Journal of Biomedical Informatics 95 (2019) 103232

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A probabilistic framework for predicting disease dynamics: A case study of psychotic depression

Marcos L.P. Bueno[a,b,*], Arjen Hommersom[c,a], Peter J.F. Lucas[d,a], Joost Janzing[e]

[a] Institute for Computing and Information Sciences, Radboud University Nijmegen, the Netherlands
[b] Department of Computer Science, Federal University of Uberlândia, Brazil
[c] Faculty of Management, Science and Technology, Open University, the Netherlands
[d] Leiden Institute of Advanced Computer Science, Leiden University, the Netherlands
[e] Department of Psychiatry, Radboud University Nijmegen Medical Center, the Netherlands

ARTICLE INFO

Keywords:
Machine learning
Psychiatry
Depression
Temporal data
Latent variables
Hidden Markov model

ABSTRACT

Unsupervised learning is often used to obtain insight into the underlying structure of medical data, but it is not always clear how to use such structure in an effective way. In this paper, we propose a probabilistic framework for predicting disease dynamics guided by latent states. The framework is based on hidden Markov models and aims to facilitate the selection of hypotheses that might yield insight into the dynamics. We demonstrate this by using clinical trial data for psychotic depression treatment as a case study. The discovered latent structure and proposed outcome are then validated using standard depression criteria, and are shown to provide new insight into the heterogeneity of psychotic depression in terms of predictive symptoms for different interventions.

1. Introduction

Much about disease processes is unknown, as often the only available information about a disease are the patient's symptoms and signs. This might result in an incomplete understanding of a medical disorder, which can in many cases be overcome by latent variable modeling. In spite of requiring extra modeling efforts, latent variables can enhance our understanding of the problem domain by capturing unmeasured quantities (e.g. related to the underlying physiology) and their relationship to observed quantities [24], and might as well provide better fitted models [25]. Hence, by using latent variables, one can try to reconstruct the underlying structure of the process at hand by using observed data.

Unsupervised learning is the machine learning task that aims to generate representations of the underlying structure of the data. Well-established usage of unsupervised learning in medical data includes, e.g., the discovery of underlying patient groups using clustering methods [15,16], which might help improve diagnosis and provide new insight into more effective treatment selection [1]. Other applications include feature selection from unlabeled data [12] where manual feature extraction might be not available or incomplete. Patient monitoring and alerting for the identification of clinical outliers has also been tackled by unsupervised techniques [7,12]. Yet, when applied to medical data, unsupervised techniques generate output that often makes experts confront themselves with questions like *what else can we do with this structure?*. This is particularly of interest in cases where it might be difficult to define hypotheses in advance to be tested, hence some form of exploratory data analysis must be conducted.

We show in this paper that unsupervised learning methods, in particular hidden Markov models (HMMs) [18], can be used not only to describe the underlying structure but also to support the formulation of meaningful medical outcomes. Previous research suggested that the formulation of clinical outcomes might be guided by latent-variable models [16,10], with the advantage of reducing the hypothesis space to be explored by inspecting model properties. By using HMMs, we claim that one can explore hypotheses on disease dynamics by inspecting model characteristics such as transition dynamics, latent states, etc.

In order to illustrate the usage of HMMs on disease dynamics, we make use of data from a clinical trial originally designed to compare pharmacological treatments to psychotic depression (PD) [22]. PD is a severe medical condition that is associated with a high burden of disease and relatively low remission rates following pharmacological treatment [19]. Although recent research has considered PD as a homogeneous subtype of major depressive disorder [23], the possibility that this subtype itself is heterogeneous should also be considered, which would stimulate the development of subgroup adjusted prognostics and treatment modifications. In this work, we apply HMMs to one of the largest pharmacological trials of patients with PD conducted

* Corresponding author at: Faculty of Science, Radboud University Nijmegen, Postbus 9010, 6500 GL Nijmegen, the Netherlands.
E-mail addresses: mbueno@cs.ru.nl (M.L.P. Bueno), arjenh@cs.ru.nl (A. Hommersom), peterl@cs.ru.nl (P.J.F. Lucas), Joost.Janzing@radboudumc.nl (J. Janzing).

https://doi.org/10.1016/j.jbi.2019.103232
Received 8 February 2019; Received in revised form 30 April 2019; Accepted 11 June 2019
Available online 13 June 2019
1532-0464/

so far [22], aiming to explore potential differences in course characteristics in the whole sample of patients and differences in sensitivity to treatment between medication groups.

The contributions of this paper are as follows. We present a procedure to guide the exploration of hypotheses on disease dynamics by means of HMMs. We then apply this methodology to yield insight into the dynamics of PD treatments by exploring clinically meaningful outcomes. The hypotheses generated using the method are then tested based on standard clinical characterization of response and remission in PD. To the best of our knowledge, this is the first effort into a more systematic, data-driven approach for exploring hypotheses on disease dynamics based on probabilistic graphical models.

The remainder of this paper is organized as follows. In Section 2 the relevant work related to this paper is discussed. In Section 3 the proposed framework for exploring insight into latent disease dynamics is introduced. In Section 4 the psychotic depression data used as case study is described together with some descriptive statistics. In Section 5 the HMM proposed for modeling PD dynamics is detailed. The experimental results are shown in Section 6. The obtained results are validated in Section 7. Section 8 summarizes the paper and gives suggestions for future work.

2. Related work

Probabilistic graphical models have been extensively used in medicine and psychiatry. Recently, network models have shown to provide new insight into depression and other disorders by exploring symptom pathways [21,4]. These models, however, do not employ latent variables and instead claim that disease complexity emerges from direct connections between symptoms. On the other hand, latent-variable models such as hidden Markov models have been also extensively used in medical domains. One advantage of HMMs is that one can easily incorporate domain knowledge into the model, e.g., by constraining model transitions and emissions [12].

When using HMMs to capture disease dynamics, it is often the case that the number of latent states is determined in advance, as researchers might be interested in a specific subset among all possible models. Hosenfeld et al. [8] have used a two-state HMM to investigate the hypothesis that patients switch between two stable states (symptom-free versus depressed) in major depressive disorder. To investigate the relationship between cognition and psychotic symptoms in Alzheimer's disease, Seltman et al. [20] have defined a four-state continuous-time HMM. By opposition, one might argue that by not imposing an *a priori* number of or already known latent states, a more ample set of possible models is considered, which can lead to more insight into disease dynamics, at the cost of a likely increased difficult to interpret such models.

The typical usage of HMMs is in prediction or as a model to describe the underlying structure of the data [17]. While prediction is self-explanatory, the description of the underlying structure is often seen as a set of clusters, and for that reason it is a more abstract and more difficult representation to get insight from. A much more specialized usage of latent variables lies in the development of data-driven outcome measures, as suggested in [16,10]. A data-driven approach to generating outcomes has the advantage that latent states might provide a more natural, compact and empirically-oriented way to measure multiple relationships between symptoms and other observables.

More recently, HMMs have been applied to electronic health records [9,13], which are much more large (and often heterogeneous) amounts of data than usually seen before. Yet, such datasets are of very different nature and thus require new methodology for using models as HMMs for the discovery of relevant knowledge.

3. A probabilistic framework for capturing disease dynamics

In this section we discuss models suitable for capturing latent disease dynamics in a probabilistic framework.

3.1. Bayesian networks and hidden Markov models

In many problems, the measured variables reflect only part of the ongoing process as it is the case with disease symptoms, which can be seen as manifestations of some unobserved underlying disorder. Latent variables can be used to capture such unmeasured quantities and the way these relate to the observed ones [24], which results in a more complete model of the problem at hand, and might also allow for a better model fit [25]. In temporal problems, such as clinical trials for patient treatment, one is also typically interested in the sequential relationship between latent states.

Hidden Markov models are models based on latent variables that are able to cope with uncertainty and sequential phenomena, which make HMMs suitable for many biomedical problems [13,8,20]. In HMMs, the observable variables typically interact only via the latent (or state) variable [17,18], which is known as the naive-structure HMM. In this work we opt for modeling the observation space as a Bayesian network (BN), which allows for much more general representations of symptom interaction. A BN is a graphical representation of probabilistic interactions between random variables, where arcs represent unconditional dependence between variables, while the absence of arcs represents conditional (in) dependences. By modeling the observation space as a BN, more insight into the problem can be obtained by a more concise latent-state representation [3].

When learning hidden Markov models, one often resorts to learning algorithms able to handle missing data due to the latent variables of HMMs. The expectation-maximization algorithm (EM) [5,18] is a well-known approach used for learning HMMs from data. In this paper, we use the EM approach tailored for structured observation spaces (see, e.g., [3,14]), as the observation space is given as a general Bayesian network.

3.2. State trajectories

Before we describe how to use HMMs to obtain insight into disease processes, we first introduce some notation and definitions. Let us denote by S the random variable representing the latent states to be modeled, where S takes values on the set $\mathrm{dom}(S) = \{s_1, \ldots, s_k\}$, such that each $s \in \mathrm{dom}(S)$ is called a latent (or hidden) state. We denote by $\{X_1, \ldots, X_m\}$ the set of observable variables, such that the ith observation X_i takes values on some set $\mathrm{dom}(X_i)$. In medical domains, each X_i will often refer to measured data such as symptoms, lab exams, medication, etc., while the latent variable S will refer to some state of the underlying disease (e.g. a disease remitting situation). The disease process of interest is assumed discrete over the time points $\{0, \ldots, T\}$, where the value of the latent variable and the observables that hold at time t will be denoted by $S^{(t)}$ and $X_i^{(t)}$ respectively. For a discrete time interval $[t_1, t_2]$, the notation $S^{(t_1:t_2)}$ will be used.

In an HMM, one typically considers a few assumptions [11]: it is assumed that the model is time homogeneous (i.e. the model parameters are the same for every time point), the symptoms from different time points interact only mediated by the state variables, and the state variables form a first-order Markov chain. In spite of being Markovian with regard to the latent states, an HMM does not imply Markovian dynamics on the observation space.

HMMs can be used to predict the hidden states associated to observations, i.e. to compute the set of states that better explain the observations. The set of most likely states depends on the optimality criterion chosen according to the intended usage of such predictions [18]. In this paper, we seek for the states which are individually most likely, as we are interested in the chances that a patient will transition to one or more states that might represent, e.g., disease recovery. Hence, the average number of times a state is predicted to occur is the quantity of interest. This differs from the so-called Viterbi path, where one seeks for

the most likely state sequence jointly taken over $\{0, ..., T\}$.

In order to predict the states which are individually most likely, one first computes the distribution of latent states at each time point t conditional on the complete patient's symptom data (i.e. the data over all the process duration):

$$\gamma_t(s) = P(S^{(t)} = s | X_1^{(0:T)}, ..., X_m^{(0:T)}) \quad (1)$$

where $\gamma_t(s)$ is the notation used in Baum-Welch algorithm for HMMs [18]. After this has been done, the sequence of states for a given patient is obtained by selecting the most likely state at each time t:

$$\hat{s}_t = \text{argmax}_{s \in \text{dom}(S)} \gamma_t(s) \quad (2)$$

for all $t \in \{0, ..., T\}$. This can be interpreted as *assigning* patients in states. For brevity sake, we do not index the predictions of Eq. (2) by patient, although it should be clear that there is a set of predictions $\hat{s}_t$, $t \in \{0, ..., T\}$, for each patient.

3.3. Exploring medical outcomes

One way to obtain insight into disease dynamics is by considering transition dynamics between latent states. This is convenient because each latent state can take into account multiple symptom dimensions at once, which makes reasoning over patient trajectory very natural. Once the states are discovered, a detailed outcome measure that provide insight into treatment dynamics can be formulated.

We propose a procedure to build outcome measures in Fig. 1. The procedure selects a set of *baseline states* $\mathscr{S}_b$ based on a selection criterion. From the remaining states, a set of *target states* $\mathscr{S}_e$ are to be selected based on its own criterion. Once $\mathscr{S}_b$ and $\mathscr{S}_e$ are obtained, *state reachabilities* from $\mathscr{S}_b$ states to $\mathscr{S}_e$ states are calculated. By varying the time interval between two given states of $\mathscr{S}_b$ and $\mathscr{S}_e$, the resulting probabilities $reach(i, j, t_1, t_2)$ indicate the temporal influence of a baseline state over a target state. Such state reachabilities can then be used to compose a rich outcome measure, e.g., by making $t_1 = 0$ and $t_2 \in \{1, ..., T\}$, which will result in a reachability trend as indicated in Fig. 1.

3.4. Selecting states

The selection of baseline states of Fig. 1 can be viewed in general terms as a function f: $\text{dom}(S) \to \{0, 1\}$, as shown in Definition 3.1.

Definition 3.1 (*Baseline state*). We say that a latent state $s \in \text{dom}(S)$ is a *baseline state* iff $f(s) = 1$. The set of baseline states is given by:

$$\mathscr{S}_b = \{s \in \text{dom}(S): f(s) = 1\} \quad (3)$$

The set of target states $\mathscr{S}_e$ of Fig. 1 can be defined analogously.

We define in the following different criteria for selecting baseline and target states either by using model parameters or predicted patient trajectories (or both). These definitions can be seen as particular instantiations of the function f from Definition 3.1. For notation convenience, we denote by D the set of patients, which typically corresponds to the data used to learn the model.

Definition 3.2 (*Baseline-state criterion 1*). We say that a latent state $s \in \text{dom}(S)$ is a baseline state iff $\hat{s}_0 = s$ holds for at least one patient of D.

In other words, Definition 3.2 labels a state as a baseline state if one or more patients are predicted to be in this state at the process start (i.e. at $t = 0$). A more strict selection criterion of baseline states would specify a *degree of uncertainty* concerning the predictions made at baseline, as shown in Definition 3.3.

Definition 3.3 (*Baseline-state criterion 2*). We say that a latent state $s \in \text{dom}(S)$ is a *baseline state* iff all the following conditions hold for at least one patient of D:

- $\hat{s}_0 = s$
- $\gamma_0(s) \geqslant \sigma$ where $0 \leqslant \sigma \leqslant 1$ is the minimal degree of uncertainty.

Definition 3.3 allows one to specify the minimal uncertainty on the state prediction that is acceptable. For example, with $\sigma = 0.95$, one imposes that the baseline state must have been predicted with low uncertainty at $t = 0$. This notion defines how strict one is for deeming a state as a baseline state. Note that parameters such as the minimal degree of certainty and the minimum number of patients (the previous definitions required at least one patient) are part of the selection criterion and may be adjusted by the user.

For target states, Definition 3.4 presents a criterion based solely on model parameters.

Definition 3.4 (*Target-state criterion*). Let $s \in \text{dom}(S) - \mathscr{S}_b$ be a non-baseline state. We say that s is a *target state* iff $P(s \to s) \geqslant \rho$, where $0 \leqslant \rho \leqslant 1$ and $P(s_i \to s_j)$ is the transition probability between states s_i and s_j.

One can use Definition 3.4 by setting, e.g., $\rho = 0.95$, which would choose non-baseline states that have a high self-transition probability. Depending on the selection criteria, the target states could act as possible final states to the process at hand by representing different patient recovery in terms of symptom severity.

4. Data

4.1. Patients

All patients had participated in the DUDG (Dutch University Depression Group) study [22], a 7 week double-blind randomized clinical trial originally designed for comparing the effectiveness of

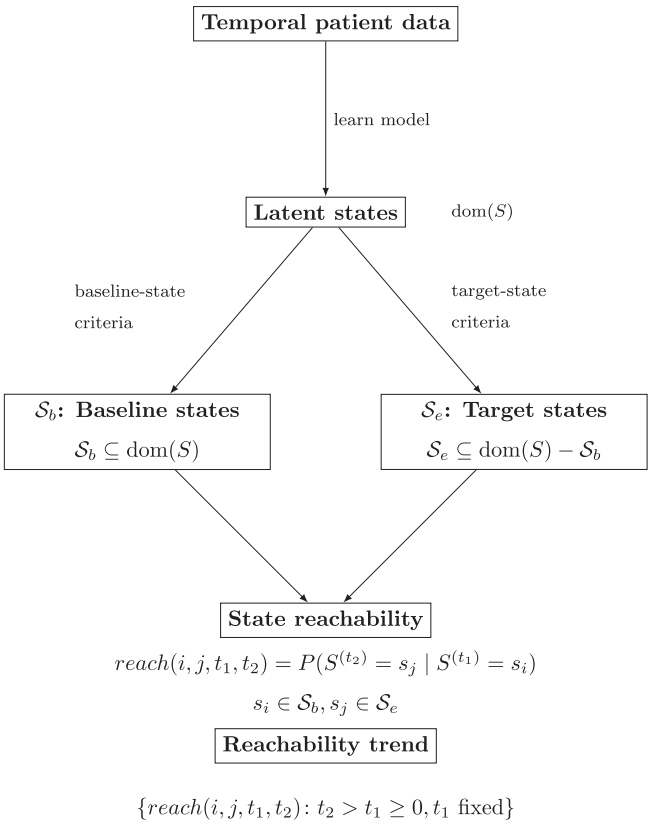


Fig. 1. Procedure to guide the generation of outcome measures based on latent-state models.

Table 1
Composition of the HAM-D score for depression. For any patient, the HAM-D score is obtained by the summing the scores of all the symptoms. The grading of items with range 0–4 is as follows: 0 – absent, 1 – mild or trivial, 2 and 3 – moderate, 4 – severe. For the other items, the grading is: 0 – absent, 1 – slight or doubtful, 2 – clearly present.

Item No.	Symptom	Score range
1	Depressed mood	0–4
2	Guilt	0–4
3	Suicide	0–4
4	Insomnia (initial)	0–2
5	Insomnia (middle)	0–2
6	Insomnia (delayed)	0–2
7	Work and interests	0–4
8	Retardation (psychomotor)	0–4
9	Agitation	0–2
10	Anxiety (psychic)	0–4
11	Anxiety (somatic)	0–4
12	Somatic symptoms (gastrointestinal)	0–2
13	Somatic symptoms (general)	0–2
14	Genital symptoms	0–2
15	Hypochondriasis	0–4
16	Loss of insight	0–2
17	Loss of weight	0–2

venlafaxine, imipramine and venlafaxine plus quetiapine (V + Q, for brevity) in PD. The dataset originally included 122 participants aged 18–65 who met DSM-IV-TR criteria for a unipolar major depressive episode with psychotic symptoms and a 17-item Hamilton Depression Rating Scale (HAM-D [6]) score of at least 18 (both at the screening visit and at baseline). Table 1 describes the symptom items used to compose the HAM-D score of each patient, which is obtained by summing the score on each item. The resultant HAMD-D score indicates severity of depression as follows: normal (0–7), mild depression (8–13), moderate depression (14–18), severe depression (19–22), and very severe depression (greater than or equal to 23).

Because of insufficient information about the specific nature of psychotic symptoms, three patients were not included in the current study resulting in a dataset with 119 patients. From the total group, 59 (49,6%) were females; the mean age was 51.1 (SD 10.9) years. Forty patients were randomized to treatment with imipramine, 38 to venlafaxine and 41 to V + Q.

4.2. Baseline and follow-up variables

Severity of depression (HAM-D, represented as a continuous variable) and the presence of psychotic symptoms (each represented as a dichotomized variable) were measured at baseline (i.e. before treatment starts) and weekly thereafter. Psychotic symptoms are delusions and hallucinations (totals at baseline, 36 and 9 in imipramine, 37 and 11 in venlafaxine, and 38 and 9 in V + Q respectively). At baseline, mean [SD] HAM-D scores were 32.5 [4.9] in imipramine, 31.7 [4.6] in venlafaxine, and 31.6 [5.4] in V + Q.

A total of 98 patients completed the trial (34 in imipramine, 30 in venlafaxine, and 34 in V + Q). Data on patients who dropped out was imputed following the last-observation-carried-forward approach, as in the original study [22].

4.3. Depression assessment

At the end of medical treatment, patients were assessed according to conventional criteria for response and remission of depression [22]. Response was defined as a reduction of at least 50% on the HAM-D score compared to baseline and a score of 14 or below, and remission as a score of 7 or below.

5. A model for psychotic depression

In this section we model the temporal latent structure of psychotic depression treatment.

5.1. General and intervention-specific model

In order to unravel treatment dynamics of the full sample of patients, as well as specific intervention-based treatment dynamics, a set of hidden Markov models are learned. The model learned from the full sample is referred to as the *general model*, while models learned from each intervention data are called *specific models*.

In order to aid comparisons of model dynamics in terms of transitioning behavior, the specific and general models share the same latent states. To this end, the general model is estimated, then each specific model is set with the obtained latent states. Then, the transition probabilities of each model are estimated using the corresponding intervention-specific data.

5.2. Model parameters and structure

The observable variables in the HMM used in this work are modeled according to the BN shown in Fig. 2, which allows for a more expressive representation than the naive-Bayes structure by connecting Hal and Del via HAM-D. By doing so, we impose less independence assumptions than the naive solution, thus the model becomes more flexible in that more dependences can be induced from data. Hence, once in a state the observables are parameterized as follows: the psychotic symptoms are encoded as binary random variables, while the depressive symptom (the HAM-D score) is parameterized as a conditional Gaussian distribution (conditioned on the state and on both psychotic symptoms, as shown in Fig. 2).

At any time point, the parameterization of each symptom is as shown in Tables 2 and 3. For a given state $s \in \mathrm{dom}(S)$, the distribution of HAM-D can be obtained by marginalizing out Del and Hal and by applying the Bayesian network factorization as follows (we omit the time index as it is equal to t):

$$P\left(\text{HAM-D} \,\middle|\, s\right) = \sum_{\text{Del,Hal}} p\left(\text{HAM-D, Hal, Del} \,\middle|\, s\right) \tag{4}$$

$$= \sum_{\text{Del,Hal}} P\left(\text{Del, Hal} \,\middle|\, s\right) p\left(\text{HAM-D} \,\middle|\, \text{Del, Hal, } s\right) \tag{5}$$

As a result, the distribution of HAM-D conditional on state s is Gaussian as it is a linear combination of the Gaussians associated to the possible configurations of Del and Hal.

Whenever the model is in a state, observations are emitted and a

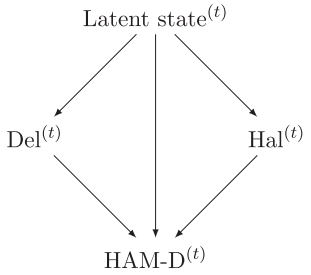


Fig. 2. Graphical structure of the HMM showing the latent variable and its direct probabilistic influence on the observables at time t. Del and Hal denote delusions and hallucinations symptoms respectively. The domain of Del and Hal is the set {absent, present}, while the domain of the state variable is a positive integer which will be determined experimentally.

Table 2
Parameterization of psychotic symptoms in the HMM. Del, Hal and S denote delusions, hallucinations and state variables respectively. Note that $P(\text{Del} = \textit{present}|S = s) = 1 - P(\text{Del} = \textit{absent}|S = s)$ and similarly for Hal.

Variable	Distribution
Del	$P(\text{Del} = \textit{absent}\|S = s)$
Hal	$P(\text{Hal} = \textit{absent}\|S = s)$

Table 3
Parameterization of the HAM-D score in the HMM. The variable HAM-D is a mixture of Gaussian distributions of the form $\mathcal{N}(\mu_i, \sigma_i)$, where μ_i and σ_i denote the mean and standard deviation of the ith combination of parents, respectively. Note that the hidden state is fixed.

Distribution of HAM-D	Parents (plus some $S = s$)
HAM-D $\sim \mathcal{N}(\mu_1, \sigma_1)$	Del = *absent*, Hal = *absent*
HAM-D $\sim \mathcal{N}(\mu_2, \sigma_2)$	Del = *absent*, Hal = *present*
HAM-D $\sim \mathcal{N}(\mu_3, \sigma_3)$	Del = *present*, Hal = *absent*
HAM-D $\sim \mathcal{N}(\mu_4, \sigma_4)$	Del = *present*, Hal = *present*

transition for the next time point is taken, and so on. The parameterization and structure discussed above are the same for all the specific models (i.e. the models obtained from each intervention data).

6. Results

6.1. Model dimension

The number of latent states was obtained by balancing model fit and interpretability. Log-likelihoods were obtained from a 10-fold cross validation procedure, where models can have from two states up to the number of states obtained prior to model overfitting (see Appendix A for more information). The selected number of states considers the mean cross-validation fit and the corresponding confidence intervals shown in Fig. A7, which is justified by the fact that in simpler models the role of latent states is more easily understood, because the states are likely more dissimilar in terms of associated symptom distribution and transition patterns. Also in favor of this procedure is the fact that the whole patient sample is split into treatment-specific data for model learning, hence models with more states would be less stable. Appendix A also shows scores of the Bayesian information criterion (BIC) which support the selection based on cross validation.

6.2. Identified states

The general model has 3 latent states, as shown in Fig. 3 (top row), where in each latent state there is one distribution for each symptom measurement (i.e., Del, Hal and HAM-D). The states can be interpreted as follows:

- The **state Hallucinations (abbreviated as state *h*)** is associated with patients with high prevalence of hallucinations and moderate prevalence of delusions. Its mean HAM-D score is the highest among all states, while it has the narrowest tail.
- The **state Delusions (abbreviated as state *d*)** is associated with patients with high prevalence of delusions and low prevalence of hallucinations. Its mean HAM-D score is moderate and has wide tail.
- The **state No Psychosis (abbreviated as state *r*)** is associated with patients with low prevalence of psychotic symptoms and moderate HAM-D score (though with wide variance).

6.3. Dynamics

Fig. 3 (bottom row) shows the transition behavior of the general model. The arcs indicate transition probabilities between latent states, e.g. the looping probability of 88.3% in state h represents the chance for reiterating in such state over two adjacent weeks. Based on Fig. 3 (top row) and on the previous characterization of the states, d and h can be seen as starting states that are primarily distinguished based on the prevalence of hallucinations in patient. Later on, depending on the patient's response to treatment, the patient will potentially move to state r. The state r can be seen as a healthier state due to the absence of psychotic symptoms, but the state does not imply depression remission or response due to its moderate mean HAM-D. In fact, the state r characterizes a wide range of no-psychosis patients, from those that still have high HAM-D to those that have achieved low HAM-D.

6.4. Comparing interventions

From the obtained latent states shown in Fig. 3, we now detail an outcome measure based on the procedure established in Section 3.3, which will also allow for comparing interventions. Based on state trajectories (Eqs. (1) and (2)), at baseline 90 patients were assigned to state d with mean (SD) probability of 100% (0), while 29 patients were assigned to state h with mean (SD) probability of 93.6% (13.2%). Hence, very little uncertainty was entailed by the model as to which initial state any given patient is predicted to be in. As a consequence, the criteria specified in Definitions 3.2 and 3.3 coincide for the PD study case, resulting in the set of baseline states $\mathcal{S}_b = \{\text{d, h}\}$. As for the set of target states $\mathcal{S}_e$, Fig. 3 shows that the state r has a self-transition probability of 98.2%, thus we set $\mathcal{S}_e = \{r\}$.

Given the sets of states $\mathcal{S}_b$ and $\mathcal{S}_e$, we define the reachability as the chances to reach the state r at time t_2 from one of the baseline states at $t_1 = 0$:

$$reach(i, j, t_1, t_2) = P(S^{(t_2)} = s_j | S^{(t_1)} = s_i) \tag{6}$$

$$reach(b, r, 0, t_2) = P(S^{(t_2)} = r | S^{(0)} = b) \tag{7}$$

where b is either the state d or the state h, and $t_2 \in \{1, \ldots, 7\}$.

In order to compare interventions, reachability values were computed from the general model (Fig. 3), as well as from the specific models (see Appendix B). The obtained reachability values were made further robust by a bagging procedure [2], where models are learned from bootstrap samples to provide more stable outcome measures. In this work, 10,000 bootstrap samples were generated, a model learned from each one and the corresponding reachability values computed. The reachability trend provided by the models learned from the bootstrap samples are then used to compute confidence intervals that indicate the variability of the reachability trend. In the following this idea will be further explored for comparing the general and specific models learned from the full sample and from each intervention data, respectively.

6.5. Reachability trend per treatment

Fig. 4 shows the reachability trends grouped by intervention. The difference between the area under the curve (AUC) of each trend was also computed. For the whole sample of patients, the 95% bootstrap confidence interval (BCI) of the AUC difference was [0.17, 2.29], while for the slope difference the AUC was [0.02, 0.17], where positive values indicate a stronger trend in favor of state d. Under venlafaxine, the AUC difference was [0.16, 3.09], whereas the slope difference was

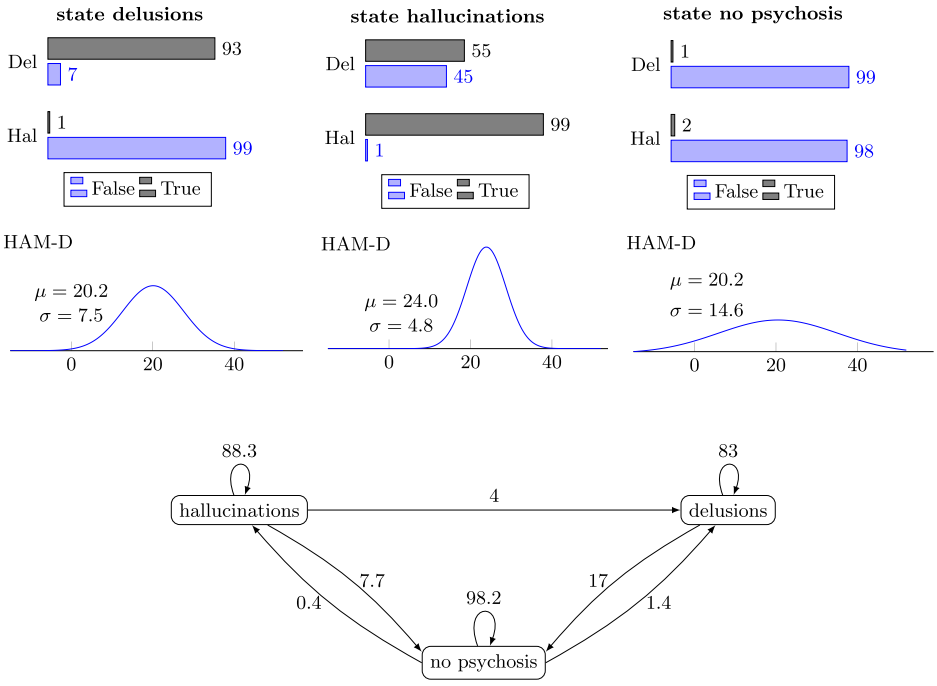


Fig. 3. Top: marginal distributions of symptoms in the latent states of the general model (Del and Hal stand for delusions and hallucinations symptoms, respectively). Bottom: dynamics of the general model. Labels indicate transition probabilities between states (in %).

[0.01, 0.23]. These results suggest that the initial state of the patient is relevant under venlafaxine in that starting in state *d* allowed for a significantly stronger reachability towards state *r* than the reachability had the patient started in state *h*.

Under imipramine, the AUC difference was [−1.62, 2.36] and the slope difference was [−0.15, 0.19]. Finally, for V+Q the AUC difference was [−0.74, 3.72], and [−0.10, 0.32] for the slope difference. Hence, starting in state *d* for imipramine and for V+Q also provided stronger trends towards *r*, but not to a significant extent. The detailed difference BCIs per week can be found in Appendix C.

6.6. Reachability trend per starting state

Fig. 5 shows the reachability trends of Fig. 4, now grouped by starting state. Patients can either start in state *d* (Fig. 5a) or in state *h* (Fig. 5b). Fig. 5a suggests that if a patient had no hallucinations at baseline (i.e. started in state *d*), then a stronger reachability trend would be achieved if treated with V+Q. For patients that had experienced hallucinations (i.e. started in state *h*), the results suggest that the strongest trend would be achieved with imipramine. Nevertheless, 95% BCIs indicate that no significant differences were found when comparing the trends starting in *h*, nor when comparing those starting in *d*.

7. Validation

In this section we investigate if aspects of the learned model and the formulated outcome can be associated to standard depression criteria computed directly from the data, as means to validate the model and the outcome.

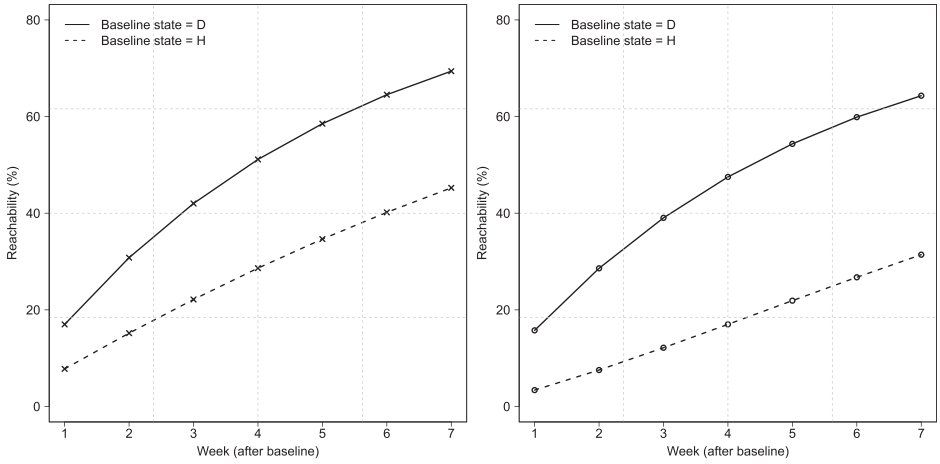
7.1. Model validation

Associations between model outputs in the form of state trajectories (see Section 3.2) and depression recovery (see Section 4.3) were computed. For each patient, we counted the number of consecutive weeks in which state *r* was predicted as the most likely state (see Eq. (2)). In case the endpoint of patient state trajectory is not predicted as state *r*, the assigned count is zero. Among the total sample, 60 patients achieved depression response, with the state *r* predicted in 4.7 weeks on the average, while the 59 patients who did not achieve response had the state *r* predicted in 1.3 weeks on the average. Fig. 6 shows a histogram of the number of patients versus the number of consecutive weeks for which state *r* was predicted. A Fisher's exact test was applied to compare the counts of the two groups from Fig. 6 (responders versus non-responders), which resulted in a p-value <0.001, suggesting that these two groups (responders and non-responders) associate significantly different to the number of weeks in the state *r* (under a 95% confidence level).

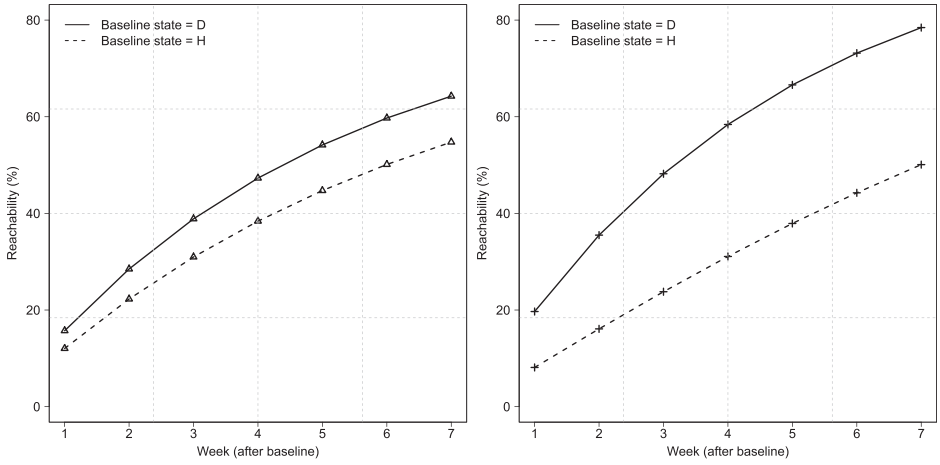
Among the total sample, 35 patients achieved depression remission, with the state *r* predicted in 5.4 weeks on the average, while the 84 who did not achieve remission had the state *r* predicted in 2.0 weeks on the average. A Fisher's exact test to compare remitters versus non-remitters resulted in a p-value <0.001 (histograms for remitters were omitted due to the small numbers). These results support the claim that the state *r* is meaningful in terms of distinguishing patients that achieved depression recovery (either response or remission) from those who did not.

7.2. Outcome validation

We now assess the claim of Section 6.5 that the state at baseline leads to significantly different state reachability for the total sample case. To this end, two distinct groups of patients were considered: patients with hallucinations at baseline (29 patients, see Section 4.2), and patients with no hallucinations at baseline (90 patients). The HAM-D scores of these groups at treatment endpoint were compared using a Mann-Whitney test for independent samples, which resulted in a p-value = 0.0007, thus suggesting that these two groups differ significantly (under a 95% confidence level). As a consequence, the psychotic symptom at baseline is predictive to depression recovery of patients in general. This evidence supports the conclusions for the model-based outcome drawn in Section 6.5, where the psychotic symptom at baseline was found to be predictive to reaching the state *r* when one considers all the patients (Fig. 4a).

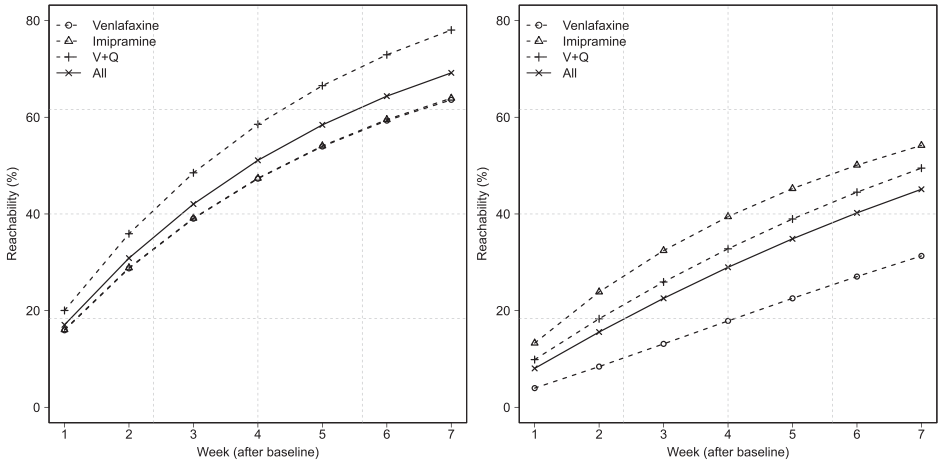


(a) All patients (b) Venlafaxine



(c) Imipramine (d) V+Q

Fig. 4. Reachability trends per intervention.



(a) At baseline: state d (b) At baseline: state h

Fig. 5. Reachability trends per latent state. The Y axis denotes the reachability at each week after baseline.

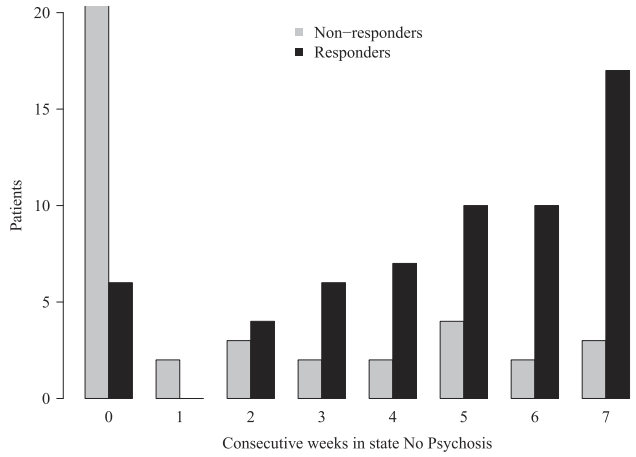


Fig. 6. Histogram of the number of times the state r was predicted in patient state trajectory. The two groups refer to patients who achieved depression response (60 patients) and those who did not (59 patients). For the sake of visualization, zero consecutive weeks for non-responders was cut down (original value was 41 patients).

8. Conclusions

This paper demonstrated that probabilistic graphical models can reveal insight into disease dynamics by considering not only the underlying structure, but also meaningful outcome measures built from such structure. We illustrated the proposed methodology by applying hidden Markov models to psychotic depression treatment data, where the models were learned in a fully data-driven way.

The identified temporal symptom structure of psychotic depression revealed that patients differed in their prognosis depending on the type of psychotic symptoms they exhibited at baseline (hallucinations versus delusions). This result was observed for the total sample and for the patients that underwent venlafaxine intervention. Hence, our methodology allowed to shed light on the heterogeneity of psychotic depression. As future work, we plan to further investigate the clinical significance of the results, as well as consider the effect of potential confounders, such as patient demographic data.

The combination of graphical models and a data-driven approach can be easily integrated into the investigation of other psychiatric disorders as well, potentially helping physicians to understand disease dynamics and may even support them in prescribing optimal pharmacological therapy. Furthermore, by applying the proposed methodology to other diseases, it should be possible to assess the method more broadly. It could of interest to perform different calculations of state trajectories that reflects the availability of only partial symptom data (e.g. to simulate an ongoing treatment), or even calculate state reachability from different starting points other than the baseline point. One could also consider adding intermediate states to the proposed framework, which could allow for greater flexibility in situations where many more latent states are obtained.

Funding

This work has been partially funded by the Netherlands Organization for Scientific Research (62001863) and by the European Regional Development Fund (NanoSTIMA). Project “NORTE-01-0145-FEDER-000016” (NanoSTIMA) is financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

The DUDG study [22] was supported by grants from AstraZeneca and Wyeth Pharmaceuticals, both of which also provided the study medication. The study was initiated by Willem Nolen. Jaap Wijkstra and Willem Nolen wrote the protocol in collaboration with the DUDG group. Jaap Wijkstra was the coordinator of the study. We are very grateful to the other DUDG collaborators: Drs. WW van den Broek, TK Birkenhäger, JA Bruijn, ML van der Loos, LM Breteler and RJ Verkes.

Declaration of Competing Interest

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

Appendix A. Model selection scores

Fig. A7 shows 10-fold cross-validation mean log-likelihoods for different number of latent states, together with 95% confidence intervals. The higher the log-likelihood of a model the better fitted such model is.

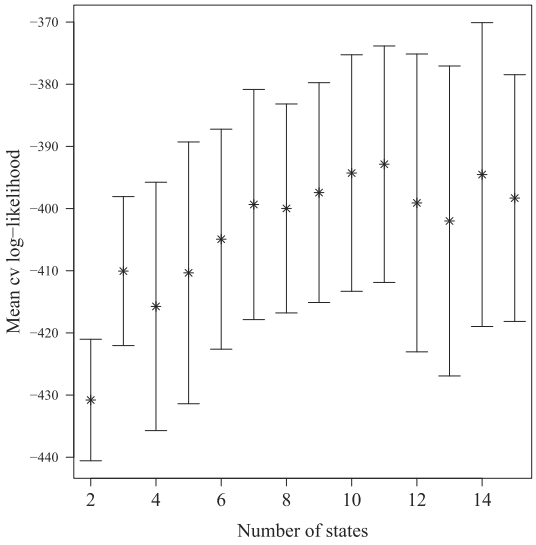


Fig. A7. 95% CIs for the mean cross-validation log-likelihoods for selecting the number of states of the general HMM.

A Mann-Whitney test was performed for comparing the cross-validation log-likelihoods shown in Fig. A7 of the 3 state model with that of other models. The resultant p-values (number of states) were: 1.0(4), 0.91(5), 0.48(6), 0.31(7), 0.35(8); the maximum p-value of the remainder cases (9 up to 15 states) was 0.08.

In addition to the 10-fold cross validation results, BIC (Bayesian Information Criterion) scores were computed for different number of states, which balances goodness of fit with a penalty based on the number of parameters and sample size. The BIC for a model M is defined as follows [11]:

$$\mathrm{BIC}(M) = -2\cdot\ln(L) + K\cdot\ln(n) \tag{A.1}$$

where L is the maximized likelihood of the model M, K is the number of parameters of M, and n is the sample size. We seek for models that minimize the BIC.

Fig. A8 shows the BIC scores for different models, suggesting the 3-state model achieves the minimal model selection score. This is in line with Fig. A7, where the overlapping confidence intervals suggest that it is likely not significant the improvement achieved by models with more than 3 states, hence a suitable dimension would be 3 states.

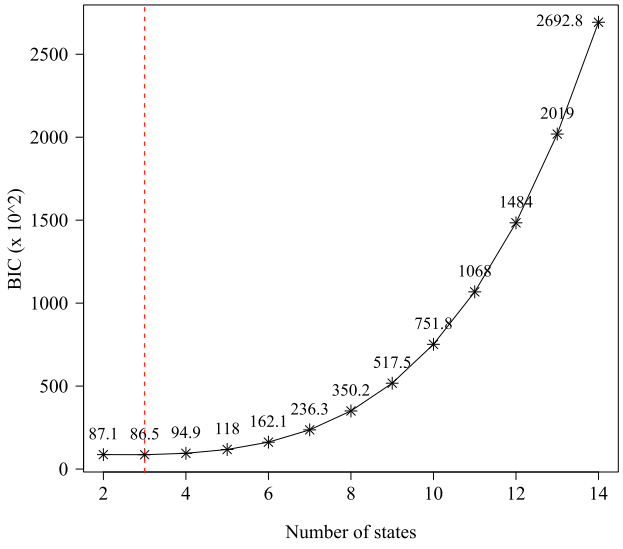
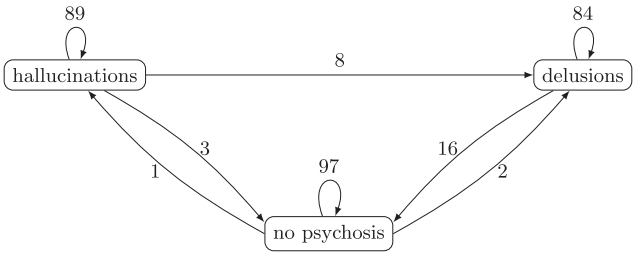


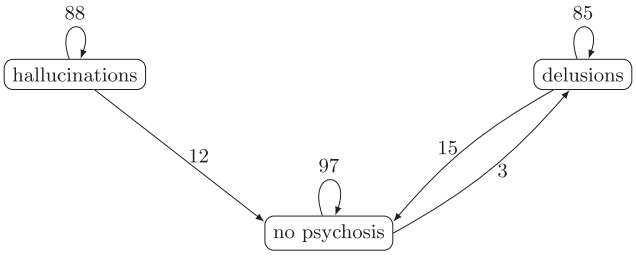
Fig. A8. BIC scores of models with different number of latent states. The vertical dashed line indicates the number of states which led to the minimal BIC.

Appendix B. Dynamics of intervention-specific models

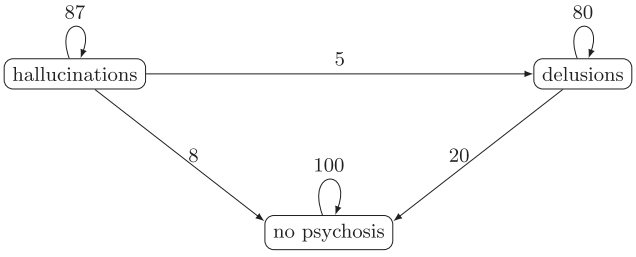
Fig. B9 shows the transition probabilities of each intervention-specific model. As described in Section 5.1, all the specific models and the general model share the same latent states, which are shown in Fig. 3 (top row).



(a) Venlafaxine



(b) Imipramine



(c) V+Q

Fig. B9. Dynamics of the intervention-specific models. Labels indicate transition probabilities between states.

Appendix C. Confidence intervals of reachability trend differences

Fig. C10 shows 95% bootstrap confidence intervals for the differences between the reachability trends of Fig. 4.

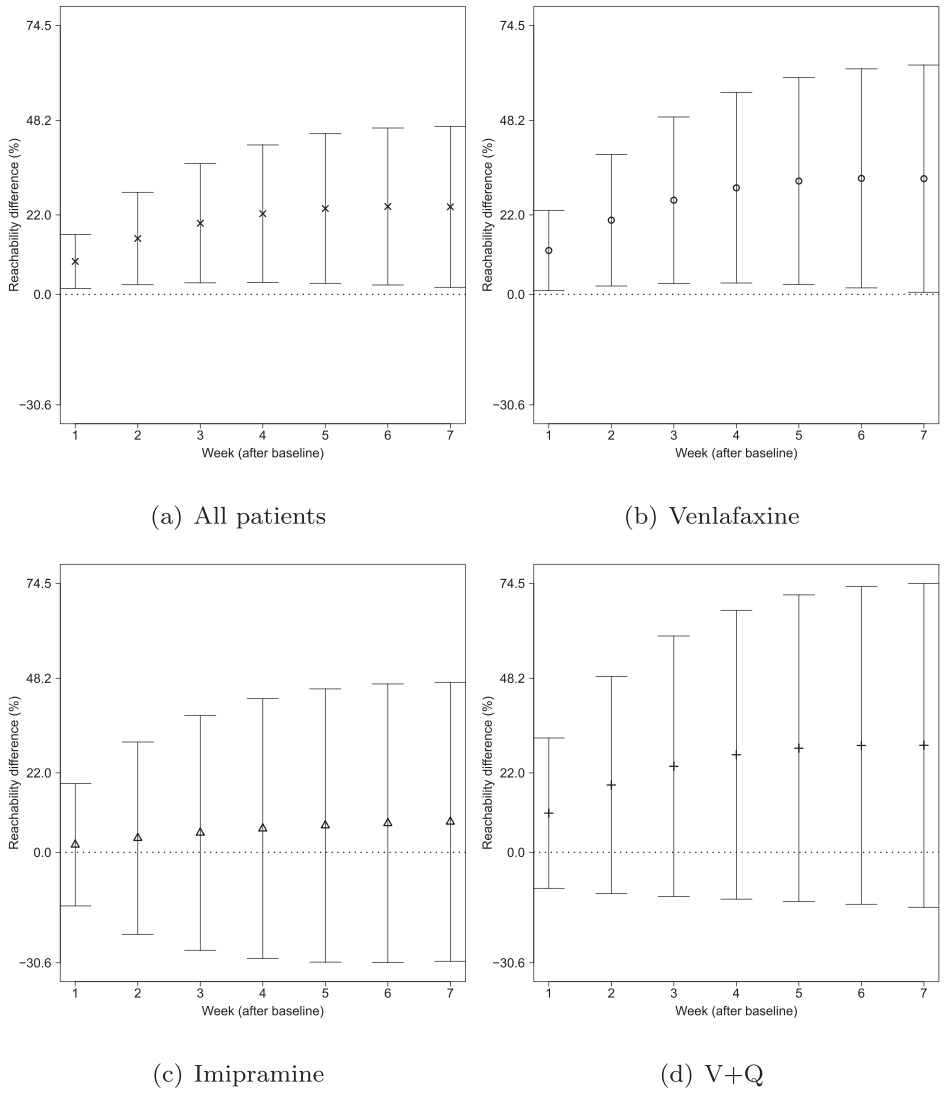


Fig. C10. 95% bootstrap confidence intervals for the differences between reachability trends. The dotted line indicates a difference equal to zero. Positive values indicate higher reachability of state *d* compared to that of state *h*.

References

[1] Emma Ahlqvist, Petter Storm, Annemarie Käräjämäki, Mats Martinell, Mozhgan Dorkhan, Annelie Carlsson, Petter Vikman, Rashmi Prasad, Dina Mansour Aly, Peter Almgren, Ylva Wessman, Nael Shaat, Peter Spegel, Hindrik Mulder, Eero Lindholm, Olle Melander, Ola Hansson, Ulf Malmqvist, Åke Lernmark, Kaj Lahti, Tom Forsén, Tiinamaija Tuomi, Anders Rosengren, Leif Groop, Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables, Lancet Diabetes Endocrinol. 6 (5) (2018) 361–369, https://doi.org/10.1016/S2213-8587(18)30051-2.

[2] Leo Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.

[3] Marcos L.P. Bueno, Arjen Hommersom, Peter J.F. Lucas, Alexis Linard, Asymmetric hidden Markov models, Int. J. Approximate Reasoning 88 (2017) 169–191, https://doi.org/10.1016/j.ijar.2017.05.011 ISSN 0888-613X.

[4] Angélique O.J. Cramer, Lourens J. Waldorp, Han L.J. van der Maas, Denny Borsboom, Comorbidity: a network perspective, Behav. Brain Sci. 33 (2-3) (2010) 137–150, https://doi.org/10.1017/S0140525X09991567.

[5] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Stat. Soc. Ser. B 39 (1) (1977) 1–38.

[6] Max Hamilton, A rating scale for depression, J. Neurol. Neurosurg. Psychiatr. 23 (1) (1960) 56–62, https://doi.org/10.1136/jnnp.23.1.56 ISSN 0022-3050.

[7] Milos Hauskrecht, Iyad Batal, Michal Valko, Shyam Visweswaran, Gregory F. Cooper, Gilles Clermont, Outlier detection for patient monitoring and alerting, J. Biomed. Inform. 46 (1) (2013) 47–55, https://doi.org/10.1016/j.jbi.2012.08.004 ISSN 1532-0464.

[8] Bettina Hosenfeld, Elisabeth H. Bos, Klaas J. Wardenaar, Henk Jan Conradi, Han L.J. van der Maas, Ingmar Visser, Peter de Jonge, Major depressive disorder as a nonlinear dynamic system: bimodality in the frequency distribution of depressive symptoms over time, BMC Psychiatr. 15 (1) (2015) 222, https://doi.org/10.1186/s12888-015-0596-5 ISSN 1471-244X.

[9] Zhengxing Huang, Zhenxiao Ge, Wei Dong, Kunlun He, Huilong Duan, Probabilistic modeling personalized treatment pathways using electronic health records, J. Biomed. Inform. 86 (2018) 33–48, https://doi.org/10.1016/j.jbi.2018.08.004 ISSN 1532-0464.

[10] Waguih William IsHak, Wes Bonifay, Katherine Collison, Mark Reid, Haidy Youssef, Thomas Parisi, Robert M. Cohen, Li Cai, The recovery index: a novel approach to measuring recovery and predicting remission in major depressive disorder, J. Affect. Disord. 208 (2017) 369–374, https://doi.org/10.1016/j.jad.2016.08.081 ISSN 0165-0327.

[11] Daphne Koller, Nir Friedman, Probabilistic Graphical Models: Principles and Techniques, The MIT Press, 2009 ISBN 0262013193, 9780262013192.

[12] Martin Längkvist, Lars Karlsson, Amy Loutfi, Sleep stage classification using unsupervised feature learning, Adv. Artif. Neu. Sys. (2012) 5:5, https://doi.org/10.1155/2012/107046 ISSN 1687-7594.

[13] Jens Meier, Andreas Dietz, Andreas Boehm, Thomas Neumuth, Predicting treatment process steps from events, J. Biomed. Inform. 53 (2015) 308–319, https://doi.org/10.1016/j.jbi.2014.12.003 ISSN 1532-0464.

[14] J. O'Connell, S. Højsgaard, Hidden semi Markov models for multiple observation sequences: the mhsmm package for R, J. Stat. Softw. 39 (2011) 1–22.

[15] Matteo Paoletti, Gianna Camiciottoli, Eleonora Meoni, Francesca Bigazzi, Lucia Cestelli, Massimo Pistolesi, Carlo Marchesi, Explorative data analysis techniques and unsupervised clustering methods to support clinical assessment of chronic obstructive pulmonary disease (copd) phenotypes, J. Biomed. Inform. 42 (6) (2009) 1013–1021, https://doi.org/10.1016/j.jbi.2009.05.008 ISSN 1532-0464.

[16] Martin P. Paulus, Murray B. Stein, Michelle G. Craske, Susan Bookheimer, Charles

T. Taylor, Alan N. Simmons, Natasha Sidhu, Katherine S. Young, Boyang Fan, Latent variable analysis of positive and negative valence processing focused on symptom and behavioral units of analysis in mood and anxiety disorders, J. Affect. Disord. 216 (2017) 17–29, https://doi.org/10.1016/j.jad.2016.12.046 ISSN 0165-0327.

[17] Jennifer Pohle, Roland Langrock, Floris M. van Beest, Niels Martin Schmidt, Selecting the number of states in hidden Markov models: pragmatic solutions illustrated using animal movement, J. Agric. Biol. Environ. Stat. 22 (3) (2017) 270–293, https://doi.org/10.1007/s13253-017-0283-8 ISSN 1537-2693.

[18] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286, https://doi.org/10.1109/5.18626 ISSN 0018-9219.

[19] Anthony J. Rothschild, Challenges in the treatment of major depressive disorder with psychotic features, Schizophr. Bull. 39 (4) (2013) 787–796, https://doi.org/10.1093/schbul/sbt046.

[20] Howard J. Seltman, Shaina Mitchell, Robert A. Sweet, A Bayesian model of psychosis symptom trajectory in Alzheimer's disease, Int. J. Geriat. Psychiatr. 31 (2) (2016) 204–210, https://doi.org/10.1002/gps.4326.

[21] C. van Borkulo, L. Boschloo, D. Borsboom, B.H. Penninx, L.J. Waldorp, R.A. Schoevers, Association of symptom network structure with the course of depression, JAMA Psychiatr. 72 (12) (2015) 1219–1226, https://doi.org/10.1001/jamapsychiatry.2015.2079.

[22] J. Wijkstra, H. Burger, W.W. Van Den Broek, T.K. Birkenhäger, J.G.E. Janzing, M.P.M. Boks, J.A. Bruijn, M.L.M. Van Der Loos, L.M.T. Breteler, G.M.G.I. Ramaekers, R.J. Verkes, W.A. Nolen, Treatment of unipolar psychotic depression: a randomized, double-blind study comparing imipramine, venlafaxine, and venlafaxine plus quetiapine, Acta Psychiatr. Scand. 121 (3) (2010) 190–200, https://doi.org/10.1111/j.1600-0447.2009.01464.x.

[23] J. Wijkstra, J. Lijmer, H. Burger, A. Cipriani, J. Geddes, W.A. Nolen, Pharmacological treatment for psychotic depression, Cochrane Database Syst. Rev. (7) (2015), https://doi.org/10.1002/14651858.CD004044.pub4 ISSN 1465-1858.

[24] Barbaros Yet, Zane Perkins, Norman Fenton, Nigel Tai, William Marsh, Not just data: a method for improving prediction with knowledge, J. Biomed. Inform. 48 (2014) 28–37, https://doi.org/10.1016/j.jbi.2013.10.012 ISSN 1532-0464.

[25] Nevin L. Zhang, Thomas D. Nielsen, Finn V. Jensen, Latent variable discovery in classification models, Artif. Intell. Med. 30 (3) (2004) 283–299, https://doi.org/10.1016/j.artmed.2003.11.004 ISSN 0933-3657. Bayesian Networks in Biomedicince and Health-Care.