A Markov model for inferring event types on diabetes patients data

Citation for published version (APA):

Document license:
CC BY

DOI:
10.1016/j.health.2022.100024

Document status and date:
Published: 01/11/2022

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
A Markov model for inferring event types on diabetes patients data

Danilo F. de Carvalho a,A, Uzay Kaymak a, Pieter Van Gorp b, Natal van Riel c

a Jheronimus Academy of Data Science, Eindhoven University of Technology, 's-Hertogenbosch, The Netherlands
b Industrial Engineering & Innovation Sciences, Eindhoven University of Technology, Eindhoven, The Netherlands
c Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

A B S T R A C T

Gathering diabetes-related data requires effort from the patient side to log specific events throughout experiments. This is an error-prone task that the patients usually handle by following a prescribed protocol. However, patients often do not follow the protocol, causing missing or imperfect data. This study investigates the possibility of generating Markov models from existing/logged data and using them for imputation. The models are used to infer information related to missing events (types/activities) in data recorded by diabetes patients, allowing for improvements in the quality and continuity of such data. Our results indicate that such an approach can help improve the quality of the data collected.

1. Introduction

Collection of relevant data is an important activity in different domains, including healthcare and, in particular, diabetes management. In this specific domain, data can be continuous and passively acquired through sensors and smart devices, but also interactively relying on actions from the contributor/patient [1–3].

Data gathering in diabetes related research receive considerable amount of attention as the use of devices containing sensors – and acquiring data – can be considered a common aspect of diabetic patients nowadays. Several takes on the development of algorithms/techniques that aim at modeling through data require as input data acquired from patients [4–7]. Thus, datasets containing such data are of fundamental importance for advancing the state of the art regarding diabetes research. During the modeling, they can be used for learning and validation, and are not limited to provide data coming from a controlled environment (e.g., specialized care during hospitalization). They can also contain data collected during patients’ daily routine by making use of wearables and leverage from the tracked/logged specific events data [8–10].

When narrowed to daily actions, patient behavior can be modeled through their own daily events (e.g., sleep, meals, and exercises), but the amount and quality of such data is critical. As a fair amount of these events are logged by the patients themselves, the data collection is error-prone since patients must log specific free-living events by interacting with devices when the events happen, or at specific times of their days. Hence, it is reasonable to assume that they may forget, resulting in incomplete logs, or even provide data at the correct time but in the wrong way. Hence, there is no guarantee that the patients will provide enough and good quality information about the occurred events.

The work we present in this paper focuses on the idea of improving the quality of data collected by patients with diabetes using wearables under free living conditions. We propose a method towards mitigating the lack of daily routine information, i.e., missing logged daily events. This is a twofold problem: on one hand it is necessary to detect that an event is missing, and on the other hand we should be able to infer which activity (event type) took place. As the problems are not easily solved when approached simultaneously, this paper makes a first step by proposing a solution for the second part of the problem: the activity inference. In particular, the paper focuses on modeling the activities performed by patients by using previously recorded information. For each patient, Markov models [11] are created able to cope with the (peculiar) non-deterministic behavior of the respective patient. The models come as a solution for the activity discovery problem. Moreover, we adapted them to not only rely on the events sequence, but also on observable factors such as blood glucose level, and time intervals associated to each event.

There is no strict focus on one specific activity. Based on historical events data, we develop Markov models able to infer activities in a patient’s day in order to identify the activities that are likely to happen in specific slots of the day. We assume that it is known that there is a missing event in the time interval between two other logged events. However, the performed activity is not known. Our method aims at providing a reliable way of filling in gaps, and/or adjusting erroneous entries in a sequence of events. With this, a step towards improving quality of data through imputation is made.
The outline of the paper is as follows. In Section 2 we give a brief summary of previous relevant works, while in Section 3 we present important preliminary concepts and aspects. The focus of Section 4 is to provide details regarding the methods used by the proposed approach. The evaluation is then described in Section 6. In Section 7, we discuss relevant facets and also limitations encountered. Finally, in Section 8, we conclude the paper and future work possibilities are presented.

2. Related work

Self-reported data have a tendency of carrying errors and low quality information, and data logged by diabetes patients bring this same issue. For instance, meal related information is critical in the diabetes field, and hence attempts on improving the quality of the associated reported events were done. As patients incorrectly informed the amount of carbohydrates for each meal, Rhyner et al. [12] developed a prototype to better estimate the carbohydrate content of meals. They also discussed its usability as patients could avoid the estimation step, not reducing the amount (and quality) of the information generated.

Another work by Zheng et al. [13], with the purpose of reminding patients to take bolus insulin after meals, developed a meal detection technique using Continuous Glucose Monitoring (CGM) data as input. However, this approach can also serve to identify meals that were not logged by patients during experiments performed in the diabetes domain. Finally, Berry et al. [14] evaluated the influence of meal related information is critical in the healthcare and diabetes domain [19–22]. Approaches such as these, can rely on stochastic models based on Markov premises and definitions [16].

Markov models are also able to fit non-deterministic behaviors due to their probabilistic nature. This aspect allows such models to be applied on detecting activity patterns in daily living conditions [17,18], as well as in the healthcare and diabetes domain [19–22].

Our approach extends the idea of identifying activity patterns on event logs by adding external observable information related to daily living conditions for diabetes patients. In this context, this factor allows for a contextualized inference of the activities.

3. Markov models

Due to the natural non-deterministic behaviors of human beings, probabilistic models can serve as an adequate framework for modeling such behaviors. Many researches in the literature focus on sequences (e.g., formed by words, phrases, and system or sensor states) to either detect patterns, infer, impute, or predict a value within the sequence [23–26]. However, they tend to only make use of the values that compose the sequence itself. In order to infer an activity in a sequence, the presented work proposes a Markov model-based inference step that considers the context around such activity. This context includes the neighbor activities in the sequence, as well as external observable factors. As a result, by translating the context into hidden states and emitted observations, the proposed approach permits us to use both layers together as causal factors in the same probabilistic model. In addition, other observable factors could be added to such context, thus giving opportunity for expanded models.

This section starts by introducing Markov chains, which consider only observable states, evolving to hidden Markov models, that make use of external factors as additional observable variables. All concepts and aspects here provided form the basis for the complete understanding of the proposed work.

3.1. Markov chain

Consider \(\{X_m, m = 0, 1, 2, \ldots\} \) a discrete-time random process. In a Markov chain \(X_{m+1}\) is conditioned to the current state \(X_m\), but it does not depend on any of the other states from previous cycles \(X_0, X_1, \ldots, X_{m-1}\). In other words, the process has no memory, the process behavior that follows any cycle depends only on its current state, and this constraint is known as the Markovian assumption or property [11].

Let \(S = \{s_1, s_2, \ldots, s_J\}\) be a finite set of possible states a probabilistic system can be in, and state \(X_m = s_j\) occurs at discrete time \(t\), and \(X_{m+1} = s_j\) at \(t+1\). In a Markov chain, \(a_{ij} = P(X_{m+1} = s_j \mid X_m = s_i)\) is the probability of going from state \(s_i\) to \(s_j\), and it is assumed that this transition does not depend on time. By generalizing the use of (1) to every possible state \(s\) \(\in S\), the transition matrix \(A\) can be defined as

\[
A = \begin{bmatrix}
a_{11} & a_{12} & \cdots & a_{1J} \\
a_{21} & a_{22} & \cdots & a_{2J} \\
\vdots & \vdots & \ddots & \vdots \\
a_{J1} & a_{J2} & \cdots & a_{JJ}
\end{bmatrix}, \text{ where } a_{ij} \geq 0
\]

(2)

\[
\sum_{j=1}^{r} a_{ij} = \sum_{j=1}^{r} P(X_{m+1} = s_j \mid X_m = s_i) = 1, \forall i
\]

(3)

where \(i\) is a row of \(A\) containing all transition probabilities of going from state \(s_i\) to any other existent state \(s_j\). In addition, an initial probability matrix \((\pi)\) can be defined as well, which specifies the probabilities of the process to start in each of the existing states.

Markov models provide a convenient way of representing a process through states and transitions, in which for each transition that is triggered, a state change occurs. In this sense, a Markov model can be visualized with a state transition diagram. To better illustrate it, if we take as example an \(A\) matrix defined as follows

\[
A = \begin{bmatrix}
0 & a_{12} & a_{13} & 0 \\
0 & 0 & 0 & a_{24} \\
a_{31} & 0 & a_{33} & a_{34} \\
a_{41} & 0 & a_{43} & 0
\end{bmatrix}
\]

(4)

the associated system can be depicted as the one in Fig. 1.

In summary, a Markov model is defined by the following:

- Set of states \(S = \{s_1, s_2, \ldots, s_J\}\).
- Matrix \(\pi\), for initial probability values.
- Matrix \(A\), defining how likely it is for the process to be in a future state given the current one.

As the process analyzed in the presented work is a real-life process, it is reasonable to say that the Markovian assumption cannot be strictly followed: given a present state, the future ones are not entirely independent of all past states. However, for the sake of simplicity, the assumption can be used when considering a limited history/memory, which suits our intention to model (and represent) a finite number of events/transition happening in a limited time window (day(s) of a diabetes patient).
Hidden Markov models (HMMs) [27] allow for an additional layer when other observable random variable exists in the same context of the states, and can be seen as an augmentation of Markov chains. They come as an answer when states are not directly observed, but can be inferred through an associated (emitted) observable value, i.e., it uses the emitted observations to evaluate the probability of a state to happen [28,29].

Given another discrete-time random process \( O_m, m = 0, 1, 2, \ldots \), and a set of possible finite observations \( O = \{ o_1, o_2, \ldots, o_k \} \), each element of \( O \) has a probability of being generated/emitted from an element of the state space \( S \). The matrix \( B \) is called emission matrix, and it has the probabilities of an observation \( o \) being generated from a state \( s \):

\[
B = \begin{bmatrix}
    b_{11} & b_{12} & \cdots & b_{1v} \\
    b_{21} & b_{22} & \cdots & b_{2v} \\
    \vdots & \vdots & \ddots & \vdots \\
    b_{v1} & b_{v2} & \cdots & b_{vv}
\end{bmatrix}, \quad \text{where } b_{iv} \geq 0,
\]

where

\[
\sum_{k=1}^{v} b_{ik} = \sum_{k=1}^{v} P(o_k | X_m = s_i) = 1, \quad \forall i.
\]

In a discrete timeline, Fig. 2 presents both related processes: one covering the hidden (non-observable) states \( X_0, X_1, \ldots, X_m \), and the other \( O_0, O_1, \ldots, O_m \) producing the observations \( \{ o_1, o_2, \ldots, o_m \} \) emitted by each state.

An essential characteristic to highlight is the output independence of each produced observation, which means that an emitted observation depends only on the state that originates it, and not on any other existing state or observation.

HMM models have the ability of inferring a sequence of states given a sequence of emitted observations, this is defined as the decoding problem [28,29], and is commonly tackled by specific dynamic programming tasks commonly relying on the Viterbi algorithm [30–32].

The diagram presented in Fig. 1 can be extended to add hidden states through the use of an emission matrix which adds another stochastic layer per state as depicted in Fig. 3.

An important difference from a chain can be noticed from the extended version: it allows for the addition of a layer able to incorporate external (observable) factors that permits the inference step to rely less on being aware of the current state. It now incorporates probabilities for transitions from one — now hidden — state to another, and emissions from each state to a possible observed value. In other words, it is defined now by both transition and emission matrices, e.g.,

\[
A = \begin{bmatrix}
    0 & a_{12} & a_{13} & 0 \\
    0 & 0 & 0 & a_{24} \\
    a_{31} & 0 & a_{33} & a_{34} \\
    0 & a_{42} & 0 & a_{44}
\end{bmatrix}, \quad B = \begin{bmatrix}
    b_{11} & 0 & b_{13} \\
    0 & b_{22} & b_{23} \\
    b_{31} & b_{32} & b_{33} \\
    0 & b_{42} & 0
\end{bmatrix}
\]

In summary, for a HMM, the following input is expected:

- Set of states \( S = \{ s_1, s_2, \ldots, s_v \} \).
- Set of observations \( O = \{ o_1, o_2, \ldots, o_v \} \).
- \( \pi, A, \) and \( B \) matrices. The latter defines how likely it is for the process to generate an observation from a specific state.

Hidden Markov models are useful for inferring events that a patient does not record, since additional data from observations regarding the context of the event sequence can be used for improved modeling. An example of such observations would be the glucose values coming from the signal provided continuously by each person — for our specific case, a diabetic patient — when using a CGM sensor.

### 4. Modeling patient events sequences through Markov models

In this section, we introduce the requirements regarding the data used in this research. It is followed by an explanation of how Markov chains are used to model event sequences, and then, how HMMs can be modeled to consider specific and related external observations. We propose three different modeling approaches applied to each patient: one relying on the chains only; and another two making use of the HMMs. The section details how each model is defined and how they are used for inference.

#### 4.1. Input data

We assume that patients collect data regarding their daily activities as well as disease related events, such glucose control measures. In addition, measurements of the glucose level are also collected in a continuous manner through CGM, which is becoming a common tool used by diabetic patients in their daily routines. The ideal gathered data would reflect in a one-to-one manner all events – from the set defined to be logged – that happened during the days of each participating patient. Usually, this ideal scenario is far from what the acquired data really reflects in the end.

Fig. 4 illustrates the events data logged by one same example diabetic patient in different days. The figure includes events such meals (breakfast, lunch, and dinner), hypo-correction (e.g., sugary drinks like juice or regular soda), snacks, and exercises, all tied to their own timestamps.

In addition, we can leverage from Fig. 4 to strengthen the argument tied to the problem tackled by the proposed work. For that, a few points can be highlighted:

- The activities in each of the displayed sequences are very similar, however the existing differences are mostly presented after lunch.
- For the day 2021-10-11, there is no dinner logged, thus the patient either have skipped such meal, or the event was not logged.
By taking the first three days, one can note that between Lunch and Dinner, either a Snack, an Exercise, or none can happen.

In the last depicted day (2021-10-22), only a Dinner was logged, which characterizes a day with very poor quality of logged data. This assumption comes due to the fact that a diabetic patient is not likely to have this kind of day.

Let us assume we come to a point where we know an event has happened between Lunch and Dinner on 2021-09-16, and we must infer the activity that took place. Using all information from Fig. 4, it is reasonable to assume it might have been a Snack or an Exercise.

However, in such scenario, not only the sequence of events can play a role, but also external observed factors as the glucose level tied to the missing event, or the associated time interval it occurred. Figs. 5 and 6 present such observable factors for days that would fit the scenario (2021-09-14 and 2021-10-04).

The Snack and Exercise events for both days differ in respect to both mentioned factors: the time they happened is different, and the glucose levels tied to each of them are also different. Such observations can be of great value on inferring an activity on scenarios such as the one previously described for day 2021-09-16, so that data correction or imputation can be done to keep the coherence of the logged data. This becomes an interesting and tangible approach to be taken in order to solve such issue, and will be detailed in the next sections.

The input data that is used in our approach can be seen as a collection of patients data following the characteristics of the aforementioned example patient. For each patient, it is expected a sequence of events recorded in multiple (and continuous) days, and glucose level signal measured in the same set of days.

### 4.2. Modeling from event sequences

According to the described input data, for each patient an event sequence is defined as follows:

**Definition 1 (Event Sequence).** Let $\mathcal{E}$ be a set of events, $\mathcal{X}$ a set of activities (or event types), and $\mathcal{T}$ the time universe. An event $e_i \in \mathcal{E}$ is denoted by $e_i = (\omega_x(e_i), \omega_t(e_i))$, where

- $\omega_x : \mathcal{E} \rightarrow \mathcal{X}$ is a surjective function linking events to activities,
- $\omega_t : \mathcal{E} \rightarrow \mathcal{T}$ is an injective function linking events to time,
- $e_i = (x, t_i) \in \mathcal{E}$ denotes that activity $\omega_x(e_i) = x \in \mathcal{X}$ happened at time $\omega_t(e_i) = t_i \in \mathcal{T}$.

An event sequence $\sigma$ consisting of $n$ events is defined by $\sigma = \langle e_1, e_2, \ldots, e_n \rangle$, with all events happening consecutively in time, where $t_1 < t_2 < \cdots < t_n$.

In the reality of the presented work we make use of a subset of event types, and each element of such subset is translated into one state, resulting in the following state space:

$$\mathcal{S} = \{\text{Breakfast, Dinner, Lunch, Exercise, Hypo-correction, Snack, Sleep-start, Sleep-end}\}.$$  

For illustrative purposes, the elements of our state space are here grouped in four different categories, forming a reduced state space:

$$\mathcal{S} = \{\text{Breakfast, Dinner, Lunch, Exercise, Hypo-correction, Snack, Sleep-start, Sleep-end}\}.$$  

Both upcoming examples (Examples 1 and 2) consider the reduced state space presented in (9). Further, these categories will be unfolded again into their contained event types (cf. Section 4.4), resulting back in the larger state space in (8) used in practice for modeling.

**Example 1.** Modeling with Events Data Only

By taking a sequence of logged events as input, first order Markov models (chains) can be created, i.e., translate event activities into states and calculate the associated probabilities of one appearing given
Example 2. Modeling with Observed BG Ranges

By following the definitions given in Section 3.2, let
\[ O = \{o_1, o_2, o_3, o_4\}. \]  
and a set of observations, and a sequence of such observations, respectively. By expanding Example 1, and using the same transitions as before, states and observations can be aligned accordingly:
\[ \langle s_4, s_1, s_2, s_2, s_2, s_2, s_3, s_3, s_3, s_2, s_1, s_1 \rangle \]
\[ \langle o_1, o_2, o_1, o_1, o_1, o_1, o_1, o_1, o_1 \rangle \]
Hence, as at each cycle an observation is generated by a state, tuples \((s_j, o_v)\) are formed (e.g., \((s_4, o_1), (s_1, o_2), (s_2, o_1), \ldots\)). The frequency of each tuple is counted, and the associated emission probability is estimated.

Let \( W \) be a multiset containing all tuples formed from the aforementioned state-observation sequence, \( \langle s_j, o_v \rangle \) a function that gives the multiplicity of a transition from \( s_j \) to \( s_j \). The estimated probability of transitioning from \( s_j \) to \( s_j \) is given by
\[ P(s_j|s_j) = \frac{\text{Count}(s_j, s_j)}{\sum_{s_k} \text{Count}(s_j, s_k)}, \text{ where } r = |S|. \]  
Applying (11) to all possible transitions leads to the following A matrix:
\[ A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} 0 & 2/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 1 & 0 & 0 & 0 \end{bmatrix} \]  
(12)

The first row of \( A \) in (12) has the probabilities of transitioning from the state \( s_1 \) (Meal) to the other existing states, thus:
\[ P(s_1|s_1) = \frac{\text{Count}(s_1, s_1)}{\sum_{s_k} \text{Count}(s_1, s_k)} = P(\text{Meal}|\text{Meal}) = 0, \]
\[ P(s_2|s_1) = \frac{\text{Count}(s_2, s_1)}{\sum_{s_k} \text{Count}(s_1, s_k)} = P(\text{Exercise}|\text{Meal}) = 2/3, \]
\[ P(s_3|s_1) = \frac{\text{Count}(s_3, s_1)}{\sum_{s_k} \text{Count}(s_1, s_k)} = P(\text{Control}|\text{Meal}) = 1/3, \]
\[ P(s_4|s_1) = \frac{\text{Count}(s_4, s_1)}{\sum_{s_k} \text{Count}(s_1, s_k)} = P(\text{Sleep}|\text{Meal}) = 0/3 = 0, \]
and \( \sum_{s_k} P(s_j|s_k) = \sum_{s_k} P(s_j|\text{Meal}) = 1 \). Similarly, the other rows contain the probabilities of transitioning from \( s_k \) (Exercise), \( s_3 \) (Control), and \( s_4 \) (Sleep). This can also be seen through the associated state transition diagram presented in Fig. 7.

Please note that, if needed, the frequency of rare (but not impossible) transitions can be artificially adjusted (e.g., using additive smoothing [33]) to avoid probability values equal to zero estimated based on the available input data.

4.3. Modeling with observed external factors

An HMM requires both transition and emission matrices. The latter is generated similarly to the former, but it relies on the link between state and observable variables to calculate the emission probabilities rather than the link between previous and next states.
It is worth noting that the amount of observations in \( O \) does not have to be the same as the amount of states in \( S \), and so the transition and emission matrices can differ in shape. However, for the example used, their sizes match.

### 4.4. Modeling approaches

Using the state space previously defined in (8), we defined three different modeling approaches to apply to the data of each patient:

- **Modeling Approach 1:** Similar to Example 1, this approach makes use solely of the transition matrix created from the events sequence.
- **Modeling Approach 2:** Makes use of both transition and emission matrix, defining an HMM based on BG ranges (Table 1) tied to each event (state) similarly to Example 2.
- **Modeling Approach 3:** Also a HMM, but now generating an emission matrix based on observed pre-defined time intervals according to Table 2, instead of BG ranges. The assumption is now that being within certain time windows would impact the probability of an event to happen, which can lead to a better probability estimation.

In our work, we use the models created by each described approach to infer one state between two others, *i.e.*, an event activity that happened between two known events. The next section describes how transition and emission matrices are used by the models to perform such inference.

### 4.5. Inferring an event activity

The concepts behind the steps taken for event inference are formalized in this section.

With each patient having their entire events sequence translated into both transition and emission matrices, we use Markov models to infer an event (state) that is more likely to happen between two others.

In the proposed inference steps, event sequences consisting of three elements (\( n = 3 \)) are used as input, and will mostly be referred in the form of triples \( (e_i, e_j, e_k) \).

---

**Definition 2 (Inference Condition).** For any triple event sequence \( \sigma \) in the form of \( (e_i, e_j, e_k) \), where \( \omega_x(e_i) \) and \( \omega_x(e_j) \) are known, the inference condition is met when \( e_j \in \sigma \) and \( \omega_x(e_j) \) is not known.

The evaluation and subsequent inference is done by defining the activity that would occur between two others, *i.e.*, the \( \omega_x(e_j) \) value for every triple where the condition in Definition 2 holds.

**Definition 3 (Inferred Activity).** Let \( \sigma \) be a triple event sequence where the inference condition holds. Let \( \Theta \) be a set of observations. \( P(\omega_x(e_j) = x) \) defines the probability of activity \( x \) to be linked to \( e_j \in \sigma \), and is denoted by

\[
P(\omega_x(e_j) = x) = P(x|\omega_x(e_j)) \cdot P(\omega_x(e_j)|x) \cdot P(\omega_y(e_j)|x)
\]

(20)

where

- \( \omega_y : E \rightarrow \Theta \) is a function linking events to observations.

The inferred activity \( \hat{x} \) is the one that \( \forall x \in X, P(\omega_x(e_j) = x) \) is maximized.

The steps covered in Definition 3 make use of preliminary concepts brought by Section 3, and the components of the probability value evaluated during the inference are taken from both transition and emission matrices of the associated (Hidden) Markov model. An important aspect to emphasize is that the Markov assumption holds for both \( P(x|\omega_x(e_j)) \) and \( P(\omega_x(e_j)|x) \), while the output independence is kept in \( P(\omega_y(e_j)|x) \), with the observation depending only on the state (activity \( x \)) that produces it.

**Example 3. Activity Inference for a Given Triple**

By taking as example the sequence \( \text{(Dinner, Exercise)} \) following Definition 1 and meeting the condition of Definition 2, we find the inferred activity \( \hat{x} \) according to Definition 3. In summary:

1. \( P(\omega_x(e_j) = x) = P(x|\text{Dinner}) \cdot P(\text{Exercise}|x) \cdot P(\omega_y(e_j)|x), \forall x \in X \) and \( \omega_y(e_j) \in \Theta \).

   - (a) The values for \( P(x|\text{Dinner}) \) and \( P(\text{Exercise}|x) \) are taken from the transition matrix.
   - (b) For the cases where observations are not used (Modeling Approach 1), \( P(\omega_y(e_j)|x) = 1 \).
   - (c) For the cases where observations are used (Modeling Approaches 2 and 3), \( P(\omega_y(e_j)|x) \) comes from the associated emission matrix.

2. The activity \( x \) with the highest estimated probability (more likely to happen) is then taken as the inferred activity \( \hat{x} \), *i.e.*, the one to be associated to \( e_j \).

---

**5. Experimental setup**

In order to assess the predictive performance of our proposed methods, we use a public dataset. In this section, we describe the dataset and explain the experimental design to assess the performance of the methods.

#### 5.1. Data description

We use in our experiments the *OhioT1DM* [37] dataset, which includes data of 12 people with diabetes mellitus type 1 (T1DM) on insulin pump therapy and using CGM devices. It contains data from continuous glucose monitoring, and also from daily reported events. It was developed and made publicly available1 to promote and facilitate research in blood glucose level prediction.

Table 3 shows the amount of data points regarding the types of events considered in the proposed approach for each of the participants (patients) included in the dataset.

---

1 smarthealth.cs.ohio.edu/OhioT1DM-dataset.html.
Table 3
General information and number of data points for events logged per patient (train and test subsets).

<table>
<thead>
<tr>
<th>ID</th>
<th>Set</th>
<th>B</th>
<th>D</th>
<th>L</th>
<th>Exercise</th>
<th>HC</th>
<th>Snack</th>
<th>Sleep</th>
</tr>
</thead>
<tbody>
<tr>
<td>540</td>
<td>Train</td>
<td>27</td>
<td>17</td>
<td>20</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>7</td>
<td>4</td>
<td>8</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>544</td>
<td>Train</td>
<td>37</td>
<td>39</td>
<td>48</td>
<td>2</td>
<td>8</td>
<td>27</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>9</td>
<td>11</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>552</td>
<td>Train</td>
<td>24</td>
<td>22</td>
<td>15</td>
<td>10</td>
<td>7</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>8</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>584</td>
<td>Train</td>
<td>30</td>
<td>36</td>
<td>20</td>
<td>3</td>
<td>0</td>
<td>9</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>7</td>
<td>10</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>596</td>
<td>Train</td>
<td>48</td>
<td>52</td>
<td>50</td>
<td>46</td>
<td>28</td>
<td>87</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>11</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>3</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>559</td>
<td>Train</td>
<td>42</td>
<td>38</td>
<td>36</td>
<td>20</td>
<td>7</td>
<td>27</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>563</td>
<td>Train</td>
<td>39</td>
<td>38</td>
<td>41</td>
<td>17</td>
<td>8</td>
<td>3</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>10</td>
<td>8</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>570</td>
<td>Train</td>
<td>37</td>
<td>39</td>
<td>30</td>
<td>25</td>
<td>1</td>
<td>29</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>11</td>
<td>10</td>
<td>8</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>575</td>
<td>Train</td>
<td>42</td>
<td>46</td>
<td>40</td>
<td>5</td>
<td>37</td>
<td>78</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>9</td>
<td>10</td>
<td>7</td>
<td>0</td>
<td>4</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>588</td>
<td>Train</td>
<td>45</td>
<td>46</td>
<td>47</td>
<td>36</td>
<td>13</td>
<td>70</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>3</td>
<td>0</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>591</td>
<td>Train</td>
<td>51</td>
<td>46</td>
<td>43</td>
<td>23</td>
<td>7</td>
<td>65</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>11</td>
<td>8</td>
<td>10</td>
<td>4</td>
<td>0</td>
<td>12</td>
<td>10</td>
</tr>
</tbody>
</table>

*Meals: Breakfast (B), Dinner (D), Lunch (L).
*HC: Hypo-correction.
*Sleep events are defined as a pair (start, end).

5.2. Experimental design

The dataset is provided already split into training and testing subsets, following splitting rules defined by the authors of the dataset with the intention to allow unbiased comparison of developed models based on it. For our experiments, the same rules were respected. Each of the subsets contains different continuous days of glucose and events data collected by the same patients. Not only the most disease directly associated data like glucose level and use of insulin are included in the dataset, but also self-reported life events. Each person is referred to by randomly selected ID numbers, and the amount of logged events vary for each one of them. In summary, the data that is relevant for the presented work:

- Every 5 min blood glucose (BG) level (CGM).
- Self-reported events: meals (breakfast, lunch, and dinner), blood glucose control (hypo-correction and snacks), exercises, and sleep.

For each patient of the dataset, the data summarized here is retrieved from both training and testing subsets, to be used in the modeling and testing steps. During the training step, transition and emission matrices are created following the methods covered in Section 4, thus models for each of the approaches (Approach 1, 2 and 3) are developed per patient. For each model, we create one normalized confusion matrix [38], and calculate both sensitivity and specificity to compare their performances. The results concerning the generated models are presented in the next section (Section 6).

6. Results

From the data described in Section 5, we create three individual models per patient following the previously described approaches. The generated matrices for a sample patient are presented in Tables 4 and 5, and 6. With the models in hand, the associated testing subset is then used for validation.

During the test step, all events – from the training subset – for a certain patient are streamlined, resulting in one sequence of events ordered in time. From the beginning to the end of this sequence, every existing triple \((e_i, e_j, e_k)\) is tested, e.g., \([(e_0, e_1, e_2), (e_1, e_2, e_3), \ldots, (e_{n-2}, e_{n-1}, e_n)]\), where \(n\) is the size of the whole sequence. For each triple, the activity \(x\) from \(e_i\) is masked, and then inferred by the trained model. When the model makes a correct inference, it is considered a success, otherwise a failure. This is done for each of the explored modeling approaches per patient separately.

To summarize the results obtained, we sum (for all tested patients) all successes and failures, and present it as a normalized multi-class confusion matrix for each modeling approach (Fig. 9). Our confusion matrix is a \(n \times n\) matrix, where \(n\) is the amount of event types, in our case, \(n = 8\). Each case in the test set has an actual class label (row), and an inferred class label (column). As a normalized matrix, each cell shows the number of cases that were inferred as the column divided by the total number of cases in the actual label (the number in parentheses for each row). For instance, Fig. 9a shows the confusion matrix for the Modeling Approach 1, which infers using a Markov chain, taking only state transitions into account; and Figs. 9b and 9c for the Modeling Approaches 2 and 3, using HMMs with observed BG ranges and time windows, respectively. The total number of inferences made are depicted between parenthesis for each label in the Actual axis.

For each confusion matrix, the diagonal shows the percentage of correct inferences made for each event type (row). Note that the percentage distribution (gradient coloring) varies for each model. The darker the cell in the diagonal is, the higher the accuracy to infer the event type associated to the row. For the cells in the diagonal that are not so dark, it is also important to check the other cells in the same row. They indicate, when trying to infer that event type (row), the percentage of times that the inference is done as another event type (column). Following these results, we use sensitivity and specificity values as measures to compare the models. We also discuss why we believe that certain models perform better when inferring a subset of event types.

Table 7 shows the calculated sensitivity and specificity for the inferences made. Again, the sum of all the results per patient is used. Although we consider both sensitivity and specificity as metrics/criteria for model selection, we give more weight to the former rather than the latter. For us positives are critical, as it is mostly important to have correct predictions, and considerations over the negatives would not lead directly to the activity discovery. Moreover, it is worth noticing that the specificity values for all three models are very similar.

In general, by using the sensitivity values, Modeling Approach 3 performed better than the others. Hence, from a macro perspective, the events have a more time bound characteristic, i.e., the patients’ events are more steered by the time of the day that they more commonly happen. However, events like Exercise, Hypo-correction, and Snack – that, unlike meals, are not strongly trussed to time intervals – demand a more in depth analysis:

- Hypo-correction is commonly triggered by a natural patient-side observation of the BG level, which leads to an expected better accuracy of Modeling Approach 2. Furthermore, for this same model, the wrong inferences regarding this event are shifted towards the other low glucose control event considered: Snack.
- By comparing the confusion matrices of the three models, one can notice that Exercise has a certain degree of dependency to event order, time, and BG level, as no model outperformed the others.

Each developed model has a peculiar characteristic as basis: Modeling Approach 1 relies solely on the events sequence order, and Modeling Approaches 2 and 3 have an expanded context that considers not only the sequence of events, but also the observed BG level and time, respectively. From the results, Meals and Sleep are better inferred by the third modeling approach, while Hypo-correction is better inferred by the second. This makes us believe that these types of events are bound to the characteristics of the observations used by each HMM, and at the same time, the inference made is heavily tied to such characteristics. This suggests that by using a hybrid or a multi-layered observation approach could improve the performance of the models. For instance, by taking Modeling Approach 3 as base and adding BG level
Table 4
Transition matrix for patient 563.

<table>
<thead>
<tr>
<th></th>
<th>B^</th>
<th>D^</th>
<th>Exercise</th>
<th>HC^</th>
<th>L^</th>
<th>Sleep-end</th>
<th>Sleep-start</th>
<th>Snack</th>
</tr>
</thead>
<tbody>
<tr>
<td>B^</td>
<td>0.0000</td>
<td>0.0769</td>
<td>0.1538</td>
<td>0.0769</td>
<td>0.6923</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>D^</td>
<td>0.1579</td>
<td>0.0526</td>
<td>0.0263</td>
<td>0.0526</td>
<td>0.0799</td>
<td>0.0263</td>
<td>0.6053</td>
<td>0.0000</td>
</tr>
<tr>
<td>Exercise</td>
<td>0.0000</td>
<td>0.5862</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.2353</td>
<td>0.0000</td>
<td>0.1765</td>
<td>0.0000</td>
</tr>
<tr>
<td>HC^</td>
<td>0.3750</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.5000</td>
<td>0.0000</td>
<td>0.1250</td>
<td>0.0000</td>
</tr>
<tr>
<td>L^</td>
<td>0.0488</td>
<td>0.5122</td>
<td>0.1951</td>
<td>0.0000</td>
<td>0.0488</td>
<td>0.0000</td>
<td>0.1220</td>
<td>0.0732</td>
</tr>
<tr>
<td>Sleep-end</td>
<td>0.8438</td>
<td>0.0000</td>
<td>0.0312</td>
<td>0.0625</td>
<td>0.0312</td>
<td>0.0000</td>
<td>0.0312</td>
<td>0.0000</td>
</tr>
<tr>
<td>Sleep-start</td>
<td>0.0033</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0033</td>
<td>0.0000</td>
<td>0.9394</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Snack</td>
<td>0.0000</td>
<td>0.6667</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.3333</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

^Meals: Breakfast (B), Dinner (D), Lunch (L).
^HC: Hypo-correction.

Table 5
Emission matrix for BG ranges for patient 563.

<table>
<thead>
<tr>
<th></th>
<th>&lt;50 and &lt;60</th>
<th>≥60 and &lt;70</th>
<th>≥70 and &lt;180</th>
<th>≥180 and &lt;250</th>
<th>≥250 and &lt;400</th>
</tr>
</thead>
<tbody>
<tr>
<td>B^</td>
<td>0.0000</td>
<td>0.1176</td>
<td>0.0000</td>
<td>0.6471</td>
<td>0.2353</td>
</tr>
<tr>
<td>D^</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0303</td>
<td>0.7576</td>
<td>0.1818</td>
</tr>
<tr>
<td>Exercise</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.7576</td>
<td>0.2424</td>
</tr>
<tr>
<td>HC^</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0256</td>
<td>0.7693</td>
<td>0.2051</td>
</tr>
<tr>
<td>L^</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1250</td>
<td>0.8750</td>
<td>0.0000</td>
</tr>
<tr>
<td>Sleep-end</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0244</td>
<td>0.8537</td>
<td>0.0732</td>
</tr>
<tr>
<td>Sleep-start</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.3333</td>
<td>0.6667</td>
</tr>
<tr>
<td>Snack</td>
<td>0.0263</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.7368</td>
<td>0.1842</td>
</tr>
</tbody>
</table>

^Meals: Breakfast (B), Dinner (D), Lunch (L).
^HC: Hypo-correction.

observations, this could improve the inference of BG bound events (e.g., Hypo-correction). Additionally, it might be that events not naturally bound to any specific characteristic (e.g., Snack, and Exercise) are also better inferred considering a combination of them.

7. Discussion

This section is dedicated to the discussion of three specific relevant points, presenting their limitations and possibilities with regard to reducing or surmounting them. First, as emphasized before, each patient has his/her own logged data, and his/her own trained model. This can be seen as a particularly welcome feature that leads to personalized models. Nonetheless, it can be taken as a limitation of the modeling, as it means that we are unable to reuse any of the existing data/model for new patients. New patients would be required to log events for a certain time period before any model could be trained. On the other hand, individual models can capture the nuances regarding the different non-deterministic behaviors. A possible take on overcoming this, is that if a profile could be set initially and more than one patient associated to this previously created profile, a more general model could be pre-trained per profile, and would be able to suit the associated patients. This could be achieved by exploring the concept of patient similarity [39–41].
Second, for the evaluation of the proposed approach, all models were trained using a real dataset. With that being said, the same – and already exposed – problems regarding missing data, are also expected to be present in this same dataset. Thus, the imperfections of the training data can also be anticipated in the trained models. However, by analyzing the generated transitions and emissions matrices, “non-usual” transitions (e.g., from Sleep to Exercise) are associated to a very low probability of happening, allowing the created models to naturally cope with such disturbances.

Finally, our approach relies on the fact that we synthetically know when a missing event occurs between two other events. We narrowed our scope by limiting our approach to tackle one part of a twofold problem, and identifying when missing events happen is an outcome that must ideally come from a solution tied to the second part. A further research is to identify the time boundaries of such missing events, giving us the opportunity to recognize a missing event slot. This future step will enforce the value of the presented approach, and can be linked seamlessly to the steps here proposed.

8. Conclusions

This paper explores the problem of dealing with and mitigating missing events data on diabetes patients related data. This is specially important due to the high effort that is expected from patients to log all required activities properly, which may not be fulfilled in practice, leading to missing events in the data streams. To solve this problem, we presented an approach based on Markov models able to infer missing activities tied to events in an existent data sequence. The approach showed that it is possible to consider not only the order of the events, but also external observations to improve the inference accuracy. Here, we presented two possible external observations: blood glucose (BG) level ranges – where the BG signal was discretized based on six recommended ranges according to clinical diagnosis and safety –, and time windows.

The choice of Markov modeling allows for getting insights in the non-deterministic behavior of a patient, which might contribute to a better and more complete patient behavior representation. For the context explored in the paper, the use of hidden states creates a link between BG signal and daily events, allowing for other BG attributes to be taken as external observations and added to the models.

The presented approach is one step towards a solution for improving quality of data from journals of diabetic patients. Models for glucose metabolism [42] and predictive decision support methods based on data from such journals can be improved with the use of better quality data with fewer missing events in them. Such solution should still take into account the identification of when missing events happen, and to what extent this can reduce the burden of logging multiple (correct) events on the patient side. Detecting missing events is a parcel of the problem intended to be tackled in future studies. In addition, future works will be devoted to extend the methodology for other possible observations, such as signals collected by accelerometer and/or heart rate monitoring sensors. This must be done not only by expanding the observation universe, but also by identifying useful ways of handling each added signal, i.e., how each additional signal must be discretized (partitioned) in order to be used by the developed models, while taking into account uncertainties in the definition of meaningful intervals and the probabilistic nature of the data [43]. Studies over the impact of the different lengths of intervals taken for such partitions – and the dynamics between them – are intended to be performed, and the consequences of such changes on the models analyzed.

Another point to be taken as future study is the generalization of the model for cases where patient profiles are a better fit than individual models. These profiles could be created by exploring the concept of patient similarity, which would allow for the identification of useful similarity functions able to link patients. Thus, similar patients could be grouped using similarities found on their data, and such clustered data used as input for the development of a per-cluster model that fits all members of such cluster.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.