Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Learning the progression patterns of treatments using a probabilistic generative model

Onintze Zaballa [a,*], Aritz Pérez [a], Elisa Gómez Inhiesto [b], Teresa Acaiturri Ayesta [b], Jose A. Lozano [a,c]

[a] BCAM-Basque Center for Applied Mathematics, Bilbao 48009, Spain
[b] Hospital Universitario Cruces, Barakaldo 48903, Spain
[c] Intelligent Systems Group, Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, Donostia 20018, Spain

ARTICLE INFO

Keywords:
Disease progression modeling
Electronic health records
Markov model
Probabilistic generative model
Unsupervised machine learning

ABSTRACT

Modeling a disease or the treatment of a patient has drawn much attention in recent years due to the vast amount of information that Electronic Health Records contain. This paper presents a probabilistic generative model of treatments that are described in terms of sequences of medical activities of variable length. The main objective is to identify distinct subtypes of treatments for a given disease, and discover their development and progression. To this end, the model considers that a sequence of actions has an associated hierarchical structure of latent variables that both classifies the sequences based on their evolution over time, and segments the sequences into different progression stages. The learning procedure of the model is performed with the Expectation–Maximization algorithm which considers the exponential number of configurations of the latent variables and is efficiently solved with a method based on dynamic programming. The evaluation of the model is twofold: first, we use synthetic data to demonstrate that the learning procedure allows the generative model underlying the data to be recovered; we then further assess the potential of our model to provide treatment classification and staging information in real-world data. Our model can be seen as a tool for classification, simulation, data augmentation and missing data imputation.

1. Introduction

Disease progression research seeks to refine, from electronic health records (EHRs), the definition of complex and heterogeneous pathologies by identifying subgroups with similar temporal evolution patterns. These repositories contain systematized collections of patient data, including demographics, procedures, diagnosis, medications, costs, medical service providers and so on. The order of occurrence of these medical events in EHRs provides valuable information on the treatment trajectory of a patient which could improve the understanding of the disease [1]. Therefore, we represent each patient using chronologically ordered sequences of medical actions. Constructing from these sequential medical events a comprehensive characterization of the treatment patterns and their temporal evolution, that is, identifying distinct subtypes of treatments and discovering their development and progression remains a major challenge in medical informatics.

This model could benefit both the clinical practice and management of healthcare. Clinically, it can help discover the associations between the shared characteristics of similar patients, reduce the uncertainty in a patient's expected outcome and identify a data-driven taxonomy of the progression of treatments associated with a disease. Thereby, improving treatment decisions. From the management viewpoint, subtyping can forecast the expected costs of care and improve the efficacy of clinical trials by enabling targeted enrollment [2].

A research area that has been used in the field of healthcare for treatment modeling is process mining [3]. The basic idea is to discover process models from event logs, where medical actions from EHRs are used as process logs. However, treatment trajectories of patients in the healthcare system are complex in part due to their variability, and this assumption leads to spaghetti-like workflow models that are very difficult to interpret [4]. In fact, there is a wide variety of activities that can typically be executed for a single disease, and, in addition, these activities are influenced by the personal preferences and characteristics of patients, physicians and other healthcare experts. Moreover, patients can respond differently to particular treatments, which may affect the order and type of activities that follow [3]. The combination of all these factors tends to make almost all cases different and becomes a problem for this type of models.

* Corresponding author.
E-mail addresses: ozaballa@bcamath.org (O. Zaballa), aperez@bcamath.org (A. Pérez), elisa.gomezinhiesto@osakidetza.eus (E. Gómez Inhiesto), mariateresa.acaiturriayesta@osakidetza.eus (T. Acaiturri Ayesta), jlozano@bcamath.org (J.A. Lozano).

https://doi.org/10.1016/j.jbi.2022.104271
Received 22 March 2022; Received in revised form 18 November 2022; Accepted 9 December 2022
Available online 15 December 2022

Machine learning techniques offer a potential solution to deal with the variability of treatment trajectories. The need to model these heterogeneous disease dynamics has been evidenced particularly by works on disease subtyping that aim to identify subgroups of patients with similar disease progression trajectories [2]. In the literature, some of the works [5–8] that assume differentiated treatment subgroups are an extension of the conventional Latent Dirichlet Allocation [9]. The limitation of these approaches is the assumption that all the individuals are at a unique treatment stage, so that their models are not able to account for the treatment progression.

Probabilistic models, and in particular, hidden Markov models (HMMs), have been widely used for disease progression due to their easy interpretability and their temporal relation assumption in data. Most existing HMMs [10–16] assume that all patients evolve through the same latent state transition dynamics, thus ignoring the heterogeneity of different subtypes of disease progression. Other probabilistic approaches that simultaneously address disease state progression and treatment subtyping [17–19] are limited to model the evolution of observed data through a latent process and do not handle the sequential dependence within medical actions. These methods actually model the number of each type of action that occurs in each stage, rather than being generative models of the sequence of actions. However, the order of occurrence of medical actions is essential to understand the progression of a disease.

Various predictive deep learning models have also been developed for healthcare settings [11,20–27]. Unfortunately, they not only ignore the variability in treatments, but also their hidden states do not correspond to clinically meaningful variables such as the treatment evolution patterns provided by our model. While these methods succeed in predicting a target outcome, they do not provide a generative model of the disease progression to identify patients with similar disease progression patterns, to understand the evolution of treatments through interpretable distributions of stage transitions, or to simulate populations of treatment trajectories.

This paper presents a method to model heterogeneous sequences of actions with a hierarchical structure of a set of latent classes of treatments and a set of latent progression stages. In summary, the key contributions of this work are as follows:

- We model EHRs using a probabilistic generative model built on Markov models to capture the order of occurrence of the events. The model discovers the subtypes of treatments by grouping the sequences of medical actions into different classes according to their evolution and identifies the progression stages of the treatments over time.
- We learn the model using the Expectation–Maximization (EM) algorithm [28], where we generalize the conventional forward–backward algorithm [29] used for HMMs to efficiently learn the parameters of our generative model.
- We evaluate the learning performance of the model in multiple simulated datasets of different sizes with the aim of demonstrating that the model underlying the data is recovered.
- We apply the model on a breast cancer dataset to represent the progression of the different classes of treatments and their phases. The results are contrasted with clinical guidelines and approved by physicians.

The remainder of this paper is organized as follows. Section 2 describes the probabilistic generative model and the learning process of the parameters by means of the EM algorithm. In Section 3, we present the results of the synthetic data experiments that evaluate the performance of the proposed method, and the application of the model on a real-world dataset. Section 4 discusses the contributions and limitations of our approach, and draws the conclusions.

2. Methodology

This section describes the proposed probabilistic generative model and the procedure for the inference and the parameter estimation.

2.1. Notation and terminology

We use health related terms throughout the paper. These are notions to guide intuition and clearly capture the idea of the model. Formally, we define them as follows:

- A *medical action* "a" represents an event in a hospital. In our case, a is depicted by the healthcare service that a patient has visited. Examples are primary care, surgery unit, hospitalization, and so on. $\mathcal{A}$ is the set of all the possible medical actions and $a \in \mathcal{A}$.
- A *treatment* is a sequence of m actions associated with a particular disease, denoted by $\mathbf{a} = (a_1, a_2, \dots, a_m)$. In fact, a treatment is a subsequence of the whole medical history of a patient, where the actions that have nothing to do with the target disease are excluded, leaving in the treatment sequence $\mathbf{a}$ those directly related to the disease.
- A *progression stage* "s" is modeled by a treatment subpattern given in the sequence of actions. That is, a sequence of actions is segmented into different stages, where each stage represents a pattern of the treatment.
- A *class of treatment* "c" represents a subgroup/subtype of treatments that share common progression stages.

2.2. The probabilistic generative model

The general idea is to develop a probabilistic generative model in order to learn the complex distribution of a set of sequences of different lengths. We assume that sequences of actions have an associated hierarchical structure of latent variables: at the top-level, we consider that sequences belong to latent classes representing the different subtypes of treatments; at the lower-level, we assume that the sequences of actions progress through a set of latent ordinal-valued stages over time, that is, each action of a sequence has an associated stage that indicates the phase of progression of the treatments at that time point. The goal, therefore, is to infer these latent classes of treatments and their progression stages. Inferring classes requires grouping the sequences of actions into different categories based on common treatment patterns. Moreover, inferring progression stages requires identifying a set of monotonous stages through which sequences of actions evolve, and afterwards, individually segmenting sequences according to the discovered stages. Both tasks have to be simultaneously considered in order to capture the heterogeneity of the sequences of medical actions.

For the definition of the generative model (Fig. 1), we consider that an action depends on the sequence's most recent action and stage within a class. Furthermore, a stage depends on the current action and the previous stage. The duration of the progression stages for each sequence is likely to be different because each patient evolves at their own rate, and consequently, the lengths of the sequences of actions vary. For that reason, we introduce the virtual end-of-treatment action a_m, which allows the length of a population of sequences of actions to be implicitly modeled. Without this end-of-treatment action, the generative model would create sequences of actions of infinite length. Besides, we consider that the sequences of actions always start in the first stage, representing the initial steps of the treatment. We assume that all the classes of treatments have the same number of stages. The definition of such stages makes it possible to segment each class of treatments into subsequences that are related to their progression. Note that the same stage values from different classes of treatments represent different subsequences, which allows the model to be more flexible and to better fit a population of sequences of actions. With these

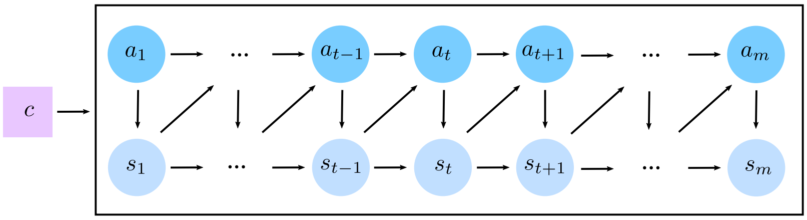


Fig. 1. Markov model defined by the conditional distributions $p(a_t|a_{t-1},s_{t-1},c)$ and $p(s_t|a_t,s_{t-1},c)$ for sequences of actions $\mathbf{a}$, latent sequences of stages $\mathbf{s}$ and latent classes c.

assumptions in mind, we develop a generative process built on Markov models that classifies and segments sequences automatically.

Let $\mathbf{a} = (a_1, \dots, a_m)$ be the sequence of medical actions representing a treatment of a patient associated with a disease. The medical actions a_i belong to a set $\mathcal{A}$ which is the set of all the possible medical actions including the virtual end-of-treatment action. Each sequence can have a different length. Let $\mathbf{s} = (s_1, \dots, s_m)$ be the sequence of latent stages of the treatment associated with the sequence of actions $\mathbf{a}$. The stages s_i belong to a set $\mathcal{S} = \{1, \dots, r\}$ that represents all the possible stages of a treatment. Finally, let c be the latent class of treatment which $\mathbf{a}$ belongs to. The classes c belong to a set $\mathcal{C} = \{1, \dots, k\}$ that represents all the possible classes, i.e., the types of treatments for a disease. Furthermore, we assume that the stages of progression of a sequence of actions are non-decreasing, that is, a sequence cannot progress backward. Therefore, $s_t \leq s_{t+1}$ for all $t = 1, \dots, m-1$.

The proposal for the probabilistic generative model is as follows:

1. Choose a class $c \sim p(c)$
2. For each sequence of actions $\mathbf{a}$:
   (a) Choose an action a_t from $p(a_t|a_{t-1},s_{t-1},c)$, the transition matrix of the Markov model conditioned on the action a_{t-1}, the stage s_{t-1} and the class c.
   (b) Choose a stage s_t from $p(s_t|a_t,s_{t-1},c)$, the transition matrix of the Markov model conditioned on the action a_t, the stage s_{t-1}, and to the class c.

Translating the generative process into a joint probability model results in the expression:

$$p(\mathbf{a},\mathbf{s},c) = p(c)\prod_{t=1}^{m} p(a_t, s_t|a_{t-1}, s_{t-1}, c) \tag{1}$$

where

$$p(a_t, s_t|a_{t-1}, s_{t-1}, c) = p(a_t|a_{t-1}, s_{t-1}, c)\cdot p(s_t|a_t, s_{t-1}, c)$$

and $p(a_1, s_1|a_0, s_0, c) = p(a_1, s_1|c)$. Furthermore, $s_1 = 1$, $a_m = end$, and $s_{t-1} \leq s_t$ for all t.

In light of the above, $p(c)$ is a discrete probability distribution that describes the probability of drawing a class from the set of classes of treatments $\mathcal{C}$. We define $\boldsymbol{\theta}_c$ as the vector of such probabilities:

$$\boldsymbol{\theta}_c = (\theta_c^1, \dots, \theta_c^k) \tag{2}$$

where $\theta_c^i = p(c = i)$ for $i = 1, \dots, k$. In addition, we define the Markov models from which the actions and stages are drawn as follows (see Fig. 1). The first conditional distributions are given by a set of $|\mathcal{C}|$ transition matrices of size $|\mathcal{A}||\mathcal{S}| \times |\mathcal{A}|$ whose model parameters are:

$$\boldsymbol{\theta}_a = \{p(a_t|a_{t-1}, s_{t-1}, c) : a_t, a_{t-1} \in \mathcal{A}, s_{t-1} \in \mathcal{S}, c \in \mathcal{C}\}. \tag{3}$$

The other conditional distributions are given by a set of $|\mathcal{C}|$ transition matrices of size $|\mathcal{A}||\mathcal{S}| \times |\mathcal{S}|$ whose model parameters are:

$$\boldsymbol{\theta}_s = \{p(s_t|a_t, s_{t-1}, c) : a_t \in \mathcal{A}, s_t, s_{t-1} \in \mathcal{S}, c \in \mathcal{C}\}. \tag{4}$$

2.3. Maximum likelihood parameter estimation

In this section we introduce the learning procedure of the parameters of the model. Let $\mathcal{D} = \{\mathbf{a}^1, \dots, \mathbf{a}^N\}$ be the set of sequences of actions to learn the model. We seek to maximize the following weighted log likelihood of the data:

$$\max_{\theta} \sum_{\mathbf{a}\in \mathcal{D}} \sum_{\mathbf{s}\in \mathcal{S}_{\mathbf{a}}} \sum_{c \in \mathcal{C}} p(\mathbf{s}, c|\mathbf{a}) \cdot \log p(\mathbf{a}, \mathbf{s}, c; \boldsymbol{\theta}) \tag{5}$$

where $\mathcal{S}_{\mathbf{a}}$ is the set of all the compatible sequences of stages for $\mathbf{a}$, $p(\mathbf{s}, c|\mathbf{a})$ is the contribution of the tuple $(\mathbf{a}, \mathbf{s}, c)$ to the model, and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_c, \boldsymbol{\theta}_a, \boldsymbol{\theta}_s\}$. The reason for weighting the log likelihood is to make each sequence $\mathbf{a}$ contribute equally to the model regardless of its length, and this is achieved because

$$\sum_{c\in \mathcal{C}} \sum_{\mathbf{s}\in \mathcal{S}_{\mathbf{a}}} p(\mathbf{s}, c|\mathbf{a}) = \sum_{c\in \mathcal{C}} \sum_{\mathbf{s}\in \mathcal{S}_{\mathbf{a}}} p(\mathbf{s}|c, \mathbf{a}) \cdot p(c|\mathbf{a}) = 1. \tag{6}$$

In order to find the parameters that maximize the log likelihood in (5), we use an EM algorithm. In the initialization of the EM, we segment the sequences of actions into equal-length intervals of stages. For the initial model of classes, we use the K-medoids method for the real EHRs. However, in the experiments with synthetic data, which have the purpose of showing the convergence to the real model underlying the data, we initialize the probability of each sequence to belong to the classes with the uniform distribution. We then add a probability $\epsilon = 0.1$ to the true class to which they belong to avoid relabeling in the results. Then, the EM algorithm that yields the following iterative algorithm is as follows:

E-step. In this step we consider, for each sequence of actions $\mathbf{a} \in \mathcal{D}$, all the compatible configurations of the latent sequences of stages $\mathbf{s} \in \mathcal{S}_{\mathbf{a}}$ of length $m_{\mathbf{a}}$ and their probability. In order to do that, we compute the probability $p(s_t = s, c|s_{t-1} = s', \mathbf{a})$ for all $s, s' \in \mathcal{S}$ and $t = 1, \dots, m_{\mathbf{a}}$ given the parameters of the current model. Notice that it requires $\binom{m-2}{r-1}$ number of configurations (the last stage is fixed), which is approximately exponential as long as $r \ll m$. Adopting the notion of the forward–backward algorithm used for learning HMMs, we develop a generalization of this dynamic programming method for the specific characteristics of our model (Appendix A), which avoids its apparent exponential complexity. The conventional algorithm does not suffice for constructing our forward/backward filtering algorithm since we need to account for the temporal relation between the observations, as well as the classes and the latent correlation structures of stages on observed actions.

M-step. In the maximization step we aim to update the parameters of the Markov model and the probability of the classes. We learn the new parameters using the probabilities calculated in the E-step. Therefore, the probability corresponding to the transition from the pair (a, s) to (a', s') given the class c where $a, a' \in \mathbf{a}$ and $s, s' \in \mathcal{S}$ is updated as follows:

$$\theta^c_{a,s,a'} = \sum_{\mathbf{a}\in \mathcal{D}} \sum_{t=1}^{m_{\mathbf{a}}} \mathbb{1}_{a,a'}(a_{t-1}, a_t) \cdot p(s_t = s, c|\mathbf{a}) \tag{7}$$

$$\theta^c_{s,a,s'} = \sum_{\mathbf{a}\in \mathcal{D}} \sum_{t=1}^{m_{\mathbf{a}}} \mathbb{1}_{a,a'}(a_{t-1}, a_t) \cdot p(s_t = s', c|s_{t-1} = s, \mathbf{a}) \tag{8}$$

where

$$\mathbb{1}_{a,a'}(a_{t-1}, a_t) = \begin{cases} 1 & \text{if } a_{t-1} = a, a_t = a' \\ 0 & \text{otherwise..} \end{cases}$$

Finally, we update the probability of the classes of treatments $c \in C$ as follows:

$$\theta_c = p(c) = \frac{\sum_{\mathbf{a} \in D} p(c|\mathbf{a})}{\sum_{c \in C} \sum_{\mathbf{a} \in D} p(c|\mathbf{a})}. \tag{9}$$

At each iteration of the algorithm, we combine the expectation and maximization steps for each sequence of actions $\mathbf{a}$ in such a way that we avoid storing, in the E-step, the exponential number of probabilities of all the possible sequences of stages and classes for all the dataset D. In addition, note that the dynamic programming based method (Appendix A) allows the EM algorithm to be solved considering the exponential number of sequences of stages with a computational complexity of $O(N \cdot m^2)$, where m is the length of the longest sequence of actions.

The large amount of possibilities in the combination of pairs of sequences of actions and stages creates problems of sparsity in the Markov models. Once the maximum likelihood estimation of the parameters assigns zero probability to some transition, there is no possibility to obtain in the subsequent step a different value for that pair of action-stages. We solve this problem by smoothing the parameters of the Markov models in each iteration of the EM algorithm.

For the sake of simplicity, we explain the classes of treatments with a fixed number of stages. This way, the notation is simplified and it is easier to understand the main idea of the model. However, it is possible to define a more flexible model in terms of stages. It may be the case that some sequences are incomplete because the treatment of a patient is still in progress by the closing date of the dataset. With this flexibility, the model manages to segment the complete sequences into the maximum number of stages r^+, but also the incomplete sequences into a lower number of stages, ranging from r^- to r^+.

2.4. Inference on latent classes and stages

Given the proposed model and the observed sequences of actions, we can efficiently make inference regarding the latent classes and stages by means of the dynamic programming based algorithm (Appendix A) in spite of their exponential number of configurations. In this way, we can compute:

- The probability of the latent classes given a sequence of actions $p(c|\mathbf{a})$ or the entire dataset $p(c)$.
- The probability of a latent sequence of stages given a sequence of actions and a class, $p(\mathbf{s}|\mathbf{a}, c)$.
- The probability of being in each latent stage of a class at each time point given the observed sequences of actions, that is, $p(s_t = s|\mathbf{a}, c)$ for $t = 1, \dots, m_{\mathbf{a}}$.
- The probability of a sequence of actions given a class, $p(\mathbf{a}|c)$.
- The probabilities $p(s_t, c|\mathbf{a})$ and $p(s_t, c|\mathbf{a}, s_{t-1})$ computed in the EM algorithm (Eq. (7) and (8)) for the parameter estimation.
- Expectations such as $\mathbb{E}_{p(\mathbf{s},c|\mathbf{a};\theta')}[\log p(\mathbf{a}, \mathbf{s}, c; \theta)]$.

Subsequently, these inferences can be used to find the most probable latent class for each sequence of actions, and group together those with common evolution patterns. In addition, in order to show the general behavior of a class, the groups can be represented by the most probable sequences of actions. All these probabilities are calculated with a polynomial time complexity using the dynamic programming based method.

3. Experimental results

In this section we empirically show two types of results. Firstly, we use synthetic datasets of different sizes to evaluate the behavior of the learning algorithm by comparing the learned models with the original generative model underlying the data. The corresponding source code is publicly available.[1] Secondly, we apply the model on real-world EHRs about breast cancer disease to classify the sequences of actions and segment them in different progression stages.

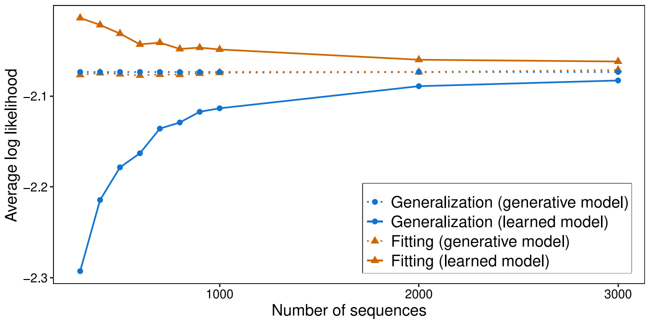


Fig. 2. Fitting and generalization of synthetic models.

3.1. Synthetic data

We firstly create a probabilistic generative model p, whose parameters are generated as follows: $p(c)$ is sampled from a uniform Dirichlet distribution with parameters $\alpha_c = 1$ for $c \in C$; $p(a'|a, s, c)$ is also sampled from a uniform Dirichlet distribution with parameters $\alpha = 1$ for $a, a' \in \mathcal{A}$, $s \in \mathcal{S}$ and $c \in C$; and $p(s'|a, s, c)$ is sampled from a Dirichlet distribution setting $\alpha = 0.7$ for the parameters whose corresponding transition stays in the same stage $(s' = s)$ and setting $\alpha = 0.3$ for those that progress to a different stage $(s' \neq s)$, for $a \in \mathcal{A}$, $s, s' \in \mathcal{S}$ and $c \in C$. The fundamental reason for setting a lower value when the transition progresses to a different stage is to generate more realistic phases by avoiding subsequences of stages which are too short.

For the sake of simplicity, we fix the total number of classes $|C| = 3$, the minimum number of stages $r^- = 3$, and the maximum number of stages $r^+ = 4$ to sample the training sets of sizes $N = \{300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000\}$ using the randomly generated model p (see Appendix B for more details about the training sets). In particular, we use 10 unique actions to generate these sequences. Apart from that, we also sample a test set of 4000 sequences from p in order to evaluate the learning process.

The objective is to show that the proposed learning algorithm is able to recover the generative model. Therefore, we fit the model on the training sets using the EM-based procedure proposed in Section 2.3 and we then analyze the evolution of the quality of the learned models as the training set size N increases. For each value of N we obtain a new model $\theta = \{\theta_c, \theta_a, \theta_s\}$ and we measure the quality of such a model by using the log likelihood of (5) normalized by N to make the datasets comparable.

The experiment is carried out five times, considering in each of them a different random generative model p, from which the N training sets and the test sets are generated. Fig. 2 shows the fitting and generalization ability of our model by means of the average log likelihood. The average log likelihood of the learned models on the training sets (solid orange line) quantifies the fitting of the models to the data, while on the test set (solid blue lines) it measures its ability of generalization. The dotted lines correspond to the average log likelihood of the 5 original generative models evaluated in the training (orange) and test (blue) datasets. We can see that as N increases, the curves that quantify the fitting and generalization of the learned models converge to the curves of the original generative models. This means that, given a sufficiently large dataset, the proposed learning algorithm recovers the original generative model underlying the data.

[1] https://github.com/onintzezaballa/ProbGenerativeModel

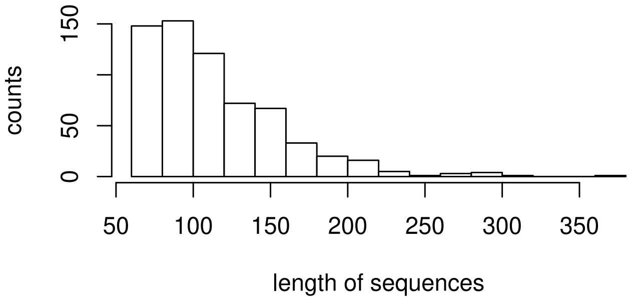


Fig. 3. Histogram of the lengths of the real sequences of actions.

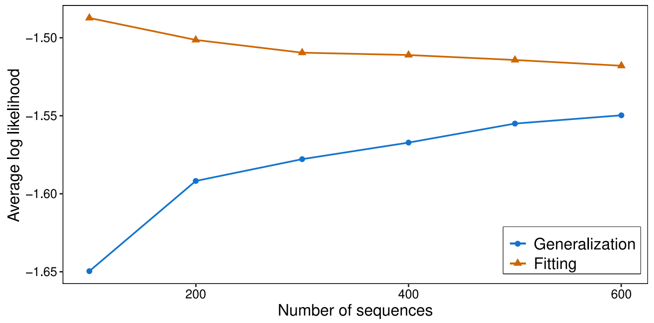


Fig. 4. Fitting and generalization of the breast cancer generative model.

3.2. Breast cancer data

Here we show the application of the model on a real-world dataset of breast cancer, where we represent the classification and stage progression of the sequences of actions associated with such pathology. The achieved results were compared with clinical guidelines [30] and discussed in detail with physicians to check their coherence and validity.

3.2.1. Dataset

We use a dataset provided by the public health care system of the Basque Country, Spain. This dataset records the sequences of actions of patients for any diagnosed disease from 2016 to 2019. As a case of use, we focus our attention on the breast cancer treatment population as in [31]. The dataset contains complete and incomplete sequences of actions. Therefore, individuals with treatments which have already started are excluded from this study, however, those that continue their treatments are included. The resulting dataset consists of 645 sequences of actions, whose average length is of 115 actions, the minimum sequence length is 63 and the maximum is 369 (see Fig. 3 for more details). They are generated by 23 unique medical actions (Appendix C), whose frequency in patients and their transition frequency are shown in Appendix D.

3.2.2. Hyperparameters

The hyperparameters (classes, minimum stage and maximum stage) of the model are set before the learning procedure. Regarding the class, we use the method developed in [31] to appropriately pick the number of different classes of treatments and initialize in the same group those with similar trajectories. We obtain a total of 5 classes of treatments and we set the minimum and maximum stages as $r^- = 3$ and $r^+ = 4$ respectively.

We replicate the experiment of Section 3.1 with the breast cancer dataset. In this case we randomly create the training sets of sizes $N = \{100, 200, 300, 400, 500, 600\}$, leaving 45 sequences of actions out to create the test set. Fig. 4 shows the results of 5 experiments where the generalization curve and the fitting curve of the models converge to the same point. Therefore, we can conclude that the size of the dataset is large enough to learn the generative model, and the hyperparameters chosen beforehand are appropriate for the breast cancer dataset, as well as the smoothing parameter with value 0.2.

3.2.3. Analysis of breast cancer treatments

The first application of the generative model is the representation of the evolution of the breast cancer disease, by classifying the different sequences of actions and identifying their multiple phases of progression over time.

Considering the hyperparameters of the previous section and randomly initializing the sequences of stages, we trained the model using the EM-based procedure described in Section 2.3. The classification of sequences of actions is carried out by associating each sequence of actions $\mathbf{a}$ with the most probable class c^* (Section 2.4), that is,

$$c^* = \underset{c}{\operatorname{argmax}}\, p(c|\mathbf{a}). \tag{10}$$

The evolution patterns of the sequences of actions of each class are characterized by a representative sequence. This is defined as the most probable sequence of actions $\mathbf{a}$ within each class (Section 2.4) normalized by the length of $\mathbf{a}$, in order to avoid the probability $p(\mathbf{a}|c)$ to exponentially decrease as long as the length of $\mathbf{a}$ increases. That is,

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmax}}\, \frac{\log p(\mathbf{a}|c)}{|\mathbf{a}|}. \tag{11}$$

Finally, the sequence of stages associated with the representative sequence $\mathbf{a}^*$ is given by the most probable stage at each time point (Section 2.4), that is,

$$s_t^* = \underset{s\in S}{\operatorname{argmax}}\, p(s_t = s|\mathbf{a}^*, c^*) \tag{12}$$

in such a way that the representative sequence of stages associated with the representative sequences of actions $\mathbf{a}^*$ is $\mathbf{s}^* = (s_1^*, \dots, s_m^*)$.

We show in Fig. 5 the five representative breast cancer treatments (sequences of actions) that characterize the progression classes and stages. The width of the horizontal lines refers to the size of the groups. The vertical lines refer to the medical actions ordered in time. To get a better insight into the behavior of the sequences of actions, we explain the major patterns of the representative treatments, which are real sequences of actions from EHRs, as follows (see Table 1).

To begin with, the diagnosis of breast cancer is based on clinical examination in combination with imaging and confirmed by pathological assessment [30]. Every class of treatments in Stage 1 includes this diagnosis process (performed on radiology, nuclear medicine and pathological anatomy medical services), and before any type of treatment is initiated, as recommended.

There exist two types of surgeries when it comes to breast cancer: breast-conserving surgery, in which the surgical team removes the tumor but tries to keep as much of the breast as possible (it is the preferred local treatment option for the majority of early breast cancer patients); or mastectomy, in which the whole breast is removed [30].

Group 1: Surgery + Chemotherapy + Radiotherapy (166 patients, 25.7%). The vast majority of these sequences of actions undergo breast-conserving surgery (Stage 1), followed by chemotherapy (Stage 2) and radiotherapy (Stage 3). According to the guideline suggestions, if both therapies are used, chemotherapy should usually precede radiotherapy, as done here. This type of treatment used after primary treatments, such as surgery, is called adjuvant treatment and its aim is to decrease the chance of cancer recurrence. Some of these patients also include adjuvant hormonal therapy in their Stage 4.

Group 2: Surgery + Radiotherapy (134 patients, 20.7%). The sequences of actions in this group begin with breast-conserving surgery (Stage 1). This is followed by radiation therapy (Stage 2), which is

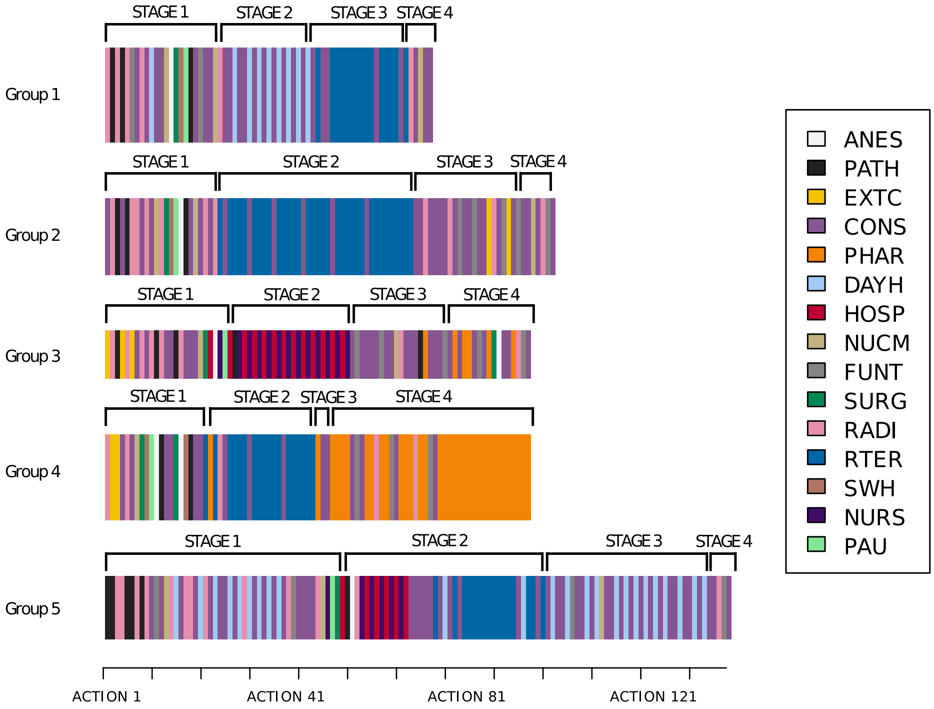


Fig. 5. Representative treatments of breast cancer segmented in the different phases of evolution. The medical actions of the legend are anesthesia (ANES), pathological anatomy (PATH), external consultation (EXTC), consultation (CONS), pharmacy (PHAR), day hospital (DAYH), hospitalization (HOSP), nuclear medicine (NUCM), functional testing (FUNT), surgery unit (SURG), radiology (RADI), radiotherapy (RTER), surgery without hospitalization unit (SWH), nursing unit (NURS), and post anesthesia care unit (PAU).

Table 1
Evolution patterns of the breast cancer treatments obtained from the learned generative model.

	N	Stage 1	Stage 2	Stage 3	Stage 4
Group 1	25.7%	Medical examinations Diagnostic tests Surgery	Chemotherapy	Radiotherapy	Medical examinations Diagnostic tests
Group 2	20.7%	Medical examinations Diagnostic tests Surgery	Radiotherapy	Medical examinations Diagnostic tests	Medical examinations
Group 3	13.1%	Medical examinations Diagnostic tests Surgery	Hospitalization	Medical examinations Diagnostic tests	Hormonal therapy Surgery
Group 4	23.3%	Medical examinations Diagnostic tests Surgery	Radiotherapy	Hormonal therapy	Hormonal therapy Medical examinations Diagnostic tests
Group 5	17.2%	Medical examinations Diagnostic tests Chemotherapy Surgery	Radiotherapy Hospitalization	Chemotherapy Diagnostic tests	Medical examinations Diagnostic tests

highly recommended after this type of surgery by the medical guidelines. Regular follow-up actions are given in Stages 3 and 4.

Group 3: Surgery + Hospitalization + Hormonal Therapy (84 patients, 13.1%). This group represents patients undergoing mastectomy (Stage 1). Hospitalization actions (Stage 2) and additional surgical events (Stage 4) are due to breast reconstruction. These patients are followed up with diagnostic tests and physical examinations in Stage 3. Finally, they have hormonal therapy as adjuvant treatment (Stage 4).

Group 4: Surgery + Radiotherapy + Hormonal Therapy (150 patients, 23.3%). Individuals in this group undergo breast-conserving surgery (Stage 1) and postoperative radiotherapy (Stage 2), as suggested. Additionally, they take hormonal therapy as adjuvant systemic treatment (Stage 3) and followed up with clinical examinations (Stage 4).

Group 5: Chemotherapy + Surgery + Radiotherapy + Chemotherapy (111 patients, 17.2%). Neoadjuvant systemic therapy is treatment administered preoperatively to reduce the extent of surgery in locally advanced and large operable cancers. This is the case for this group of patients, who receive neoadjuvant chemotherapy before breast-conservative surgery or mastectomy (Stage 1). Afterwards, they complete their adjuvant treatment with radiotherapy (Stage 2) and chemotherapy (Stage 3). They are followed up in Stage 4.

See Appendix D for more details about the behavior of the medical actions within each class of treatments.

4. Discussion

The main contribution of this paper is the development of a novel probabilistic generative model, which characterizes the progression of the treatment trajectories of a disease. State-of-the-art disease progression approaches [5–8,11–21] partially adopt the main properties of our model, which we consider essential in order to describe and understand the behavior of the treatment trajectories. In particular, our model simultaneously classifies the heterogeneous sequences of actions based on their treatment evolution over time, segments the sequences of actions in different progression stages of the disease, and captures the sequential dependence between medical actions.

Another contribution of this work is the proposal of an efficient learning process of the parameters of the model to make the computation of the EM algorithm feasible. Exact inference often requires high computational cost for learning, in fact, an *ad hoc* algorithm would require an exponential complexity. We propose a generalization of the forward–backward algorithm for the learning process to reduce this complexity to be polynomial.

Experiments on synthetic datasets validate that our model converges to the original model underlying the data. On the other hand, the breast cancer experiment shows the ability of the model to discover different treatment progression patterns and their temporal evolution by means of explainable treatment stages. Nevertheless, we intend to validate the model on an external dataset to increase the robustness of the results. Treatment subtyping and phase identification are useful to extract potential information, such as essential or critical treatment behaviors and their causal dependencies in treatment sequences, as well as to understand disease mechanisms and health practices. Apart from classification and segmentation of treatment trajectories, another benefit of our model is the simulation of fictitious sequences of actions that resemble original treatments. This can have an economic impact on the healthcare system by assisting in resource management and anticipating the expected costs of treatments [2]. The model can be also regarded as a data augmentation tool when little information is available, for example, for rare diseases. In addition to this, since healthcare datasets are frequently incomplete and the removal of missing values may result in a dataset that is too small or induce statistical bias [1], the model has the ability to impute such missing values in the trajectories of patients or reconstruct incomplete sequences of actions. In terms of interpretability, our model provides easier comprehension and explanation for end-users than other approaches developed in the healthcare setting [22–27].

Let us also mention some limitations of our approach. The stages are defined as ordered discrete values of progression and in their evolution only two steps are allowed: to be increased in one stage with respect to the previous stage; or be maintained in the same one. In a more realistic scenario, diseases with recurrent stages would be considered, and, consequently, the sequences of actions could pass through the same stage more than once or move from one stage to another without setting an ordered progression. However, this assumption requires a modification in the dynamic programming procedure that would exponentially increase the complexity of the model. On the other hand, as in many other classification machine learning methods, the number of classes is not a flexible parameter and has to be chosen beforehand. Despite this, we solved this problem by initializing the classes of treatments with a previous clustering of sequences, where the number of classes that best fits the data was selected. For the minimum and maximum stages, we could estimate their value by including them in the learning process of the model, assuming again an increase in its complexity.

Note that this is a memoryless model due to Markovianity assumption. That is, the actions and stages only depend on the previous time instead of depending on the whole or part of the medical history. In some cases the duration in a stage matters, or a medical event could be a result of more than one previous action. In the future we plan to include this relevant information about the past of patients in the model.

For future work, we propose an extension of the generative model by including new features. For instance, irregular timing between medical actions is an interesting task due to its important role in the progression of a disease [32]. Another line of effort is related to the assignment of the stages. In our model the stages of a class are independently defined from the stages of the other classes. However, there may exist a relationship between the stages even if they belong to different classes of treatments. The sequences of actions would be separated into multiple fragments in such a way that the fragments would refer to the common patterns of all the treatments of the disease.

In conclusion, we introduce a model to characterize the treatment variability of a disease. We demonstrate the potential of our approach as a treatment classification and stage identification tool in breast cancer patients. We further validate the proposed learning process by a simulation experiment, where the original model is recovered. Definitely, the proposed method has the capability to make substantial clinical impact and is readily applicable to any progressive disease, such as other types of cancers, respiratory diseases or neurodegenerative diseases.

CRediT authorship contribution statement

Onintze Zaballa: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Funding acquisition. **Aritz Pérez:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Elisa Gómez Inhiesto:** Conceptualization, Resources, Writing – review & editing. **Teresa Acaiturri Ayesta:** Conceptualization, Resources, Writing – review & editing. **Jose A. Lozano:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition.

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.

Acknowledgments

This research is supported by the Basque Government through the BERC 2022–2025 program and BMTF projects, and by the Ministry of Science, Innovation and Universities: BCAM Severo Ochoa accreditation SEV-2017-0718 and PID2019-104966GB-I00. Onintze Zaballa also holds a predoctoral grant (EJ-GV 2019) from the Basque Government.

Appendix A. Efficient inference based on dynamic programming

Training a generative model is a challenging task. In fact, the exact learning of the parameters of the model is computationally expensive for large datasets and long sequences. The forward–backward algorithm proposes a dynamic programming based method to compute the posterior marginal distribution of hidden states given a sequence of observations in HMMs. Following the same strategy, we develop a dynamic programming method for the specific characteristics of our generative model, reducing the number of computations, and thus, the complexity of our approach. This inference plays an important role in the learning procedure of the model, particularly in the E-step. These results are then used in the M-step to solve (7), (8) and (9) so that we update the parameters of the model $\theta = \{\theta_c, \theta_a, \theta_s\}$.

Let us assume that we have a training set $\mathcal{D}$ of sequences of actions $\mathbf{a} = (a_1, \ldots, a_m)$, a latent variable of stages $\mathbf{s} = (s_1, \ldots, s_m)$ and a latent variable of classes c. Remember that we aim to estimate the maximum likelihood parameters θ of the model in each iteration of the EM algorithm. Hence, we aim to learn a model $p(a|a', s', c)$ and $p(s|a, s', c)$ for any value of $a, a' \in \mathcal{A}$ $s, s' \in \mathcal{S}$ and $c \in C$, using the set of sequences of actions in $\mathcal{D}$. Suppose that, for a sequence of actions $\mathbf{a}$, we observe the transition $a_{t-1} = a$ to $a_t = a'$ in the training set. Now, we shall calculate the sum of the probabilities of all the possible sequences of stages for which $s_t = s$ in each class c. That is, the probability of all the sequences of stages with the form $(s_1, \ldots, s_{t-2}, s_{t-1}, s', s_{t+1}, \ldots, s_m)$ in c.

Let us assume that $f_c(t,s)$ is the sum of the probabilities of all the sequences of stages $(s_1, \ldots, s_t)$ in the class c that ends at $s_t = s$, and $g_c(t,s)$ is the sum of the probabilities of all the sequences of stages $(s_{t+1}, \ldots, s_m)$ that starts at $s_t = s'$ in the class c. Then,

$$f_c(t,s) = \sum_{\mathbf{s}_{1\ldots t-2}} p(\mathbf{s}_{1\ldots t-2}, s|\mathbf{a}_{1\ldots t-1}, c) \cdot p(s'|a_t, a_{t-1}, s, c)$$

$$g_c(t,s) = \sum_{\mathbf{s}_{t+1,\ldots,m}} p(\mathbf{s}_{t+1\ldots m}|\mathbf{a}_{t,\ldots,m}, s', c),$$

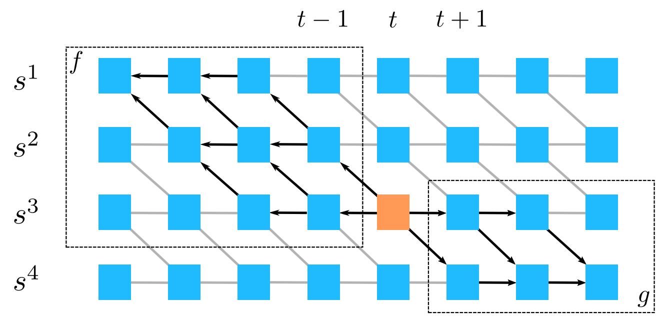


Fig. A.6. Dynamic programming procedure developed to learn the parameters of the Markov model. In this case, the orange box represents (A.1), and f and g correspond to the recursive functions. The black arrows generate all the possible sequences of stages that pass through the orange box. Note that in this example the maximum stage r^+ is the same as the minimum stage r^-.

where $\mathbf{a}_{i,\dots,j} = (a_i, \dots, a_j)$ and $\mathbf{s}_{i,\dots,j} = (s_i, \dots, s_j)$. Note that f_c includes the term $p(s'|a_t, a_{t-1}, s, c)$ and the probability in g_c is conditioned to $(\mathbf{a}_{t,\dots,m}, s', c)$.

Now, we can express the sum of the probabilities of the sequences for which $\mathbf{s}_{t-1,t} = (s, s')$ as

$$\begin{aligned} &p(\mathbf{s}_{1,\dots,t-2}, s_{t-1} = s, s_t = s', \mathbf{s}_{t+1,\dots,m} | \mathbf{a}_{1,\dots,m}, c) \\ &\quad = f_c(t-1, s) \cdot p(s'|a_t, a_{t-1}, s, c) \cdot g_c(t, s') \\ &\quad = \sum_{\mathbf{s}_{1,\dots,t-2,t+1,\dots,m}} p(\mathbf{s}_{1,\dots,t-2}, s_{t-1} = s | \mathbf{a}_{1,\dots,t-1}, c) \cdot \\ &\quad\quad \cdot p(s_t = s' | a_t, a_{t-1}, s_{t-1} = s, c) \cdot \\ &\quad\quad \cdot p(\mathbf{s}_{t+1,\dots,m} | \mathbf{a}_{t,\dots,m}, s_t = s', c) \end{aligned} \tag{A.1}$$

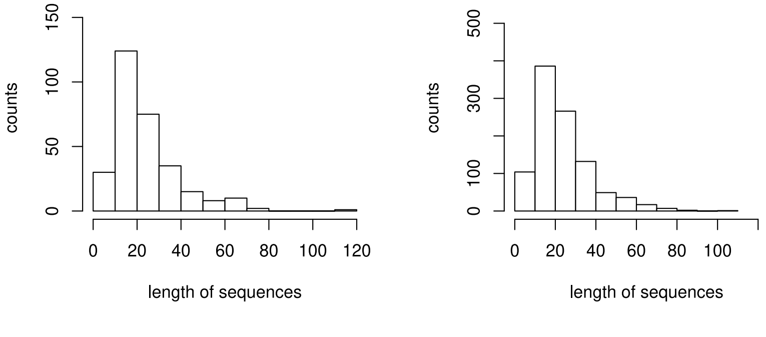
With this in mind, we propose to create a matrix associated with each function f and g. These functions are defined as recursive functions (Fig. A.6):

$$\begin{aligned} f_c(t, s) =& p(s|a_t, a_{t-1}, s, c) \cdot f_c(t-1, s) + \\ & p(s|a_t, a_{t-1}, s-1, c) \cdot f_c(t-1, s-1) \\ g_c(t, s) =& p(s+1|a_t, a_{t+1}, s, c) \cdot g_c(t+1, s+1) + \\ & p(s|a_t, a_{t+1}, s, c) \cdot g_c(t+1, s) \end{aligned}$$

The functions f_c and g_c are defined in such a way that the stages are non-decreasing. By means of dynamic programming, we complete the matrices f and g and, consequently, reduce the number of computations for the parameter estimation. Intuitively, instead of calculating the probability of all the possible $(\mathbf{a}, \mathbf{s}, c)$ independently one by one, dynamic programming reuses those probabilities of transition that the sequences of actions-stages share.

Appendix B. Heterogeneity on synthetic sequences

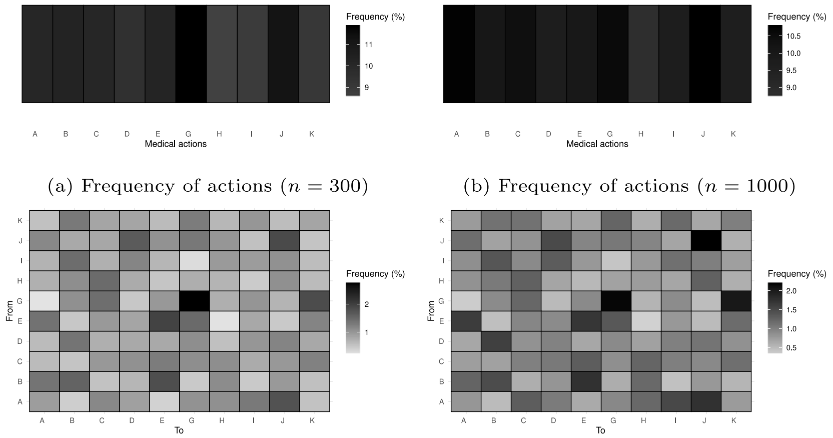
In this appendix we aim to show the variability of the synthetic sequences generated for the experiments in Section 3.1. For each experiment we represent the distribution of the lengths of the sequences, the frequency of actions and the frequency of the transition between actions for two different sizes of the dataset (see Figs. B.7–B.16).



(a) $n = 300$

(b) $n = 1000$

Fig. B.7. Experiment 1: histogram of the lengths of the sequences of actions.



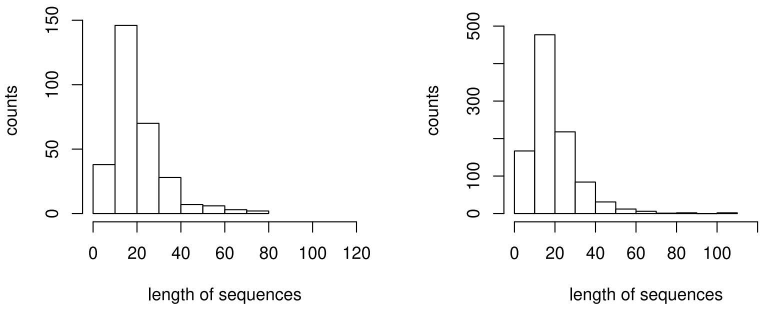
(a) Frequency of actions ($n = 300$)

(b) Frequency of actions ($n = 1000$)

(c) Frequency of transitions ($n = 300$)

(d) Frequency of transitions ($n = 1000$)

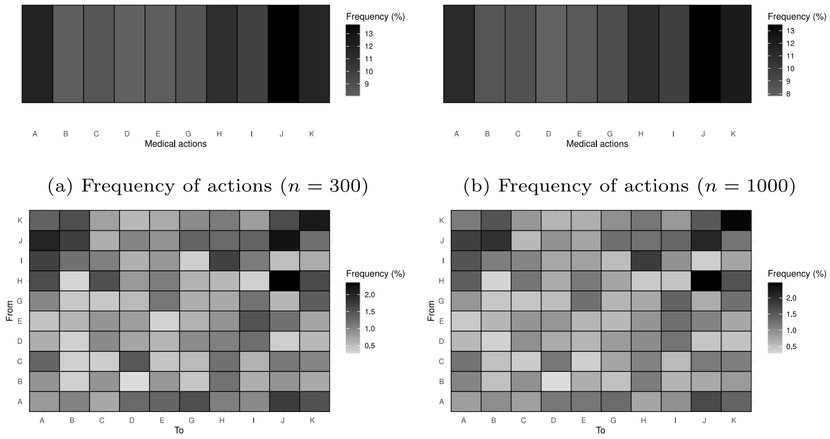
Fig. B.8. Experiment 1: frequency of actions and their transitions.



(a) $n = 300$

(b) $n = 1000$

Fig. B.9. Experiment 2: histogram of the lengths of the sequences of actions.



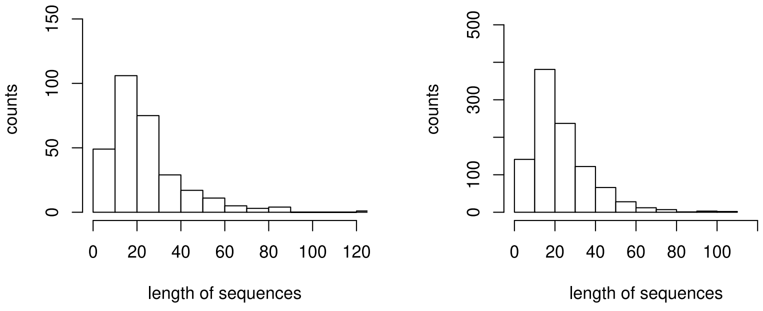
(a) Frequency of actions ($n = 300$)

(b) Frequency of actions ($n = 1000$)

(c) Frequency of transitions ($n = 300$)

(d) Frequency of transitions ($n = 1000$)

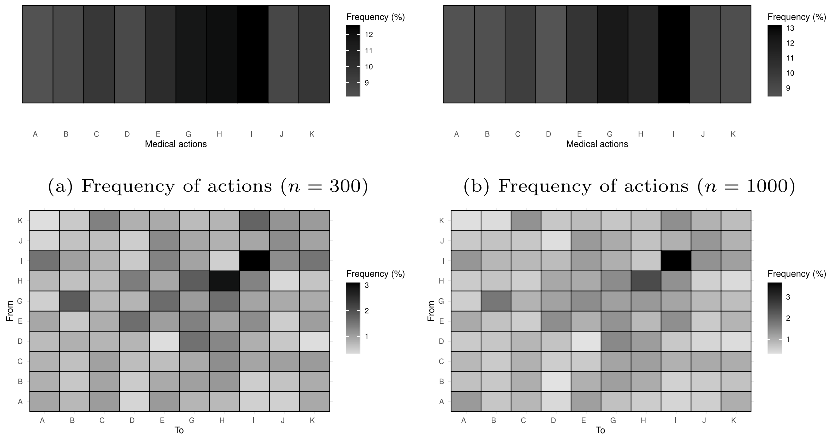
Fig. B.10. Experiment 2: frequency of actions and their transitions.



(a) $n = 300$

(b) $n = 1000$

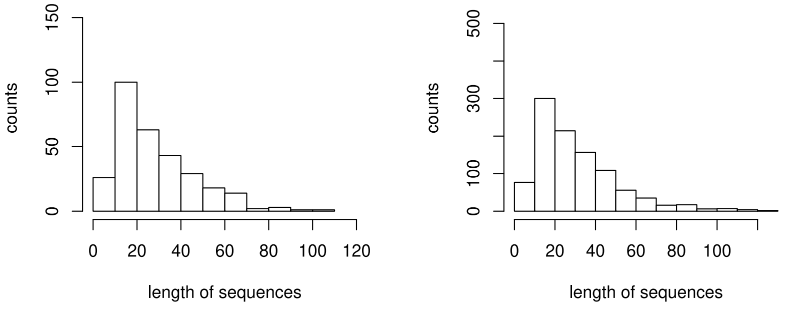
Fig. B.11. Experiment 3: histogram of the lengths of the sequences of actions.



(a) Frequency of actions ($n = 300$) (b) Frequency of actions ($n = 1000$)

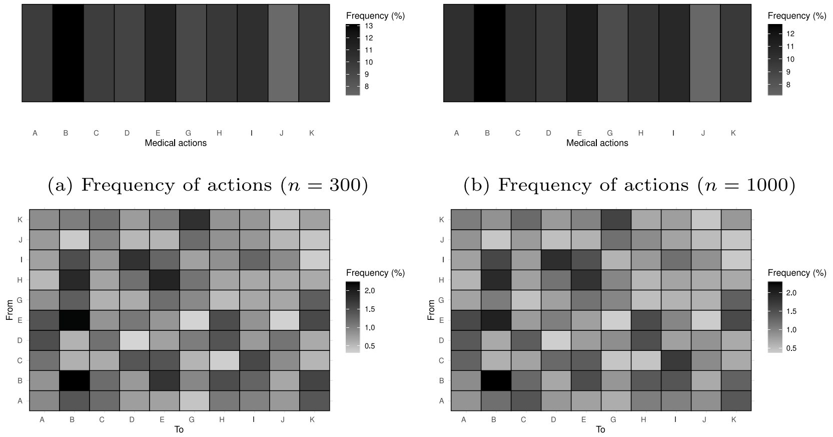
(c) Frequency of transitions ($n = 300$) (d) Frequency of transitions ($n = 1000$)

Fig. B.12. Experiment 3: frequency of actions and their transitions.



(a) $n = 300$ (b) $n = 1000$

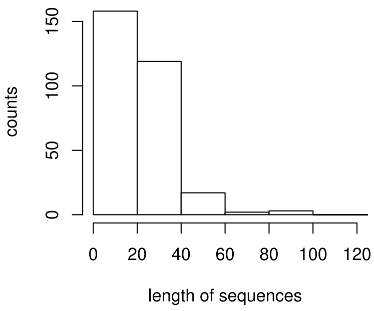
Fig. B.13. Experiment 4: histogram of the lengths of the sequences of actions.



(a) Frequency of actions ($n = 300$) (b) Frequency of actions ($n = 1000$)

(c) Frequency of transitions ($n = 300$) (d) Frequency of transitions ($n = 1000$)

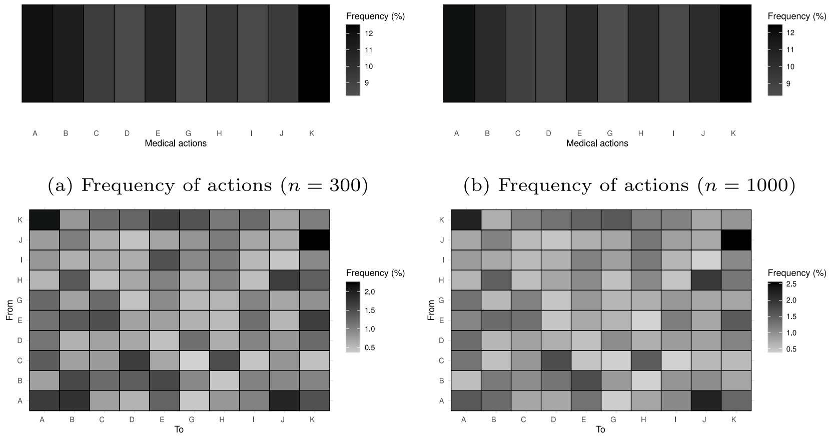
Fig. B.14. Experiment 4: frequency of actions and their transitions.



(a) $n = 300$

(b) $n = 1000$

Fig. B.15. Experiment 5: histogram of the lengths of the sequences of actions.



(a) Frequency of actions ($n = 300$)

(b) Frequency of actions ($n = 1000$)

(c) Frequency of transitions ($n = 300$)

(d) Frequency of transitions ($n = 1000$)

Fig. B.16. Experiment 5: frequency of actions and their transitions.

Appendix C. Description of the medical actions of real EHRs

In this appendix we explain the abbreviation and description of each medical action of the real breast cancer data (see Table C.2).

Table C.2
Description of the medical actions.

Abbreviated form	Full form
ANES	Anesthesia
CONS	Consultation
DAYH	Day Hospital
EXTC	External Consultation
FUNT	Functional Testing
HCRI	Critical Care Hospitalization
HOMEH	Home Hospitalization
HOSP	Hospitalization
INCO	Interconsultation
LABO	Laboratory
NUCM	Nuclear Medicine
NURS	Nursing Unit
OSAT	Osatek (Magnetic Resonance Service)
PATH	Pathological Anatomy
PAU	Post Anesthesia Care Unit
PHAR	Pharmacy
PHARAMB	Hospital Pharmacy Services
RADI	Radiology
REHA	Rehabilitation
RTER	Radiotherapy
SURG	Surgery Unit
SWH	Surgery without Hospitalization
UCRI	Nursing Critical Care Unit

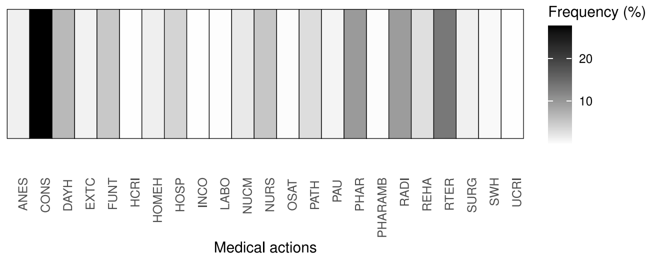


Fig. D.17. Frequency (%) of medical actions in real EHRs.

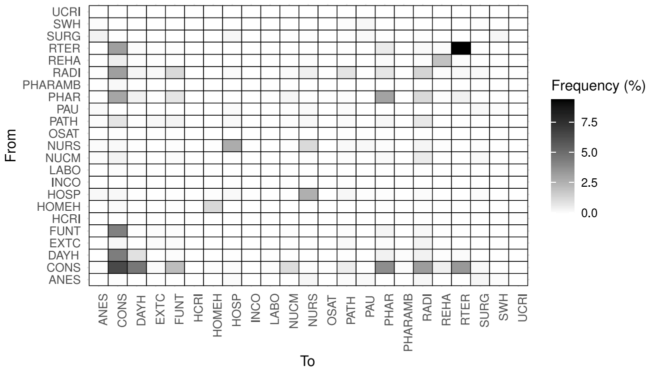
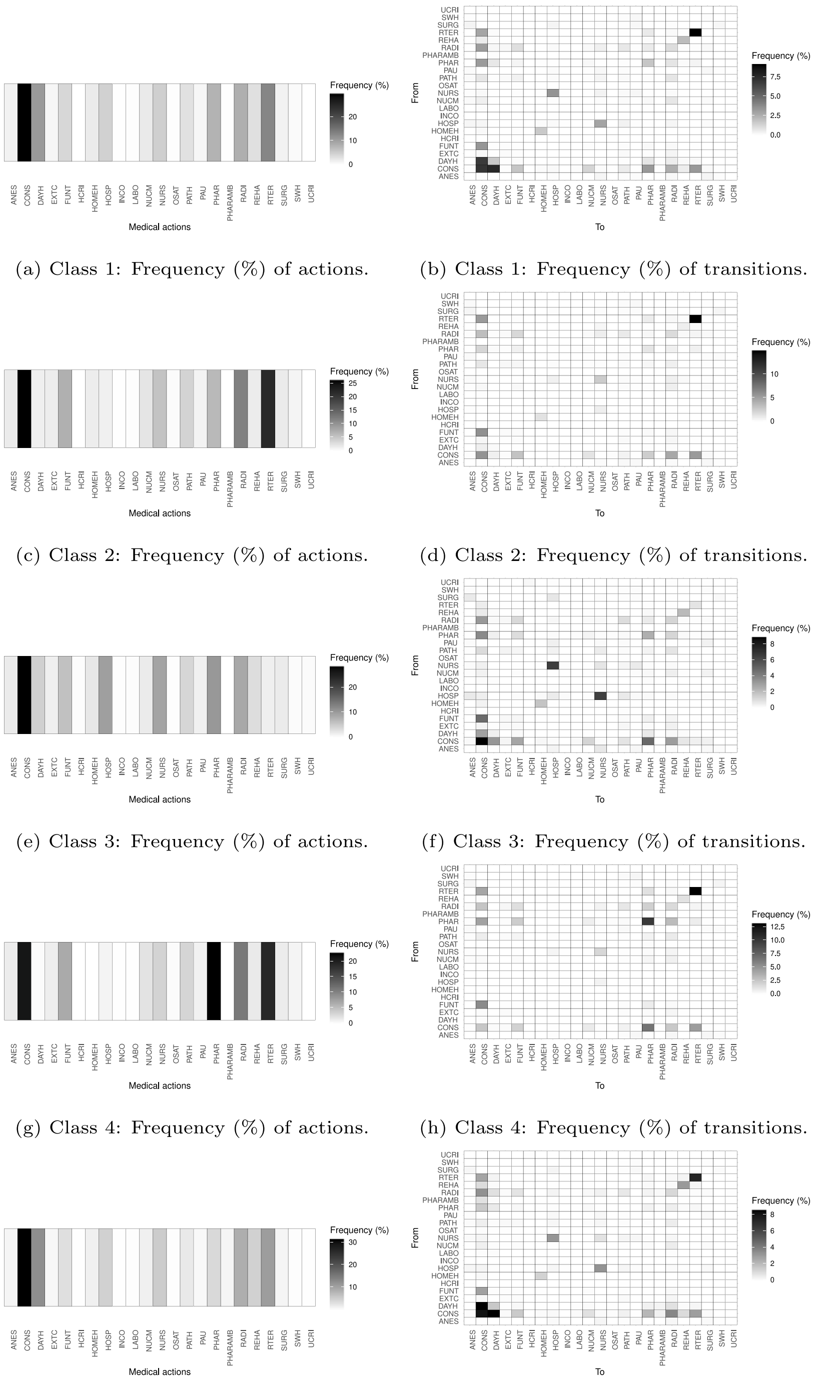


Fig. D.18. Frequency (%) of the transitions between medical actions in real EHRs.



(a) Class 1: Frequency (%) of actions.

(b) Class 1: Frequency (%) of transitions.

(c) Class 2: Frequency (%) of actions.

(d) Class 2: Frequency (%) of transitions.

(e) Class 3: Frequency (%) of actions.

(f) Class 3: Frequency (%) of transitions.

(g) Class 4: Frequency (%) of actions.

(h) Class 4: Frequency (%) of transitions.

(i) Class 5: Frequency (%) of actions.

(j) Class 5: Frequency (%) of transitions.

Fig. D.19. Frequency (%) of actions and their transitions of each class.

Appendix D. Heterogeneity in sequences of real EHRs

In this appendix we firstly show the frequency of medical actions and their transitions in real EHRs (see Figs. D.17 and D.18). Secondly, we represent these frequencies within each class of treatments that we obtained in the experiment of Section 3.2.3.

D.1. Inter-class heterogeneity

The objective of Fig. D.19 is to show the variety of medical actions that can typically be executed for each class, as well as the transition between them.

References

[1] T. Sarwar, S. Seifollahi, J. Chan, X. Zhang, V. Aksakalli, I. Hudson, K. Verspoor, L. Cavedon, The secondary use of Electronic Health Records for data mining: Data characteristics and challenges, ACM Comput. Surv. 55 (2) (2022) 1–40.

[2] S. Saria, A. Goldenberg, Subtyping: What it is and its role in precision medicine, IEEE Intell. Syst. 30 (4) (2015) 70–75.

[3] E. Rojas, J. Munoz-Gama, M. Sepúlveda, D. Capurro, Process mining in healthcare: A literature review, J. Biomed. Inform. 61 (2016) 224–236.

[4] J. Munoz-Gama, N. Martin, C. Fernandez-Llatas, O.A. Johnson, M. Sepúlveda, E. Helm, V. Galvez-Yanjari, E. Rojas, A. Martinez-Millana, D. Aloini, et al., Process

mining for healthcare: Characteristics and challenges, J. Biomed. Inform. 127 (2022) 103994.

[5] Y. Wang, Y. Zhao, T.M. Therneau, E.J. Atkinson, A.P. Tafti, N. Zhang, S. Amin, A.H. Limper, S. Khosla, H. Liu, Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records, J. Biomed. Inform. 102 (2020) 103364.

[6] Z. Huang, W. Dong, L. Ji, C. Gan, X. Lu, H. Duan, Discovery of clinical pathway patterns from event logs using probabilistic topic models, J. Biomed. Inform. 47 (2014) 39–57.

[7] Z. Huang, W. Dong, P. Bath, L. Ji, H. Duan, On mining latent treatment patterns from electronic medical records, Data Min. Knowl. Discov. 29 (4) (2015) 914–949.

[8] J. Chen, L. Sun, C. Guo, Y. Xie, A fusion framework to extract typical treatment patterns from electronic medical records, Artif. Intell. Med. 103 (2020) 101782.

[9] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.

[10] K.A. Severson, L.M. Chahine, L. Smolensky, K. Ng, J. Hu, S. Ghosh, Personalized input-output hidden markov models for disease progression modeling, in: Machine Learning for Healthcare Conference, PMLR, 2020, pp. 309–330.

[11] A.M. Alaa, M. van der Schaar, Attentive state-space modeling of disease progression, Adv. Neural Inf. Process. Syst. 32 (2019).

[12] X. Wang, D. Sontag, F. Wang, Unsupervised learning of disease progression models, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 85–94.

[13] R. Sukkar, E. Katz, Y. Zhang, D. Raunig, B.T. Wyman, Disease progression modeling using hidden Markov models, in: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2012, pp. 2845–2848.

[14] C.H. Jackson, L.D. Sharples, S.G. Thompson, S.W. Duffy, E. Couto, Multistate Markov models for disease progression with classification error, J. R. Stat. Soc. Ser. D (the Statistician) 52 (2) (2003) 193–209.

[15] Z. Huang, Z. Ge, W. Dong, K. He, H. Duan, Probabilistic modeling personalized treatment pathways using electronic health records, J. Biomed. Inform. 86 (2018) 33–48.

[16] Y.-Y. Liu, S. Li, F. Li, L. Song, J.M. Rehg, Efficient learning of continuous-time hidden markov models for disease progression, Adv. Neural Inf. Process. Syst. 28 (2015).

[17] N. Galagali, M. Xu-Wilson, Patient subtyping with disease progression and irregular observation trajectories, 2018, arXiv preprint arXiv:1810.09043.

[18] J. Yang, J. McAuley, J. Leskovec, P. LePendu, N. Shah, Finding progression stages in time-evolving event sequences, in: Proceedings of the 23rd International Conference on World Wide Web, 2014, pp. 783–794.

[19] T. Ceritli, A.P. Creagh, D.A. Clifton, Mixture of input-output hidden Markov models for heterogeneous disease progression modeling, in: Workshop on Healthcare AI and COVID-19, PMLR, 2022, pp. 41–53.

[20] X. Teng, S. Pei, Y.-R. Lin, Stocast: Stochastic disease forecasting with progression uncertainty, IEEE J. Biomed. Health Inf. 25 (3) (2020) 850–861.

[21] T. Pham, T. Tran, D. Phung, S. Venkatesh, Predicting healthcare trajectories from medical records: A deep learning approach, J. Biomed. Inform. 69 (2017) 218–229.

[22] P. Nguyen, T. Tran, N. Wickramasinghe, S. Venkatesh, Deepr: a convolutional net for medical records, IEEE J. Biomed. Health Inf. 21 (1) (2016) 22–30.

[23] E. Choi, M.T. Bahadori, A. Schuetz, W.F. Stewart, J. Sun, Doctor ai: Predicting clinical events via recurrent neural networks, in: Machine Learning for Healthcare Conference, PMLR, 2016, pp. 301–318.

[24] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Sci. Rep. 6 (1) (2016) 1–10.

[25] A. Rajkomar, E. Oren, K. Chen, A.M. Dai, N. Hajaj, M. Hardt, P.J. Liu, X. Liu, J. Marcus, M. Sun, et al., Scalable and accurate deep learning with electronic health records, NPJ Digit. Medicine 1 (1) (2018) 1–10.

[26] E. Choi, M.T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart, RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, 29, Curran Associates, Inc., 2016.

[27] Z. Che, S. Purushotham, G. Li, B. Jiang, Y. Liu, Hierarchical deep generative models for multi-rate multivariate time series, in: International Conference on Machine Learning, PMLR, 2018, pp. 784–793.

[28] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1) (1977) 1–22.

[29] P.A. Devijver, Baum's forward-backward algorithm revisited, Pattern Recognit. Lett. 3 (6) (1985) 369–373.

[30] F. Cardoso, S. Kyriakides, S. Ohno, F. Penault-Llorca, P. Poortmans, I.T. Rubio, S. Zackrisson, E. Senkus, Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up, Ann. Oncol. 30 (8) (2019) 1194–1220, http://dx.doi.org/10.1093/annonc/mdz173.

[31] O. Zaballa, A. Pérez, E. Gómez Inhiesto, T. Acaiturri Ayesta, J.A. Lozano, Identifying common treatments from Electronic Health Records with missing information. An application to breast cancer, PLoS One 15 (12) (2020) e0244004.

[32] B. Shickel, P.J. Tighe, A. Bihorac, P. Rashidi, Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis, IEEE J. Biomed. Health Inf. 22 (5) (2017) 1589–1604.