emotional state	sensitive	0.380 (0.378,0.381)	0.269 (0.266,0.272)	1.411 (1.395,1.426)
Food cravings	carbs craving	0.334 (0.332,0.336)	0.238 (0.234,0.242)	1.403 (1.378,1.429)
Energy level	high energy	0.214 (0.213,0.215)	0.153 (0.150,0.155)	1.400 (1.377,1.423)
Social behavior	conflict	0.208 (0.206,0.210)	0.149 (0.145,0.153)	1.399 (1.362,1.438)
Vaginal discharge type	creamy	0.314 (0.312,0.316)	0.224 (0.220,0.228)	1.399 (1.372,1.427)
Social behavior	sociable	0.444 (0.442,0.446)	0.320 (0.315,0.325)	1.388 (1.365,1.411)
Sexual health	withdrawal sex	0.159 (0.158,0.160)	0.115 (0.112,0.117)	1.386 (1.358,1.415)
Energy level	exhausted	0.235 (0.234,0.236)	0.170 (0.167,0.172)	1.382 (1.361,1.403)
Stool health	normal	0.475 (0.473,0.477)	0.344 (0.339,0.349)	1.381 (1.361,1.402)
Digestive health	gassy	0.400 (0.398,0.402)	0.290 (0.285,0.294)	1.381 (1.358,1.405)
Hair health	oily hair	0.244 (0.242,0.246)	0.178 (0.173,0.183)	1.368 (1.332,1.407)
Physical maladies	fever	0.119 (0.116,0.121)	0.087 (0.082,0.092)	1.368 (1.285,1.458)
Emotional state	pms	0.160 (0.159,0.161)	0.117 (0.115,0.119)	1.367 (1.342,1.393)
Food cravings	chocolate craving	0.263 (0.261,0.264)	0.194 (0.190,0.198)	1.357 (1.329,1.386)
Motivation level	productive	0.266 (0.264,0.267)	0.197 (0.193,0.201)	1.347 (1.318,1.376)
Physical maladies	injury	0.105 (0.102,0.107)	0.078 (0.073,0.083)	1.346 (1.260,1.442)
Type of medication taken	antibiotic	0.123 (0.120,0.126)	0.092 (0.086,0.097)	1.345 (1.264,1.433)
Party-related experiences	hangover	0.200 (0.198,0.203)	0.149 (0.144,0.155)	1.343 (1.293,1.397)
Physical maladies	allergy	0.236 (0.233,0.239)	0.176 (0.169,0.183)	1.343 (1.289,1.402)
Party-related experiences	big night party	0.215 (0.212,0.217)	0.160 (0.154,0.166)	1.342 (1.293,1.393)
Party-related experiences	cigarettes	0.290 (0.287,0.293)	0.217 (0.211,0.223)	1.337 (1.297,1.379)
Stool health	diarrhea	0.225 (0.223,0.226)	0.169 (0.165,0.173)	1.330 (1.298,1.363)
Food cravings	salty craving	0.331 (0.330,0.333)	0.249 (0.245,0.253)	1.330 (1.307,1.353)
Energy level	low energy	0.489 (0.488,0.491)	0.376 (0.373,0.379)	1.302 (1.290,1.314)
Social behavior	withdrawn	0.397 (0.395,0.399)	0.307 (0.302,0.312)	1.294 (1.272,1.317)
Sexual health	high sex drive	0.224 (0.223,0.226)	0.174 (0.171,0.176)	1.292 (1.271,1.313)
Digestive health	bloated	0.502 (0.500,0.504)	0.390 (0.385,0.395)	1.287 (1.270,1.305)
Food cravings	sweet craving	0.527 (0.526,0.529)	0.411 (0.406,0.416)	1.283 (1.268,1.299)
Mental state	stressed	0.353 (0.351,0.354)	0.276 (0.272,0.280)	1.277 (1.257,1.298)
Sexual health	unprotected sex	0.354 (0.353,0.356)	0.279 (0.276,0.282)	1.271 (1.256,1.286)
Motivation level	unmotivated	0.446 (0.444,0.448)	0.352 (0.347,0.356)	1.270 (1.251,1.288)
Hair health	good hair	0.421 (0.419,0.424)	0.336 (0.331,0.342)	1.253 (1.231,1.276)
Period flow	light	0.250 (0.249,0.251)	0.203 (0.200,0.205)	1.233 (1.219,1.248)
Skin health	acne skin	0.489 (0.487,0.491)	0.400 (0.395,0.405)	1.222 (1.207,1.237)
Vaginal discharge type	egg white	0.298 (0.297,0.300)	0.244 (0.240,0.249)	1.222 (1.199,1.245)

Category	Symptom	High variability group	Low variability group	Odds ratio
Type of pain experienced	cramps	0.529 (0.528,0.530)	0.442 (0.439,0.445)	1.198 (1.189,1.206)
Sexual health	protected sex	0.219 (0.218,0.220)	0.183 (0.181,0.186)	1.196 (1.178,1.215)
Physical exercise	biking	0.129 (0.128,0.131)	0.109 (0.105,0.113)	1.188 (1.144,1.235)
Hours of sleep	6-9	0.474 (0.473,0.476)	0.400 (0.396,0.403)	1.188 (1.177,1.198)
Physical exercise	yoga	0.262 (0.260,0.265)	0.223 (0.217,0.228)	1.179 (1.151,1.209)
Physical maladies	cold/flu	0.529 (0.525,0.533)	0.453 (0.444,0.462)	1.169 (1.144,1.194)
Method for period collection	pad	0.583 (0.581,0.585)	0.505 (0.499,0.511)	1.155 (1.141,1.170)
Physical exercise	running	0.563 (0.560,0.566)	0.490 (0.484,0.496)	1.149 (1.133,1.164)
Period flow	medium	0.388 (0.387,0.389)	0.345 (0.342,0.347)	1.126 (1.117,1.136)
Party-related experiences	drinks party	0.635 (0.632,0.638)	0.594 (0.587,0.602)	1.069 (1.055,1.084)
Type of medication taken	pain	0.597 (0.593,0.601)	0.561 (0.552,0.571)	1.063 (1.044,1.082)
Method for period collection	tampon	0.210 (0.209,0.212)	0.218 (0.213,0.223)	0.967 (0.943,0.991)
Period flow	heavy	0.078 (0.077,0.079)	0.096 (0.094,0.097)	0.817 (0.802,0.833)
Method for period collection	menstrual cup	0.075 (0.074,0.076)	0.100 (0.096,0.103)	0.755 (0.726,0.785)

The following figures showcase the empirical cumulative distributions of the proportion of cycles with symptom out of cycles with category between different user groups — the consistently highly variable group is indicated in orange, and the consistently not highly variable group is indicated in teal. Figures are organized based on their Kolmogorov-Smirnov test value, in descending order. The mean (dotted line) and 95% confidence interval (shaded region) of the bootstrapped CDF with 100,000 samples is also shown.



(a) Heavy period flow. (b) Tender breasts pain. (c) Spotting period flow.

Figure A.2: Empirical CDFs of proportion of cycles with symptom out of cycles with category between different user groups for ‘heavy’, ‘tender breasts’, and ‘spotting’.

Appendix B

Supplementary information for Chapter 4

B.1 Supplementary Information: Methods

B.1.1 Simulated data

In order to assess the ability of our model to recover skipped cycles, we separately train our model on simulated cycle length data for 10,000 users (with $C = 10$ cycles each), generated from our proposed generative process. We then take two cohorts of users: those who have never skipped a cycle in their history, and those who have skipped a cycle in their history. Note that we have access to ground truth cycle length and skipping information in this simulated case. For a sample user from each of these cohorts, we predict their probabilities of possible cycle skips $p\left(s^*|\hat{u}, d_i, d^* > d_{current}\right)$ for the 11th cycle, utilizing the inferred population-wide hyperparameters $\hat{u}$ and individual cycle length histories d_i.

B.1.2 Implementation details

We optimize the negative log-likelihood $-\ln(p(d|u)) = -\ln(\sum_i p(d_i|u))$ with $p(d_i|u)$ as in Eqn. (4.6) with respect to hyperparameters u via stochastic gradient descent. Specifically, we utilize Adam [106], an adaptive gradient method. All models have been implemented using

PyTorch, and trained with minibatches of size 100. All neural network-based models are trained (with dropout) on the observed cycle lengths for the whole cohort. Predictions are based on each per-user available cycle lengths.

Since we sequentially predict next cycle length, our train-test split is over the number of cycle lengths available, i.e., we train the models on C cycles and predict the $C+1$th cycle, where $C = \{2, \cdots, 10\}$.

For reproducibility, we provide the settings for priors, learning rate, and other details for each of the models below:

- `CNN`: number of layers = 1, kernel size = 3, stride = 1, padding = 0, dilation = 1, nonlinearity = $tanh$, dropout = 0.9, training criterion = MSE, epoch convergence criteria as maximum number of epochs = 1000, loss convergence criteria $\epsilon_{loss} = 1e-3$, optimizer = Adam, learning rate = 0.01.

- `RNN`: number of layers = 1, hidden size = 3, nonlinearity = $tanh$, dropout = 0.9, epoch convergence criteria as maximum number of epochs = 1000, loss convergence criteria $\epsilon_{loss} = 1e-3$, optimizer = Adam, learning rate = 0.01.

- `LSTM`: number of layers = 1, hidden size = 3, nonlinearity = $tanh$, dropout = 0.9, epoch convergence criteria as maximum number of epochs = 1000, loss convergence criteria $\epsilon_{loss} = 1e-3$, optimizer = Adam, learning rate = 0.01.

- `Proposed model`: $u_0 = [\kappa_0 = 180, \gamma_0 = 6, \alpha_0 = 2, \beta_0 = 20]$, $S = 100$ (for both inference and prediction), $M = 1000$ (for both inference and prediction), epoch convergence criteria as maximum number of epochs = 1000, loss convergence criteria $\epsilon_{loss} = 1e-3$, optimizer = Adam, learning rate = 0.01.

- `Proposed model (s=0)`: same as above, with $S = 100$ in inference but $S = 0$ for next-cycle length prediction.

B.2 Supplementary Information: Results

B.2.1 Performance stability across different priors

For the results presented in the main text, we utilize a prior $u_0 = [\kappa_0 = 180, \gamma_0 = 6, \alpha_0 = 2, \beta_0 = 20]$, from which we draw our initial $\theta = [\lambda, \pi]$. This is informed by expert knowledge about average cycle length (around 30 days) and the likelihood of skipping (relatively low) in our dataset.

In order to assess the impact of the prior, we also test training the model on different ones, namely a uniform prior on π (no prior knowledge on skipping likelihood), as well as a less informative (i.e., flatter) prior on both λ and π. We showcase the prediction RMSE results on day 0 of the next cycle for both priors in Figures B.1 and B.2, where the blue line represents results for $s \geq 0$ and the green line represents results for $s = 0$. Note that these results look similar in magnitude and spread as the prior we have chosen, and we therefore conclude that our method is stable to different choices of priors.

B.2.2 Performance stability across different dataset sizes and ordering of cycles

To demonstrate our model's robustness across different dataset sizes, we showcase prediction RMSE results across different numbers of individuals, I (left) and training cycles, C (right) in Figure B.3. We see that our model performance is robust to different I and C values – our model's prediction RMSE remains around 7.5 even with relatively small I or C.

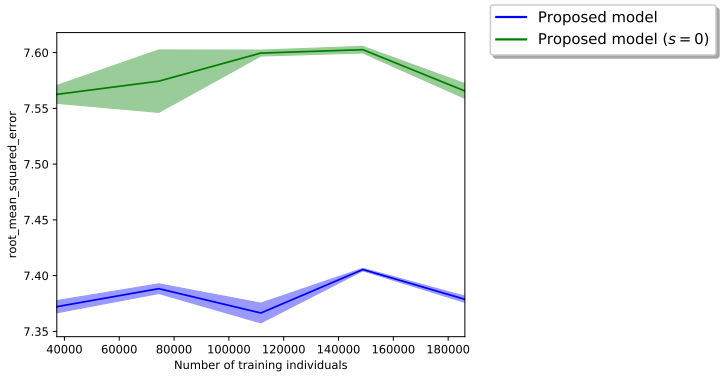


Figure B.1: Prediction RMSE over number of training individuals for a less informative (i.e., a more uncertain) prior on λ and π, $u_0 = [60, 2, 0.01, 0.1]$.

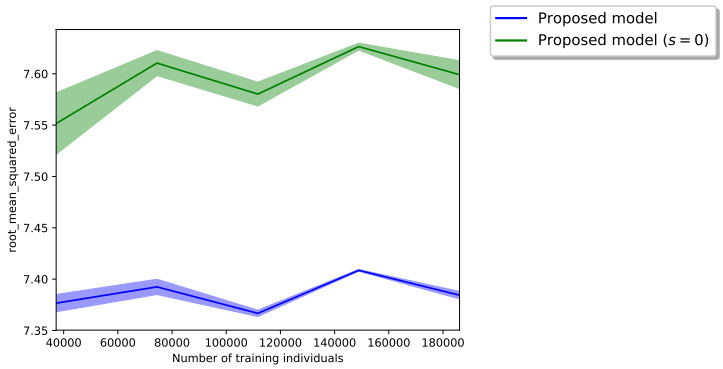


Figure B.2: Prediction RMSE over number of training individuals for a less informative prior on λ and a completely uninformative (i.e., uniform) one on π, $u_0 = [60, 2, 1, 1]$.

While our model performance is generally stable to dataset size as in Figure B.3, we note also that there is some very small magnitude fluctuation in performance with $C = 10$. This is due to data randomness – that is, since we utilize the first C cycles in each training subset, there may be users who happened to have less adherent tracking near the end of their tracking history (i.e., with $C = 10$), resulting in a small uptick in prediction RMSE. To showcase this, we perform an experiment utilizing $I = 10,000$ users across 10 runs of our model; for each run, we randomly draw $I = 10,000$ users from the full dataset, train our model, and compute predictions. The results of this experiment averaged over the 10 runs are shown in

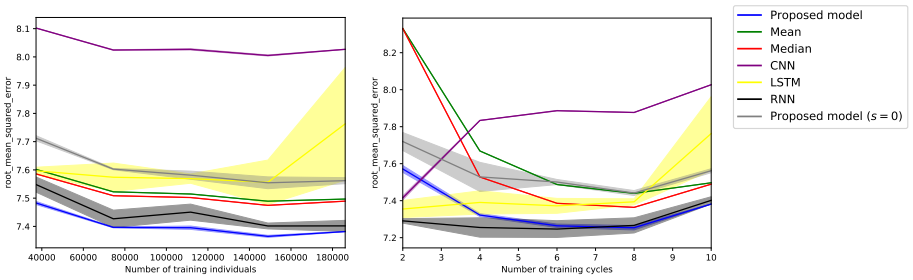


Figure B.3: Prediction RMSE for proposed model and baselines on day 0 over number of individuals, I (**a**) and number of training cycles, C (on the full set of I) (**b**). $C = 2$ means 2 input cycles were used to predict the third and so on. **(a)** Our model outperforms summary statistic-based and neural network-based baselines on day 0 when we account for skipped cycles (blue line), across all subsets of I. In addition, our model produces sharper estimates (lower variance) and is stable across I – with less than $40,000$ users, we have an RMSE less than 7.5. **(b)** Our model is robust to different C, as shown by consistent RMSE with at least 4 training cycles. Note that all models experience some fluctuations in RMSE depending on number of training cycles; this is due to data randomness, see Figure B.4.

Figure B.4, where we see that there is some fluctuation in prediction RMSE across C (not just for $C = 10$), verifying that the small fluctuation for $C = 10$ on the full dataset is an artifact of data randomness.

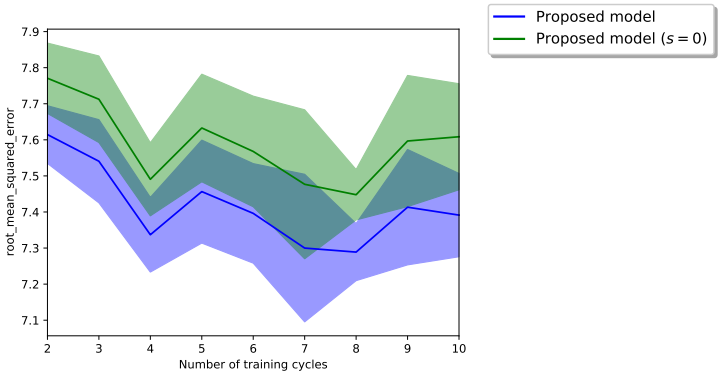


Figure B.4: Prediction RMSE over number of training cycles, averaged over 10 runs of different randomly-drawn datasets of $I = 10,000$ users.

To further test the dependency of model predictive performance on the ordering of the observed training cycles, we also run the same experiment with a random shuffling of a user's cycle history before selecting the first C cycles for training. We showcase these results in Figure B.5 and see again that there are small fluctuations in performance across C, verifying further the impact of data randomness. This also showcases the negligible effect of choosing to either take the first C cycles without shuffling (as in Figure B.4) or with shuffling (as in Figure B.5).

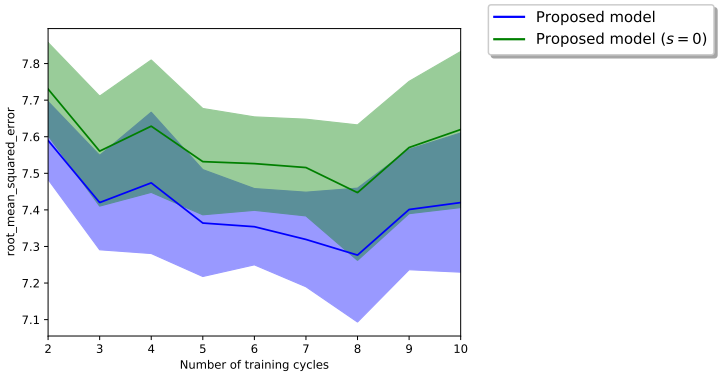


Figure B.5: Prediction RMSE over number of training cycles, averaged over 10 runs of different randomly-drawn datasets of $I = 10,000$ users. Here, before we take the first C cycles from each user, we randomly shuffle them.

B.2.3 Baseline results with different neural network settings

In the results of our main text, we utilize neural network-based baselines with one layer and a kernel size or hidden size of 3. To assess the performance of neural network-based baselines with different settings, we test (i) different numbers of layers and (ii) different kernel and hidden sizes (using a kernel or hidden size equal to the number of training cycles C instead of fixed at 3). Figures B.6, B.7, and B.8 showcase the performance RMSEs across I for 1, 2, 5,

and 10-layer CNNs, LSTMs, and RNNs, respectively (with fixed kernel or hidden size of 3). Figures B.6, B.7, and B.8 showcase the performance RMSEs across I for 1, 2, 5, and 10-layer CNNs, LSTMs, and RNNs, respectively (with kernel or hidden size of $C = 10$). We see that across the number of layers and kernel or hidden size of 3 or $C = 10$, the prediction RMSE is stable, with average differences of at most 0.5 between different settings. Therefore, we conclude that one-layer neural networks, with fixed kernel or hidden size of 3, are reasonable baselines.

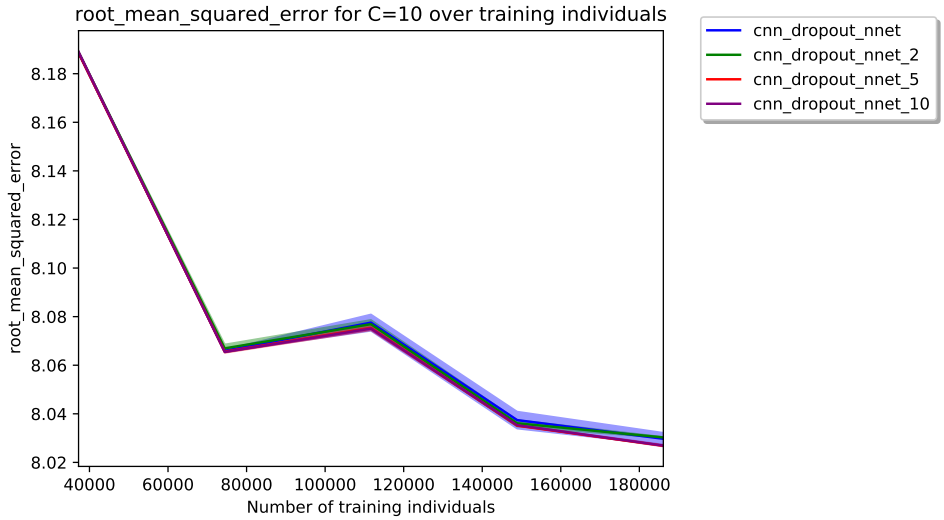


Figure B.6: Prediction RMSE over number of individuals for CNNs with 1, 2, 5, and 10 layers (blue, green, red, and purple lines, respectively) and a kernel size of 3.

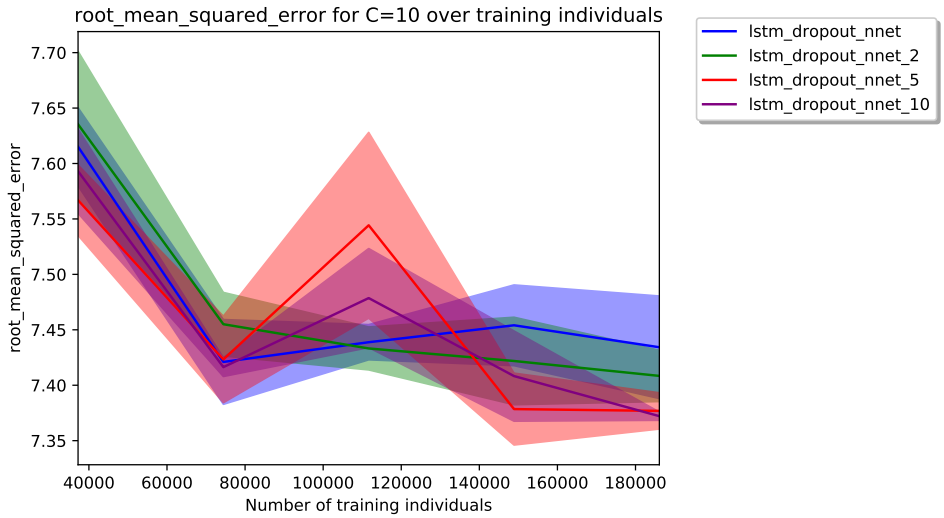


Figure B.7: Prediction RMSE over number of individuals for LSTMs with 1, 2, 5, and 10 layers (blue, green, red, and purple lines, respectively) and a hidden size of 3.

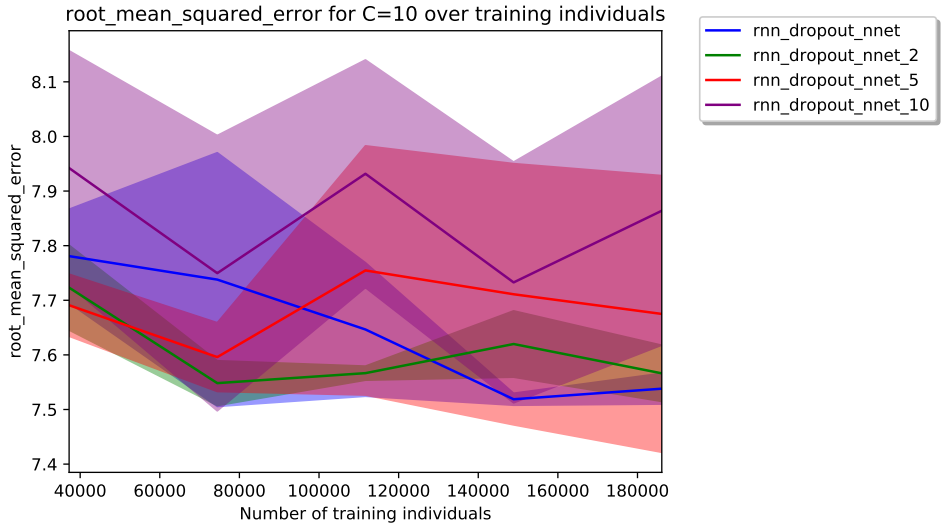


Figure B.8: Prediction RMSE over number of individuals for RNNs with 1, 2, 5, and 10 layers (blue, green, red, and purple lines, respectively) and a hidden size of 3.

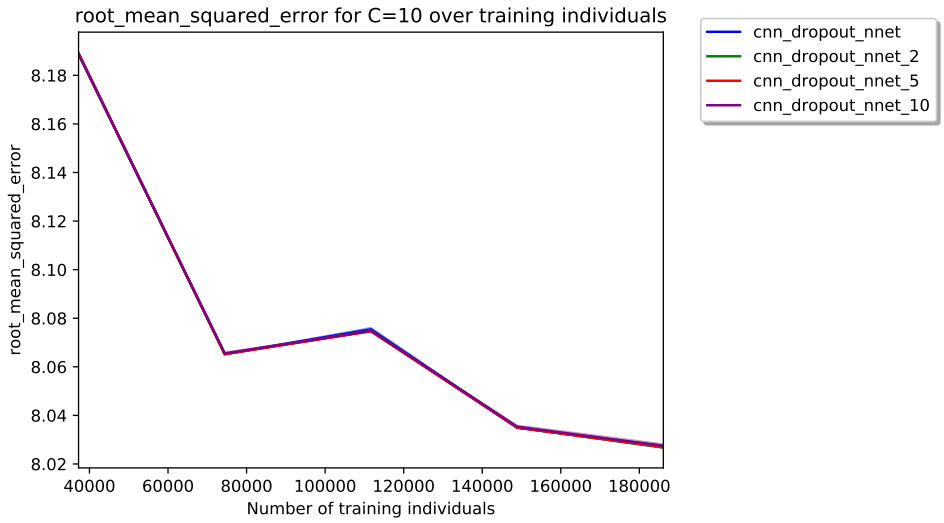


Figure B.9: Prediction RMSE over number of individuals for CNNs with 1, 2, 5, and 10 layers (blue, green, red, and purple lines, respectively) and a kernel size of $C = 10$.

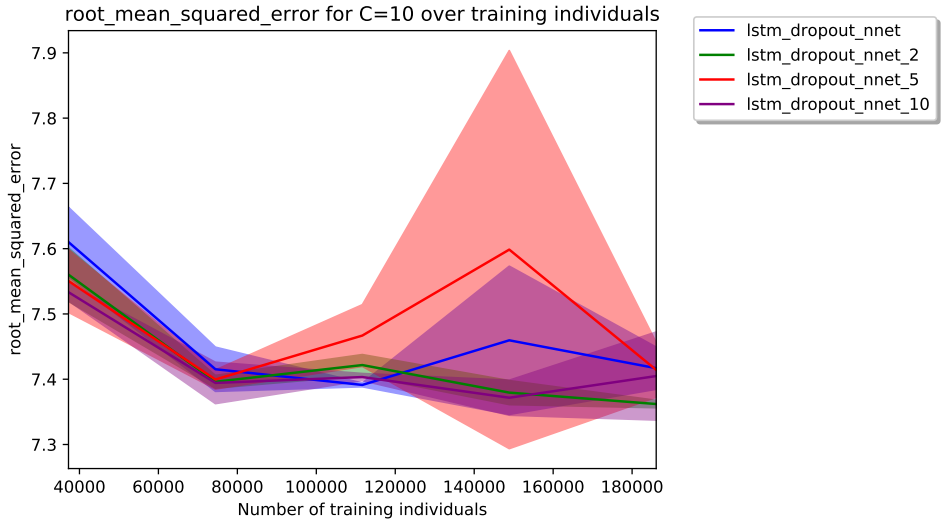


Figure B.10: Prediction RMSE over number of individuals for LSTMs with 1, 2, 5, and 10 layers (blue, green, red, and purple lines, respectively) and a hidden size of $C = 10$.

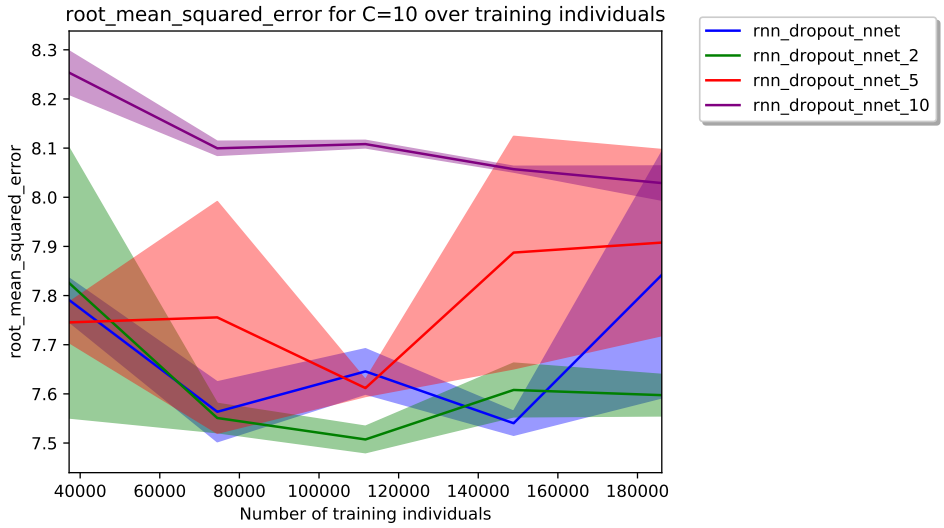


Figure B.11: Prediction RMSE over number of individuals for RNNs with 1, 2, 5, and 10 layers (blue, green, red, and purple lines, respectively) and a hidden size of $C = 10$.

Appendix C

Supplementary information for Chapter 5

C.1 Supplementary Information: Methods

C.1.1 Computing the MLE of b

We want to compute the MLE of b, given true latent $z_{1:T}$ and data $x_{1:T}$, for which we will maximize the likelihood of data $x_{1:T}$ given $z_{1:T}$:

$$p(x_{1:T}|z_{1:T}, b) = \left(\prod_{t=1}^{T} p(x_t|z_t, b)\right) \tag{C.1}$$

$$= \prod_{t=1}^{T} ([x_t = 0, z_t = 0]p(x_t|z_t, b) + [x_t = 1, z_t = 0]p(x_t|z_t, b) \tag{C.2}$$

$$+[x_t = 0, z_t = 1]p(x_t|z_t, b) + [x_t = 1, z_t = 1]p(x_t|z_t, b)) \tag{C.3}$$

$$= p(x_t = 0|z_t = 0, b)^{n_{[x_t=0,z_t=0]}} + p(x_t = 1|z_t = 0, b)^{n_{[x_t=1,z_t=0]}} \tag{C.4}$$

$$+ p(x_t = 0|z_t = 1, b)^{n_{[x_t=0,z_t=1]}} + p(x_t = 1|z_t = 1, b)^{n_{[x_t=1,z_t=1]}} \tag{C.5}$$

$$\tag{C.6}$$

For the values of $p(x_t|z_t, b)$, we marginalize adherence indicators g_t as follows:

$$p(x_t|z_t, b) = \sum_{g_t} p(x_t, g_t|z_t, b) = \sum_{g_t} p(x_t|g_t, z_t)p(g_t|b) \tag{C.7}$$

$$= p(x_t|g_t = 0, z_t)p(g_t = 0|b) + p(x_t|g_t = 1, z_t)p(g_t = 1|b) \tag{C.8}$$

$$= p(x_t|g_t = 0, z_t) \cdot (1 - b) + p(x_t|g_t = 1, z_t) \cdot b \tag{C.9}$$

which result in the following 4 cases

$$p(x_t = 0|z_t = 0, b) = p(x_t = 0|g_t = 0, z_t = 0) \cdot (1 - b) + p(x_t = 0|g_t = 1, z_t = 0) \cdot b \tag{C.10}$$

$$= 1 \cdot (1 - b) + 1 \cdot b \tag{C.11}$$

$$= 1 \tag{C.12}$$

$$p(x_t = 1|z_t = 0, b) = p(x_t = 1|g_t = 0, z_t = 0) \cdot (1 - b) + p(x_t = 1|g_t = 1, z_t = 0) \cdot b \tag{C.13}$$

$$= 0 \cdot (1 - b) + 0 \cdot b \tag{C.14}$$

$$= 0 \tag{C.15}$$

$$p(x_t = 0|z_t = 1, b) = p(x_t = 0|g_t = 0, z_t = 1) \cdot (1 - b) + p(x_t = 0|g_t = 1, z_t = 1) \cdot b \tag{C.16}$$

$$= 1 \cdot (1 - b) + 0 \cdot b \tag{C.17}$$

$$= (1 - b) \tag{C.18}$$

$$p(x_t = 1|z_t = 1, b) = p(x_t = 1|g_t = 0, z_t = 1) \cdot (1 - b) + p(x_t = 1|g_t = 1, z_t = 1) \cdot b \tag{C.19}$$

$$= 0 \cdot (1 - b) + 1 \cdot b \tag{C.20}$$

$$= b \tag{C.21}$$

and therefore, we are dealing with a likelihood of

$$p(x_{1:T}|z_{1:T},b) = p(x_t=0|z_t=0,b)^{n_{[x_t=0,z_t=0]}} + p(x_t=1|z_t=0,b)^{n_{[x_t=1,z_t=0]}} \tag{C.22}$$

$$+ p(x_t=0|z_t=1,b)^{n_{[x_t=0,z_t=1]}} + p(x_t=1|z_t=1,b)^{n_{[x_t=1,z_t=1]}} \tag{C.23}$$

$$= 1^{n_{[x_t=0,z_t=0]}} + 0^{n_{[x_t=1,z_t=0]}} \tag{C.24}$$

$$+ (1-b)^{n_{[x_t=0,z_t=1]}} + (b)^{n_{[x_t=1,z_t=1]}} \tag{C.25}$$

$$= 1^{n_{[x_t=0,z_t=0]}} + (1-b)^{n_{[x_t=0,z_t=1]}} + (b)^{n_{[x_t=1,z_t=1]}} \tag{C.26}$$

$$\tag{C.27}$$

To compute the MLE, we maximize with respect to b:

$$\hat{b} = \arg\max_b \log p(x_{1:T}|z_{1:T},b) \tag{C.28}$$

$$= \arg\max_b \left(n_{[x_t=0,z_t=1]}\log(1-b) + n_{[x_t=1,z_t=1]}\log(b)\right) \tag{C.29}$$

which only depends on $n_{x_t=0,z_t=1}$ and $n_{x_t=1,z_t=1}$, i.e., it's based on the ratio between when $x_t = 0$ or $x_t = 1$ ONLY when $z_t = 1$. More precisely, by looking at gradients

$$\frac{\partial \log p(x_{1:T}|z_{1:T},b)}{\partial b} = \frac{\partial\left(1n_{[x_t=0,z_t=0]} + 0n_{[x_t=1,z_t=0]} + n_{[x_t=0,z_t=1]}(1-b) + n_{[x_t=1,z_t=1]}(b)\right)}{\partial b} \tag{C.30}$$

$$= n_{[x_t=0,z_t=1]}\frac{\partial \log(1-b)}{\partial b} + n_{[x_t=1,z_t=1]}\frac{\partial \log(b)}{\partial b} \tag{C.31}$$

$$= n_{[x_t=0,z_t=1]}\frac{-1}{(1-b)} + n_{[x_t=1,z_t=1]}\frac{1}{b} \tag{C.32}$$

By equating gradients to 0

$$0 = n_{[x_t=0,z_t=1]} \frac{-1}{(1-b)} + n_{[x_t=1,z_t=1]} \frac{1}{b} \tag{C.33}$$

$$n_{[x_t=0,z_t=1]} b = (1-b) n_{[x_t=1,z_t=1]} \tag{C.34}$$

$$b(n_{[x_t=0,z_t=1]} + n_{[x_t=1,z_t=1]}) = n_{[x_t=1,z_t=1]} \tag{C.35}$$

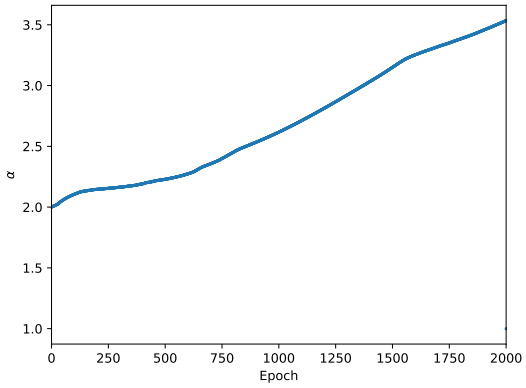
$$\tag{C.36}$$

Resulting in

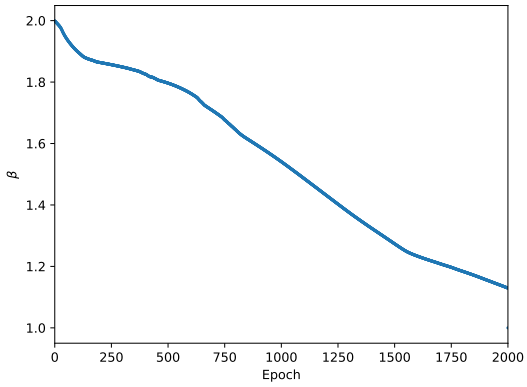
$$\hat{b} = \frac{n_{[x_t=1,z_t=1]}}{n_{[x_t=0,z_t=1]} + n_{[x_t=1,z_t=1]}} \tag{C.37}$$

C.2 Supplementary Information: Results

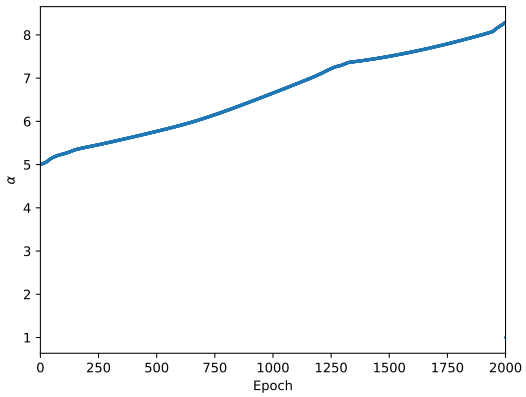
C.2.1 Learned α and β values



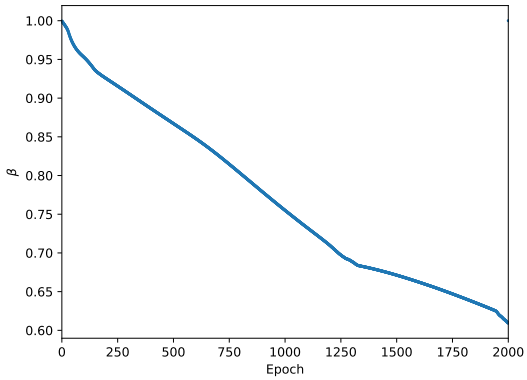
(a) Learned α over epochs, initialized at $(2, 2)$.



(b) Learned β over epochs, initialized at $(2, 2)$.



(c) Learned α over epochs, initialized at $(5, 1)$.



(d) Learned β over epochs, initialized at $(5, 1)$.

Figure C.1: Learned α and β values over epochs for bleeding only model for a particular seed, across different initializations of $(2, 2)$ and $(5, 1)$.

C.2.2 RMSE over prediction day

Table C.1: RMSE of predicted next cycle start over prediction day (aligned at day 0 per user) for model using bleeding only.

Prediction day	Mean RMSE	SD RMSE	Number of users
1	19.73	0.0	59.67
2	20.48	0.0	53.67
3	20.68	0.0	49.67
4	20.77	0.0	41.67
5	22.16	1.05	44.0
6	23.48	1.64	32.0
7	24.12	1.66	15.67
8	25.3	3.43	9.0
9	12.15	7.22	9.0
10	7.26	7.24	9.0
11	6.81	6.6	9.67
12	7.11	6.55	13.67
13	7.19	5.9	11.0
14	5.77	4.89	16.33
15	4.83	3.27	18.33
16	3.75	2.39	26.33
17	3.59	2.11	35.0
18	3.31	1.38	49.0
19	3.14	1.28	56.67
20	2.63	0.6	84.67
21	2.62	0.51	108.67
22	2.47	0.31	153.67
23	2.41	0.3	244.33

Prediction day	Mean RMSE	SD RMSE	Number of users
24	2.23	0.32	449.67
25	2.1	0.3	842.67
26	2.14	0.19	1546.33
27	2.18	0.12	2512.67
28	2.32	0.1	3620.0
29	2.45	0.07	4443.67
30	2.64	0.04	4655.67
31	2.85	0.02	4377.0
32	2.99	0.04	3773.33
33	3.09	0.06	3048.33
34	3.15	0.06	2445.33

Table C.2: RMSE of predicted next cycle start over prediction day (aligned at day 0 per user) for model using bleeding and energy.

Prediction day	Mean RMSE	SD RMSE	Number of users
1	21.89	0.0	5.0
2	17.87	3.87	6.33
3	20.72	0.0	5.0
4	19.3	2.27	7.67
5	21.54	0.82	20.67
6	23.45	1.78	21.33
7	24.34	0.38	12.33
8	25.37	2.59	7.0
9	15.91	9.89	10.0
10	14.45	8.8	13.0

Prediction day	Mean RMSE	SD RMSE	Number of users
11	14.28	8.72	12.33
12	12.96	7.76	16.0
13	10.97	5.85	16.33
14	7.78	4.08	16.33
15	6.1	2.2	17.33
16	3.59	1.02	21.33
17	3.1	1.08	28.67
18	3.14	0.65	40.33
19	3.29	0.6	46.0
20	2.9	0.38	62.33
21	2.81	0.47	68.67
22	2.63	0.4	96.0
23	2.37	0.21	157.67
24	2.32	0.26	273.0
25	2.25	0.24	515.0
26	2.2	0.21	932.0
27	2.25	0.11	1524.0
28	2.41	0.1	2190.67
29	2.57	0.1	2636.33
30	2.78	0.08	2785.67
31	2.95	0.07	2646.33
32	3.1	0.08	2230.67
33	3.33	0.17	1796.67
34	3.43	0.11	1412.67

Table C.3: RMSE of predicted next cycle start over prediction day (aligned at day 0 per user) for model using bleeding and emotion.

Prediction day	Mean RMSE	SD RMSE	Number of users
1	19.18	0.18	103.67
2	20.13	0.18	102.67
3	19.79	0.35	95.33
4	20.97	0.15	91.33
5	21.52	0.32	80.33
6	22.5	0.57	49.67
7	22.9	0.17	26.67
8	22.05	0.65	18.0
9	20.61	0.23	20.0
10	19.56	0.35	24.0
11	18.96	0.74	25.67
12	18.85	0.44	25.0
13	16.25	0.45	23.33
14	13.36	0.35	24.67
15	9.86	0.51	26.33
16	6.87	0.61	33.0
17	5.95	0.13	45.33
18	3.97	0.35	61.33
19	3.64	0.01	68.0
20	3.3	0.06	89.67
21	3.44	0.06	90.33
22	3.18	0.02	121.67
23	2.95	0.03	191.33

Prediction day	Mean RMSE	SD RMSE	Number of users
24	2.56	0.07	318.67
25	2.57	0.01	561.0
26	2.37	0.04	1057.0
27	2.35	0.04	1696.67
28	2.5	0.02	2434.33
29	2.66	0.03	2930.0
30	2.85	0.05	3088.33
31	3.05	0.03	2945.0
32	3.18	0.02	2503.0
33	3.48	0.01	2012.67
34	3.54	0.03	1593.0

Table C.4: RMSE of predicted next cycle start over prediction day (aligned at day 0 per user) for model using bleeding and pain.

Prediction day	Mean RMSE	SD RMSE	Number of users
1	19.36	0.0	34.0
2	19.73	0.0	34.67
3	19.51	0.0	30.33
4	20.86	0.93	23.67
5	22.01	0.71	34.33
6	23.27	0.9	28.0
7	23.91	0.64	16.0
8	21.67	1.38	8.67
9	17.81	3.38	14.33
10	14.64	5.49	12.33

Prediction day	Mean RMSE	SD RMSE	Number of users
11	11.49	6.64	13.33
12	14.4	1.05	13.67
13	9.81	5.38	13.0
14	8.64	4.69	15.0
15	6.86	3.22	16.67
16	5.01	1.59	26.33
17	4.24	1.49	38.67
18	3.04	0.69	49.67
19	3.19	0.62	55.33
20	2.79	0.48	77.67
21	2.7	0.53	91.33
22	2.47	0.24	121.67
23	2.36	0.3	201.67
24	2.1	0.16	346.0
25	2.08	0.18	651.33
26	2.12	0.09	1205.67
27	2.16	0.06	1938.0
28	2.34	0.03	2827.0
29	2.51	0.03	3453.67
30	2.71	0.03	3657.33
31	2.9	0.04	3461.0
32	3.07	0.05	2935.0
33	3.3	0.08	2357.0
34	3.32	0.12	1852.0

C.2.3 AUC of predicting bleeding on days $4-7$ of test set

Table C.5: AUC of predicting bleeding on day 4 of the test set over prediction day (aligned at day 0) for bleeding only model. The average and SD are computed across 3 seeds.

Prediction day	Mean AUC	SD AUC
0	0.61	0.04
1	0.67	0.08
2	0.66	0.1
3	0.72	0.06

Table C.6: AUC of predicting bleeding on day 5 of the test set over prediction day (aligned at day 0) for bleeding only model. The average and SD are computed across 3 seeds.

Prediction day	Mean AUC	SD AUC
0	0.59	0.04
1	0.66	0.08
2	0.65	0.1
3	0.68	0.07
4	0.74	0.05

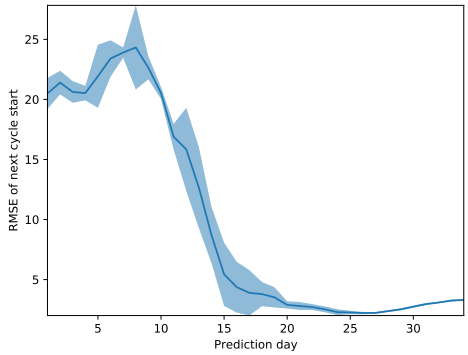
Table C.7: AUC of predicting bleeding on day 6 of the test set over prediction day (aligned at day 0) for bleeding only model. The average and SD are computed across 3 seeds.

Prediction day	Mean AUC	SD AUC
0	0.58	0.04
1	0.62	0.06
2	0.6	0.06
3	0.61	0.06
4	0.65	0.06
5	0.76	0.03

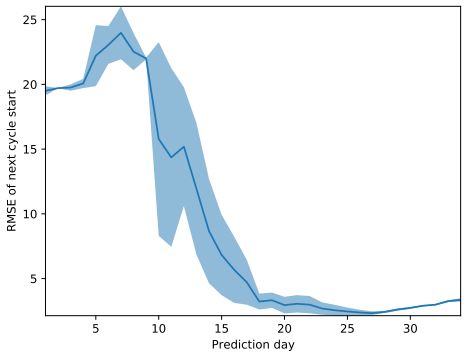
Table C.8: AUC of predicting bleeding on day 7 of the test set over prediction day (aligned at day 0) for bleeding only model. The average and SD are computed across 3 seeds.

Prediction day	Mean AUC	SD AUC
0	0.55	0.04
1	0.64	0.03
2	0.59	0.04
3	0.59	0.04
4	0.57	0.09
5	0.57	0.18
6	0.83	0.04

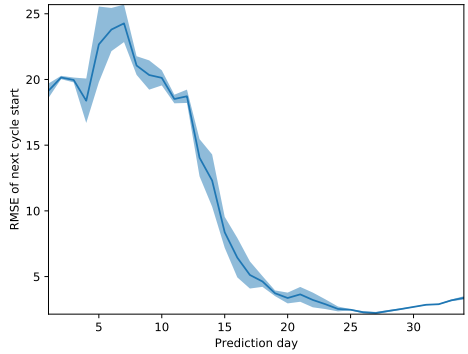
C.2.4 RMSE of predicting next cycle start for $(2, 2)$ initialization



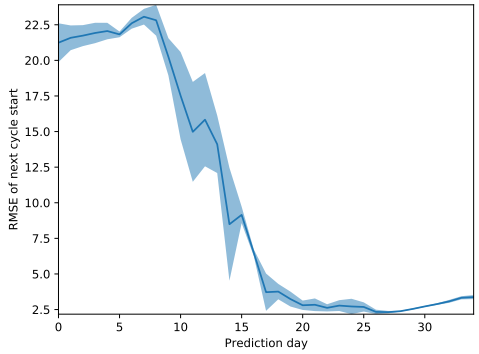
(a) RMSE of predicted next cycle start, using bleeding only.



(b) RMSE of predicted next cycle start, using bleeding and energy.



(c) RMSE of predicted next cycle start, using bleeding and emotion.



(d) RMSE of predicted next cycle start, using bleeding and pain.

Figure C.2: RMSE of predicting next cycle start across models using $(2, 2)$ initialization for α and β.

C.2.5 Normalized histogram of events per day

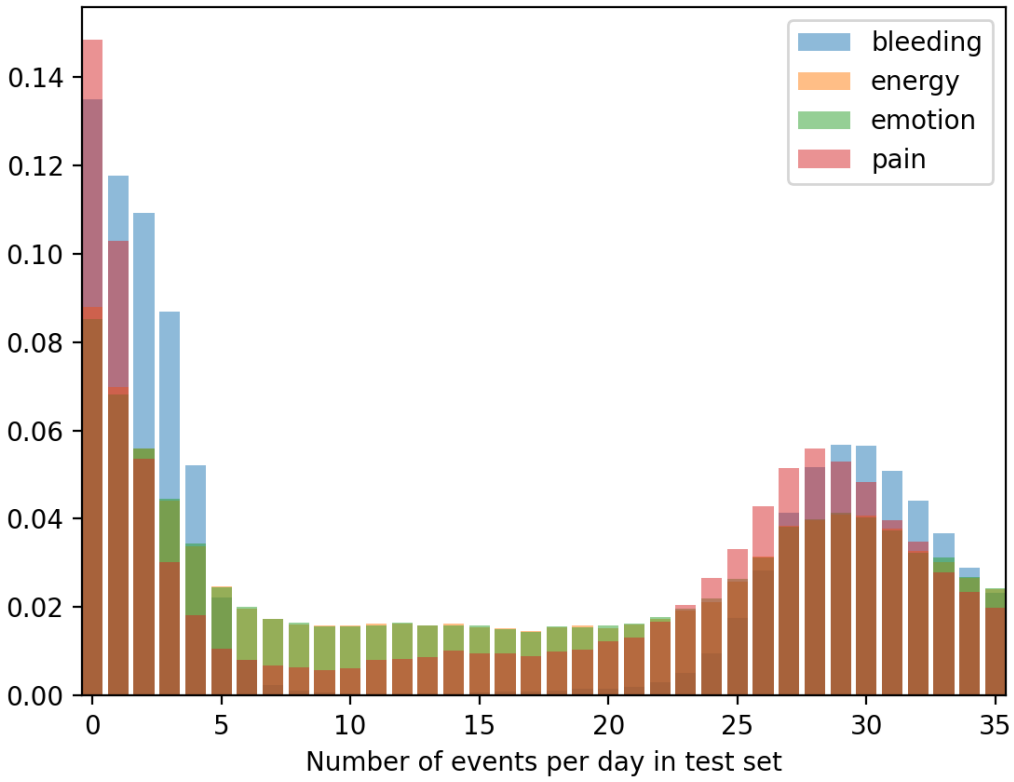


Figure C.3: Histogram of observed number of events per symptom on each day of the test set, normalized by total number of events per symptom, i.e., the proportion of tracking events per symptom on each day.

Bibliography

[1] Treloar, A. E., Boynton, R. E., Behn, B. G. & Brown, B. W. Variation of the human menstrual cycle through reproductive life. *International journal of fertility* **12**, 77–126 (1967). URL http://europepmc.org/abstract/MED/5419031.

[2] Chiazze, L., Brayer, F. T., John J. Macisco, J., Parker, M. P. & Duffy, B. J. The Length and Variability of the Human Menstrual Cycle. *The Journal of the American Medical Association* **203**, 377–380 (1968).

[3] Ferrell, R. J. *et al.* Monitoring reproductive aging in a 5-year prospective study: aggregate and individual changes in steroid hormones and menstrual cycle lengths with age. *Menopause* **12**, 567–757 (2005).

[4] Vitzthum, V. J. The ecology and evolutionary endocrinology of reproduction in the human female. *American Journal of Physical Anthropology* **140**, 95–136 (2009). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ajpa.21195. https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajpa.21195.

[5] Harlow, S. D. *et al.* Executive Summary of the Stages of Reproductive Aging Workshop + 10: Addressing the Unfinished Agenda of Staging Reproductive Aging. *The Journal of Clinical Endocrinology & Metabolism* **97**, 1159–1168 (2012). URL https://doi.org/10.1210/jc.2011-3362.

[6] Shands, K. N. *et al.* Toxic-shock syndrome in menstruating women. *New England Journal of Medicine* **303**, 1436–1442 (1980). URL https://doi.org/10.1056/NEJM198012183032502. PMID: 7432402, https://doi.org/10.1056/NEJM198012183032502.

[7] Rubinsky, V., Gunning, J. & Cooke-Jackson, A. "i thought i was dying:" (un)supportive communication surrounding early menstruation experiences. *Health Communication* **35** (2018).

[8] Schmitt, M. L. *et al.* The intersection of menstruation, school and family: Experiences of girls growing up in urban cities in the u.s.a. *International Journal of Adolescence and Youth* **26**, 94–109 (2021). URL https://doi.org/10.1080/02673843.2020.1867207. https://doi.org/10.1080/02673843.2020.1867207.

[9] Mason, L. *et al.* 'we keep it secret so no one should know' – a qualitative study to explore young schoolgirls attitudes and experiences with menstruation in rural western kenya. *PLOS ONE* **8**, 1–11 (2013). URL https://doi.org/10.1371/journal.pone.0079132.

[10] Schoep, M. E., Nieboer, T. E., van der Zanden, M., Braat, D. D. & Nap, A. W. The impact of menstrual symptoms on everyday life: a survey among 42,879 women. *American Journal of Obstetrics and Gynecology* **220**, 569.e1–569.e7 (2019). URL https://www.sciencedirect.com/science/article/pii/S0002937819304272.

[11] Jordan, J., Craig, K., Clifton, D. K. & Soules, M. R. Luteal phase defect: the sensitivity and specificity of diagnostic methods in common clinical use. *Fertility and Sterility* **62**, 54 – 62 (1994). URL http://www.sciencedirect.com/science/article/pii/S0015028216568150.

[12] Crawford, N. M., Pritchard, D. A., Herring, A. H. & Steiner, A. Z. Prospective evaluation of luteal phase length and natural fertility. *Fertility and Sterility* **107**, 749 – 755 (2017). URL http://www.sciencedirect.com/science/article/pii/S0015028216630224.

[13] Prior, J. C. Perimenopause: The Complex Endocrinology of the Menopausal Transition. *Endocrine Reviews* **19**, 397–428 (1998). URL http://dx.doi.org/10.1210/edrv.19.4.0341.

[14] Landgren, B.-M. *et al.* Menopause Transition: Annual Changes in Serum Hormonal Patterns over the Menstrual Cycle in Women during a Nine-Year Period Prior to Menopause. *The Journal of Clinical Endocrinology & Metabolism* **89**, 2763–2769 (2004). URL http://dx.doi.org/10.1210/jc.2003-030824.

[15] Prior, J. C. & Hitchcock, C. L. The endocrinology of perimenopause: need for a paradigm shift. *Frontiers in bioscience (Scholar edition)* **3**, 474 – 486 (2011). URL https://doi.org/10.2741/s166.

[16] Solomon, C. G. *et al.* Menstrual cycle irregularity and risk for future cardiovascular disease. *The Journal of Clinical Endocrinology & Metabolism* **87**, 2013–2017 (2002).

[17] American College of Obstetricians and Gynecologists. Menstruation in Girls and Adolescents: Using the Menstrual Cycle as a Vital Sign. *Obstetrics & Gynecology* **126**, 143–6 (2015). URL https://www.acog.org/Clinical-Guidance-and-Publications/Committee-Opinions/Committee-on-Adolescent-Health-Care/Menstruation-in-Girls-and-Adolescents-Using-the-Menstrual-Cycle-as-a-Vital-Sign.

[18] Bobel, C. Beyond the Managed Body: Putting Menstrual Literacy at the Center. In Bobel, C. (ed.) *The Managed Body: Developing Girls and Menstrual Health in the Global South*, 281–321 (Springer International Publishing, Cham, 2019). URL https://doi.org/10.1007/978-3-319-89414-0_8.

[19] Lippe Taylor, Inc. Scientific forum addresses menstrual cycle as vital sign (2004). URL http://www.eurekalert.org/pub_releases/2004-09/lti-sfa092004.php.

[20] of Pediatrics, A. A., of Obstetricians, A. C., Gynecologists *et al.* Menstruation in girls and adolescents: using the menstrual cycle as a vital sign. *Pediatrics* **118**, 2245–2250 (2006).

[21] Critchley, H. O. *et al.* Menstruation: science and society. *American Journal of Obstetrics and Gynecology* **223**, 624 – 664 (2020). URL `http://www.sciencedirect.com/science/article/pii/S0002937820306190`.

[22] As-Sanie, S. *et al.* Assessing research gaps and unmet needs in endometriosis. *American Journal of Obstetrics and Gynecology* **221**, 86 – 94 (2019). URL `http://www.sciencedirect.com/science/article/pii/S0002937819303850`.

[23] Abdul, G. Pantone's newest color is a nod to menstruation: Period red (2020). URL `https://www.nytimes.com/2020/09/30/business/pantone-color.html`.

[24] Hripcsak, G. *et al.* Characterizing treatment pathways at scale using the OHDSI network. *Proceedings of the National Academy of Sciences* **113**, 7329–7336 (2016). URL `https://www.pnas.org/content/113/27/7329`. `https://www.pnas.org/content/113/27/7329.full.pdf`.

[25] Li, I., Dey, A. & Forlizzi, J. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 557–566 (ACM, 2010).

[26] Kohane, I. S. Ten things we have to do to achieve precision medicine. *Science* **349**, 37–38 (2015). URL `http://science.sciencemag.org/content/349/6243/37`. `http://science.sciencemag.org/content/349/6243/37.full.pdf`.

[27] Fox, S. & Duggan, M. Tracking for Health. Tech. Rep., Pew Research Center (2013). URL `http://www.pewinternet.org/2013/01/28/tracking-for-health/`.

[28] Krebs, P. & Duncan, D. T. Health app use among US mobile phone owners: a national survey. *JMIR mHealth and uHealth* **3** (2015).

[29] Althoff, T. Population-Scale Pervasive Health. *IEEE Pervasive Computing* **16**, 75–79 (2017).

[30] Althoff, T. *et al.* Large-scale physical activity data reveal worldwide activity inequality. *Nature* **547** (2017).

[31] Chan, Y.-F. Y. *et al.* The Asthma Mobile Health Study, a large-scale clinical observational study using ResearchKit. *Nature biotechnology* **35**, 354 (2017).

[32] Webster, D. E. *et al.* The Mole Mapper Study, mobile phone skin imaging and melanoma risk data collected using ResearchKit. *Scientific Data* **4** (2018).

[33] Egger, H. L. *et al.* Automatic emotion and attention analysis of young children at home: a ResearchKit autism feasibility study. *Nature Digital Medicine* **1** (2018).

[34] Bot, B. M. *et al.* The mPower study, Parkinson disease mobile data collected using ResearchKit. *Scientific data* **3**, 160011 (2016).

[35] Dagum, P. Digital biomarkers of cognitive function. *Nature Digital Medicine* **1** (2018).

[36] Smets, E. *et al.* Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *Nature Digital Medicine* **1** (2018).

[37] Byambasuren, O., Sanders, S., Beller, E. & Glasziou, P. Prescribable mHealth apps identified from an overview of systematic reviews. *Nature Digital Medicine* **1** (2018).

[38] Ata, R. *et al.* Clinical validation of smartphone-based activity tracking in peripheral artery disease patients. *Nature Digital Medicine* **1** (2018).

[39] Torous, J. *et al.* Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia. *Nature Digital Medicine* **1** (2018).

[40] Urteaga, I., McKillop, M., Lipsky-Gorman, S. & Elhadad, N. Phenotyping Endometriosis through Mixed Membership Models of Self-Tracking Data. In *2018 Machine Learning for Healthcare (MLHC)* (2018). URL `https://www.mlforhc.org/s/27.pdf`.

[41] Wartella, E., Rideout, V., Montague, H., Beaudoin-Ryan, L. & Lauricella, A. Teens, health and technology: A national survey. *Media and Communication* **4**, 13–23 (2016).

[42] Fox, S. & Duggan, M. Mobile Health 2012. Tech. Rep., Pew Research Center (2012). URL `http://www.pewinternet.org/2012/11/08/mobile-health-2012/`.

[43] Clue by BioWink GmbH, Adalbertstraße 7-8, 10999 Berlin, Germany. https://helloclue.com/ (2019).

[44] Dot: A fertility tracker app. https://www.dottheapp.com/ (2019).

[45] Glow: An App for Fertility & Beyond. https://glowing.com/glow (2019).

[46] Spot On: A Birth Control And Period Tracker App powered by Planned Parenthood. https://shortyawards.com/9th/spot-on (2019).

[47] Natural Cycles: Digital Birth Control. https://www.naturalcycles.com (2019).

[48] Pierson, E., Althoff, T., Thomas, D., Hillard, P. & Leskovec, J. The menstrual cycle is a primary contributor to cyclic variation in women's mood, behavior, and vital signs. *bioRxiv* (2019). URL `https://www.biorxiv.org/content/early/2019/03/20/583153`. `https://www.biorxiv.org/content/early/2019/03/20/583153.full.pdf`.

[49] Bull, J. R. *et al.* Real-world menstrual cycle characteristics of more than 600,000 menstrual cycles. *Nature Digital Medicine* **2** (2019).

[50] Symul, L., Wac, K., Hillard, P. & Salathé, M. Assessment of menstrual health status and evolution through mobile apps for fertility awareness. *Nature Digital Medicine* **2** (2019).

[51] Epstein, D. A. *et al.* Examining Menstrual Tracking to Inform the Design of Personal Informatics Tools. *Proceedings of the SIGCHI conference on human factors in computing systems. CHI Conference* **2017**, 6876–6888 (2017). URL http://europepmc.org/articles/PMC5432133.

[52] Moglia, M. L., Nguyen, H. V., Chyjek, K., Chen, K. T. & Castaño, P. M. Evaluation of Smartphone Menstrual Cycle Tracking Applications Using an Adapted Applications Scoring System. *Obstetrics and Gynecology* **127**, 1153–1160 (2016).

[53] The belmont report: Ethical principles and guidelines for the protection of human subjects of research (1979). URL https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html.

[54] Koller, D. & Friedman, N. *Probabilistic graphical models: principles and techniques* (MIT Press, 2009).

[55] Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

[56] Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016). http://www.deeplearningbook.org.

[57] Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324 (1998).

[58] Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).

[59] Graves, A. *Supervised Sequence Labelling with Recurrent Neural Networks* (Springer, Berlin, Heidelberg, 2012).

[60] Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997). URL https://doi.org/10.1162/neco.1997.9.8.1735.

[61] Tansey, W. *et al.* Dose-response modeling in high-throughput cancer drug screenings: an end-to-end approach. *Biostatistics* **23**, 643–665 (2021). URL ttps://doi.org/10.1093/biostatistics/kxaa047.

[62] Dusenbery, M. *Doing Harm: The Truth* (HarperOne, 2018).

[63] Ju, H., Jones, M. & Mishra, G. The Prevalence and Risk Factors of Dysmenorrhea. *Epidemiologic Reviews* **36**, 104–113 (2013). URL https://doi.org/10.1093/epirev/mxt009. https://academic.oup.com/epirev/article-pdf/36/1/104/16731532/mxt009.pdf.

[64] Unsal, A., Ayranci, U., Tozun, M., Arslan, G. . & Calik, E. Prevalence of dysmenorrhea and its effect on quality of life among a group of female university students. *Upsala Journal of Medical Sciences* **115**, 138–145 (2010). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2853792/.

[65] Evans, S. F., Brooks, T. A., Esterman, A. J., Hull, M. L. & Rolan, P. E. The comorbidities of dysmenorrhea: a clinical survey comparing symptom profile in women with and without endometriosis. *Journal of Pain Research* **11**, 3181–3194 (2018). URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6300370/`.

[66] Culley, L. *et al.* The social and psychological impact of endometriosis on women's lives: a critical narrative review. *Human Reproduction Update* **19**, 625–639 (2013). URL `https://doi.org/10.1093/humupd/dmt027`. `https://academic.oup.com/humupd/article-pdf/19/6/625/2498455/dmt027.pdf`.

[67] Moradi, M., Parker, M., Sneddon, A., Lopez, V. & Ellwood, D. Impact of endometriosis on women's lives: a qualitative study. *BMC Womens Health* **14** (2014). URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4287196/`.

[68] Shea, A. A. & Vitzthum, V. J. The extent and causes of natural variation in menstrual cycles: Integrating empirically-based models of ovarian cycling into research on women's health. *Drug Discovery Today: Disease Models* **32**, 41–49 (2020).

[69] Arey, L. B. The Degree of Normal Menstrual Irregularity. *American Journal of Obstetrics & Gynecology* **37**, 12–29 (1939).

[70] Münster, K., Schmidt, L. & Helm, P. Length and variation in the menstrual cycle — a cross-sectional study from a Danish county. *The British Journal of Obstetrics and Gynaecology* **99**, 422 – 429 (1992).

[71] Belsey, E. M. & Pinol, A. P. Menstrual bleeding patterns in untreated women. Task Force on Long-Acting Systemic Agents for Fertility Regulation. *Contraception* **55**, 57–65 (1997).

[72] Burkhart, M. C., de Mazariegos, L., Salazar, S. & Hess, T. Incidence of irregular cycles among Mayan women who reported having regular cycles: implications for fertility awareness methods. *Contraception* **59**, 271 – 275 (1999).

[73] Vitzthum, V. J., Spielvogel, H., Caceres, E. & Gaines, J. Menstrual patterns and fecundity among non-lactating and lactating cycling women in rural highland Bolivia: implications for contraceptive choice. *Contraception* **62**, 181 –187 (2000).

[74] Creinin, M. D., Keverline, S. & Meyn, L. A. How regular is regular? An analysis of menstrual cycle regularity. *Contraception* **70**, 289–292 (2004).

[75] Williams, S. R. Menstrual cycle characteristics and predictability of ovulation of Bhutia women in Sikkim, India. *Journal of physiological anthropology* **25**, 85–90 (2006).

[76] Cole, L. A., Ladner, D. G. & Byrn, F. W. The normal variabilities of the menstrual cycle. *Fertility and Sterility* **91**, 522–527 (2009). URL `https://www.fertstert.org/article/S0015-0282(07)04138-6/fulltext`.

[77] Kolmogorov, A. N. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italianodegli Attuari* **4**, 83–91 (1933).

[78] Druet, A. What is an "irregular" menstrual cycle? https://helloclue.com/articles/cycle-a-z/what-is-an-irregular-menstrual-cycle (2018). Clue by BioWink GmbH, Adalbertstraße 7-8, 10999 Berlin, Germany.

[79] Johannisson, E., Landgren, B.-M., Rohr, H. P. & Diczfalusy, E. Endometrial morphology and peripheral hormone levels in women with regular menstrual cycles. *Fertility and Sterility* **48**, 401 – 408 (1987). URL `http://www.sciencedirect.com/science/article/pii/S0015028216594060`.

[80] Fehring, R. J., Schneider, M. & Raviele, K. M. Variability in the phases of the menstrual cycle. *Journal of obstetric, gynecologic, and neonatal nursing : JOGNN* **35 3**, 376–84 (2006).

[81] Lenton, E. A., Landgren, B.-M. & Sexton, L. Normal variation in the length of the luteal phase of the menstrual cycle: identification of the short luteal phase. *BJOG: An International Journal of Obstetrics & Gynaecology* **91**, 685–689 (1984). URL `http://dx.doi.org/10.1111/j.1471-0528.1984.tb04831.x`.

[82] Lenton, E. A., Landgren, B.-M., Sexton, L. & Harper, R. Normal variation in the length of the follicular phase of the menstrual cycle: effect of chronological age. *BJOG: An International Journal of Obstetrics & Gynaecology* **91**, 681–684 (1984). URL `https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1111/j.1471-0528.1984.tb04830.x`.

[83] Pierson, E., Althoff, T. & Leskovec, J. Modeling Individual Cyclic Variation in Human Behavior. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, 107–116 (International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2018). URL `https://doi.org/10.1145/3178876.3186052`.

[84] Jones, E. K., Jurgenson, J. R., Katzenellenbogen, J. M. & Thompson, S. C. Menopause and the influence of culture: another gap for Indigenous Australian women? *BMC Women's health* **12** (2012).

[85] Ayobi, A., Marshall, P., Cox, A. L. & Chen, Y. Quantifying the body and caring for the mind: Self-tracking in multiple sclerosis. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, 6889–6901 (Association for Computing Machinery, New York, NY, USA, 2017). URL `https://doi.org/10.1145/3025453.3025869`.

[86] Desai, P. M. *et al.* Personal health oracle: Explorations of personalized predictions in diabetes self-management. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–13 (Association for Computing Machinery, New York, NY, USA, 2019). URL `https://doi.org/10.1145/3290605.3300600`.

[87] McKillop, M., Mamykina, L. & Elhadad, N. Designing in the dark: Eliciting self-tracking dimensions for understanding enigmatic disease. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 1–15 (Association for Computing Machinery, New York, NY, USA, 2018). URL `https://doi.org/10.1145/3173574.3174139`.

[88] Costa Figueiredo, M. *et al.* Self-tracking for fertility care: Collaborative support for a highly personalized problem. *Proc. ACM Hum.-Comput. Interact.* **1** (2017). URL https://doi.org/10.1145/3134671.

[89] Consolvo, S., Klasnja, P., McDonald, D. W. & Landay, J. A. Designing for healthy lifestyles: Design considerations for mobile technologies to encourage consumer health and wellness. *Foundations and Trends® in Human–Computer Interaction* **6**, 167–315 (2014). URL http://dx.doi.org/10.1561/1100000040.

[90] Epstein, D. A., Ping, A., Fogarty, J. & Munson, S. A. A lived informatics model of personal informatics. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, 731–742 (Association for Computing Machinery, New York, NY, USA, 2015). URL https://doi.org/10.1145/2750858.2804250.

[91] Shaw, R. J. *et al.* Mobile health devices: will patients actually use them? *Journal of the American Medical Informatics Association* **23**, 462–466 (2016). URL https://doi.org/10.1093/jamia/ocv186.

[92] Vaghefi, I. & Tulu, B. The continued use of mobile health apps: Insights from a longitudinal study. *JMIR mHealth and uHealth* **7** (2019). URL https://doi.org/10.2196/12983.

[93] Choe, E. K., Lee, N. B., Lee, B., Pratt, W. & Kientz, J. A. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, 1143–1152 (Association for Computing Machinery, New York, NY, USA, 2014). URL https://doi.org/10.1145/2556288.2557372.

[94] Pierson, E., Althoff, T., Thomas, D., Hillard, P. & Leskovec, J. Daily, weekly, seasonal and menstrual cycles in women's mood, behaviour and vital signs. *Nature Human Behaviour* 1–10 (2021).

[95] Li, K. *et al.* Characterizing physiological and symptomatic variation in menstrual cycles using self-tracked mobile health data. *Nature Digital Medicine* **3** (2020). 1808.02932.

[96] Soumpasis, I., Grace, B. & Johnson, S. Real-life insights on menstrual cycles and ovulation using big data. *Human Reproduction Open* **2020** (2020). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7164578/.

[97] Fox, S. & Epstein, D. A. Monitoring menses: Design-based investigations of menstrual tracking applications. *The Palgrave Handbook of Critical Menstruation Studies* 733–750 (2020).

[98] Bishop, C. M. Model-based machine learning. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **371**, 20120222 (2013). URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2012.0222. https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2012.0222.

[99] Bortot, P., Masarotto, G. & Scarpa, B. Sequential predictions of menstrual cycle lengths. *Biostatistics* **11**, 741–755 (2010). URL https://doi.org/10.1093/biostatistics/kxq020. https://academic.oup.com/biostatistics/article-pdf/11/4/741/17738013/kxq020.pdf.

[100] Harlow, S. D. & Zeger, S. L. An application of longitudinal methods to the analysis of menstrual diary data. *Journal of Clinical Epidemiology* **44**, 1015 – 1025 (1991). URL http://www.sciencedirect.com/science/article/pii/089543569190003R.

[101] Harlow, S. D., Lin, X. & Ho, M. Analysis of menstrual diary data across the reproductive life span applicability of the bipartite model approach and the importance of within-woman variance. *Journal of Clinical Epidemiology* **53**, 722 – 733 (2000). URL http://www.sciencedirect.com/science/article/pii/S0895435699002024.

[102] Guo, Y., Manatunga, A. K., Chen, S. & Marcus, M. Modeling menstrual cycle length using a mixture distribution. *Biostatistics* **7**, 100–114 (2005). URL https://doi.org/10.1093/biostatistics/kxi043. https://academic.oup.com/biostatistics/article-pdf/7/1/100/669902/kxi043.pdf.

[103] Oliveira, T., Bruinvels, G., Pedlar, C. & Newell, J. Modelling menstrual cycle length in athletes using state-space models (2020).

[104] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, 8024–8035 (Curran Associates, Inc., 2019). URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[105] Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag, Berlin, Heidelberg, 2006).

[106] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization (2017). 1412.6980.

[107] Zou, K. H., O'Malley, A. J. & Mauri, L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation* **115**, 654–657 (2007).

[108] Zeiler, M. D. Adadelta: An adaptive learning rate method (2012). URL https://arxiv.org/abs/1212.5701.