# Unraveling temporal processes using probabilistic graphical models

de Paula Bueno, M.L.

# 5

# UNDERSTANDING MULTIMORBIDITY THROUGH CLUSTERS OF HIDDEN STATES

*Nowadays, a significant portion of the population has more than one chronic disease at the same time, which is known as the problem of* multimorbidity. *Better understanding multimorbidity is hindered by the fact that most available clinical research datasets are small in size, making it harder to investigate interactions between diseases. The current availability of large volumes of routinely collected health care data is a promising source for learning about disease interaction. In this chapter, we propose a latent or hidden variable-based approach to understand patient evolution in temporal electronic health records, which can be uninformative due to the fact that it contains little detailed information. We introduce the notion of* clusters of hidden states *which may allow for an expanded understanding of the multiple dynamics that underlie events in such data. Clusters are defined as part of hidden Markov models learned from such data, where the number of hidden states is not known beforehand. We evaluate the proposed approach based on a large dataset from Dutch practices of patients that had events on medical conditions related to atherosclerosis. The discovered clusters are further correlated to medical outcomes in order to show the usefulness of the proposed method.*

## 5.1 INTRODUCTION

With the availability of large volumes of health care data, promising new data sources have come to the disposal of the research community to investigate health care problems that require much data. A typical example is the study of interactions among diseases as done in *multimorbidity* research, i.e. when multiple diseases occur at the same time in people [6, 140, 159]. Influenced by factors such as the aging of the population, multimorbidity is the rule, not the exception. Mutimorbidity research is not really feasible with typical clinical research datasets, which are small in size and usually only deal with a single disease. More recently, machine learning techniques applied to electronic health records (EHRs, for short) in the order of billion data points have been able to provide accurate predictions [142], which shows that it is possible to take advantage of such datasets, despite their low quality compared to research datasets such as those from clinical trials.

In spite of its volume-related advantages, health care data are noisy, incomplete, and usually not directly suitable for research purposes, making analysis hard. One source of data used for investigating multimorbidity and disease interaction is data collected from visits to general practitioners [106], where each patient visit is often assigned a single diagnosis code meant for administrative and billing purposes. It is, however, possible that patients have additional conditions at the time of the visit (some of which might be chronic conditions, such as hypertension or Alzheimer's disease), which would mean the existence of multimorbidity in patient. It is also often the case that symptoms and signs are not available in such health care data. As a result, one cannot directly detect multimorbidity by simply looking at GP visits individually.

With health care data, one can resort to investigating sequential disease interaction in order to partially overcome the discussed limitations of such data. By doing so, one could ultimately obtain insight on multimorbidity. Uncertainty also plays a central role because future events are typically not completely determined by the current patient status. Much research has been dedicated to the analysis of health care data, but most of it tends to focus on managerial aspects such as patient flow, hospital resources, etc. [45, 120] more often than on understanding diseases dynamics [92, 126].

In this chapter, we hypothesize that using latent information next to the diagnostic data can increase our understanding of disease interaction dynamics. By using as a basis hidden Markov models [141], multiple latent states can be associated to a given diagnostic event (where an event could be a visit due to, e.g., type 2 diabetes mellitus or a myocardial infarction). Based on this, we introduce the notion of *clusters of hidden states*, where a cluster contains all the states that produce the same observation (i.e. the same event). Although apparently simplistic, states within a cluster can have quite different dynamics in terms of transitioning patterns (i.e. how a state can be reached by or left from). By looking at these transition patterns, we will be able to give multiple roles to each event, which sheds light on the influence of such event on disease interaction. Besides the structural differences of states within a cluster, we show that these states are associated in different ways to medical outcomes. The identification of latent information has been shown valuable for gaining a better understanding of health care data [91, 92], although we pursue a different angle on what to cluster than previous research.

The contributions of this chapter are as follows. We first define the notion of clusters of states from the perspective of electronic health records. This is followed by the identification of general transition patterns that might emerge in clusters of hidden states. We then introduce a case study based on data collected from Dutch practices amounting to 32,227 patients that had visits related to atherosclerosis. Atherosclerosis is a medical condition that can be seen as an umbrella term of many other diseases, thus it is suitable for illustrating clusters and the role of their states in real-world data. Once an HMM is learned from the atherosclerosis data, we provide application-oriented interpretation to the

clusters of states by looking at a medical outcome (the number of total diseases that were registered in patients) correlated to states of clusters.

This chapter is organized as follows. Section 5.2 describes the structure of EHRs and modeling assumptions. Section 5.3 defines clusters of states and transition patterns associated to them. Section 5.4 describes the data used as case study, while in Section 5.5 the results of applying the proposed notions of state clusters to such data are discussed. Section 5.6 discusses the related work, while Section 5.7 summarizes the chapter and discusses future work.

## 5.2 HEALTH-CARE EVENT DATA

### 5.2.1 *Representation*

Let us suppose that there are $n$ possible diagnoses, each one represented by a random variable $X_i$ taking values from the domain $\{0,1\}$, with $X_i = 1$ indicating *presence* and $X_i = 0$ *absence* of diagnosis $i$. The full set of diagnosis variables is denoted by $\mathbf{X} = \{X_1, \ldots, X_n\}$. This representation allows one to represent the occurrence of multiple conditions in patients at each time point. In the considered EHRs, however, patient visits to their general practitioner are recorded such that each patient visit is typically assigned a single diagnosis code (sometimes called the *main diagnosis*), which means that effectively only one disease is registered at each time point. The main diagnosis code in patient visits can be related, e.g., to a chronic condition (e.g. diabetes mellitus) or not (e.g. a fracture).

By taking the single diagnosis assumption into account, each event can be represented by an instantiation of $\mathbf{X}$, such that $X_i = 1$ and $X_1 = \cdots = X_{i-i} = X_{i+1} = \ldots = X_n = 0$, where $X_i$ corresponds to the main diagnosis associated to the event. The time interval between any two visits is often arbitrary. Next to the diagnosis data, additional data might be available, such as medication prescription and results of lab exams.

An alternative representation would use a single variable taking values on a domain with $n$ values, which could be seen as the state space of a Markov chain. However, we prefer using individual diagnosis variables because it is more general and flexible enough for easily allowing one to add more patient information into event data if such information is available. For example, if it is known that a chronic condition previously diagnosed still occurs in the patient, one could mark the corresponding variable as active in addition to the main diagnosis of the current visit. However, additional assumptions or patient data would be required in order to confirm such previous diagnoses, as there is always some degree of uncertainty as to whether previous conditions are indeed chronic. As a consequence, we did not make such assumptions.

### 5.2.2 *Modeling*

Health care data from EHRs is often fine grained, in the sense that each event will likely reflect only information that is limited to the current patient visit. This differs, e.g., from longitudinal clinical trials [175], which are often characterized by repeated measurements of symptoms and signs associated to one or more conditions. As a consequence, data from such clinical trials normally allows for a more complete assessment of patient evolution, as opposed to health care data. This suggests that one could capture unmeasured patient information in such EHR data by including *latent variables*, such that it could provide a richer characterization of patients when combined with observable data.

In this work, hidden Markov models are used to capture the sequential interaction between observable and latent variables. In the multimorbidity context, the diagnosis variables $\mathbf{X}$ correspond to the observable variables, and we assume that there is a latent variable $S$. The usage of hidden states attempts to compensate for the mentioned difficulties present in temporal EHRs. We consider the family of independent HMMs for modeling (see Chapter 2 for details on HMMs). This choice is justified by the large amount of data in EHR datasets and the low number of observable variables (as shall be discussed in Section 5.4).

In order to comply with the event data representation, we further assume that the emission distributions of the HMM are deterministic such that only one observable variable $X_i$ is *active*, i.e. for every $S$ there is some $X_i$ such that:

$$P\left(X_j^{(t)} = 1 \mid S^{(t)}\right) = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

## 5.3 IDENTIFYING TRANSITION PATTERNS

### 5.3.1 *Clusters of states*

The events constructed from health care data imply that in order to fully comply with the data concerning $n$ diagnoses, the hidden states should be constrained to emit one out of $n$ different observations at each moment, as defined in Equation 5.1. In spite of this apparent simplicity, the underlying process being modeled could still be quite complex (e.g. by having multiple stages at different moments). In order to properly capture such distribution, more states could be needed, which can lead to the situation where multiple states are associated to the same diagnosis (e.g. if one decides to model more states than observable variables). From these considerations, we define a *cluster of states* as a set of states that have the same emission distribution.

### 5.3.2 *Transition patterns*

Modeling state transitions in a probabilistic way, e.g. as in Markov chains, implies that a state can often be reached in different ways and can lead to different future

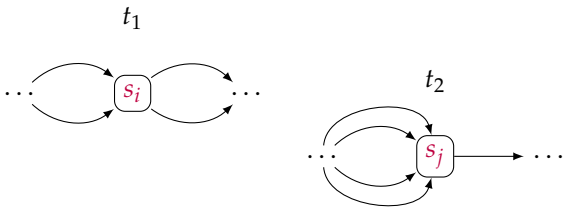



Figure 5.1: Cluster of states $C = \{s_i, s_j\}$, where $s_i$ can be reached from two states and can transition to two states, while $s_j$ can be reached from four states and can transition to a single state.

states. As we show next, by considering clusters of states such dynamics are further enriched, because such past-present-future transitioning can occur in multiple ways. For example, consider two states $s_i$ and $s_j$ belonging to a cluster $C$, as shown in Figure 5.1. This suggests that $s_i$ will likely be reached earlier for the first time than $s_j$, and it also suggests that both states can lead to quite different incoming and outgoing states. Of course, such multiple *roles* of a given diagnosis (represented by the cluster $C$) stem from the complexity of the underlying process, where a given diagnosis could be associated to different medical situations when one looks at the whole care process. For example, the states of a cluster could be associated to different levels of severity or worsening of patient health that could happen at different moments.

In order to better understand the roles of states in clusters, we discuss transition patterns that might arise. This characterization involves states and transitions from and to them, and is provided at a high level, because it is intuitively unfeasible to anticipate all the possible ways by which the states of clusters can interact.

#### 5.3.2.1 *Internal patterns*

A state is associated to an *internal transition pattern* if most of the probability mass of its incoming and outgoing probabilities associates to states from the same cluster. The most trivial internal pattern occurs when a state has a loop probability close to 1, which we call a *recurrent pattern*. A more formal description is that a state $s$ has a recurrent pattern if $s$ has a transition probability $P(S^{(t+1)} = s \mid S^{(t)} = s) \geq \alpha$, where $\alpha$ will typically be close to 1.

A more complex internal pattern would occur when there is a cycle involving two or more states from the same cluster. In this case, at any moment it is very likely that the system (e.g. a patient) is switching between the same diagnosis represented by different states. We call such patterns *internal feedback patterns*.

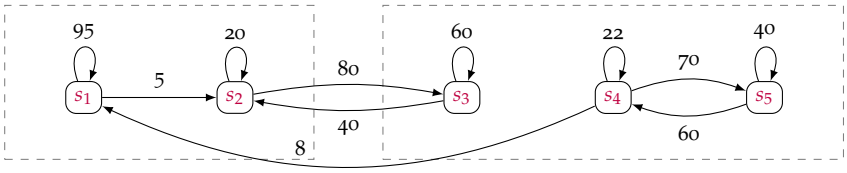



Figure 5.2: An example with two clusters of states $C1$ (left) and $C2$ (right) for depicting patterns of state transition. Probabilities are given by percentages.

#### 5.3.2.2 *External patterns*

*External transition patterns* involve states from two or more clusters. One type of such patterns are the *external feedback patterns*, which involve states from two or more clusters such that most of the incoming and outgoing probabilities stay in the cluster.

In the context of disease interaction, external patterns occur when transitions involve different diagnoses, as opposed to internal patterns. Hence, if a cluster is involved in both an internal and an external pattern, then the same diagnosis could lead to different future events. In other words, the same diagnosis could play distinct roles.

**Example 5.1.** *Suppose two clusters of states* $C1 = \{s_1, s_2\}$ *and* $C2 = \{s_3, s_4, s_5\}$*, where* $C1$ *and* $C2$ *are associated to two different diagnosis codes, as shown in Figure 5.2. It holds that state* $s_1$ *is involved in a recurrent pattern due to its high self-transition probability (for* $\alpha = 0.95$*). States* $s_4$ *and* $s_5$ *are involved in an internal feedback pattern, while states* $s_2$ *and* $s_3$ *are involved in an external pattern.*

## 5.4 CASE STUDY

In order to illustrate the value of the proposed methods, we consider the Primary Care Database from the NIVEL institute (Netherlands Institute for Health Services Research), a Dutch institute that maintains routinely electronic health records from health care providers to monitor health in Dutch patients [127]. In the NIVEL data, patient visits are assigned an ICPC code (International Classification of Primary Care) indicating a diagnosis for the visit.

### 5.4.1 *Variables and observations*

We focus on variables related to atherosclerosis, which is a cardiovascular condition that has complex associations to a number of other conditions. Although in the literature atherosclerosis has been known to be associated to chronic diseases like diabetes [95], there is still active research on its implications and associations [125, 129, 164]. In our data pre-processing steps, we first selected ICPC codes related to atherosclerosis, then groups of codes that refer to a given medical symptom or condition were built based on medical experts. As a result, each

| ICPC code, description | Variable (model) |
|---|---|
| K02.00, Pressure/tightness of heart<br>K74.00, Angina pectoris<br>K74.02, Stable angina pectoris<br>K76.01, Coronary sclerosis | *Angina* |
| K75.00, Acute myocardial infarction<br>K76.02, Previous myocardial infarction (> 4 weeks earlier) | *Myocardial infarction* |
| K89.00, Transient cerebral ischemia/TIA<br>K90.00, Cerebrovascular accident<br>K90.03, Cerebral infarct | *Cerebrovascular accident* |
| K92.01, Intermittent claudication | *Claudication* |
| K99.01, Aortic aneurysm | *Aortic aneurysm* |
| K91.00, Atherosclerosis | *Atherosclerosis* |

Table 5.1: ICPC codes related to atherosclerosis, and their mapping into variables of the model.

group of codes gave rise to an observable variable, as shown in Table 5.1. The variables constructed based on Table 5.1 can be seen as comorbidities that might occur in patients with atherosclerosis.

In order to construct the event data from the raw NIVEL data, we first ordered the raw data in ascending dates. Then, whenever a patient visit having as diagnosis one of the ICPC codes from Table 5.1 was found, a new observation was created, where the variable associated to the ICPC code was instantiated as the value 1 and the remaining variables were assigned zeros. The visits that were not associated to any of such ICPC codes were ignored.

### 5.4.2 *Sample*

We considered a sample of 32,227 patients that had visits between 1st of January, 2003 and 31st of December, 2011. To be included, a patient must have had at least one visit related to one of the diagnoses listed in Table 5.1. The data construction procedure previously discussed resulted in a dataset with 216,580 observations, where the average number of observations per patient is 6.7 (StDv = 10.9). A total of 11,932 patients have only one observation, whereas 20,295 patients have two or more.

### 5.4.3 *Number of hidden states*

In order to select an appropriate number of states when learning HMMs, the Akaike Information Criterion (AIC, for short)

$$\mathrm{AIC}(M) = 2\log K - 2\log \hat{\mathcal{L}}(M) \tag{5.2}$$

was used, where $M$ is a candidate model, $K$ is the number of parameters of $M$, and $\hat{\mathcal{L}}(M)$ is the likelihood of $M$ based on maximum likelihood estimates of the parameters.

The AIC is a less conservative model selection functions than scoring functions as BIC (see Section 2.14). This is justified in this situation because there are large amounts of data in the case study, which allows us to model more latent states by using the AIC score. The AIC score is supposed to be minimized. Models are evaluated by increasing their number of states until the addition of states does not improve the score substantially, which is an strategy to combat overfitting.

For learning of HMMs the Baum-Welch algorithm is used (see Section 2.6.2), which is sensitive to its initial parameters, especially with larger number of states. In order to reduce such effect, the best initial model was selected out of 30 candidates randomly generated.

### 5.4.4 *Clinical interpretation of clusters*

If clusters of states are identified in the learned model, one would expect that states within a cluster are indeed necessary, i.e. they should not be replaced by a single state, at the cost of, e.g, worsening model fit. The clusters of states and associated transition patterns also give insight in the *structural* role played by the states. In order to further understand the role of states of a cluster, we consider measures used in multimorbidity research. Multimorbidity measures can be used to look at patients from different angles, which is related to the notion of complexity of patient [117].

The most common way to measure multimorbidity impact in a population is by means of *disease counts* [94], in which single diseases are added resulting in a total number of diseases per patient. The count of diseases is related to the functional status and quality of life [94], thus it can be used to provide additional significance to the HMM states learned from the EHRs data. In this case study, the disease counts were calculated as the total number of distinct diagnoses that were registered for each patient, which might include other events than those listed in Table 5.1. This provides an approximation to the number of diseases that have occurred in the patient. We detail next the manner by which disease counts are associated to the latent states.

Let us consider a latent state $s_j \in \mathrm{dom}(S)$ and the $i$th patient in the data. We first compute the chances that this patient is in state $s_j$ at some instant $t$ based on the full observations of the patient, which is denoted by:

$$\gamma_t[i](j) = P(S^{(t)} = s_j \mid \mathbf{X}[i]^{(0:T_i)}) \tag{5.3}$$

where $T_i$ refers to the last observation of the $i$th patient (see Section 2.6.2 for HMM notation). When the patient has more than one observation, this will result

in a sequence of probabilities for a state $s_j$. As we will associate the states to the total number of diseases, the average of such probabilities is taken:

$$\overline{\gamma}[i](j) = \frac{1}{T_i + 1} \sum_{t=0}^{T_i} \gamma_t[i](j) \tag{5.4}$$

From Equation 5.3, if the latent variable has $k$ states $\{s_1, \ldots, s_k\}$, then each patient will be associated to $k$ average state probabilities, one for each state $s_j$. It is straightforward to see that these average probabilities sum to 1.

Once the quantities in Equation 5.4 are computed, a further analysis is performed based on the total number of diseases. In particular, we are interested in how the average occurrence of states of Equation 5.4 changes when the total number of diseases changes. To facilitate the visualization of results, such average probabilities are grouped per total number of diseases, so that we calculate the *group average* of state $s_j$ for the patients with exactly $r$ diseases, denoted by $g_r(j)$, as follows:

$$g_r(j) = \frac{1}{|D_r|} \sum_{i \in D_r} \overline{\gamma}[i](j) \tag{5.5}$$

where $D_r$ is the set of patients with exactly $r$ diseases. As a result, pairs with number of diseases and group averages are obtained, between which associations, e.g., by the Pearson correlation coefficient, are computed.

## 5.5 EXPERIMENTAL RESULTS

### 5.5.1 *Model dimension*

Figure 5.3 shows the model selection scores, which served as a basis for selecting an HMM with 9 states as the suitable model. All the states of the model were associated to fully deterministic emission distributions, such that only one diagnosis variable had a probability equal to 1 in each state, while the other variables had probabilities equal to zero. This means that the property discussed in Section 5.2.1 by which the learned model should emit events with only one active variable (representing the main diagnosis) was met.

### 5.5.2 *Clusters*

Figure 5.4 shows the learned HMM, where each state is named according to the observable that is active (i.e. the observable that has probability equal to 1). Figure 5.4 shows that three non-unitary clusters were obtained, suggesting that patient visits associated to angina, myocardial infarction and cerebrovascular accident were suitably represented by 2 states each. Intuitively, it is relevant to model a visit to, e.g., angina by means of 2 different states, hence such diagnosis could lead to two different patient courses. As expected, determining which of the two states a visit is associated to depends, e.g., on what is known so far about the patient in terms of past visits.

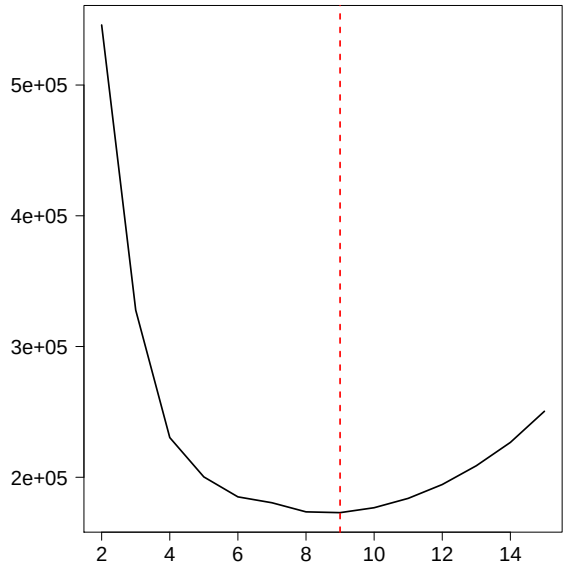



Figure 5.3: Model selection scores. X axis: number of hidden states, Y axis: AIC score. The vertical dashed line indicates the number of states where the AIC was minimal.

### 5.5.3 *Transition patterns*

Based on the state transitions of Figure 5.4, there is clearly a state in each cluster that will very likely be involved in a self-transition. These states are CVA6, Angina7 and MI3. Such states associate, therefore, to internal patterns in the form of internal recurrent patterns.

The HMM of Figure 5.4 suggests external patterns as well. In particular, angina seems to be a central event in this model: when moving from either the CVA cluster or the MI cluster, it is likely that this transition will reach the Angina cluster (in particular, the Angina5 state). Once in the Angina cluster, a transition to the other clusters is also possible, with probability larger than 0.05. Hence, such external patterns can be thought of as external feedback patterns.

### 5.5.4 *Clinical interpretation of clusters*

The average probabilities defined in Equation 5.4 are summarized by histograms in Figure 5.5. Each bar corresponds to the number of patients in which a state $s_j$ achieved some average probability. For example, the first bar of CVA2 state means that in around 30,000 patients CVA2 had an average probability between 0 and 11.1%, while for CVA6 the same mean probability was achieved in around 22,500 patients. The histograms allows one to conclude that the CVA6 state was more likely than CVA2 in most patients.

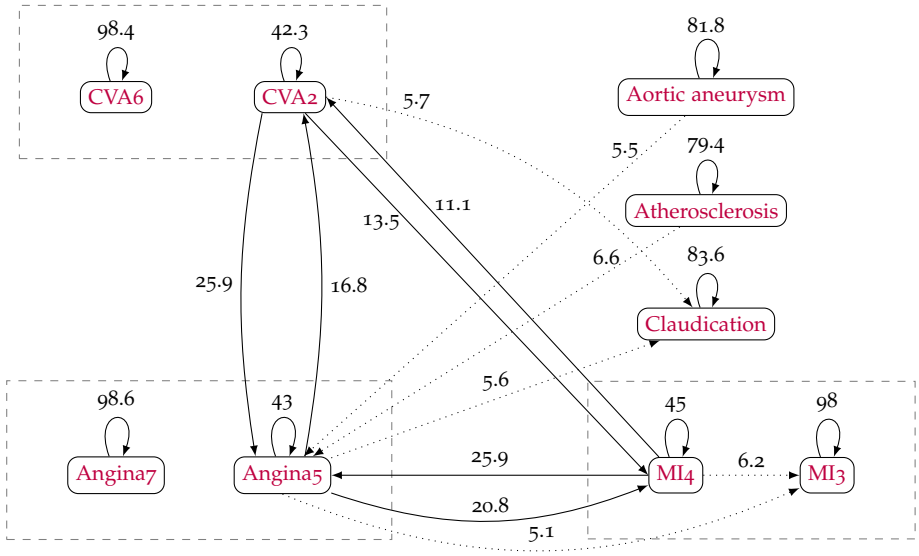
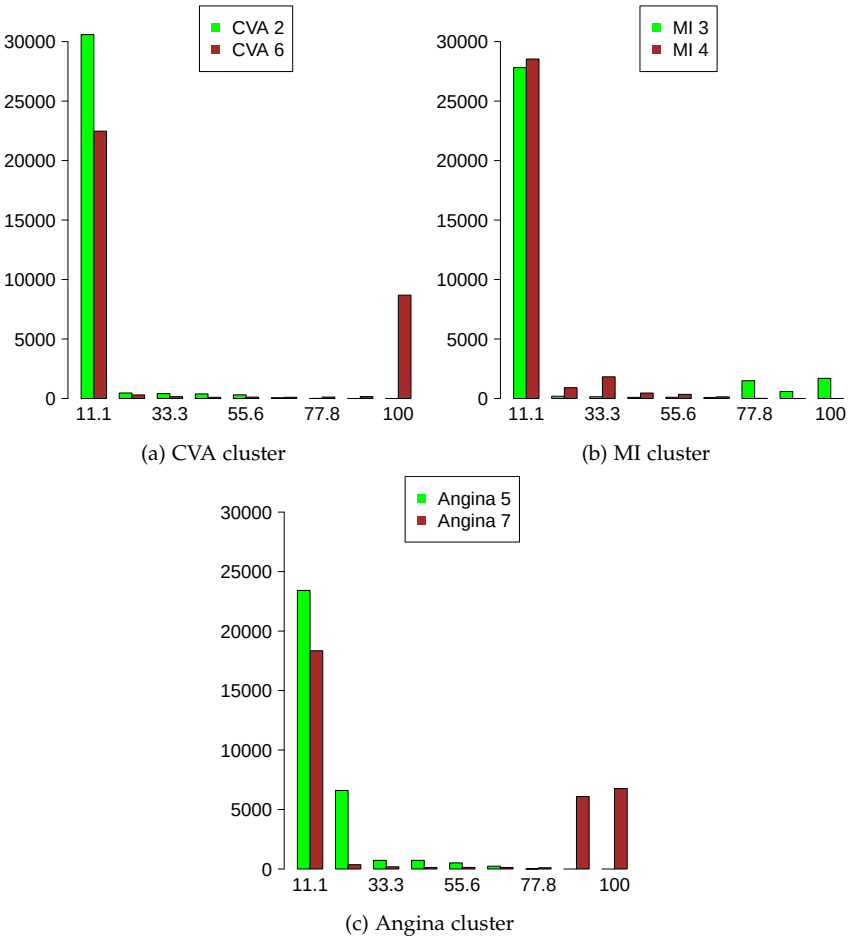



Figure 5.4: Clusters of hidden states, denoted by dashed rectangles. Arcs denote state transitions, with labels indicating probability (in %). For the sake of visualization, transitions with probability between 5 and 10% are shown by dotted lines, and only transitions with probability greater than or equal to 5% are shown.

In general, the histograms of Figure 5.5 suggest that within each cluster there are states that are substantially more prevalent than others, and such separation is more or less uniform depending on the cluster. In general, recurrent-pattern states were more likely than the non-recurrent pattern states, which might suggest that patients likely had several visits due to the same diagnosis before a diagnosis associated to a different comorbidity was registered.

For the second analysis described in Section 5.4.4, Figure 5.6 shows the total number of diseases in patients against the group probabilities. Visual inspection shows that up to 50 diagnoses the trend is substantially more stable than that of all the groups. As around 97% of the patients had at most 50 distinct diagnoses, we will focus on such groups for obtaining a better understanding of the general trend.

Figure 5.6 suggests that, in general, the states of clusters are correlated to the number of diseases in different ways. For the CVA case, patients with only a few diseases are more likely in state CVA6 (internal patterns) rather than CVA2 (external patterns). However, as the number of diseases increases, the chances to be in CVA6 decreases while the chances to be in CVA2 increases, although such trends occur at different paces. Analogously, for an MI event, it is likely the patient will be in state MI3 (internal patterns) if the patient has involves only a few diseases, but a probability decrease is expected for when more diseases are involved. On the other hand, not much can be said about MI4, as the correlation is very low. Intuitively, one would indeed expect that patients with more diseases




(a) CVA cluster

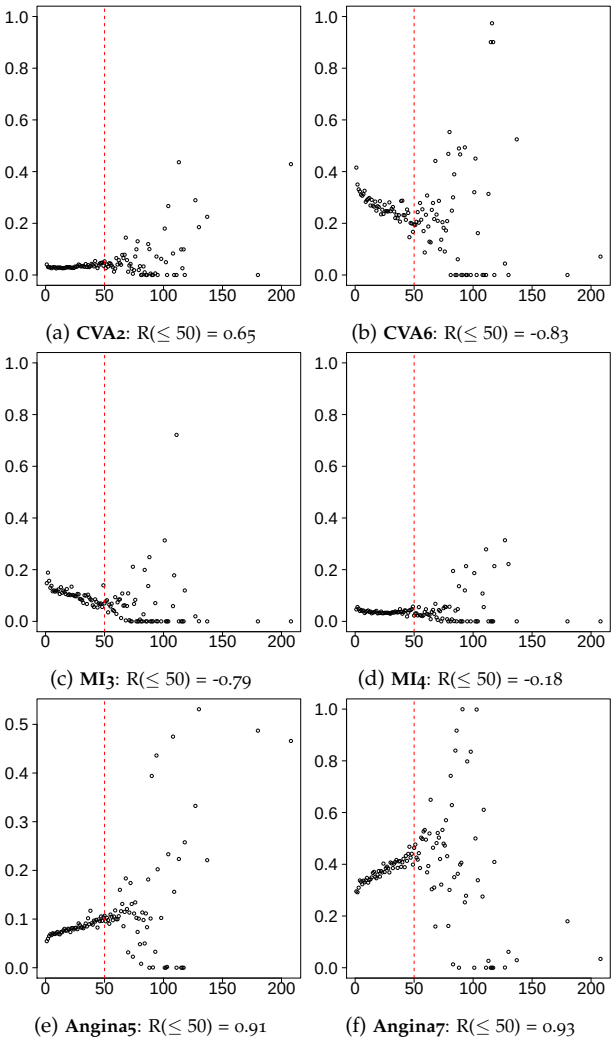
(b) MI cluster

(c) Angina cluster

Figure 5.5: Histograms of average probabilities of states (in %). X axis: average probability of state $s_j$ in the $i$th patient, i.e. the values $\overline{\gamma}[i](j)$ defined in Equation 5.4. Y axis: number of patients. For example, the first green bar in (a) means that in around 30,000 patients the state CVA2 had an average probability between 0 and 11.1%.

will be related to more transitions between the clusters, which helps explain the observed trends of the CVA and MI clusters.

As opposed to the previous clusters, Figure 5.6 suggests that the dynamics of the Angina cluster has a less straightforward association to the number of diseases. In this cluster, both of its states become more prevalent as the number of diseases increases (up to 50), which might suggest the increasing importance of angina by acting as a proxy for the comorbidities considered in this case study, as well as for other chronic and non-chronic diagnoses not explicitly considered but included in the total number of diseases.



(a) **CVA2**: R($\leq$ 50) = 0.65 (b) **CVA6**: R($\leq$ 50) = -0.83

(c) **MI3**: R($\leq$ 50) = -0.79 (d) **MI4**: R($\leq$ 50) = -0.18

(e) **Angina5**: R($\leq$ 50) = 0.91 (f) **Angina7**: R($\leq$ 50) = 0.93

Figure 5.6: Association of cluster states to clinical outcome (total number of distinct diagnoses). X axis: number of distinct diagnoses, Y axis: group averages $g_r(j)$ (Equation 5.5). The vertical line is drawn at $X = 50$. R indicates the Pearson coefficient, calculated considering only the groups with at most 50 diagnoses (which amounts to 97% of all the patients).

### 5.5.5 *Are the clusters needed? A comparison to Markov chains*

The need for the clusters learned in the HMM can be assessed by comparing the model fit of the HMM with that of a Markov chain. The state space of such MC is **X**, i.e., the six comorbidities listed in Section 5.4.1, hence learning this MC amounts to estimating the initial and transition probabilities involving the variables in **X**. This comparison can illustrate whether the multiple states

associated to a given comorbidity (in this chapter, the multiple states of CVA, MI and Angina) are indeed necessary for delivering a better model.

Table 5.2 shows the AIC scores computed for the 9-state HMM and for the MC, which indicates a superior model fit for the HMM. Besides such advantage, with the MC it is no longer possible to identify that the occurrence of a certain event such as angina, can be correlated to different patient characteristics (we used in this chapter the total amount of diseases, but other medical outcomes could be devised as well).

| **Model** | **State clusters** | **AIC** |
|---|---|---|
| 9-state HMM | 3 clusters | 172,942.8 |
| Markov chain | No clusters | 185,013.5 |

Table 5.2: AIC scores of the HMM and the Markov chain learned from the health care data. The smaller the AIC, the better the model fit is.

## 5.6 RELATED WORK

The notion of clustering states in hidden Markov models has not been investigated so far to the best of our knowledge. A related approach is clustering applied to timed automata [82, 180], where state sequences are clustered based on their distance by means of hierarchical clustering methods. Based on Bayesian HMMs that use topic modeling, clustering of patient journeys has been proposed [91], which uses the full set of events associated to unstable angina. In contrast, in our case the clusters are determined based on the states, which shifts the focus towards the dynamics that involve states within clusters. Despite their differences, our methods and those from the literature share the goal of moving towards explainable artificial intelligence [80, 114], as we aimed not only to obtain a model with suitable fit, but also to understand more about the patient situation by looking at the structure of the HMM. An example in our case is the deterministic emissions, which can facilitate interpreting models like HMMs to a great extent, at the same time obeying constraints of the multimorbidity problem.

In the context of electronic health records of multimorbidity, a cohort of the NIVEL data used in this chapter had been used for learning graphical models based on Bayesian networks, in static [106] and temporal [107] contexts. In those cases, however, the goal was to model differences in practices, hospitals, or regions, without taking into account latent variables.

## 5.7 CONCLUSIONS

In this chapter we proposed a modeling methodology for health care data from EHRs. Due to the fine-grained nature of such event data, we used HMMs for capturing latent information that is not directly measured. A first step towards

capturing clusters of latent states was taken, which are states associated to the same emission distribution. In the context of EHR data, the states of a cluster are associated to the same diagnosis code. The states of a cluster can, however, be associated to very different transitioning patterns. Based on this, we defined the notion of transition patterns.

We illustrated the proposed ideas by means of a case study with data from atherosclerosis patients collected by Dutch general practitioners. The learned HMM had 9 states, in which clusters involving angina, myocardial infarction and cerebrovascular accident were identified. This suggests that these diagnoses are too complex to be managed by a single latent state, hence a model with better fit was obtained when such diagnoses were allowed to be represented by multiple states (or roles), as we did with the obtained HMMs.

Suggestions for future work include a complementary analysis to the correlations computed between average state probabilities and the total number of diseases. Instead of computing separate correlations, one could consider regression models to predict the average probabilities for different number of diseases and states. In terms of model class, we also would like to investigate the effect of adding medication and lab exams, which are available to some patients in the NIVEL data. These could be added as model inputs (i.e. covariates), which would allow to capture switching regimes for the transitions.

Further research might also benefit from a more formal definition of clusters of hidden states allowing one to capture more general transition patterns. This could make the patterns more explainable. One could also add criteria to help decide which states are part of a cluster in a more general way, which could be of interest if the emissions are not fully deterministic (e.g. when there is a second diagnosis available in the data).