A Probabilistic Generative Model to Discover the Treatments of Coexisting Diseases with Missing Data

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, Bizkaia, Spain*
[b]*Hospital Universitario Cruces, Barakaldo, 48903, Bizkaia, Spain*
[c]*Intelligent Systems Group, Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, Donostia, 20018, Gipuzkoa, Spain*

Abstract

Background and Objective: Comorbidities, defined as the presence of co-existing diseases, progress through complex temporal patterns among patients. Learning such dynamics from electronic health records is crucial for understanding the coevolution of diseases. In general, medical records are represented through temporal sequences of clinical variables together with their diagnosis. However, we consider the specific problem where most of the diagnoses are missing. We present a novel probabilistic generative model with a three-fold objective: (i) identify and segment the medical history of patients into treatments associated with comorbidities; (ii) learn the model associated with each identified disease treatment; and (iii) discover subtypes of patients with similar coevolution of comorbidities. **Methods:** To this end, the model considers a latent structure for the sequences, where patients are modeled by a latent class defined by the evolution of their comorbidities, and each observed medical event of their clinical history is associated with a latent disease. The learning process is performed using an Expectation-Maximization algorithm that considers the exponential number of configurations of the latent variables and is efficiently solved with dynamic programming. **Results:** The evaluation of the method is carried out both on synthetic and real world data: the experiments on synthetic data show that the learning procedure allows the generative model underlying the data to be recovered; the experiments on real medical data show accurate results in the segmentation of sequences into different treatments, subtyping of patients and diagnosis imputation. **Conclusion:** We present an interpretable

generative model that handles the incompleteness of EHRs and describes the different joint evolution of coexisting diseases depending on the active comorbidities of the patient at each moment.

Keywords: Comorbidity Modeling, Electronic Health Records, Latent Variable Model, Markov Model, Probabilistic Generative Model

1. Introduction

Electronic health records (EHRs), which contain large amounts of information about patients and their treatments over time, provide the opportunity to build models which are able to extract knowledge about diseases and their evolution. EHRs can be represented as temporal sequences of clinical variables (e.g., diagnosis, procedures, medical services), where each sequence chronologically collects the information of medical events from a single patient. Due to the increasing availability of medical data, disease progression modeling has attracted a great deal of interest in two broad directions: the discovery of meaningful patterns and intelligible representations of disease dynamics through unsupervised learning [1, 2]; and the prediction of outcomes with labeled information [3, 4, 5, 6].

In this work, we address the specific problem of modeling the joint progression of coexisting diseases when most of the diagnoses in EHRs are missing. Probabilistic models are a practical solution to face this challenge, not only because they can handle missing data, but also because they account for temporal relationships in data and are interpretable models that can extract clinically meaningful representations from the inferred latent variables. In the literature, most probabilistic models developed for disease progression are based on variants of Hidden Markov models [2, 7, 8, 9, 10, 11, 12] or are extensions of latent Dirichlet allocation [13, 14] that capture the evolution of disease trajectories through latent states. While medical events are time-dependent variables, these models generally ignore the direct stochastic dependence between such observations and are limited to modeling sequential correlations of data only through latent states [11].

In general, existing models describe the evolution of single-disease trajectories instead of their evolution in multiple co-existing diseases (comorbidities) settings [8, 10, 15, 16]. Including comorbidities in the structure of the methods is crucial for a detailed insight into the co-occurrence patterns of diseases, and in this sense, there still remains a need for developing an inter-

pretable framework to capture and explain their joint progression patterns [17]. The works that model the coexistence of diseases [2, 9, 13, 14, 18, 19] assume that diagnosis labels are available at each patient visit, which might not be true in reality in the EHRs. Moreover, diagnostic information is recorded at the specific time the diagnosis is reported, however, in the records which follow it might not be specified.

There also exist some comorbidity progression approaches based on deep learning techniques that have been specifically built for predicting future outcomes [20]. Some of them construct comorbidity networks or learn multilevel embeddings of hospital visits to predict the onset of new diseases without providing insights into disease coevolution patterns over time [18, 19, 21, 22]. The main purpose of these latter models is to recognize the underlying structure within each hospital visit rather than identifying the hidden diagnosis of most of the visits based on the dynamics of the clinical history. Some other works have attempted to create interpretable Recurrent Neural Network-based models [4, 5, 23] using attention mechanism to interpret hidden disease dynamics and provide an explanation of their discriminative predictions. In general, these methods are not motivated from a generative perspective and do not face common challenges in the healthcare setting, such as limited data availability, missingness or uncertainty in medical data [16, 17].

We propose a novel probabilistic generative model to address the challenges posed by EHRs, paying special attention to missing data. The objective of such a model is threefold: (i) identify and segment the medical history of patients into treatments associated with each disease they suffer from; (ii) learn the model associated with each identified disease treatment; and (iii) discover subtypes of patients with similar patterns of coevolution of comorbidities. For this purpose, the model considers a latent structure for temporal sequences, where patients are modeled by a latent class defined by the evolution patterns of their comorbidities, and each observed medical event of their clinical histories is associated with a latent diagnosis. In other words, we seek to extract diagnosis-associated subsequences from the complete sequence of medical events (i.e., from the clinical history), where classes represent similar coevolution of these subsequences of latent diagnoses.

The main contributions of this work are as follows:

- We propose a probabilistic generative model of medical treatments for patients suffering from several comorbidities. The model builds on Markov models to capture the transitions between medical events.

- The generative model is trained on EHRs that are characterized by a significant amount of missing data related to the diagnosis variable.
- The model allows different subtypes of patients to be identified according to their evolution patterns of comorbidities and includes a generative submodel for the treatment of each comorbidity.
- We propose an Expectation-Maximization (EM) scheme with a dynamic programming-based method as an efficient learning algorithm for the parameters of the model.

We use synthetic and real-world data to demonstrate the validity and practical significance of the model. The experiments show the ability of our method to model the progression of coexisting diseases and to extract meaningful and individualized representations of the different treatments.

The remainder of this paper is organized as follows. Section II describes the problem and the probabilistic generative model. In Section III, we present the results of the synthetic data experiments that evaluate the performance of the proposed method and the application of the model to real-world EHRs. Section IV discusses the contributions and limitations of our approach and draws the conclusions.

2. Methods

2.1. Problem formulation

A patient's clinical history, denoted by $\mathbf{h}$, is a sequence of medical data collected during repeated hospital visits. Let A be the set of medical activities and D the set of diagnoses, we define a patient's EHRs as

$$\mathbf{h} = (h_1, ..., h_m),$$

where $h_t = (a_t, d_t)$ represents the t-th medical event of the patient, $a_t \in A$ is the medical specialty (for instance, oncology, hematology, cardiology, etc.) attended and $d_t \in D$ the diagnosis/disease, for $t = 1, ..., m$. The sequence of medical specialties $\mathbf{a} = (a_1, ..., a_m)$ is an observable variable, while the sequence of diagnoses $\mathbf{d} = (d_1, ..., d_m)$ is partially observed since it often presents missing values.

The ultimate objective is to capture the different subtypes of joint evolution of comorbidities in EHRs. For that, we first seek to identify and

segment the medical history of patients into treatments associated with each comorbidity $d \in D$. This is not a straightforward task as $\mathbf{d}$ is incomplete (Fig. 1), and therefore, requires to estimate the diagnosis $d_t \in \mathbf{d}$ for each medical specialty visit $a_t \in \mathbf{a}$. Furthermore, the priority of treating a disease or ongoing medical therapies often involves the modification or interruption of other treatments. For instance, the majority of anticancer therapies are associated with some cardiovascular toxicities, ranging from asymptomatic and transient to more clinically significant and long-lasting cardiac events [24]. Depending on the previous existence of cardiovascular diseases and their progression, patients are at higher risk for the development of subsequent cardiovascular injuries (e.g., heart failure), which would lead to closer and more intense monitoring of such pathology and may affect the cancer treatment. This means that the transition dynamics of comorbidities depends not only on the subtype of patient, but also on the coexisting diseases of the patient at each moment.

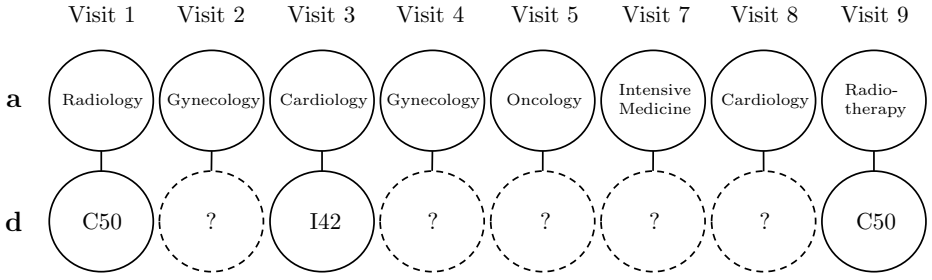


Figure 1: Example of EHRs with missing diagnosis information. ICD-10 code C50 corresponds to breast cancer diagnosis and I42 is a cardiomyopathy diagnosis.

The problem can be seen as an unsupervised classification of a set of treatments with different progression dynamics of their comorbidities.

2.2. Model definition

Our proposed generative model is built on a Markov model that enables the description of the sequential evolution of data through a series of transitions between medical events (see Fig 2). Let $\mathbf{a} = (a_1, ..., a_m)$ be the observed sequence of medical actions that describe a patient's trajectory, where a_t belongs to the set of medical specialties A. We assume that $\mathbf{a}$ has an associated hidden structure of comorbidities that relates medical actions to diseases.

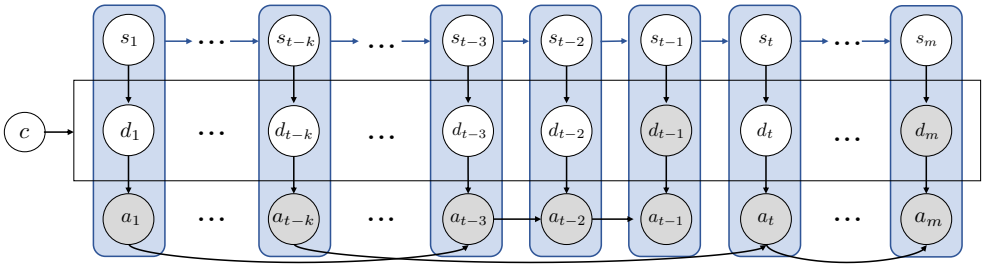


Figure 2: Proposed comorbidity model defined by the conditional distributions $p(a_{d_t:t}|a_{d_t:t'}, d_t)$, $p(d_t|c, s_t)$ and $p(s_t|s_{t-1}, d_{t-1}, a_{t-1})$ for observed sequences of actions $\mathbf{a}$, latent sequence of active disease states $\mathbf{s}$, latent sequences of diseases $\mathbf{d}$ and latent classes c.

This means that a patient trajectory consists of subsequences of medical actions associated with different diseases, $\mathbf{a}_d$ for $d \in D$, and these subsequences are mixed in a way that constitutes the clinical history $\mathbf{h}$. However, extracting the subsequences $\mathbf{a}_d$ for $d \in D$ is not trivial since most of the diagnosis are missing.

In this hidden structure, the presence or absence of comorbidities over time is captured by a sequence of active disease states $\mathbf{s} = (s_1, ..., s_m)$ associated with $\mathbf{a}$, where each state s_t is the set of active diseases at each time $t = 1, ..., m$ and represents the comorbidity patterns of a patient in t. The set of active disease states is defined as $S = \{0, 1\}^{|D|}$ where 1 indicates that the disease $d \in D$ is active at a specific time and 0 means that disease is not active in the patient at that time. The transition dynamics of these active disease states define the activation and deactivation of diseases, and therefore, the possible occurrence of diseases over time. Let $\mathbf{d} = (d_1, ..., d_m)$ be the latent sequence of diseases, where d_t belongs to the set of diagnoses $D = \{1, ..., r\}$ for $t = 1, ..., m$. The active disease states determine the distribution of such diseases over time, since the dynamics of the diseases depend on which comorbidities are active at the same time. Therefore, when a comorbidity is activated or deactivated, the distribution of the remaining active diseases changes. We further consider that once an active disease is deactivated, it cannot be present in the patient again.

Finally, let c be the latent class which $\mathbf{a}$ belongs to. The class c belongs to a set $C = \{1, ..., k\}$, which represents the subtypes of similar coevolution patterns of comorbidities among patients. The role of this latent variable is

to capture the heterogeneity among the clinical histories based on the joint evolution of diseases. By doing so, it enables the classification of patients into distinct groups characterized by diverse comorbidity patterns over time. The classes influence the distribution of diseases but do not affect the transition dynamics of medical actions. That is, the generative model assumes that the stochastic model of the treatment of a disease is common to all patients, while it is the evolution of diseases over time that creates the different subgroups of patients.

The proposal for the generative model is as follows (see Fig. 2):

a) Draw a class $c \sim Mult(\boldsymbol{\theta}_C)$

b) Sample the initial active disease state (set of potential comorbidities), the initial disease, and the initial medical action,

$$s_1 \sim \ Cat(\boldsymbol{\pi}_S),$$

$$d_1|s_1, c \sim \ Cat(\boldsymbol{\pi}_D^{s_1,c}), \quad a_1|d_1 \sim \ Cat(\boldsymbol{\pi}_A^{d_1})$$

c) For each time t:

i) Sample an active disease state from $p(s_t|s_{t-1}, d_{t-1}, a_{t-1})$, that is,

$$s_t|s_{t-1}, d_{t-1}, a_{t-1} \sim Cat(\boldsymbol{\theta}_S^{s_{t-1},d_{t-1},a_{t-1}})$$

ii) Sample a disease d_t from $p(d_t|s_t, c)$,

$$d_t|s_t, c \sim Cat(\boldsymbol{\theta}_D^{s_t,c})$$

iii) Sample an action $a_{d_t:t}$ from $p(a_{d_t:t}|d_t, a_{d_t:t'})$, that is,

$$a_{d_t:t}|d_t, a_{d_t:t'} \sim Cat(\boldsymbol{\theta}_A^{d_t,a_{d_t:t'}})$$

where $a_{d:t}$ is the t-th action associated with the disease d and $a_{d:t'}$ the previous action associated with the same disease d, so that $\mathbf{a}_d$ is the treatment of the disease d.

Translating the generative process into a joint probability model results in the following expression (Fig. 2):

$$p(\mathbf{a}, \mathbf{s}, \mathbf{d}, c) = p(c) \prod_{t=1}^{m} p(s_t|s_{t-1}, d_{t-1}, a_{t-1}) \cdot p(d_t|c, s_t) \cdot p(a_{d:t}|d_t, a_{d:t'}) \quad (1)$$

where $p(s_1|s_0, d_0, a_0) = p(s_1)$ and $p(a_{d:t}|d, a_{d:0}) = p(a_{d:t}|d)$ for any value of $t = 1, ..., m$.

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 C. We define $\boldsymbol{\theta}_C$ as the set of such probabilities that we have to learn:

$$\boldsymbol{\theta}_C = \{p(c) : c \in C\}.$$

The active disease states determine the coexisting diseases at each time t. The probability of transition from a state s to s' is defined by a Markov model, whose parameters are:

$$\boldsymbol{\theta}_S = \{\boldsymbol{\theta}_S^{s,d,a} : s \in S, d \in D, a \in A\} = \{p(s'|s, d, a) : s, s' \in S, d \in D, a \in A\}.$$

Diseases follow a categorical distribution conditioned to the set of coexisting diseases $s_t \in S$ at time t and the class of patient $c \in C$. Thus, for each active disease state $s \in S$ and each class $c \in C$, we have the following parameters:

$$\boldsymbol{\theta_D} = \{\boldsymbol{\theta}_D^{s,c} : s \in S, c \in C\} = \{p(d|s, c) : d \in D, s \in S, c \in C\}.$$

In addition, we define a Markov model from which the medical actions are drawn. The conditional distributions of this model are given by a set of $|D|$ transition matrices of size $|A| \times |A|$ whose model parameters are:

$$\boldsymbol{\theta}_A = \{\boldsymbol{\theta}_A^{d,a} : d \in D, a \in A\} = \{p(a'|a, d) : a, a' \in A, d \in D\}.$$

Finally, $\boldsymbol{\pi}_S$, $\boldsymbol{\pi}_D^{s,c}$ and $\boldsymbol{\pi}_A^d$ are the parameters of the initial model for the active disease states, diseases and medical actions, respectively.

2.3. Maximum likelihood parameter estimation

In this section, we introduce the learning procedure of the model parameters. Let $\mathcal{A} = (\mathbf{a}^1, ..., \mathbf{a}^N)$ be the set of observed sequences of actions and let $\mathcal{S} = (\mathbf{s}^1, ..., \mathbf{s}^N)$ be the associated set of sequences of active disease states. As we mentioned in Section 2.1, the sequence of diseases $\mathbf{d}$ is partially observed, providing an intuition about the onset and end of the diseases, and therefore, about their activation and deactivation timestamps. However, note that the activation time of a disease tends to be inherently unobservable in EHRs since the first and last records of a diagnosis may not reliably indicate the

real time of disease onset and end. We define a time parameter τ to determine the time interval in which a disease is active $(t_{init} - \tau, t_{end} + \tau)$, where t_{init} and t_{end} are the first and last time a diagnosis is observed in EHRs, respectively. Thus, we determine the sequence of active disease states for each sequence of actions $\mathbf{a} \in \mathcal{A}$ in such a way that the corresponding set of sequences of diseases, $\mathcal{D}_{\mathbf{a}}$, is limited to all the possible sequences of diseases that coherently fit the existing diagnoses in EHRs.

To learn the distribution underlying the sequences, we seek to maximize the following weighted log-likelihood of the data:

$$\max_{\mathbf{\Theta}} \sum_{\substack{\mathbf{a}\in\mathcal{A} \\ \mathbf{s}\in\mathcal{S}}} \sum_{\mathbf{d}\in\mathcal{D}_{\mathbf{a}}} \sum_{c\in C} p(\mathbf{s}, \mathbf{d}, c|\mathbf{a}) \cdot \log p(\mathbf{a}, \mathbf{s}, \mathbf{d}, c; \mathbf{\Theta}) \tag{2}$$

where $\mathcal{D}_{\mathbf{a}}$ is the set of sequences of diseases for $\mathbf{a}$, $p(\mathbf{s}, \mathbf{d}, c|\mathbf{a})$ is the contribution of the tuple $(\mathbf{a}, \mathbf{s}, \mathbf{d}, c)$ to the model, and $\mathbf{\Theta} = \{\boldsymbol{\theta}_A, \boldsymbol{\theta}_D, \boldsymbol{\theta}_S, \boldsymbol{\theta}_C\}$. The reason for weighting the log-likelihood is to make each sequence $\mathbf{a} \in \mathcal{A}$ contribute to the model regardless of its length, and this is achieved because

$$\sum_{c\in C} \sum_{\mathbf{d}\in\mathcal{D}_{\mathbf{a}}} p(\mathbf{s}, \mathbf{d}, c|\mathbf{a}) = 1. \tag{3}$$

Note that the maximum size of the set $\mathcal{D}_{\mathbf{a}}$ is $|D|^{|\mathbf{a}|}$ and exponentially increases with the length of the sequence $\mathbf{a}$. Indeed, the parameters depend on the number of diseases we jointly model, and, in this work, we assume that the number of coexisting diseases at a specific time, s_t, is moderate even though the total number of diseases $|D|$ can be large.

To find the parameters that maximize the log-likelihood in Eq. 2, we use the iterative EM algorithm.

E-step: In this step, we have to consider all the possible configurations of latent diseases $\mathbf{d} \in \mathcal{D}_{\mathbf{a}}$ for the observed sequences of actions $\mathbf{a} \in \mathcal{A}$ and compute their probability for all the classes in C (Eq. 1). A brute force approach would require an exponential number of computations due to the exponential size of $\mathcal{D}_{\mathbf{a}}$, hence, we propose an algorithm based on dynamic programming (Appendix 4) that allows us to solve it in polynomial time. Instead of computing the probability of all the configurations $(\mathbf{d}, c)$ for each $\mathbf{a}$ one by one, the proposed learning method computes given a sequence of actions $\mathbf{a}$, a sequence of co-occurrence states $\mathbf{s}$ and a class c, the probability of all the sequences of diseases that have the disease d at time t, $p(d_t = d|c, s_t, \mathbf{a})$.

M-step: In the maximization step we have to update the parameters of the model with the probabilities computed in the previous E-step. If $\theta_a^{d,a'}$, $\theta_s^{s',d,a}$, $\theta_d^{s,c}$, θ_c denote a component in $\boldsymbol{\theta}_A^{d,a'}$, $\boldsymbol{\theta}_S^{s',d,a}$, $\boldsymbol{\theta}_D^{s,c}$, $\boldsymbol{\theta}_C$, respectively, the parameters of the model are updated as follows:

$$\theta_a^{d,a'} = \frac{\sum_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^{|\mathbf{a}|} \sum_{t'<t} \mathbb{1}_{a',a}(a_{d:t'}, a_{d:t}) \cdot p(d_t = d|\mathbf{a}, \mathbf{s})}{\sum_{a\in A} \sum_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^{|\mathbf{a}|} \sum_{t'<t} \mathbb{1}_{a',a}(a_{d:t'}, a_{d:t}) \cdot p(d_t = d|\mathbf{a}, \mathbf{s})} \tag{4}$$

where

$$\mathbb{1}_{a',a}(a_{d:t'}, a_{d:t}) = \begin{cases} 1 & \text{if } a_{d:t'} = a', a_{d:t} = a \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

$$\theta_s^{s',d,a} = \frac{\sum_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^{|\mathbf{a}|} \mathbb{1}_{a,s',s}(a_{t-1}, s_{t-1}, s_t) \cdot p(d_{t-1} = d|\mathbf{a}, \mathbf{s})}{\sum_{s\in S} \sum_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^{|\mathbf{a}|} \mathbb{1}_{a,s',s}(a_{t-1}, s_{t-1}, s_t) \cdot p(d_{t-1} = d|\mathbf{a}, \mathbf{s})} \tag{6}$$

where

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

$$\theta_d^{s,c} = \frac{\sum_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^{|\mathbf{a}|} \mathbb{1}_s(s_t) \cdot p(d_t = d, c|\mathbf{a}, \mathbf{s})}{\sum_{d\in D} \sum_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^{|\mathbf{a}|} \mathbb{1}_s(s_t) \cdot p(d_t = d, c|\mathbf{a}, \mathbf{s})} \tag{7}$$

$$\theta_c = \frac{\sum_{\boldsymbol{a}\in\mathcal{A}} p(c|\mathbf{a}, \mathbf{s})}{\sum_{c\in C} \sum_{\boldsymbol{a}\in\mathcal{A}} p(c|\mathbf{a}, \mathbf{s})} \tag{8}$$

The proposed learning algorithm based on dynamic programming allows the E-step to be polynomially solved, where the exponential number of configurations of diseases and classes for a given sequence of actions is considered. Furthermore, the complexity of the M-step is of order $\mathcal{O}(\sum_{\mathbf{a}\in\mathcal{A}} |\mathbf{a}|)$, that is, the total number of medical actions of the set $\mathcal{A}$.

A large amount of configurations of diseases, classes, and actions creates problems of sparsity in the parameters of the model. Once a parameter reaches a value of 0, that parameter cannot obtain a different value in the subsequent iterations. We add a smoothing parameter to the model in each iteration of the EM algorithm to prevent this sparsity problem.

3. Experimental evaluation

We present two sets of experiments to validate the model. The goal of the first set of experiments is to evaluate the ability of our learning algorithm to recover the original generative model underlying the data, for which we use synthetic data. The second set of experiments show some applications of the generative comorbidity model on real-world data, such as the segmentation of the medical history of a patient into different treatments, the identification of the different classes based on the joint progression of comorbidities, and the imputation of missing diagnoses. The corresponding source code is publicly available[1].

3.1. Results on synthetic experiments

We perform experiments on synthetic data to show the behavior of the learning algorithm in controlled environments. In these experiments the diagnoses are considered unknown in the learning process. Since this is an artificial domain, the evaluation of the learned model is carried out using the log-likelihood in training and test data so that we can quantify the fitting and generalization abilities, respectively.

To this end, the first step of the experiment is to build a original generative model. In order to do that, we consider random parameters. For simplicity, we perform experiments with 2 and 3 comorbidities. In both cases, we set 2 classes and 10 medical actions. The parameters of the generative model are created as follows: $p(c)$, $p(a'|a,d)$ and $p(d|s,c)$ are sampled from a uniform Dirichlet distribution for $c \in C$, $a, a' \in A$ and $d \in D$; and $p(s'|s,d,a)$ is also sampled from a Dirichlet distribution with $\alpha = 1$ but limiting the active disease states to only activate or deactivate a single disease in each transition. To avoid the generative model taking values too close to 0, we smooth the sufficient statistics $p(c)$, $p(a'|a,d)$ and $p(d|s,c)$ by adding 10^{-2}, and $p(s|s',d,a)$ by adding 10^{-3}.

From the generative model we sample training sets of sizes $N = \{100, 300, 500, 800, 1000, 1200, 1500\}$ and a test set of size 1500. We learn the parameters of the model $\boldsymbol{\Theta}^n = \{\boldsymbol{\theta}_C^n, \boldsymbol{\theta}_S^n, \boldsymbol{\theta}_D^n, \boldsymbol{\theta}_A^n\}$ for each training set of size $n \in N$ using the EM algorithm proposed in Section 2.3. At each iteration of the EM algorithm the sufficient statistics are smoothed by adding 10^{-2} to $p(c)$,

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

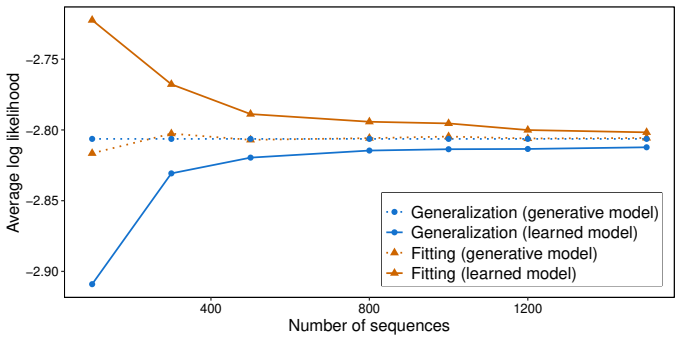


Figure 3: Fitting and generalization of synthetic generative models with 2 comorbidities.

$p(a|a', d)$ and $p(d|s, c)$, and 10^{-3} to $p(s|s', a, d)$. Once the model has converged, we measure the quality of these learned models with the log-likelihood of the data (Eq. 2) normalized by the total number of actions in each dataset of size $n \in N$ to make the results comparable.

This experiment is repeated five times, considering, for each of them, a different original generative model. Fig. 3 and Fig. 4 show the fitting and generalization ability of the method through the average log-likelihood of 2 and 3 comorbidities. The average log-likelihood of the learned models on the training sets (orange solid line) quantifies the fitting of the models to the data, while on the test set (blue solid 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 the sample size 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 can reach the original generative model underlying the data.

3.2. Results on real data

We show the utility of the generative model on patients with breast cancer and cardiovascular diseases, which are highly related comorbidities [24]. We use the generative model in two different applications: we first perform an experiment to show the segmentation of individual clinical histories into disease treatments; and then, a population-level experiment to obtain the co-evolution patterns of these two comorbidities. We further assess the results of these experiments by predicting the diagnosis of unseen instances.

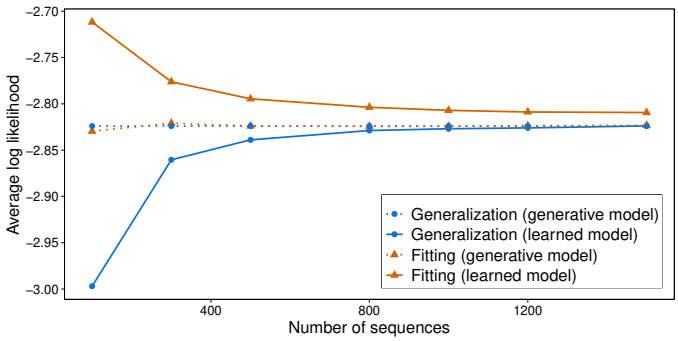


Figure 4: Fitting and generalization of synthetic generative models with 3 comorbidities.

3.2.1. Data description

We use a dataset provided by the public healthcare system of the Basque Country, Spain. These EHRs cover every hospital visit of patients from 2016 to 2019. As a use case, we focus our attention on the comorbidities of the breast cancer population, specifically on cardiovascular diseases. These diseases are biologically connected through some deleterious effects of cancer treatment on cardiovascular health [24]. The resulting dataset consists of 90 clinical histories, whose average length is 140 medical actions, and they are generated by 29 unique medical specialties (selected following the process in [25]). The percentage of missing diagnoses of these EHRs is 81%.

3.2.2. Hyperparameters and model specifications

We consider breast cancer patients with any diagnosis related to cardiovascular diseases, that is, $|D| = 2$. According to clinical guidelines [24], patients evolve according to their severity of short-term cardiotoxic effects caused by anticancer therapies. In order to have a sufficient number of patients per class and after conducting experiments for different values of the latent class, we have concluded that $|C| = 2$ is appropriate for the available data.

Besides, since we are in a realistic scenario, we include prior diagnosis knowledge in the model, so that we can obtain more accurate results and reduce the model complexity. Since 19% of the diagnoses are available, we force them to remain fixed in their original time position in the latent sequences of diseases. Varying the value of τ can have a significant influence on both accuracy and computational efficiency. Through experiments conducted with different values of $\tau = \{90, 180, 360, 720, 1080, 1440\}$, we observed that

setting τ to 720 days gets a good balance between computational efficiency and model performance. Therefore, to establish the active disease states, we assume that the transition between two medical actions of the same disease may occur within a maximum interval of $\tau = 720$ days.

	ACTION 1	ACTION 2	ACTION 3	ACTION 4	ACTION 5	ACTION 6	ACTION 7	ACTION 8	ACTION 9	ACTION 10	ACTION 11
BREAST CANCER	Radiology		Oncology	**Oncology**	**Oncology**	Oncology		Oncology	**Oncology**	Oncology	Radiology
CARDIOVASCULAR DISEASE		Cardiology					Cardiology				

	ACTION 12	ACTION 13	ACTION 14	ACTION 15	ACTION 16	ACTION 17	ACTION 18	ACTION 19	ACTION 20	ACTION 21	ACTION 22
BREAST CANCER		Gynecology	Oncology	Anesthesia	**Anesthesia**	**Surgery**	**Oncological gynecology**		Pathological anatomy	Gynecology	**Radiotherapy**
CARDIOVASCULAR DISEASE	**Cardiology**							Radiology			

	ACTION 23	ACTION 24	ACTION 25	ACTION 26	ACTION 27	ACTION 28	ACTION 29	ACTION 30	ACTION 31	ACTION 32	ACTION 33
BREAST CANCER	Oncology				Radiotherapy	**Radiotherapy**	Cardiology	**Radiotherapy**	Radiotherapy	Radiotherapy	Radiotherapy
CARDIOVASCULAR DISEASE		**Rehabilitation**	Rehabilitation	Rehabilitation							

Figure 5: Disentangle of a partial clinical history of a patient with the diagnosis of breast cancer and cardiovascular disease. The bold medical specialties represent the real diagnosis collected in EHRs. The results are obtained from the model learned in Section 3.2.3.

3.2.3. Individualized segmentation of clinical histories

The first objective of the model is to segment the sequence of actions, $\mathbf{a}$, into subsequences associated with the different comorbidities. This is useful, for instance, for understanding the evolution of a single disease in a patient, extracting its associated treatment dynamics from the clinical history, or even for an informing forecast of expected costs of care and medical resources for specific diseases and patients by simulating trajectories from each disease related model.

In this experiment we train the model with the whole dataset. Then, the association between medical specialties and diagnosis at each time t of the sequence $\mathbf{a} \in \mathcal{A}$ is given by the diagnosis of maximum probability at time t, that is,

$$\max_{d \in D} p(d|\mathbf{a}, s_t) \tag{9}$$

where s_t is the set of active diseases at time t.

Thus, we can extract the subsequence associated with each diagnosis from a patient's clinical trajectory **h**. An example of that is the segmentation of a partial clinical history of a real patient that we show in Fig. 5. Although in Fig. 5 we attribute a diagnosis to each medical event, the model allows us to assign to each medical action the probability of belonging to any disease. In reality, a fundamental aspect of caring for a patient undergoing potentially cardiotoxic anticancer therapy is to be treated by a multidisciplinary team of oncologists, cardiologists, and other healthcare professionals [24]. This means that a medical event may not be the consequence of a single disease, but is caused by a set of diseases that co-exist over time.

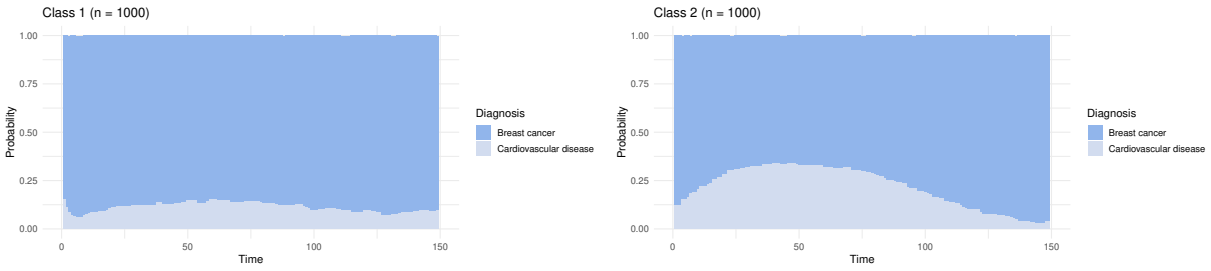
3.2.4. Representation of the joint progression of comorbidities at population-level

The learned generative model enables knowledge to be extracted about comorbidity evolution patterns at population-level regarding the subtypes of treatments. This is a simulation experiment to provide a representation of the different joint evolution of breast cancer and cardiovascular diseases.

Following the generative process in Section 2.2, we randomly sample a set of 1000 clinical histories for each class from the learned model in the previous paragraph. The clinical histories are of variable length and we set the maximum number of actions to be 140. We show the joint evolution of comorbidities by calculating the probability of a disease-related event occurring at each time point, that is,

$$p(d_t = d|c), \quad \text{for all } t. \tag{10}$$

In Fig. 6 we show the joint evolution of the breast cancer and cardiovascular diseases for the 2 classes. Although breast cancer treatment clearly dominates in both classes, the occurrence of cardiovascular treatment is different depending on the class. The probability of treating cardiovascular diseases remains constant in the first class (Fig. 6a), while it increases in the initial part of the medical records in the second class (Fig 6b). Therefore, class 1 may refer to patients with pre-existing cardiovascular disease or cardiovascular risk factors undergoing potentially cardiotoxic anticancer therapy that requires routine monitoring [24]. On the contrary, class 2 may indicate more severe cardiovascular complications as a consequence of the harmful effect of anticancer therapies on the cardiovascular system [24].



(a) Class 1 of coevolution treatments.

(b) Class 2 of coevolution treatments.

Figure 6: Joint evolution of comorbidities at a population-level.

3.2.5. Imputation of diagnosis

Another application of our generative model is the imputation of missing diagnosis values of EHRs. In other words, we seek to label a new patient's medical events with a diagnosis for each timestep. To assess the diagnosis assignment of the model, we carry out a 10-fold cross-validation, where we split the dataset into training and test sets in a 90:10 proportion. We train the model as in previous experiments, including the diagnoses collected in the EHRs. However, in this experiment we propose the most complex scenario for the test set, considering every diagnosis to be unknown. The problem consists of setting a diagnosis label for each medical specialty of the test set with Eq. 9, and afterward, checking them with 19% of available diagnoses as ground truth.

We replicated the cross-validation experiment using two simplified versions of the model to demonstrate the significance and utility of the latent class and activation state variables in the assignment of diagnoses to medical events. On the one hand, the first simplification we carry out to our original model is to delete the class information. In this sense, we assume that there are no subtypes of progression in comorbidities, and therefore, there are no patients with higher probability of developing one disease over another. Then,

$$p(\mathbf{a}, \mathbf{s}, \mathbf{d}) = \prod_{t=1}^{m} p(s_t|s_{t-1}, d_{t-1}, a_{t-1}) \cdot p(d_t|s_t) \cdot p(a_{d:t}|d_t, a_{d:t'}).$$

On the other hand, in the second baseline model we eliminate the activation states from the original model, while still considering class information. This model assumes that comorbidities are always active throughout the en-

tire medical history of a patient. The joint probability of this model is defined as

$$p(\mathbf{a}, \mathbf{d}, c) = p(c) \prod_{t=1}^{m} p(d_t|c) \cdot p(a_{d:t}|d_t, a_{d:t'}).$$

Model	AUC	Accuracy	F1-score	
			Breast cancer	Cardiovasc.
Complete model	0.81	0.84	0.90	0.75
Model without classes	0.76	0.80	0.85	0.71
Model without activation states	0.75	0.78	0.83	0.68

Table 1: Comparative evaluation of the models.

We can observe in Table 1 the improved assignment performance of our model, which achieves higher AUC, accuracy, and F1-score values. These results highlight the significance of including both latent classes and activation states in our model. In addition, this experiment not only supports the quality of the segmentation of clinical histories into treatments of individual patients (Section 3.2.3), but also the comorbidity evolution dynamics captured in the simulation experiment (Section 3.2.4).

4. Discussion and conclusion

This paper proposes a novel probabilistic generative model for patients with co-existing diseases. Modeling comorbidity dynamics from EHRs is not straightforward and involves addressing challenges such as small datasets, uncertainty, and missingness [17, 26]. We face the particular problem where the diagnosis is missing in most of the EHRs. Hence, the model is specifically focused on the identification of the diagnoses associated with medical events and the discovery of subtypes of similar disease coevolution patterns. To the best of our knowledge, this is the first method to learn the dynamics of underlying comorbidity without observing the entire clinical history of diagnoses.

Experiments show that the generative model can accurately estimate the diagnosis of medical records. These results emphasize the model's ability to extract treatment subsequences from EHRs and capture the main subtypes of comorbidity evolution dynamics based on medical specialties. This correct diagnosis imputation is of great interest for training models that require

complete EHRs or avoiding loss of information observed in other imputation methods [26].

Although we recognize that other real-world EHR data contain many other types of information, our current scenario is based on a limited administrative dataset with missing values. However, in future work, we plan to incorporate additional clinical data such as medical procedures, medications, and laboratory results. One approach we plan to explore is taking the Cartesian product of these variables, which would involve learning a larger number of parameters, and therefore, a sufficient number of instances will be necessary. The inclusion of this additional data would result in a more informative model, particularly in tasks such as simulating clinical histories.

We build our model on Markov models to ensure interpretability. These models are comprehensible because they can be summarized in probability matrices that easily describe the transition rates between diseases and medical events. However, a limitation of Markov models is their memoryless assumption, where an individual's current action depends on the previous medical action, instead of considering his/her entire previous clinical history. Future work will be focused on relaxing this memoryless characteristic.

Another limitation of our model is its complexity when the number of diseases is too large. The number of parameters of the disease distribution $\boldsymbol{\theta}_D$ to be learned is $2^{|D|}$. Nevertheless, the number of coexisting comorbidities that we consider is not so large as to become an intractable problem. An alternative to deal with a larger amount of diseases would be to simplify the model by assuming the same distribution of diseases through the whole clinical history instead of being dependent on the active diseases at each time.

In conclusion, this is an interpretable generative model to understand comorbidity dynamics that handles the incompleteness of EHRs. We demonstrate the success of the model on both synthetic and real-world datasets. The model is well-suited to the scenario where coexisting diseases evolve differently depending on the active comorbidities of the patient and achieves the objective of modeling such progression.

Conflict of interest

The authors declare that they have no competing interests.

Acknowledgements

This research has been 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 CEX2021-001142-S. Onintze Zaballa also holds a predoctoral grant (EJ-GV 2019) from the Basque Government.

Appendix A. Efficient learning of the model parameters with a dynamic programming-based algorithm.

A brute force learning of the parameters of the model is computationally expensive for large datasets and long sequences. We propose an alternative learning algorithm based on dynamic programming to considerably reduce the number of computations, and thus, the complexity of the model from exponential to polynomial. This inference plays an important role in the learning procedure of the model, particularly in the E-step. The computations obtained in this step are then used in the M-step to update the parameters of the model $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_C, \boldsymbol{\theta}_S, \boldsymbol{\theta}_D, \boldsymbol{\theta}_A\}$.

Let us assume that we have a training set $\mathcal{A} = \{\mathbf{a}^i\}_{i=1}^N$ of sequences of actions, where $\mathbf{a}^i = (a_1^i, ..., a_m^i)$ for all i. Suppose that each $\mathbf{a}$ has an associated latent sequence of diagnoses $\mathbf{d} = (d_1, ..., d_m)$, where $d_t \in D$ for all $t = 1, ..., m$, and belongs to a latent class $c \in C$. Let $\mathbf{s} = (s_1, ..., s_m)$ be the sequence of active disease states associated with $\mathbf{a}$, where s_t denotes the active diseases at time $t = 1, ..., m$ and belongs to the set of active disease states S. The aim is to estimate the maximum likelihood parameters $\boldsymbol{\Theta}$ of the model in each iteration of the EM algorithm. Hence, our objective is to learn the parameters of the model $p(c)$, $p(s|s', d, a)$, $p(a|a', d)$ and $p(d|c, s)$ for any value of $a, a' \in A$, $d \in D$, $s, s' \in S$ and $c \in C$, using the set of sequences of observed actions in $\mathcal{A}$. Suppose that, for a sequence of actions $\mathbf{a}$, we observe the transition from $a_{t'} = a'$ to $a_t = a$ in the training set between two any time points t' and t, $t' < t$. We shall calculate the sum of the probabilities of all the possible sequences of diseases for which $d_{t'} = d$, $d_t = d$ and $d_{t'+1}, ..., d_{t-1} \neq d$ in each class c. That is, the probability of all the sequences of diseases with the form

$$(d_1, ...d_{t'-1}, d, d_{t'+1}, ..., d_{t-1}, d, d_{t+1}, ..., d_m),$$

where $d_{t'+1}, ..., d_{t-1} \neq d$.

Let us define $f_c(t_1, ..., t_r)$ as the function that computes the sum of probabilities of all the possible sequences of diseases $\mathbf{d} = (d_1, ..., d_t)$ in the class c, where $(t_1, ..., t_r)$ $(r = |D|)$ indicates the last time that the diseases in D appear in the sequence $\mathbf{d}$. We compute the probability of all the sequences of diseases that have the disease $d \in D$ at time t as follows:

$$f_c(t_1, ..., t_r) = \sum_{\mathbf{d}_{1,...,t-1}} p(\mathbf{a}_{1,...,t}, \mathbf{s}_{1,...,t}, \mathbf{d}_{1,...,t-1}, d_t = d|c)$$

where $\mathbf{d}_{1,...,t-1} = (d_1, ..., d_{t-1})$, $\mathbf{s}_{1,...,t} = (s_1, ..., s_t)$ and $\mathbf{a}_{1,...,t} = (a_1, ..., a_t)$.

On the other hand, let us define $g_c(t_1, ..., t_r)$ as the function that computes the sum of probabilities of all the possible sequences of diseases $\mathbf{d} = (d_{t+1}, ..., d_m)$ in c, where $(t_1, ..., t_r)$ $(r = |D|)$ indicates the first time each disease $d \in D$ appears in the sequence $\mathbf{d}$. That is,

$$g_c(t_1, ..., t_r) = \sum_{\mathbf{d}_{t+1,...,m}} p(\mathbf{a}_{t+1,...,m}, \mathbf{s}_{t+1,...,m}, \mathbf{d}_{t+1,...,m}|c, d_t = d).$$

Using these functions, we can express the sum of the probabilities of all the sequences for which $d_t = d$ as follows:

$$\begin{aligned} p(\mathbf{d}_{1,...,t-1}, d_t = d, \mathbf{d}_{t+1,...,m}, \mathbf{a}_{1,...,m}, \mathbf{s}_{1,...,m}|c) = \\ = f_c(t_1, ..., t_i = t', ..., t_r) \cdot p(s_t|s_{t-1}, d_{t-1}, a_{t-1}) \\ \cdot p(d|s_t, c) \cdot p(a_{d:t}|a_{d:t'}, d) \cdot g_c(t_1, ..., t_i = t, ..., t_r) \end{aligned} \tag{11}$$

where t' is the previous time where the same disease d is allocated.

With this in mind, we propose to create a matrix of size $|D| \times |D|$ associated with each function f_c and g_c, each of them calculated with the recursive functions in Algorithm 1 and Algorithm 2.

Algorithm 1 Computation of f_c matrix

Input: $\{t_1, ..., t_r\}$: set of the last time we saw each disease
$\boldsymbol{\Theta} = \{\boldsymbol{\theta}_C, \boldsymbol{\theta}_S, \boldsymbol{\theta}_D, \boldsymbol{\theta}_A\}$: model parameters.
Output: $f_c(t_1, ..., t_r)$
$t_i \leftarrow \max\{t_1, ..., t_r\}$
$t_j \leftarrow \max\{t_1, ..., t_{i-1}, t_{i+1}, ..., t_r\}$
$t \leftarrow t_i$
if $t_i - t_j > 1$ **then**
$f_c(t_1, ..., t_r) \leftarrow p(s_t|s_{t-1}, d_{t-1}, a_{t-1}) \cdot p(d^i|c, s_t) \cdot p(a_{d^i:t}|a_{d_i:t-1}, d^i) \cdot f_c(t_1, ..., t_{i-1}, t-1, t_{i+1}, ..., t_r)$
else
$f_c(t_1, ..., t_r) \leftarrow p(s_t|s_{t-1}, d_{t-1}, a_{t-1}) \cdot p(d^i|c, s_t) \cdot \sum_{t'=0}^{t_j-1} p(a_{d^i:t}|a_{t'}, d^i) \cdot f_c(t_1, ..., t_{i-1}, t', t_{i+1}, ..., t_r)$
end if

Notice that in Algorithm 1 the statement $t_i - t_j > 1$ means that the action at time $t_i = t$ comes from the same disease as the action in the previous time $t-1$, while the statement $t_i - t_j = 1$ means that we do not know from which previous action (or time) the action at time t comes.

Algorithm 2 Computation of g_c matrix

Input: $\{t_1, ..., t_r\}$: set of the first time we saw each disease
$\boldsymbol{\Theta} = \{\boldsymbol{\theta}_C, \boldsymbol{\theta}_S, \boldsymbol{\theta}_D, \boldsymbol{\theta}_A\}$: model parameters.
Output: $g_c(t_1, ..., t_r)$
$t_i \leftarrow max\{t_1, ..., t_r\}$
$t \leftarrow t_i + 1$
$g_c(t_1, ..., t_r) \leftarrow \sum_i p(s_t|s_{t-1}, d_{t-1}, a_{t-1}) \cdot p(d^i|c, s_t) \cdot p(a_t|a_{d^i:t_i}, d^i) \cdot g_c(t_1, ..., t_{i-1}, t_i = t, t_{i+1}, ...t_r)$

We can now compute Eq. 11 that allows us to update the parameters of the model using Eq. 4, 6, 7 and 8. We have to calculate the probability of the transitions between any two actions, that is, from a' to a. If a appears at time t in the sequence $\mathbf{a}$, let t' be the set of times such that we can find the action a' in the subsequence $\mathbf{a}_{1,...,t-1}$, that is, $t' = \{y < t : a_y = a'\}$. We learn the parameters of the model $\boldsymbol{\Theta}$, specifically for the transition from an action a' to the action a at time t, considering the different possibilities that may occur that depend on the time position of a'.

Let $T = (t_1, ..., t_r)$ be the vector that indicates the last time each type of disease $d^i \in D$, $i = 1, ..., r$, appears in the sequence $\mathbf{d} = (d_1, ..., d_t)$, and let $h = \max t' = \max\{y < t : a_y = a'\}$.

For each time t where the action a is observed, and for each disease $d^i \in D$, $i = 1, ..., r$, the two following options can occur:

1. If $t - h > 1$:

   1.1) If at least one disease has already finished before t:

   For the set of finished diseases before t, $d^f \in D$, we use their endpoint in the sequence $\mathbf{d}$ to set in the vector T the last time we have seen that disease.

   For the set of unfinished diseases that are already initialized, we fix each disease, $d^j \in D$, at $t_j = t - 1$ $(j \neq f, i)$ while we set $t_{r'} = 0, ..., t - 2$ in the rest of the unfinished diseases ($d^{r'} \in D$, $r' \neq j, i, f$). Take into account that if any disease's endpoint is fixed at $t - 1$, we do not have to set any unfinished disease t_j in $t - 1$, rather they are all fixed at $t_{r'} = 0, ..., t - 2$ $(r' \neq i, f)$.
   For those diseases that have not already been initialized their position in T is fixed at 0.

   1.2) If no disease has finished before t, we fix for each disease $d^j \in D$ their last position in T as $t_j = t - 1$ and the rest of the initialized diseases' position at $t_{r'} = 0, ..., t - 2$ $(r' \neq j, i)$.

   Let $J = \{1, ..., i - 1, i + 1, ..., r\}$, then we can compute $p(d_t = d^i, \mathbf{a}, \mathbf{s}|c)$ as

$$\sum_{y \in t'} \sum_{j \in J} \sum_{\substack{t_1,...,t_r \\ t_j = t-1 \\ t_{r'}, r' \neq j,i,f}} f_c(t_1, ..., t_i = y, ..., t_r) \cdot p(s_t|s_{t-1}, d_{t-1}, a_{t-1}) \cdot p(d_t = d^i|c, s_t) \cdot p(a_t|a_{d^i:t'} = a', d_t = d^i) \cdot g_c(t_1, ..., t_i = t, ..., t_r)$$

2. If $t - h = 1$:
   We fix for each disease $d^j \in D$ their position t_j at the maximum position t', that is, $t_j = h$. In addition, $t_{r'} = 0, ..., t - 2$ for all $r' \neq i, j$. Then,

let $J = \{1, ..., i-1, i+1, ..., r\}$, then we can compute $p(d_t = d^i, \mathbf{a}, \mathbf{s}|c)$ as

$$\sum_{y \in t'} \sum_{j \in J} \sum_{\substack{t_1,...,t_r \\ t_j = h \\ t_{r'}, r' \neq j}} f_c(t_1, ..., t_i = y, ..., t_r) \cdot$$
$$p(s_t|s_{t-1}, d_{t-1}, a_{t-1}) \cdot p(d_t = d^i|c, s_t)$$
$$\cdot\, p(a_t|a_{d^i:t'} = a', d_t = d^i) \cdot g_c(t_1, ..., t_i = t, ..., t_r)$$

References

[1] A. L. Young, R. V. Marinescu, N. P. Oxtoby, M. Bocchetta, K. Yong, N. C. Firth, D. M. Cash, D. L. Thomas, K. M. Dick, J. Cardoso, et al., Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference, Nature communications 9 (1) (2018) 1–16.

[2] 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.

[3] 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.

[4] 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, Advances in neural information processing systems 29 (2016).

[5] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, J. Gao, Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks, in: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017, pp. 1903–1911.

[6] C. Lee, M. Van Der Schaar, Temporal phenotyping using deep predictive clustering of disease progression, in: International Conference on Machine Learning, PMLR, 2020, pp. 5767–5777.

[7] O. Zaballa, A. Pérez, E. G. Inhiesto, T. A. Ayesta, J. A. Lozano, Learning the progression patterns of treatments using a probabilistic generative model, Journal of Biomedical Informatics 137 (2023) 104271.

[8] Y.-Y. Liu, S. Li, F. Li, L. Song, J. M. Rehg, Efficient learning of continuous-time hidden markov models for disease progression, in: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS'15, MIT Press, Cambridge, MA, USA, 2015, p. 3600–3608.

[9] B. Maag, S. Feuerriegel, M. Kraus, M. Saar-Tsechansky, T. Züger, Modeling longitudinal dynamics of comorbidities, in: Proceedings of the Conference on Health, Inference, and Learning, 2021, pp. 222–235.

[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] I. Stanculescu, C. K. Williams, Y. Freer, Autoregressive hidden markov models for the early detection of neonatal sepsis, IEEE journal of biomedical and health informatics 18 (5) (2013) 1560–1570.

[12] 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.

[13] 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, Journal of biomedical informatics 102 (2020) 103364.

[14] H.-M. Lu, C.-P. Wei, F.-Y. Hsiao, Modeling healthcare data using multiple-channel latent dirichlet allocation, Journal of biomedical informatics 60 (2016) 210–223.

[15] X. Teng, S. Pei, Y.-R. Lin, Stocast: Stochastic disease forecasting with progression uncertainty, IEEE Journal of Biomedical and Health Informatics 25 (3) (2020) 850–861.

[16] G. Martí-Juan, G. Sanroma-Guell, G. Piella, A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in alzheimer's disease, Computer methods and programs in biomedicine 189 (2020) 105348.

[17] P. B. Jensen, L. J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nature Reviews Genetics 13 (6) (2012) 395–405.

[18] E. Choi, N. Du, R. Chen, L. Song, J. Sun, Constructing disease network and temporal progression model via context-sensitive hawkes process, in: 2015 IEEE International Conference on Data Mining, IEEE, 2015, pp. 721–726.

[19] Z. Qian, A. Alaa, A. Bellot, M. Schaar, J. Rashbass, Learning dynamic and personalized comorbidity networks from event data using deep diffusion processes, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 3295–3305.

[20] 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 journal of biomedical and health informatics 22 (5) (2017) 1589–1604.

[21] E. Choi, C. Xiao, W. Stewart, J. Sun, Mime: Multilevel medical embedding of electronic health records for predictive healthcare, Advances in neural information processing systems 31 (2018).

[22] E. Choi, Z. Xu, Y. Li, M. Dusenberry, G. Flores, E. Xue, A. Dai, Learning the graphical structure of electronic health records with graph convolutional transformer, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 34, 2020, pp. 606–613.

[23] A. M. Alaa, M. van der Schaar, Attentive state-space modeling of disease progression, Advances in neural information processing systems 32 (2019).

[24] G. Curigliano, D. Lenihan, M. Fradley, S. Ganatra, A. Barac, A. Blaes, J. Herrmann, C. Porter, A. Lyon, P. Lancellotti, et al., Management of cardiac disease in cancer patients throughout oncological treatment:

Esmo consensus recommendations, Annals of Oncology 31 (2) (2020) 171–190.

[25] 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.

[26] 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 Computing Surveys (CSUR) 55 (2) (2022) 1–40.