Labeling Self-Tracked Menstrual Health Records With Hidden Semi-Markov Models

Laura Symul and Susan Holmes

***Abstract*—Globally, millions of women track their menstrual cycle and fertility via smartphone-based health apps, generating multivariate time series with frequent missing data. To leverage this type of data for studies of fertility or studies of the effect of the menstrual cycle on symptoms and diseases, it is critical to have methods for identifying reproductive events, such as ovulation, pregnancy losses or births. Here, we present a hierarchical approach relying on hidden semi-Markov models that adapts to changes in tracking behavior, explicitly captures variable– and state–dependent missingness, allows for variables of different type, and quantifies uncertainty. The accuracy on simulated data reaches 98% with no missing data and 90% with realistic missingness. On our partially labeled real-world time series, the accuracy reaches 93%. Our method also accurately predicts cycle length by learning user characteristics. Its implementation is publicly available (`HiddenSemiMarkov R` package) and transferable to any health time series, including self-reported symptoms and occasional tests.**

***Index Terms*—Hidden semi-Markov models (HSMM), digital health, gynecology, fertility, mobile applications, unsupervised learning, semisupervised learning, statistical learning.**

I. Introduction

HEALTH tracking apps have become increasingly popular and self-reported health records collected via apps or connected devices are progressively adopted by the scientific community for personalized health or epidemiological research [1]. Menstrual cycle and fertility tracking apps are among the most used health apps [2]. These apps are now used by millions of women worldwide, generating very large datasets of self-reports related to the menstrual cycle and reproductive events. Users of these apps typically report their period bleeding along with physical or psychological symptoms and/or fertility-related body-signs.

These large datasets have already been used to characterize the duration of the menstrual cycle and the follicular (before ovulation) and luteal (after ovulation) phases [3]–[5], to evaluate the association between sexually transmittable infections (STI) and pre-menstrual symptoms [6], and to evaluate the association between cycle length irregularities and reported symptoms [7]. In addition to these findings, this data indubitably holds additional information on fertility, pregnancy losses and menstrual health in general. This information can be used to tackle scientific challenges and address unanswered questions about the human reproductive biology. For example, this data can be used to evaluate whether seasonal and geographical variations of fertility [8] is due to changes in ovulation or loss rates or to study, at large scale, the predictability of mental health variations throughout the menstrual cycle [9], [10]. Beyond the potential of these large retrospective datasets, apps and/or connected devices also provide a scalable way to prospectively collect longitudinal data of menstrual-health related body signs and symptoms for a large population size over a long period of time without requiring in-person visits to a clinic. The prospective digital collection of data related to fertility and menstrual health provides an opportunity to evaluate the association between the menstrual cycle or reproductive status and other health variables at large scale.

A first challenge in using such self-reported data is the contextualization of each observation within the reproductive timeline of an individual. The interpretation of a reported symptom varies greatly if the individual is pregnant or going through a long anovulatory phase. This contextualization requires the labeling of users' time series with biologically-relevant states such as "pregnant" or "ovulating".

Labeling self-tracked datasets can be a challenging process given the multivariate nature of the datasets, the prevalence of missing data, and the lack of available ground-truth. To our knowledge, there are no available labeled datasets for menstrual self-tracked data. Thus, supervised labeling methods such as Long-Short-Term-Memory (LSTM) models [11], [12], or Transformers [13], cannot be used. Fortunately, this lack of available labeled samples is balanced by a good knowledge of the underlying reproductive biology. This knowledge can be translated into statistical priors and inform the design of unsupervised or generative models.

For example, it has been well documented that cervical mucus properties and quantities are controlled by cycling reproductive hormones [14]–[16] and that these changes can be observed and reported by app users [3], [5]. Body temperature at wake-up has been shown to increase after ovulation and in early pregnancy [15], [17]. Concentration in luteinizing hormone (LH) surges before ovulation [15], [18] and this surge can be detected

Manuscript received March 9, 2021; revised July 23, 2021; accepted August 29, 2021. Date of publication September 8, 2021; date of current version March 7, 2022. The work of Laura Symul was supported by "Stanford Clinical Data Science Fellowship" and the work of Susan Holmes was supported by NSF under Grant DMS RTG 1501767. *(Corresponding author: Susan Holmes.)*

The authors are with the Department of Statistics, at Stanford University, Stanford, CA 94035 USA (e-mail: lsymul@stanford.edu; susan@stat.stanford.edu).

Digital Object Identifier 10.1109/JBHI.2021.3110716

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

by cheap at-home urine kits [19]. Bleeding, the most obvious body-sign to report in a menstrual-cycle tracking app, is highly correlated with menses (periods), pregnancy losses or births. Light bleeding may also be indicative of ovulation or be reported in early pregnancy [20], [21].

Hidden state models are appropriate for labeling self-tracked time series because the underlying biological states can be matched to the model's hidden states. In the medical literature, the menstrual cycle is frequently split into successive phases (menses, early follicular phase, peri-ovulatory phase, early and late luteal phases) and pregnancies are frequently divided into trimesters. Given that these phases have been well characterized, they can be naturally translated into a discrete state model: each latent state matches one of the menstrual or pregnancy phases. In previous work, hidden Markov models (HMM), the most common discrete hidden state model for time series, have been used to label menstrual-cycle time series [3]. However, the Markovian property imposes a geometrical distribution for the duration of each state, which does not accurately model the menstrual or pregnancy phases. Hidden Markov models only performed well in labeling single cycles whose start and ends were already identified and where users had reported enough data to constrain the duration of each phase. Others have proposed cyclic HMM (CyHMM) to recover cycle characteristics from menstrual cycle app data [22]. While this framework is successful in identifying cycles, it did not include prior biological knowledge beyond the average cycle length. Consequently, the hidden states can not directly be matched to and interpreted as biological states. Additionally, because that framework assumed cycles with relatively small variations in length, it was suitable for identifying menstrual cycles but not pregnancies or post-partum states, preventing the labeling of such events.

Hidden semi-Markov models (HSMM) allow for non-geometric distributions of state duration, called "state sojourn" in the semi-Markov context. HSMMs can be approximated, often exactly, by HMMs in which HSMM states are divided into chains of sub-states with specific transition probabilities [23]. The HMM approach is especially efficient when the sojourns of each state remain relatively short. However that approach loses its benefits if some states are very long or if one wishes to impose constraints on the type of sojourn distribution. Given the duration of states, such as pregnancy or breast-feeding, and the prior knowledge available on the duration of pregnancies, the HSMM approach is more suited to the task.

While hidden semi-Markov models have been used in a large variety of applications, ranging from biological sequence analyses [24] to modeling financial market variations [25], [26], there are, to our knowledge, no previous implementations that fulfill the requirements of our task. In particular, the 'hsmm' package by Bulla *et al.* [27] did not allow for decoding of sequences with missing observations. The 'mhsmm' package by O'Connell *et al.* [28] allows for missing data-points and for users to define their own functions for various emission distributions. However, four features were critically lacking for our task. First, while it allows for missing time-points, it only enables all variables to be missing at a given time-point. If only one variable is missing, the values of the other variables were not taken into account. In our case, given the sparsity of our dataset, this implied losing over 90% of our data. In addition, this package did not allow users to define state-dependent censoring probabilities. However, users of fertility apps modify their tracking behavior depending on their biological state and their reproductive objectives. We thus wanted a method which took advantage of this "informative missingness". Third, while the 'mhsmm' package allows for multivariate time-series, its current implementation relies on multivariate Gaussian variables. Finally, the 'mhsmm' package does not allow users to define different sojourn distributions for each state. In our case, reproductive states might be best described by different distributions. Consequently, we developed a new package, which addressed all of our task requirements, offers more flexibility in terms of variable distributions, and provides a suite of visualization and interactive labeling tools to facilitate its use.

Our contributions are (a) the adaptation of hidden semi-Markov models to decode censored multivariate time series, (b) the implementation of these changes in a publicly available Rpackage (`HiddenSemiMarkov`), (c) the definition of a HSMM describing the reproductive biology and (d) a hierarchical method relying on HSMMs which accounts for long-term changes in tracking behavior.

We evaluate the performances of our method on a real-world dataset and compare them to those of a HMM and of a HSMM with weak priors. In order to quantify the sensitivity of the decoding accuracy to increased levels of sparsity, we simulated a synthetic dataset with varying amounts of missing data. Finally, we evaluate the ability of our model to learn individuals' cycle characteristics by quantifying the error on cycle length prediction.

Our real-world data is a de-identified dataset provided by the menstrual cycle and fertility tracking app Kindara (see Materials and Methods). This dataset was composed of the self-tracked logs of 64 long-term users of the app. The features reported by users were (1) their bleeding flow (none, spotting, light, medium, heavy), (2) the consistency of their cervical mucus (none, creamy, egg-white, watery, sticky) and the quantity (little, medium, lots) when it was not missing, (3) their body temperature, in Fahrenheit, and whether they marked their temperature measurement as "questionable," which is recommended by the app if the value is oddly low or high or if the user did not sleep enough hours, (4) the results of LH tests (positive or negative) and (5) the results of pregnancy tests (Fig. 1). Each of these features can be reported daily by users. However, users do not report all of these features every day and there is a large variability in tracking frequency between users [3]. Missing data are very frequent. The average tracking frequency is just above 50%, which means that on average users open and log a feature in the app approximately every other day, but it may be as low as 16% or as high as 100% for some users (see Table I). Fig 1(b) provides two examples of time series logged by app users.

Given the generative nature of our model, a synthetic dataset was simulated from our HSMM with various amounts of missing data so that the effect of tracking frequency on accuracy could be evaluated (Methods).

Fig. 1. Data acquisition and time series examples. (a) Snapshot of the tracking screen of the Kindara app. (b) Examples of time series tracked by two users of the app. For all features, the absence of vertical line indicates missing data for that variable. In the bleeding line (top), gray lines indicate 'no bleeding,' orange/red lines mean that bleeding was reported. Darker reds indicate heavier bleeding flow. In the LH and pregnancy test lines (2nd and last lines), red lines indicate negative tests, blue lines positive ones. Temperature is depicted by a gradient ranging from blue for temperatures below the user's median value to red for temperatures above the median value. No mucus (3rd line) is depicted by gray lines, while fertile mucus is indicated in blue, sticky mucus in yellow and creamy one in beige.

TABLE I
STATISTICS ON USERS' TRACKING BEHAVIOR

metric	mean	median	min	percentiles 5th	percentiles 95th	max
tracking frequencies						
Overall	0.55	0.47	0.16	0.20	0.99	1.00
Temperature	0.30	0.14	0.00	0.00	0.92	1.00
Mucus	0.18	0.06	0.00	0.00	0.73	0.99
LH test	0.01	0.00	0.00	0.00	0.06	0.26
Pregnancy test	0.00	0.00	0.00	0.00	0.02	0.04
Longest consecutive						
missing days	183.5	55.0	1	7	712	1386
tracked days	333.9	141.5	7	11	1474	2604

II. MATERIALS AND METHODS

In the first part of this section, we detail the adaptations brought to hidden semi-Markov models for multivariate time series with state– and variable– dependent missingness. The second part of this section details our HSMM of the female reproductive biology. The third one describes our hierarchical approach to account for changes in tracking behavior. Fourth, we describe the datasets used to assess performances, and finally we define the experiments and metrics used to assess the performances of our model.

A. Hidden Semi-Markov Models for Multivariate Time Series With State– and Variable– Dependent Missingness

Our task is to label time series with a sequence of hidden states. In the context of hidden (Semi-)Markov models, two algorithms may be used for this purpose: the Viterbi and the Forward-Backward algorithms. The Viterbi algorithm returns the most likely sequence of hidden states, *i.e.* the sequence of hidden states that maximizes the likelihood of the sequence of observations (see below). The Forward-Backward algorithm returns the probability of each state at each time-point, *i.e.* $P(S_t = j|\mathbf{X})$, where S_t is the state at time-point t, j is one of the J hidden states of the model and $\mathbf{X}$ is the sequence of observations. Efficient versions of these algorithms for hidden semi-Markov models have been proposed by Guédon *et al.* [29] and implemented in `C` by O'Connel and Hojsgaard [28]. This `C` implementation is used in our `R` package, with a minor correction of the backtracking step for the Viterbi algorithm. Below, we introduce the HSMM notation and methods and describe the adaptations introduced to decode censored multi-variate time-series.

1) Notation and Model Parameters: In general, hidden semi-Markov models are defined by the following set of parameters:

J is the number of states, π are the initial probabilities ($\pi_j = P(S_1 = j)$), T are the transition probabilities ($T_{j,k} = P(S_t = j|S_{t-1} = k)$), $\{d_j(u)\}_{j=1...J}$ are the sojourn distributions for each state, *i.e.* the distributions of the time spent in a given state (u is the relative time variable since the last state transition), and $e_{x,j}$ are the emission probabilities for each state, i.e. $P(X = x|S = j)$. This set of parameters is represented by θ. X is a random variable measured at a sequence of time-points and may be discrete, continuous or categorical. It is either observed, taking a value x, or missing, taking the value $\varnothing$. We use the shorthand notation $\mathbf{X}$ for a sequence of observations of length N: $\mathbf{X} = (x_1, x_2, \ldots, x_N)$. A sequence of hidden state is written as $\mathbf{s} = (s_1, s_2, \ldots, s_N)$; s_i or x_i is the shorthand notation for $s_{t=i}$ or $x_{t=i}$.

2) Likelihood of a Sequence of Observation and Hidden State Predictions: The likelihood of a sequence of observations given a sequence of hidden states and the model parameters is given by:

$$P(\mathbf{X}|\mathbf{s};\theta) = \pi_{s_1} d_{s_1}(u_1)\, T_{s_{R-1},s_R} D_{s_R}(u_R) \left(\prod_{r=2}^{R} T_{s_{r-1},s_r}\, d_{s_r}(u_r)\right) \prod_{i=1}^{N} P(x_i|s_i)$$

where π_{s_1} is the probability associated with the first state of the sequence, $d_{s_1}(u_1)$ is the sojourn probability of the first state, with u the relative time spent within a state, r is an index going through the sequence of states (regardless of their duration) while i is an index running along the observation sequence (time-points), s_r is the rth state in the state sequence, T_{s_{r-1},s_r} is the transition probability between the state preceding the rth state and the rth state, $d_{s_r}(u_r)$ is the sojourn probability of the rth state, R is the length of the state sequence and $D_{s_R}(u_R)$ is the "survivor" sojourn probability of the sequence's last state, *i.e.* is it the probability that the state lasts u_R or longer.

3) Fitting Model Parameters to Observations: HSMMs are usually fitted to sequences of observations using an *Expectation-Maximization* (EM) approach which alternates between E-step, in which the hidden states sequence is estimated and the likelihood is computed given the current parameter values, and an M-step, in which the parameters are updated to maximize the likelihood. These two steps are repeated until the gain in likelihood is smaller than a given threshold. The fitted model is the set of parameters maximizing the likelihood of the observed sequences, *i.e.* the maximum likelihood estimator $\hat{\theta}^* = \arg\max_\theta\ P(\mathbf{X}|\theta)$

The Forward-Backward algorithm proposed by Guédon *et al.* [29] is used in the E-step to obtain the probability of each state at each time-point. In the M-step, the emission parameters are updated using these probabilities as weights on the observations.

4) Multivariate Data: In the case of multivariate data, each time-point i is associated to a random vector of length K: $(X_i^1, X_i^2, \ldots, X_i^K)$. In our case, the first variable is bleeding, the second is mucus, the third is temperature, etc. To adapt for multivariate data, we need to specify the *joint emission probabilities* at time-point i, *i.e.* $P(X_i^1 = x_i^1, X_i^2 = x_i^2, \ldots, X_i^K = x_i^K|S_i = j)$ and define how potential within-state dependencies between the variables are accounted for.

Past research in reproductive biology has mostly focused on experimentally measuring marginal probabilities. For example, variations in temperature and in cervical mucus have been described separately and there is no available literature to inform us about potential correlations within a particular hormonal state. We thus assume conditional independence of the variables given the hidden state when initializing the joint emission probabilities. Each variable is specified as a non-parametric distribution or by a distribution family and set of parameters. For example, temperature, a continuous variable, may be specified as a normal distribution. Cervical mucus is a categorical variable and may be described by a non-parametric distribution. LH and pregnancy tests results are binary variables (positive or negative results) and may be described as Bernoulli variables.

These initial specifications are used to list all possible observable combinations of values and initialize the probability associated with each combination as the product of the marginal probabilities assuming independence conditionally on the states. As the model parameters are fitted to sequences of observations, potential within-state dependencies may be learned as the joint emission probabilities are updated without assuming independence. In the online supplementary material (see the *code and data availability* section), we show how a model is able to learn such dependencies when the direction of the correlation between two variables is the only difference between two states. Computationally, within-state dependencies can be learned because continuous variables are discretized into a given number of bins so that all possible combinations of variable values can be stored in a table. The number and/or size of the bins can be specified as one of the model parameter.

5) Missing Data: The Censoring Model: Self-reported health records are subject to a high level of missingness with large inter-subject variations, and the tracking frequency of a user may also change over time. Missing observations may be modeled as a two-step process: first, users must open the app on a given day, and second, they must measure and report a specific variable on that day. Both processes can be modeled as a Bernoulli events with state-dependent probabilities. The probability that a user does not open the app on a given day is p_j. The probability that a user did not report a specific variable k after opening the app is $q_{j,k}$.

Altogether, when a hidden semi-Markov is specified, joint emission probabilities are initialized as:

$$P(X^1, X^2, \ldots, X^K|S = j) = \begin{cases} p_j + (1 - p_j) \prod_{k=1}^{K} q_{j,k} \\ \quad \text{if all variables are missing, } \textit{i.e.}\ \forall_k X^k = \varnothing \\ (1 - p_j) \prod_{k \in M} q_{j,k} \prod_{o \in O} (1 - q_{j,o}) P(X^o = x^o|S = j) \\ \quad \text{otherwise} \end{cases}$$

TABLE II
STATES OF THE HSMM FOR REPRODUCTIVE EVENTS

#	Abbr.	Names
1	**M**	Menses
2	**lE**	Early Follicular (low estradiol)
3	**hE**	Late Follicular (high estradiol)
4	**preO**	Day before ovulation
5	**O**	Ovulation
6	**postO**	Two days following ovulation
7	**Lut**	Luteal phase
8	**Ano**	Anovulatory cycle
9	**AB**	Anovulatory with bleeding
10	**P**	Implantation (Pregnancy)
11	**PL**	Pregnancy with Loss
12	**L**	Loss
13	**lEpL**	Low estradiol phase following a loss
14	**PB1**	Pregnancy with birth (1st trimester)
15	**PB2**	Pregnancy with birth (2nd trimester)
16	**PB3**	Pregnancy with birth (3rd trimester)
17	**B**	Birth
18	**PP**	Postpartum
19	**BF**	Breastfeeding

with M and O being the set of missing/observed variables.

The previous equation reflects that all variables may be missing because a user did not open the app on a given day (with probability p_j) or because the user opened the app but (with probability $(1-p_j)\prod_{k=1}^{K} q_{j,k}$) neither measured nor reported any of the variables. If at least one variable is reported, that implies that the user opened the app on that day (with probability $1-p_j$), that all missing variables were missing with probability $q_{j,k}$ and those not missing were reported with probability $(1-q_{j,k})$ multiplied by their specific emission probability.

These initial probabilities are updated in the M-step of the fitting procedure so that potential dependencies between variables, including missingness dependencies, may be learned from sequences of observations.

B. Generative Models of the Female Reproductive Cycles

We defined the simplest hidden semi-Markov model that would as accurately as possible reflect our current knowledge of the menstrual cycle and pregnancy. States are listed in Table II. Fig.s. 2(a–b) show the model graph and the prior sojourn distributions of most states.

We specified a 19-state model composed of 2 main loops (Fig.. 2(a)). The first loop is a 7-state chain describing the successive phases of the menstrual cycles while the second loop describes the successive events following a conception. The conception loop further splits into two sub-loops: one in the event of a pregnancy loss and one in the event of a birth. The birth branch splits into two scenarios depending on whether or not the mother breastfeed their newborn since breastfeeding typically delays the return of menstrual cycles. In addition to these main loops, two states capture anovulatory phases. The first one corresponds to cycles in which quasi-constant bleeding (light or heavy) is reported and in which no signs of ovulation, such as a positive LH test or a rising temperature, would be reported. The second one corresponds to the scenario in which a low temperature is reported consistently between two bleeding episodes without abnormal bleeding being reported.

All state transitions are uni-directional except for the transition between the 'high estrogen' state and the 'low estrogen' state. This transition is initialized with a low probability and allows the description of cycles typically experienced by users suffering of poly-cystic ovary syndrome (PCOS).

The model parameters (sojourn and emission distributions) were specified to match the observed biological ranges and typical values. Menses last between 2 and 8 days [20]. The early follicular is the most variable phase of the menstrual cycle. Its typical duration is 3 to 8 days but can last longer in individuals with long cycles [30]. We thus specified its sojourn with a long tail. This phase is characterized by low, slowly increasing estradiol levels, medium-high FSH levels, and low progesterone levels. Consequently, cervical mucus, whose production depends on estradiol [31], is rarely observed [3], and temperatures are low as progesterone levels are low [32]. In the late follicular phase, estradiol levels are rising sharply, leading to mucus production, while FSH levels are decreasing. This phase has been reported to last 2-5 days [30]. We defined a pre-ovulatory state (pre-O) with a fixed sojourn of one day, distinct from the late follicular state because the probability of a positive LH test is higher in 'pre-O' since LH starts pulsing in the day leading to ovulation. The ovulation state has a fixed sojourn of one day as ovulation is a brief event and that the temporal resolution of our data is of 1 d. The duration of the luteal phase, which starts after ovulation, is known to vary less inter– and intra–individually than the follicular phase [3], [5], [33]–[36]. In the luteal phase, given elevated progesterone levels, the basal body temperature is higher than in the follicular phase. However, past studies have shown that it takes a few days before the temperature reaches its highest plateau [3]. Thus, we divided the luteal phases into two states. The first one (post-O), of fixed duration (two days), follows the ovulation state. The second one (Luteal), lasts about 11 days with a slight skew for shorter durations.

Although anovulatory cycles are not well described in the literature, owing to the difficulty of assessing the absence of ovulation, we included an 'Ano' state which follows the early follicular phase and is characterized by both frequent observation and low temperatures. We specified it with a mean duration of 15 days and a 6-day variance, based on the results from Malcolm *et al.* [37] and Prior *et al.* [38].

Anovulation may also be occurring in individuals experiencing prolonged periods of light to heavy uterine bleeding. We thus defined the 'Anovulatory with bleeding' state. This state is not very well characterized in duration from the existing literature but has been reported by patients [20] and users of the app. We thus specified a sojourn distribution ranging from height to a hundred days for this state.

When conception happens, the 7-8 days following fertilization (ovulation) are very similar to a luteal phase when no conception happens. However, once the fertilized egg implants, this initiates the production of the HCG hormone, which can be

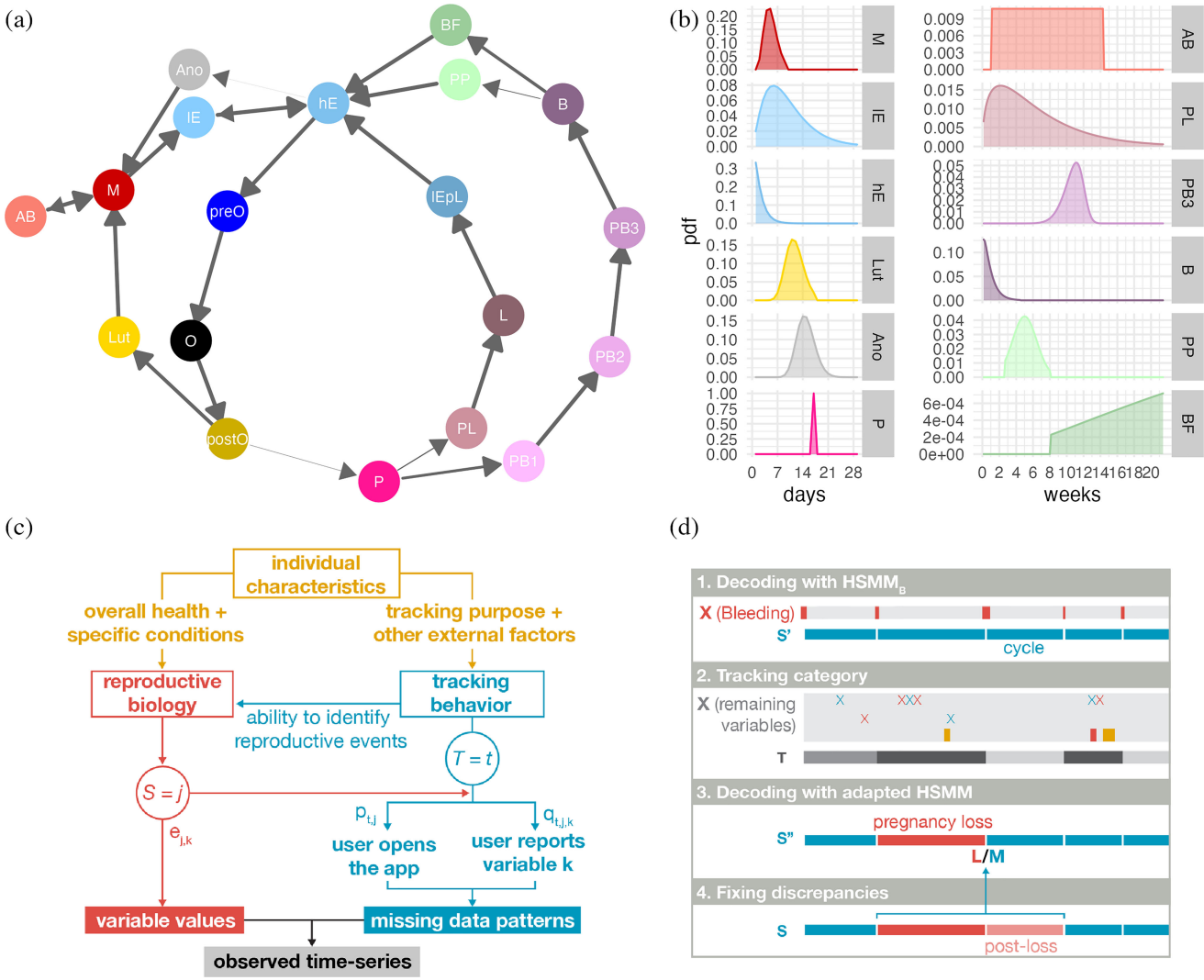

Fig. 2. Modeling the data generation process. (a) Graph of the specified HSMM for modelling reproductive events. Arrows indicate possible transitions, their width is proportional to the transition probability. This graph should be read starting from the red circle ('M' for menses). From 'M,' a first loop matches the 6 states defining ovulatory cycles ('IE,' 'hE,' 'pre-O' are follicular phase states, 'O' stands for ovulation, and 'post-O' and 'Lut' are luteal phase states). After ovulation, a pregnancy may start ('P) and end-up in a loss ('PL'-'L'-'lEpL' loop) or in a birth ('PB1'-'PB2'-'PB3'-'B'-'PP' (post-partum without breast-feeding) or 'BF' (breast-feeding) loops). Finally, two anovulatory states are defined: 'AB' for anovulatory with bleeding and 'Ano' for anovulatory without bleeding. See Methods and Supplementary Material for state definition and descriptions. (b) Prior and initial sojourn distributions for states which do not have a fixed duration. (c) Graph of the generative model assumed to lead to the observed sequences. (d) Schematic illustrating the hierarchical approach to account for long-term changes in tracking behavior.

detected in urine by pregnancy tests. Additional progesterone production leads to an increase or sustained plateau of high temperatures. After implantation, pregnancies may be interrupted (spontaneous or induced pregnancy loss) or continue, leading to a birth. Consequently, we designed our model such that, from an implantation state (P), the model allows two transitions: ones towards the 'PL' state (pregnancy with loss) or one towards the first trimester of a pregnancy without loss (PB1). We fixed the duration of the implantation state to 17 days, which is longer than the longest reported luteal phases. In that state, temperatures are high, positive pregnancy tests are likely, and censoring probabilities are lower than in the subsequent states. Indeed, once a blood test has confirmed the pregnancy, users are less likely to keep tracking their temperature or to report pregnancy test results.

The 'Pregnancy with Loss' state has a highly skewed sojourn distribution as losses occurring early in pregnancy are much more common than late losses [39], [40]. The following state, the loss (L) state, is associated to the moment when the loss occurs. Losses often lead to uterine bleeding for a few days, which the app users may report. After a loss, individuals usually return to ovulatory cycles [41]. However, pregnancy test results may remain positive for a few days after the loss, likely due to the residual presence of HCG hormone in urine. To account for that, we created an additional state, 'lEpL' (for low-Estrogen post-loss) with the same sojourn and emission distributions than

TABLE III
ADAPTATION OF OUR HSMM DEPENDING ON TRACKED VARIABLES. B: BLEEDING, LH: LH TESTS, P: PREGNANCY TESTS, T: TEMPERATURE, M: MUCUS. PARENTHESES INDICATE OPTIONAL TRACKING

Tracked variables	$T_{.,\text{Ano}} = 0$	$d_{\text{hE}} = \delta(2)$	$d_{\text{Lut}} = \delta(11)$	$T_{\text{hE, LE}} = 0$
B	X	X	X	X
B, P	X	X	X	
B, LH, (P)	X	X		
B, T, (LH, P)		X		
B, M, (LH, P)	X			
B, T, M, (LH, P)				

the 'lE' state, except that positive pregnancy tests are more likely in that state.

If there is no loss, the model progresses through the three trimesters of pregnancy. We fixed the duration of the first two trimesters so that the sojourn distribution of the third trimester embeds the whole observed variability in pregnancy duration. Indeed, pregnancies last about 38 weeks but preterm birth rates reach 4-10% depending on countries [42]. Post-term births are less frequent as births are usually induced when past term. Consequently, the sojourn duration of the third trimester of pregnancy is specified as a skewed normal distribution with a heavier tail for shorter duration.

Finally, following birth, the mother may or not breastfeed her newborn child. In the absence of breastfeeding, menses return 6-8 weeks after delivery [43], which means that estradiol levels rise 4-6 weeks after delivery. The duration of the 'post-partum' state (PP, when mothers do not breastfeed) is thus described by a normal distribution of mean 5 weeks and a standard deviation of 10 days. If the mother breastfeeds, this usually delays the return of ovulatory cycles [44]. Given that the breastfeeding duration is highly variable [45], the sojourn for that state is specified as a flat distribution ranging from 7 weeks to over two years.

C. Hierarchical Approach to Adapt for Changes in Tracking Behavior

In principle, the tracking behavior does not affect an individual's biology. However, it affects our ability to detect specific reproductive events (Fig. 2(c)). For example, if a user only tracks their bleeding, it may be impossible to differentiate early pregnancy losses from long cycles or to pinpoint the day of ovulation. To lift these identifiability issues, we adapt our HSMM of reproductive events described in the previous section to the tracking behavior by fixing the sojourn of some states or preventing specific state transitions (Table III).

We proceed in four steps (Fig. 2(d)). First, we decode the time-series of reported bleeding to roughly identify cycles and pregnancies. Menses are ideal sub-sequences boundaries as they are the most likely reproductive event that users would report in a menstrual cycle tracking app. Second, we determine the tracking behavior category of each sub-sequence by examining which variables are reported with sufficient frequency. Third, we decode each sub-sequence of the multivariate time-series with the appropriate model, i.e. with the model for which reported variables within that sub-sequence will allow identifiability. Finally, because the decoding at the first step might have contained mistakes since based on a single variable, we look for discrepancies in predicted states at the transitions between sub-sequences. If any discrepancy is found, we decode the time-series from the last menses preceding the problematic time-points to the next menses following it.

The four ways that our HSMM for reproductive events needs to be adapted are as follow (see summary in Table III). First, if temperature is not reported throughout the cycle, anovulatory cycles cannot be detected from the other variables. The transition to the "Ano" state is thus removed. Second, if mucus is not reported, the sojourn of the "hE" state is fixed to two days since mucus is the only identifying variable of that state. Third, if there are no variable allowing for the identification of ovulation, the sojourn of the luteal phase is fixed. Finally, if only bleeding is reported, there is no information to differentiate long cycles from early pregnancy losses. The transition probability from the "hE" to "lE" state is set to zero.

D. Datasets

1) Real-World Dataset: A de-identified dataset was provided by Prima-Temp (Boulder, Colorado), the company owning the menstrual cycle and fertility tracking app *Kindara*. This study was exempted by the Stanford IRB given the de-identified nature of the dataset. The dataset was lightly pre-processed before being labeled using our hidden semi-Markov models we define in the next section. Temperature reports were transformed into temperature differences from their median value because the inter-individual variations in temperature are larger than the within-cycle variations. Specifically, for each user, temperatures marked by the app users as questionable were removed from the time series and their median temperature, computed on the remaining values, was subtracted from their reported values. Additionally, the 13 different possible mucus readings were grouped into 5 categories (none, creamy, fertile, very fertile, sticky) because mucus readings are subjective and the literature is too sparse to allow the distinction between these categories (see online supplementary material). Finally, because bleeding is the most remarkable body-sign associated with the menstrual cycle, if bleeding was missing on days when other features were reported, it was assumed to have the value 'none'.

2) Synthetic Dataset: To evaluate the robustness of our framework to varying amount of missing data, we generated N_u synthetic time series mimicking actual observations by reverting our decoding approach. We first specified the censoring probabilities p_j and $q_{j,k}$ such that they matched those of diligent app users. These probabilities were then scaled up or down to create five levels of tracking assiduity (Fig. 4, online supplementary material). In addition, slight modifications are brought to the HSMM to reflect person-specific characteristics such as a typical luteal phase duration, a specific temperature shift, etc. Once time-series are simulated, the set of reported variables within each cycle is randomly selected to reflect changes in tracking behaviors.

The synthetic dataset is decoded using our hierarchical approach relying on HSMM (h-HSMM, see below). Performance

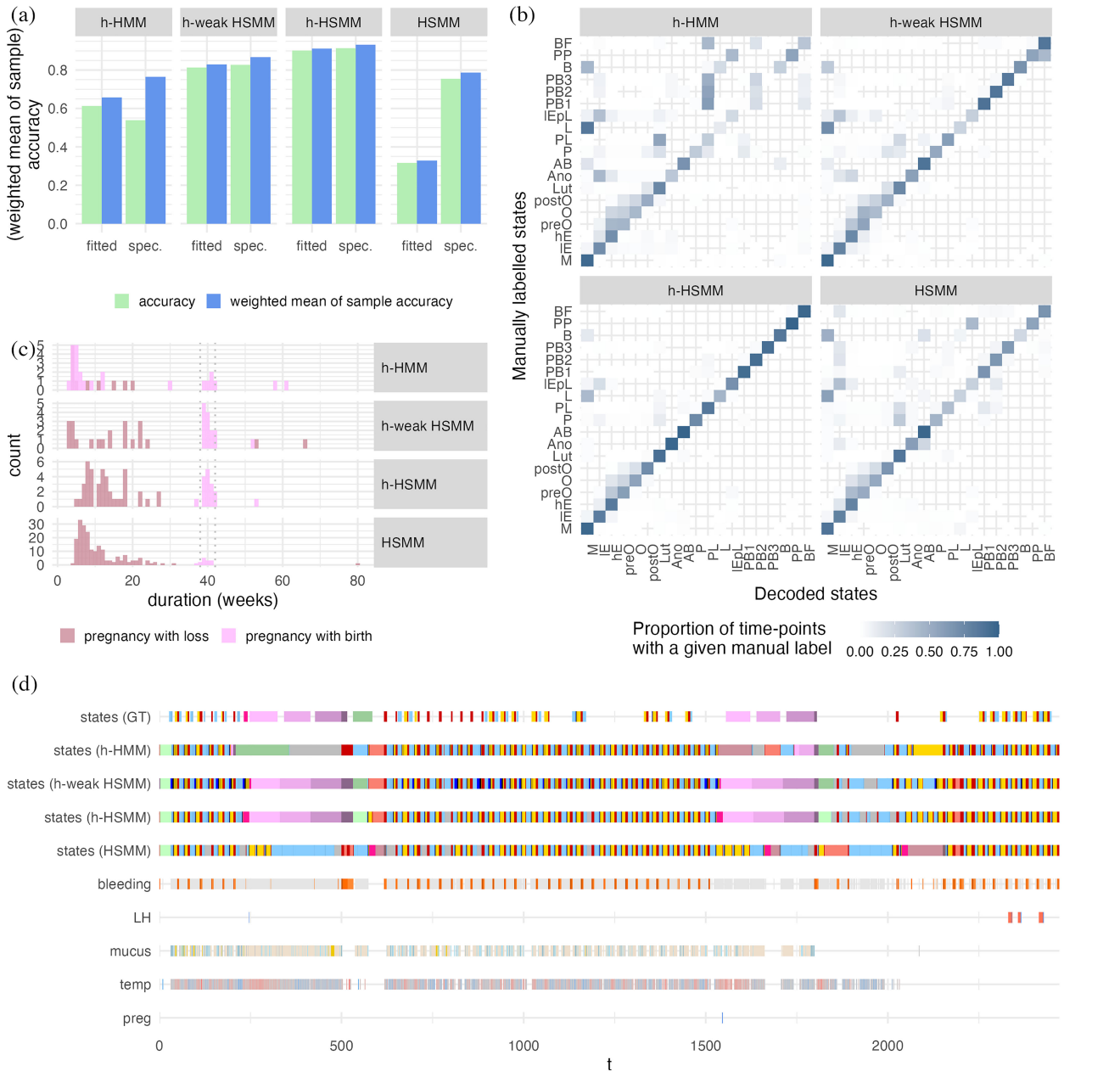

Fig. 3. Performances on the Kindara dataset. (a) Accuracy (light green) and weighted mean of sample accuracy (blue) for our proposed approach, the hierarchical HSMMs (h-HSMM) with several baseline methods. (b) Confusion matrix normalized by the number of time-point in each ground-truth state (the sum across rows is equal to 1). (c) Distribution of the duration of pregnancies with loss (dark purple) and with birth (light pink) as detected by our method and the three baseline methods. (d) Example of a time series from a Kindara users with the manual labels (first row), and the predicted labels by the different methods (second to fifth rows).

metrics (see below) are computed as a function of the tracking frequency and the set of reported variables.

E. Performance Metrics

1) Labeling Performance Evaluation: We evaluate the performances of our framework by measuring its ability to recover simulated or manually labeled ground-truth. We compute the accuracy A defined as $A = \frac{1}{N}\sum_{i=1}^{N}\delta(\hat{\mathbf{s}}_i - \mathbf{s}_i^*)$ and the state-specific accuracy which is defined as $A_j = \frac{1}{N_j}\sum_{i=1}^{N_j}\delta(\hat{\mathbf{s}}_i - \mathbf{s}_i^*)\,\delta(j - \mathbf{s}_i^*)$.

In order to evaluate whether the model provides a higher uncertainty on time-points with labeling mistake, we compute the weighted mean of sample accuracy as $A_w = (\sum_{i=1}^{N} w_i\delta(\hat{\mathbf{s}}_i - \mathbf{s}_i^*))/(\sum_{i=1}^{N} w_i)$ where the sample weights w_i are the posterior state probabilities for the most likely state, *i.e.*

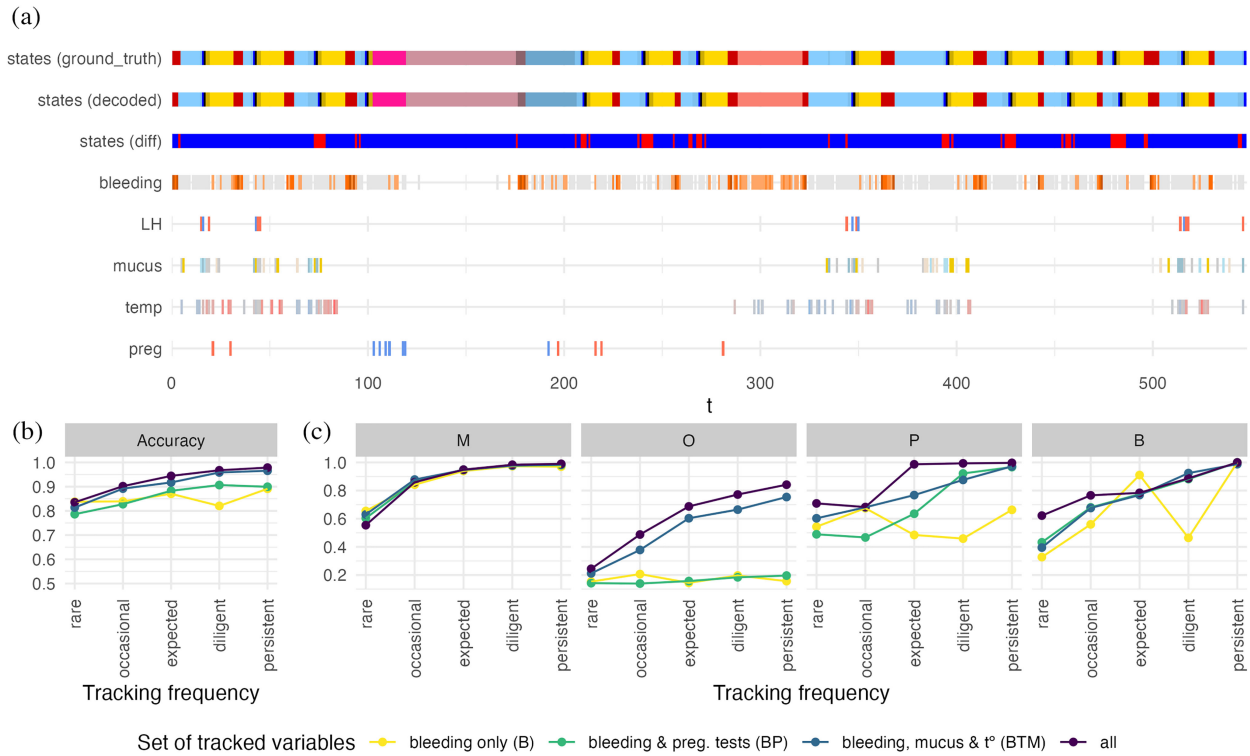

Fig. 4. Performance on synthetic dataset. (a) Example of a simulated sequence of states (ground truth, first row) and observations (last five rows) and the sequence of hidden states predicted by our method (second row). The third row shows the difference between the two rows (blue for agreement, red for labeling errors). (b) Accuracy for different tracking frequency (x-axis) and for different sets of tracked variables (colors). (c) State-specific accuracy for the periods (menses - M), ovulation (O), implantation (pregnancies - P) and births (B).

$w_i = P(S(t=i) = \hat{\mathbf{s}}_i|\mathbf{X};\theta)$. Samples in which states are predicted with a high probability have more weight than samples in which the model predicted several states with a low probability.

We compare the performances of our approach (h-HSMM) with those of three baseline methods: h-HMM, h-weak HSMM and HSMM. The first one (h-HMM) combines our hierarchical approach (h-) with a HMM for the modeling of the reproductive events. The HMM has the same number of state as our HSMM and the same emission probabilities. Transition probabilities are specified such that the associated geometric distribution best fit the HSMM's sojourns. The second baseline method (h-weak HSMM) is similar to the proposed approach but the sojourns are weakly specified; they have a much broader distribution than the those proposed in our h-HSMM. Finally, to evaluate the benefits of the hierarchical approach, we compare the performances of our h-HSMM with those of the HSMM alone. We also evaluate the performances of our method and baseline methods by computing the predicted duration of pregnancies and comparing them to empirical distributions.

2) Predicting the Next Period Date in Ovulatory Cycles: In addition to labeling user's time series, our model can also be used to predict the date of the next period. While most individuals have regular cycles (each cycle has approximately the same duration), many individuals have irregular cycles and it may be difficult to predict the date of their next period.

To evaluate the ability of our method to learn individual-specific characteristics, such as their typical temperature in the follicular or luteal phase or the length of their luteal phase, we selected users in our dataset with irregular cycles and compared the predictions from our method with a baseline method which uses the average cycle length of users.

Specifically, we selected stretches of five consecutive ovulatory cycles without pregnancy, fitted our reproductive event model on the first four cycles and made the prediction for the length of the fifth cycles. To evaluate if the prediction improves as the fifth cycle progresses in time, we perform the prediction from each day since the beginning of the cycle. We decode the fifth cycle up to a given day, detect the last state transition and use the fitted sojourn distributions of the remaining states to predict the total cycle length.

Given that the most variable phase of the cycle is the phase preceding ovulation, we expected our prediction to improve once ovulation is detected. We report the mean square error (MSE) between the predicted and the actual length of the fifth cycle and compare it with the MSE when using the average cycle length of the four previous cycle to predict the length of the fifth cycle.

III. Results

A. Labeling Performance on the Kindara Dataset

To quantify the performances on our dataset, we used the interactive app embedded in our `HiddenSemiMarkov`package to manually label about 11% of our dataset. These labels were independently validated by a fertility awareness methods expert

(see Acknowledgments) and are shown for the full dataset in the online supplementary material. Fig. 3(d) provides an example of a real-world labelled time series.

Overall, the proposed method (h-HSMM) reaches higher accuracy and weighted accuracy than alternative methods (Fig. 3(a)). Semi-Markov models perform better than the HMM, demonstrating the advantage of using non-geometric distributions. Strong priors, i.e. priors closely following the empirical distributions of biological state duration also contributes to better performances. Finally, our results highlight the benefit of the hierarchical approach to account for user's tracking habits and long-term changes in tracking behavior. The state-specific accuracy is also better with the h-HSMM compared to other methods (Fig. 3(b)). Specifically, the semi-Markov property allows a more accurate detection of pregnancies and following events. In particular, Fig. 3(c) (and the example in Fig. 3(d)) shows that the duration of pregnancies detected by the h-HMM are outside biological ranges.

The weighted mean of sample accuracy, *i.e.* the accuracy weighted by the uncertainty on the labels at each time-point (see Methods), is higher than the accuracy (Fig. 3(a)). This indicates that, as desired, uncertainty is higher on labels that differ from the ground-truth. In other words, our method is able to warn against potential labeling mistakes.

One interesting observation is that the accuracies are higher for the specified models than for the fitted models (Fig. 3(a)). When examining the decoded sequences, the differences appear to originate from sequences with pregnancies during which users logged few features. For these sequences, only biologically realistic sojourn distribution for these states allows to differentiate between pregnancies with births or with losses.

B. Labeling Performance on Synthetic Data

Our results on a synthetic dataset (Fig. 4) show that our method is able to recover the ground truth with an accuracy of 98% when variables are always reported (persistent tracking, no missing data). This provides an approximate upper-bound on our ability to decode real-world time series.

As expected, we observe a higher accuracy when variables are reported more frequently (less missing data, see tracking categories on the x-axis of Fig.. 4(b–c)) and when more variables are reported, *e.g.* tests results are reported in addition to bleeding (see colored lines in Fig. 4(b) and Methods for the definition of tracking categories and the specification of missing patterns). With time series mimicking the expected tracking behavior of a user whose purpose is to identify their fertile window and pregnancies early on, the accuracy is 92%. The accuracy is of 89% when the tracking behavior is "occasional," i.e. with an average tracking frequency of about 10% (Fig. 4(b) and online supplementary material).

States recovered with the highest accuracy are the menses and pregnancy states (Fig. 4(c), online supplementary material). The states surrounding ovulation suffer the most from a low tracking frequency; without a high tracking frequency, it is impossible to pin-point the day of ovulation. A low accuracy is expected for these states when tracking frequency is low.

TABLE IV
MSE ON THE PREDICTED CYCLE LENGTH

Prediction day	MSE Baseline	MSE HSMM
At cycle start	17.56	16.99
After ovulation (10 days before next cycle)	17.56	6.10

C. Predicting the Next Period

Table IV provides the mean square error (MSE) on the cycle length prediction for the baseline method (*i.e.* average cycle length of the past four cycles) and for our method, using a HSMM fitted to the user's past four cycles data. The table shows the MSE for predictions done at two different moment of the on-going cycle. The first row provides the MSE when the prediction is made on the first cycle day. The second row provides the MSE when the prediction is made after ovulation, 10 days before the next period. While both method perform similarly at the beginning of the cycle, our method is able to detect ovulation and learn the typical luteal phase duration for that user, providing a much more precise estimate (MSE is 2.88 times smaller) as one progresses through the cycle.

IV. DISCUSSION

Unsupervised labeling of self-reported health records with biologically-relevant states is a challenging, multivariate problem given the high frequency of missing data and the changes in tracking behavior. Here, we presented a hierarchical generative method based on hidden semi-Markov models. Our results on synthetic data and real-world data, here self-reported fertility body-signs, show accurate recovery of the hidden states sequence. This framework returns the likelihood at each time-point of this most likely sequence in addition to the most likely sequence of hidden states. Our results show that the decoding accuracy is higher when the likelihood is high which implies that our model is able to adequately quantify uncertainty. In contrast, most medical or psychological studies currently use methods which are unable to quantify the uncertainty or the likelihood of their estimates, such as manual labeling or deterministic rules, to identify the timing of reproductive events such as ovulation day. Because our method, *i.e.* hierachical HSMMs, is able to capture biological states of specific duration by adequate initialization of the sojourn distribution and adapt to long-term changes in tracking behavior, its accuracy is much higher (93%) than that of a hierarchical HMM (61%) or of a single HSMM (75%). In addition, our method predicts cycle lengths of ongoing cycles with a 2.88 times lower error on average than the baseline method for users with irregular cycles.

Beyond modeling reproductive events, our adaption of hidden semi-Markov models allows (i) for missingness patterns that may differ between variables, (ii) for censoring probabilities that may differ between states, (iii) for variables of different types (continuous, discrete, categorical), and (iv) for continuous variables specified from different marginal distributions (*e.g.* poisson and gaussian). We have implemented our method in a publicly available `R` package (`HiddenSemiMarkov`) which

builds upon the existing implementation of the Viterbi and Forward-Backward algorithms from the `mhsmm` package [28].

The proposed model is ideal for decoding any self-reported time series such as physical activity patterns, or time series of incomplete diagnosis data. As an example from the current pandemic, our hidden semi-Markov model could be fitted to datasets of covid-19 test results and reported symptoms to identify the different phases of infection from "uninfected" to "recovered" over "incubating" and "infectious". Another example could consist in inferring someone's mental health state over time from various self-reported symptoms in a tracking app or from the tone of their messages.

In addition to time series labeling, HSMMs are also used to detect outliers in time series, *i.e.* values which may be in the expected variable range of value but that would be unexpected at that particular moment in the time series. Consequently, our implementation of HSMM could help detect abnormal missing data and the failure of a measurement/reporting process.

One limitation of our framework is that within-state dependencies between variables cannot be specified when initializing the model. However, our simulations show that these dependencies are successfully learned when the model is fitted to a sequence of observations where these correlations are present (see online supplementary material, link in the *code and data availability* section). In addition to expected functions of a hidden semi-Markov package, *i.e.* functions to specify and fit censored hidden semi-Markov models, simulate time series, and predict sequences of hidden states using the Viterbi or the Forward-Backward algorithm, we also provide several visualization functions for inspecting labeled time series and model parameters. Finally, we implemented an interactive app which can be used to manually label time series and/or confirm predicted labels. This interactive app can be used to create some ground-truth for time series or to use an *interactive boosting* approach to accelerate the fitting process.

This package, in combination with the proposed reproductive model presented here, provide ready-to-use off-the-shelf tools that any scientist interested in studying health and biological variations associated with the menstrual cycle can use. For example, the labeling method presented here can be used to label large retrospective dataset from menstrual cycle tracking app and evaluate the changes in reported symptoms at specific phases of the cycle before and after pregnancies. And, while users in our dataset were naturally cycling, the proposed reproductive model could be extended to allow for the detection of birth-control changes. Therefore, reported symptoms could be compared before and after birth control transitions. Our model will also facilitate the study of associations between the menstrual cycle and the course of chronic conditions [46], [47]. Indeed, several studies have already shown that patients which chronic conditions such as inflammatory bowel disease [48], asthma [49] or systemic lupus erythematosus [50] report different level of pains or symptoms at different phases of the menstrual cycle. Our framework could thus encourage medical researchers to partner with a tracking app or include a few questions related to participants' fertility, such as contraceptive use, and their menstrual cycle, such as daily report of their bleeding. This would ensure a more comprehensive understanding of the effect of sex as a biological variable on the course of chronic diseases, the efficacy of treatments or in epidemiological studies.

Altogether, this study has demonstrated the accuracy of a hierarchical hidden semi-Markov models for labeling multivariate time series with many missing data-points. Our statistical model is especially suited for applications in which the hidden states may impact the frequency of missing data.

Acknowledgment

The authors thank the fertility tracking app Kindara and their users. They also warmly thank Valentina Salonna, fertility awareness counselor certified by the Symptotherm foundation (Switzerland), for validation of the manual labels on the Kindara data. **Competing interests**: none. **Code and data availability**: Code for the experiments presented here and for additional experiments referred to in the manuscript, additional implementation details, justifications for the reproductive events models and synthetic dataset can be found at[1]. Kindara's users data are not publicly available. The `HiddenSemiMarkov` package can be found at[2].

References

[1] E. J. Topol, "A decade of digital medicine innovation," *Sci. Transl. Med.*, vol. 11, no. 498, pp. 1–4, 2019.
[2] S. Fox and M. Duggan, "Tracking for health," *Pew Res. Center*, 2013.
[3] L. Symul, K. Wac, P. Hillard, and M. Salathé, "Assessment of menstrual health status and evolution through mobile apps for fertility awareness," *npj Digit. Med.*, vol. 2, no. 64, 2019.
[4] J. R. Bull, S. P. Rowland, E. B. Scherwitzl, R. Scherwitzl, K. G. Danielsson, and J. Harper, "Real-world menstrual cycle characteristics of more than 600,000 menstrual cycles," *NPJ Digit. Med.*, vol. 2, no. 83, 2019.
[5] L. Faust *et al.*, "Findings from a mobile application-based cohort are consistent with established knowledge of the menstrual cycle, fertile window, and conception," *Fertility Sterility*, vol. 112, no. 3, pp. 450–457, 2019.
[6] A. Alvergne, M. V. Wheeler, and V. H. Tabor, "Do sexually transmitted infections exacerbate negative premenstrual symptoms? Insights from digital health," in *Proc. Evol., Med. Public Health*, no. 1, pp. 138–150, 2018.
[7] K. Li *et al.*, "Characterizing physiological and symptomatic variation in menstrual cycles using self-tracked mobile health data," *NPJ Digital Med.*, vol. 3, no. 1, pp. 1–13, 2019.
[8] L. Symul, P. Hsieh, A. Shea, D. J. Skene, S. Holmes, and M. Martinez, "Unmasking seasonal cycles in human fertility: How holiday sex and fertility cycles shape birth seasonality," *MedRxiv*, 2020.
[9] T. A. Eisenlohr-Moul *et al.*, "Are there temporal subtypes of premenstrual dysphoric disorder?: Using group-based trajectory modeling to identify individual differences in symptom change," *Psychol. Med.*, vol. 50, no. 6, pp. 964–972, 2019.
[10] E. Pierson, T. Althoff, D. Thomas, P. Hillard, and J. Leskovec, "The menstrual cycle is a primary contributor to cyclic variation in women ' s mood, behavior, and vital signs," *BioRxiv*, vol. 5, pp. 716–725, 2019.
[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, 1997.
[12] B. Liu *et al.*, "Predicting pregnancy using large-scale data from a women's health tracking mobile application," *The Web Conf. - Proc. World Wide Web Conf.*, WWW 2019, 2019.
[13] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Conf. Neural Inf. Process. Syst.*, 2017.

[1][Online]. Available: https://github.com/lasy/semiM-Public-Repo
[2][Online]. Available: https://lasy.github.io/HiddenSemiMarkov

[14] E. L. Billings, J. B. Brown, J. J. Billings, and H. G. Burger, "Symptoms and hormonal changes accompanying ovulation," *Lancet*, vol. 299, no. 7745, pp. 282–284, 1972.

[15] K. S. Moghissi, F. N. Syner, and T. N. Evans, "A composite picture of the menstrual cycle," *Amer. J. Obstet. Gynecol.*, vol. 114, no. 3, pp. 405–418, 1972.

[16] J. L. Bigelow, D. B. Dunson, J. B. Stanford, R. Ecochard, C. Gnoth, and B. Colombo, "Mucus observations in the fertile window: A better predictor of conception than timing of intercourse," *Hum. Reproduction*, vol. 19, no. 4, pp. 889–892, 2004.

[17] C. L. Buxton and W. B. Atkinson, "Hormonal factors involved in the regulation of basal body temperature during the menstrual cycle and pregnancy," *J. Clin. Endocrinol. Metab.*, vol. 8, no. 7, pp. 544–549, 1948.

[18] C. Pauerstein, C. Eddy, H. Croxatto, R. Hess, T. Siler-Khodr, and H. Croxatto, "Temporal relationships of estrogen, progesterone, and luteinizing hormone levels to ovulation in women and infrahuman primates," *Amer. J. Obstet. Gynecol.*, vol. 130, no. 8, pp. 876–886, 1978.

[19] H.-W. Su, Y.-C. Yi, T.-Y. Wei, T.-C. Chang, and C.-M. Cheng, "Detection of ovulation, a review of currently available methods," *Bioeng. Transl. Med.*, vol. 2, no. 3, pp. 238–246, 2017.

[20] I. S. Fraser, H. O. Critchley, M. Broder, and M. G. Munro, "The FIGO recommendations on terminologies and definitions for normal and abnormal uterine bleeding," *Seminars Reprod. Med.*, vol. 29, no. 5, pp. 383–390, 2011.

[21] E. W. Harville, A. J. Wilcox, D. D. Baird, and C. R. Weinberg, "Vaginal bleeding in very early pregnancy," *Hum. Reproduction*, vol. 18, no. 9, pp. 1944–1947, 2003.

[22] E. Pierson, T. Althoff, and J. Leskovec, "Modeling individual cyclic variation in human behavior," in *Proc. Web Conf.*, 2018.

[23] W. Zucchini, I. L. MacDonald, and R. Langrock, *Hidden Markov Models for Time Series: An Introduction Using R.* Boca Raton, FL, USA: CRC Press, ch. 12, 2017, pp. 165–185.

[24] S. Z. Yu, "Hidden semi-Markov models," *Artif. Intell.*, vol. 174, no. 2, pp. 215–243, 2010.

[25] J. Bulla and I. Bulla, "Stylized facts of financial time series and hidden semi-Markov models," *Comput. Statist. Data Anal.*, vol. 51, no. 4, pp. 2192–2209, 2006.

[26] G. D'Amico, G. Di Biase, J. Janssen, and R. Manca, "Semi-Markov backward credit risk migration models compared with Markov models," in *Proc. 3rd Int. Conf. Appl. Math., Simul., Modelling,* 2009, pp. 112–116.

[27] J. Bulla, I. Bulla, and O. Nenadić, "HSMM - AN R package for analyzing hidden semi-Markov models," *Comput. Statist. Data Anal.*, vol. 54, no. 3, pp. 611–619, 2010.

[28] J. O'Connell and S. Hojsgaard, "Hidden semi Markov models for multiple observation sequences: The MHSMM package for R," *J. Stat. Softw.*, vol. 39, no. 4, pp. 1–22, 2011.

[29] Y. Guédon, "Estimating hidden semi-Markov chains from discrete sequences," *J. Comput. Graphical Statist.*, vol. 12, no. 3, pp. 604–639, 2003.

[30] C. J. Munro, G. H. Stabenfeldt, J. R. Cragun, L. A. Addiego, J. W. Overstreet, and B. L. Lasley, "Relationship of serum estradiol and progesterone concentrations to the excretion profiles of their major urinary metabolites as measured by enzyme immunoassay and radioimmunoassay," *Clin. Chem.*, vol. 37, no. 6, pp. 838–844, 1991.

[31] K. S. Moghissi, "Cyclic changes of cervical mucus in normal and progestin-treated women.," *Fertility Sterility*, vol. 17, no. 5, pp. 663–675, 1966.

[32] N. Charkoudian and N. Stachenfeld, "Sex hormone effects on autonomic mechanisms of thermoregulation in humans," *Autonomic Neurosc.,: Basic Clin.*, vol. 196, pp. 75–80, 2016.

[33] S. D. Harlow and S. A. Ephross, "Epidemiology of menstruation and its relevance to women's health," *Public Health*, vol. 17, no. 2, pp. 265–286, 1995.

[34] L. A. Cole, D. G. Ladner, and F. W. Byrn, "The normal variabilities of the menstrual cycle," *Fertility Sterility*, vol. 91, no. 2, pp. 522–527, 2009.

[35] E. A. Lenton, B. M. Landgren, L. Sexton, and R. Harper, "Normal variation in the length of the follicular phase of the menstrual cycle: Effect of chronological age," *Brit. J. Obstet. Gynaecol.*, vol. 91, no. 7, pp. 681–684, Jul. 1984.

[36] E. A. Lenton, B. M. Landgren, and L. Sexton, "Normal variation in the length of the luteal phase of the menstrual cycle: Identification of the short luteal phase," *Brit. J. Obstet. Gynaecol.*, vol. 91, no. 7, pp. 685–689, Jul. 1984.

[37] C. E. Malcolm and D. C. Cumming, "Does anovulation exist in eumenorrheic women?," *Obstet. Gynecol.*, vol. 102, no. 2, pp. 317–318, 2003.

[38] J. C. Prior, M. Naess, A. Langhammer, and S. Forsmo, "Ovulation prevalence in women with spontaneous normal-length menstrual cycles—A population-based cohort from HUNT3, Norway," *PLoS ONE*, vol. 10, no. 8, 2015.

[39] A. J. Wilcox, D. D. Baird, and C. R. Weinberg, "Time of implantation of the conceptus and loss of pregnancy," *New England J. Med.*, vol. 340, no. 23, pp. 1796–1799, 1999.

[40] L. M. Rossen, K. A. Ahrens, and A. M. Branum, "Trends in risk of pregnancy loss among U.S. women, 1990-2011," *Paediatric Perinatal Epidemiol.*, vol. 32, no. 1, pp. 19–29, 2018.

[41] *ACOG*, "FAQ: Early pregnancy loss".

[42] M. Delnord and J. Zeitlin, "Epidemiology of late preterm and early term births—An international perspective," *Seminars Fetal Neonatal Med.*, vol. 24, no. 1, pp. 3–10, 2019.

[43] E. Jackson and A. Glasier, "Return of ovulation and menses in postpartum nonlactating women," *Obstet. Gynecol.*, vol. 117, no. 3, pp. 657–662, 2011.

[44] P. T. Ellison, "Breastfeeding, fertility, and maternal condition," *Breastfeed.*, 1995, ch. 11.

[45] L. Adair, B. Popkin, and D. Guilkey, "The duration of breast-feeding: How is it affected by biological, sociodemographic, health sector, and food industry factors?," *Demography*, vol. 30, no. 1, pp. 63–80, 1993.

[46] S. Oertelt-Prigione, "Immunology and the menstrual cycle," *Autoimmunity Rev.*, vol. 11, no. 6–7, pp. A486–A492, 2012.

[47] A. M. Case and R. L. Reid, "Effects of the menstrual cycle on bipolar disorder," *Bipolar Disord.*, vol. 158, pp. 1405–1412, 2014.

[48] S. Bharadwaj, G. Kulkarni, and B. Shen, "Menstrual cycle, sex hormones in female inflammatory bowel disease patients with and without surgery," *J. Dig. Dis.*, vol. 16, no. 5, pp. 245–255, 2015.

[49] E. Ridolo, C. Incorvaia, I. Martignago, M. Caminati, G. W. Canonica, and G. Senna, "Sex in respiratory and skin allergies," *Clin. Rev. Allergy Immunol.*, vol. 56, no. 3, pp. 322–332, 2019.

[50] K. Colangelo, S. Haig, A. Bonner, C. Zelenietz, and J. Pope, "Self-reported flaring varies during the menstrual cycle in systemic lupus erythematosus compared with rheumatoid arthritis and fibromyalgia," *Rheumatol.*, vol. 50, no. 4, pp. 703–708, 2011.