Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM)

Truyen Tran[a,b,*], Tu Dinh Nguyen[a], Dinh Phung[a], Svetha Venkatesh[a]

[a] *Center for Pattern Recognition and Data Analytics, Deakin University, Geelong, Victoria, Australia*

[b] *Department of Computing, Curtin University, Perth, Western Australia, Australia*

[*] *Corresponding author. E-mail address: truyen.tran@deakin.edu.au*

Abstract

Electronic medical record (EMR) offers promises for novel analytics. However, manual feature engineering from EMR is labor intensive because EMR is complex - it contains temporal, mixed-type and multimodal data packed in irregular episodes. We present a computational framework to harness EMR with minimal human supervision via restricted Boltzmann machine (RBM). The framework derives a new representation of medical objects by embedding them in a low-dimensional vector space. This new representation facilitates algebraic and statistical manipulations such as projection onto 2D plane (thereby offering intuitive visualization), object grouping (hence enabling automated phenotyping), and risk stratification. To enhance model interpretability, we introduced two constraints into model parameters: (a) nonnegative coefficients, and (b) structural smoothness. These result in a novel model called eNRBM (EMR-driven nonnegative RBM). We demonstrate the capability of the eNRBM on a cohort of 7,578 mental health patients under suicide risk assessment. The derived representation not only shows clinically meaningful feature grouping but also facilitates short-term risk stratification. The F-scores, 0.21 for moderate-risk and 0.36 for high-risk, are significantly higher than those obtained by clinicians

and competitive with the results obtained by support vector machines.

Keywords: Electronic medical records, vector representation, medical objects embedding, feature grouping, suicide risk stratification.

---

1. Introduction

Modern electronic medical records (EMRs) have changed the landscape of clinical data collecting and sharing, facilitating efficient care delivery [1]. The data in EMR offers insights into key questions: What are the comorbidity patterns? [2] What are the relationships between diseases and interventions under multimorbidity? What is the risk of adverse events for this patient? [3] However, it remains an open problem in formulating efficient mining techniques to discover these answers [4]. This is partly due to the complexity of the EMR data. The EMR contains a mixture of static, temporal, type-specific data packed in irregular episodes. Huge effort is required for extracting meaningful features [4] and developing prognostic models from EMR [5].

We hypothesize that the answers lie in *unsupervised learning* of EMR representations [4, 6]. Unsupervised learning lets clinical patterns emerge through the learning process. We approach the problem by utilizing a recent advancement in deep learning [7, 8]. In particular, we adopt restricted Boltzmann machines (RBM) [9] as a *generative model of EMR*. RBM has a bipartite structure, in which an input layer is connected to a representation layer. The input layer consists of observed clinical variables over multiple periods of time. The representation layer is composed of unobserved binary factors, which act as the underlying aspects of illness and healthcare processes. These aspects jointly generate clinical observables. The RBM transforms raw, high-dimensional and mixed-type EMR data into a homogeneous representation. Clinical objects such as disease, procedure and health trajectory are *embedded* in the same vector space. The

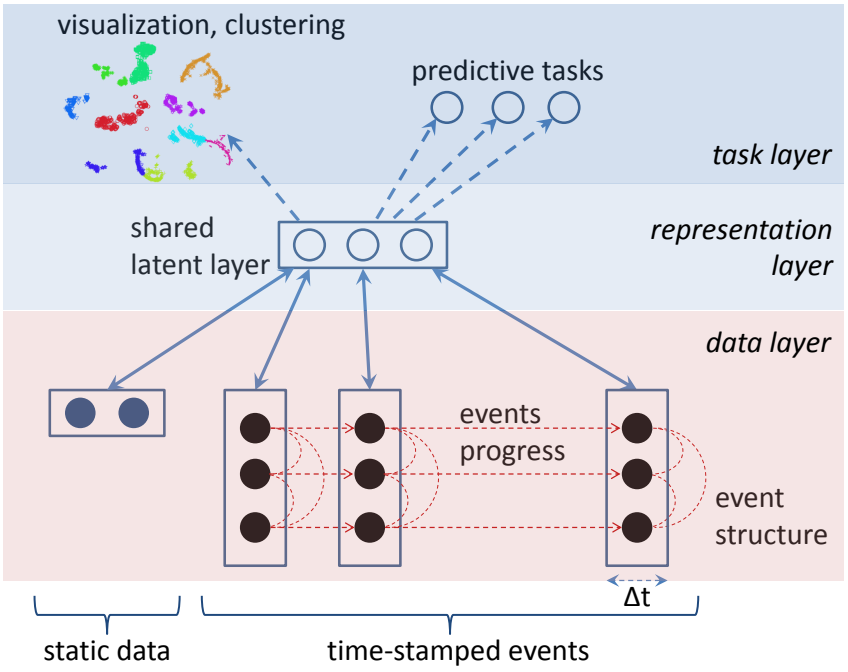

Figure 1: *e*NRBM for EMR modeling, visualization and prognosis. The data layer represents raw information extracted from EMR; the representation layer exhibits higher-level semantics; and the task layer makes use of the derived representation for tasks of interest. The connections between the data and representation layers are undirected, letting patterns emerge through information passing in both directions. Filled nodes represent observed variables, empty nodes the hidden. Boxes represent groups of variables that share the same property (e.g., time interval). Event structures and progression (represented as thin dashed lines and curves) are implicitly captured through regularization in the learning process (Sec. 3.2)

embedding facilitates visualization, manipulation and risk prognosis. See Fig. 1 for a graphical illustration of the RBM-based framework.

The standard RBM, however, suffers from two key limitations that hinder its usability in the clinical context. First, the embedding coefficients can be either positive or negative, making interpretation of group membership difficult. Second, the RBM assumes unstructured inputs but ignores explicit structures inherent in the EMR, leading to incoherent grouping.

We modify the RBM to overcome these limitations. First, the embedding coefficients are constrained to be nonnegative. This leads to model sparsity

where only a few embedding coefficients are non-zeros. Each latent factor corresponds to a small group of features which potentially play the role of a derived phenotype. Second, model learning is guided by clinical structures derived from the disease taxonomy, the procedure hierarchy and the temporal progression of illness and care. These two modifications result in a novel model called *EMR-driven nonnegative RBM* (*e*NRBM).

We validate the proposed *e*NRBM on a large cohort of 7,578 mental health patients in several tasks, including disease/procedure embedding and visualization, comorbidity grouping, and short-term suicidal risk stratification. We demonstrate that *e*NRBM-based embedding leads to meaningful grouping of diseases and interventions. The merit of the proposed method is highlighted by comparing the predictive performance on risk stratification against support vector machines.

The rest of the paper is organized as follows. Sec. 2 introduces restricted Boltzmann machines. Sec. 3 presents the main contributions of the paper: (a) an introduction of the RBM as a generative model of the EMR; (b) introducing medical object embedding; (c) introducing nonnegative coefficients into the RBM leading to coherent feature grouping and more compact representations; and (d) adding structural constraints into the RBM by exploiting inherent structures in the EMR. This is followed by an experimental section which demonstrates the capacity of the proposed methods on a large cohort of mental health patients. Finally, Sec. 5 discusses findings, limitations and future work.

2. Preliminaries

Restricted Boltzmann machine (RBM) is a type of neural networks. As illustrated in Fig. 2, a RBM is a bipartite graph consisting of: (i) an input layer of *visible units* that encode the observables (e.g., disease occurrences), (ii) a la-

tent layer of *hidden units*, and (iii) weighted connections between every visible unit to hidden units [8, 9]. The RBM differs from standard neural networks in important ways. First, it is stochastic rather than deterministic: Variables are randomly distributed according to a joint distribution specified by the model. Second, the network is undirected allowing information to propagate in both directions (feedforward and feedback modes). And finally, learning is unsupervised without labels.

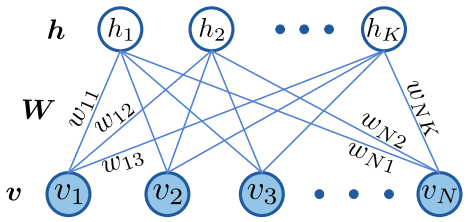

Figure 2: Graphical illustration of a RBM representing connections between input observations given through the N visible units (shaded) with K hidden units (clear). The connections are undirected and the weights represent the strength of connections.

Let $\boldsymbol{v}$ denote the set of visible variables: $\boldsymbol{v} = (v_1, v_2, ..., v_N) \in \{0,1\}^N$ and $\boldsymbol{h}$ the set of hidden factors: $\boldsymbol{h} = (h_1, h_2, ..., h_K) \in \{0,1\}^K$. Let $\boldsymbol{W} \in \mathbb{R}^{N \times K}$ be the weight matrix connecting the hidden and visible units. The connection weight W_{nk} measures the association strength between the visible unit i and the hidden unit k, that is the tendency of these two units being co-active. The interaction between variables defines an *energy function*:

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\left(\boldsymbol{a}^\top \boldsymbol{v} + \boldsymbol{b}^\top \boldsymbol{h} + \boldsymbol{v}^\top \boldsymbol{W} \boldsymbol{h}\right) \tag{1}$$

where $\boldsymbol{a}, \boldsymbol{b}$ are the bias coefficients of hidden and visible units, respectively. The model admits the Boltzmann distribution:

$$P(\boldsymbol{v}, \boldsymbol{h}) \quad \propto \quad e^{-E(\boldsymbol{v}, \boldsymbol{h})} \tag{2}$$

The RBM is a generative model of data whose density is $P(\boldsymbol{v}) = \sum_{\boldsymbol{h}} P(\boldsymbol{v}, \boldsymbol{h})$.

The parameters are often estimated by maximizing the data likelihood $P(\boldsymbol{v})$. For example, an update rule for mapping weights is

$$W_{ik} \leftarrow W_{ik} + \eta \left(\langle v_i \rho_k \rangle_{\tilde{P}} - \langle v_i h_k \rangle_P \right) \tag{3}$$

where ρ_k represents $P(h_k = 1 \mid \boldsymbol{v})$, $\tilde{P}$ denotes empirical distribution of the visible data, $\langle \cdot \rangle_P$ denotes expectation with respect to distribution P, and η is learning rate. The data expectation $\langle v_i \rho_k \rangle_{\tilde{P}}$ is easy to evaluate. The model expectation $\langle v_i h_k \rangle_P$ is computationally difficult but can be efficiently approximated by short Markov chains starting from the observations $\boldsymbol{v}$ in a procedure known as "contrastive divergence" [10].

3. *e*NRBM: A framework for EMR modeling

3.1. High-level representation of abstracted trajectories

The EMR data broadly consist of two types: static information (such as gender, ethnic background) and healthcare trajectory. The trajectory is recorded as a series of time-stamped events (such as admission, diagnosis or intervention)[1]. We are mainly interested in discrete events and assume that continuous and real-valued data such as EEG signals and blood sugar readings have been discretized through existing methods such as temporal abstraction [11]. Static elements naturally form a vector. The entire trajectory is divided into disjoint intervals of predefined lengths. Events occurring within each interval are aggregated and arranged as a sparse vector. All intervals form a temporal matrix, as illustrated in the data layer of Fig. 1.

[1]Demographic factors such as age, location and income do change over time, but they might be considered as static at the present time if their interaction with clinical variables are not obvious.

3.1.1. RBM-based modeling

In RBM-based modeling of EMRs, as illustrated in Fig. 1, all data elements share the same hidden representation layer. The hidden layer is utilized in the tasks of interest (e.g., visualization of patients, diagnosis of a present disease, or prognosis of future risk). Thus, the hidden layer is a mediator between history (recorded illness), present (diagnosis) and future (prognosis). It "explains" the data through:

$$P(v_i^1 \mid \boldsymbol{h}) = \sigma\left(a_i + \sum_k W_{ik} h_k\right) \tag{4}$$

where v_i^1 represents $v_i = 1$, and $\sigma(x) = [1 + e^{-x}]^{-1}$. As all hidden units jointly represent the data, the representation is said to be *fully distributed*. This makes the representation highly compact: The model can be considered as a giant mixture of 2^K components with only $KN + K + N$ parameters.

This mixture view is attractive because healthcare is a complex process, and the recorded events are the result of interaction between multiple processes (e.g., the underlying illness, comorbidity, diagnostic decision and intervention), each of which can be captured by one or more hidden units.

3.1.2. Object embedding

The RBM embeds medical objects (e.g., diagnosis codes) and health trajectories into a vector space. Each object i is represented by a row vector $\boldsymbol{W}_{i\bullet}$ in $\mathbb{R}^K$. The vector embedding facilitates algebraic manipulations such as similarity calculation and retrieval, translation and rotation, and 2D projection for visualization. See Fig. 4 for an example of diseases embedded in 2D. An entire health trajectory can also be represented in the same space through probabilistic projection:

$$\rho_k = P\left(h_k = 1 \mid \boldsymbol{v}\right) = \sigma\left(b_k + \sum_i W_{ik} v_i\right) \tag{5}$$

where $\sigma(x)$ is the sigmoid function defined in Eq. (4). The posterior vector $\boldsymbol{\rho} = (\rho_1, \rho_2, ..., \rho_K)$ represents the entire patient trajectory. This can then be used for classification and prognosis (see Sec. 4.5 for a demonstration).

For a typical EMR, a practical issue arises since the input features are not binary but counts. We employ a simple solution: features are normalized into the range $[0, 1]$ and treated as empirical probability. A more theoretical drawback is that the RBM is not effective in organizing features, and does not takes the inherent structures of the EMR into account. In what follows, we show how to modify RBM to tackle these problems.

3.2. Structure discovery

This subsection presents modifications to RBMs for promoting the grouping of features and enhancing interpretability. We introduce two constraints into the parameter structure: *nonnegative weights* and *EMR-driven smoothness*, resulting in a novel model called *EMR-driven nonnegative RBM* (eNRBM).

3.2.1. Enforcing nonnegativity

The first modification is to constrain the connection weights $\{W_{ik}\}$ to be nonnegative. To enforce nonnegativity, we augmented the data log-likelihood $\log P(\boldsymbol{v})$ with a barrier function $B(W_{ik}) = W_{ik}^2$ if $W_{ik} < 0$ and 0 otherwise. Minimizing the augmented log-likelihood would drive negative weights toward zeros.

This leads to several interesting properties. First, the mapping matrix $\boldsymbol{W}$ is sparse, that is, only few elements are non-zeros. Second, hidden factors must "compete" to generate data, and thus creating an "explaining away" effect (where only a few latent factors are plausible explanation of the data). The result is a parts-based representation where each hidden unit is responsible to explain a part of the EMR [12].

The "explaining away" effect also leaves some hidden units unused (with near-zero mapping weight vectors $\boldsymbol{W}_{\bullet k}$). Thus it offers a natural way to estimate the *intrinsic dimensionality* of the data. A hidden unit k is declared "dead" if $\|\boldsymbol{W}_{\bullet k}\|_1^1 N^{-1} \leq \tau$ for small τ. This capacity is not seen in standard RBMs.

3.2.2. Promoting structural smoothness

The other modification is based on the inherent structures in the EMR. Due to the progressive nature of health, events often repeat over time. Thus, a disease occurring in consecutive time-intervals results in related features. Other structures are in the hierarchical organization of diseases and interventions, including the disease taxonomy ICD-10[2] and the procedure cube ACHI[3]. For example, two diseases that share the same parent in the taxonomy, by definition, possess similar characteristics.

Here we introduce a novel regularization scheme to realize these structures. Assume that the structures can be encoded into a feature graph G whose edges indicate the relatedness between features. Let $\gamma_{ij} > 0$ be the relation strength between feature i and j, the relatedness can be realized by minimizing the following smoothness objective:

$$\Omega(\boldsymbol{W}) \quad = \quad \sum_{ij} \gamma_{ij} \sum_{k} \left(W_{ik} - W_{jk}\right)^2 \tag{6}$$

In model estimation, this objective is added to the data log-likelihood, in addition to the nonnegativity constraint mentioned above. The details are presented in Appendix A.

In our implementation, we construct the feature graph as follows. An edge is created if any of the following requirements are met:

[2] http://apps.who.int/classifications/icd10
[3] https://www.aihw.gov.au/procedures-data-cubes/

- Two codes share the same two-character prefix. In particular, we use the first two numbers or letters (using ICD-10 for diseases, and ACHI for procedures). For example, F10 (mental disorder due to alcohol) and F17 (mental disorder due to tobacco) are linked since they are children of F1 (Mental disorders due to psychoactive substance use). However, F10 and F20 (schizophrenia) do not share a direct relation. We feel that this balances well between the relatedness and specificity of the disease classification.

- A code is recorded in consecutive intervals. For example, if F10 is recorded in [0-3] months and [3-6] months prior to a specified date, this constitutes an edge. This is because two close events of the same type would behave similarly.

4. Case study: Suicide risk stratification

4.1. Experiment setup

4.1.1. Data

Our focus is on mental health patients who were under assessment for suicidal risk. Mental health is a global burden that accounts for 14% of the world health expenditure [13]. Among mental health problems, suicidal risk is devastating: suicidal thoughts occur in 10% of the population in their lifetime [14], and suicide attempts happen in 0.3% of the population each year [15]. The risk of suicide has led to mandatory assessments. However, suicide risk assessments are often inaccurate leading to concern over practicality [16, 17].

We used a mental health cohort previously extracted from Barwon Health, a large regional hospital in Australia [18, 19]. Data was collected between January 2009 and March 2012. The dataset contains $7,578$ patients (49.3% male, 48.7% under 35) who underwent collectively $17,566$ assessments. Any patient who

had at least one encounter with the hospital services and one risk assessment was included. Most patients had one assessment (62%), but 3% of patients had more than 10 assessments. Diagnoses are coded using ICD-10. More details are described in [19].

4.1.2. Risk stratification task

Each assessment was considered as a data point from which a prediction would be made. The future outcomes within 3 months following an assessment were categorized into three ordinal levels of risk according to [18]: no-risk, moderate-risk (non-fatal consequence), and high-risk (fatal consequence). The risk classes were decided using a look-up table from the ICD-10 codes. If there were more than one outcome classes, the highest risk class would be chosen. There were $15,272$ (86,9%) no-risk outcomes, $1,436$ (8.2%) moderate-risk and 858 (4.9%) high-risk.

4.1.3. Implementation details

Following [18, 19], we split the 48-month history prior to each risk assessment into non-overlapping intervals: $(0-3)$, $(3-6)$, $(6-12)$, $(12-24)$ and $(24-48)$. The increasing interval widths toward the far past are based on the assumption that events in the far past have less influence on current outcomes. Each interval has the same set of time-stamped variables: 201 diagnoses, 657 procedures, 31 Elixhauser comorbidities, diagnosis related groups (DRG), emergency attendances and admissions. Infrequent diagnoses and procedures were grouped into rare categories. Together with demographic variables (ages in 10-year intervals and gender), there were totally $5,267$ input variables.

The posterior vector $\boldsymbol{\rho}$ (Eq. 5) was used as input for logistic regression classifiers (LR) for predicting outcomes. For robustness, the LR was equipped with elastic net regularization [20]. Besides the standard RBM, we employed support

vector machines (SVM) which ran on normalized features and PCA-derived features. We used the implementation of SVM in LIBSVM package [21]. As the LR and the SVM are binary classifiers, the *one-versus-all* strategy was used for this 3-class problem.

For risk stratification, we used 10-fold validation. For each fold, parameters were learnt on the training set and hyperparameters were turned for the best performance on the validation set. Results were reported as an average across folds. For the SVM, we used the linear kernel. For both the RBM and the eNRBM, the numbers of hidden units were set to $K = 200$. The learning rate was scheduled as $^{0.1}/_{\sqrt{t}}$ at epoch t. This weight decay helped stabilize the parameter updates towards the end of the learning process. The weights were initialized randomly from $\mathcal{N}(0; 0.1)$, and the biases were from zeros. Parameters were then updated after every "mini-batch" of 100 data points. Learning was terminated after 100 epochs. Hyperparameters of the eNRBM were empirically tuned to obtain accurate data reconstruction and high group coherence, while keeping the F-measure competitive.

4.2. Intrinsic dimensionality and group coherence

To estimate the number of hidden units, we examined the intrinsic dimensionality of data, as described in Sec. 3.2.1. Fig. 3 plots the number of used hidden units against the total number for an eNRBM estimated on $1,005$ diagnosis codes. The curves were averaged over a set of thresholds ($\tau \in \{0.01; 0.02; ...; 0.06\}$). The dimensionality stays around 250. To obtain a compact representation, we used $K = 200$ hidden units in subsequent experiments.

To quantify the coherence of feature group, we borrowed the concept from topic modeling [22]. For each group, we kept T member features with largest mapping weights. Let $D\left(v_i^{(k)}\right)$ and $D\left(v_i^{(k)}, v_j^{(k)}\right)$ be occurrences of feature i

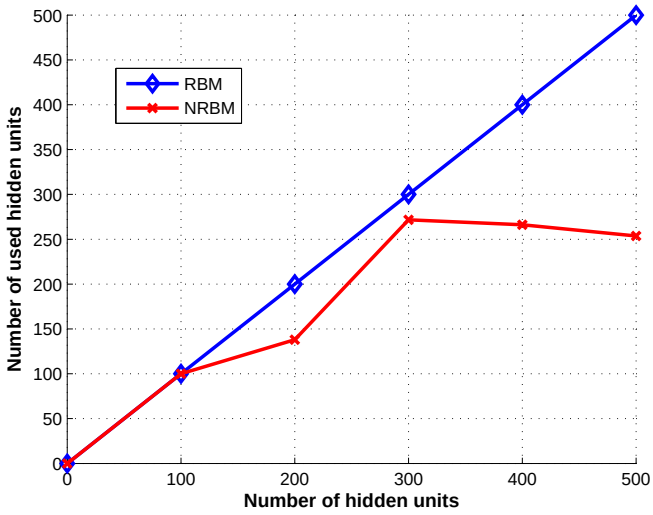

Figure 3: Intrinsic dimensionality of the disease space (1,005 variables).

and feature pair (i, j) under factor k, respectively. The group coherence was defined as:

$$C(k) = \sum_{i=1}^{T-1} \sum_{j=i+1}^{T} \log \frac{1 + D\left(v_i^{(k)}, v_j^{(k)}\right)}{1 + D\left(v_i^{(k)}\right)} \tag{7}$$

Intuitively, the coherence of a group is large if its members co-occur frequently, relative to the popularity of each member. With $T = 10$, the eNRBM had a coherence of -130.88, higher than that of the standard RBM (-173.3).

4.3. Disease and procedure embedding and clustering

Here we validate the effectiveness of object embedding (Sec. 3.1.2). Two eNRBMs were created, one using only diagnoses (called model *DIAG*), the other using both diagnoses and procedures (called model *DIAG+PROC*). A RBM was learned using diagnosis codes for comparison.

For each model, the mapping weight matrix $\boldsymbol{W}$ was examined. Elements of row vector $\boldsymbol{W}_{i\bullet}$ are coordinates of the object i in the embedding space of K dimensions. Objects were projected onto 2D using t-SNE [23]. As shown

in Fig. 4, diseases naturally form coherent groups (colored by k-means). Note that t-SNE is a visualization method and it was not involved in computing the embedding of codes.

Similarly, Fig. 5 presents the embedding/clustering of both diseases and procedures. Since diseases and procedures are jointly embedded in the same space, their relations can be directly assessed. For several groups, we plotted the top 5 procedures and 5 diagnoses, where the font size was proportional to inverse distances to the group centers. The grouping is meaningful, for example:

- *Group 1*: Diagnosis C34 (Malignant neoplasm of bronchus and lung) is associated with procedures 543 (Examination procedures on bronchus) and 536 (Tracheostomy).

- *Group 2*: Diagnosis C78 (Secondary malignant neoplasm of respiratory and digestive organs) and C77 (Secondary and unspecified malignant neoplasm of lymph nodes) are associated with procedures 392 (Excision procedures on tongue) and 524 (Laryngectomy).

- *Group 3*: Diagnosis K35 (Acute appendicitis) is associated with procedure 926 (Appendicectomy).

In contrast, the groups produced by RBM in Fig. 6 are less coherent and their diagnosis codes do not clearly explain suicide risks.

We compared the discovered groups with the risk factors found in previous work [18]. The relevance of a group is the number of matches in the top 10 risk factors under the group. On average, 4.4 out of 10 risk factors per group found by the eNRBM matched those in [18]. This is higher than the matching rate by the RBM, which was 1.6.

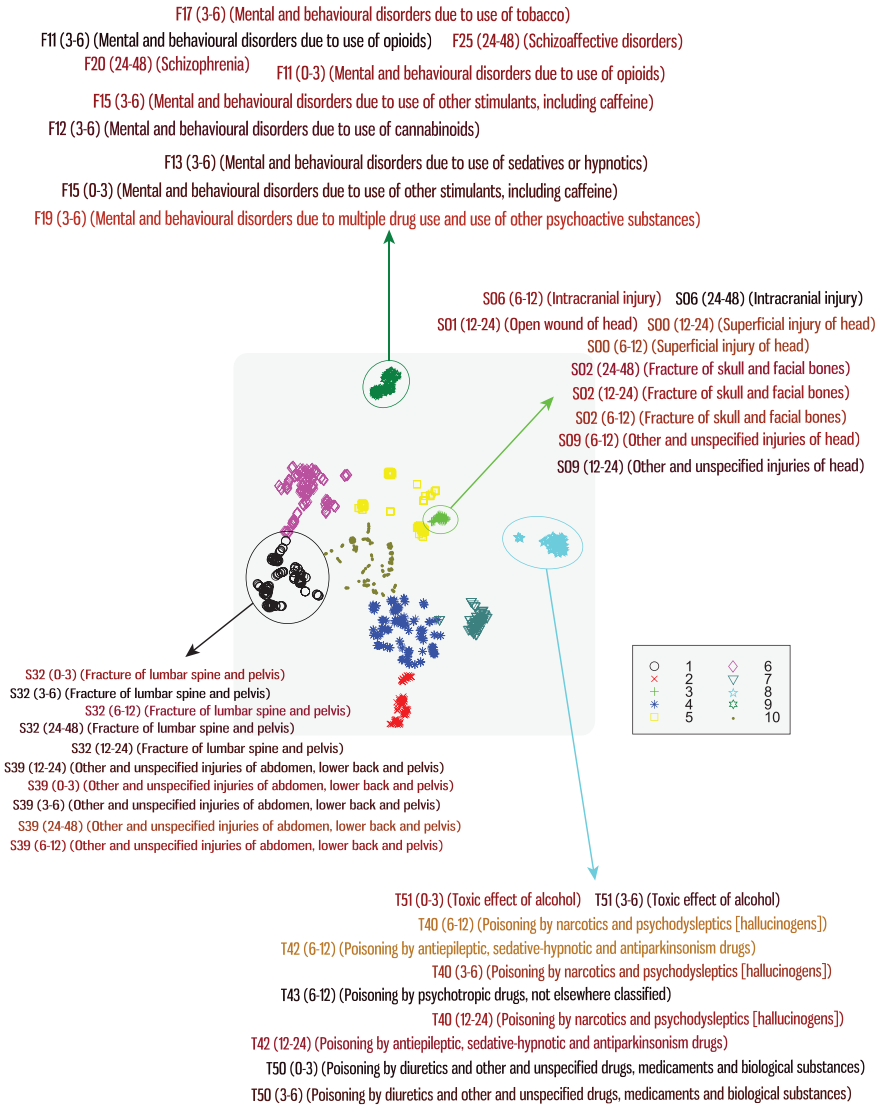

Figure 4: Disease embedding (model *DIAG*). Diseases were first embedded into 200 dims using *e*NRBM, then projected onto 2D using t-SNE [23]. Note that t-SNE did not contribute to original embedding or clustering. Color shows disease clusters discovered by k-means with 10 clusters.

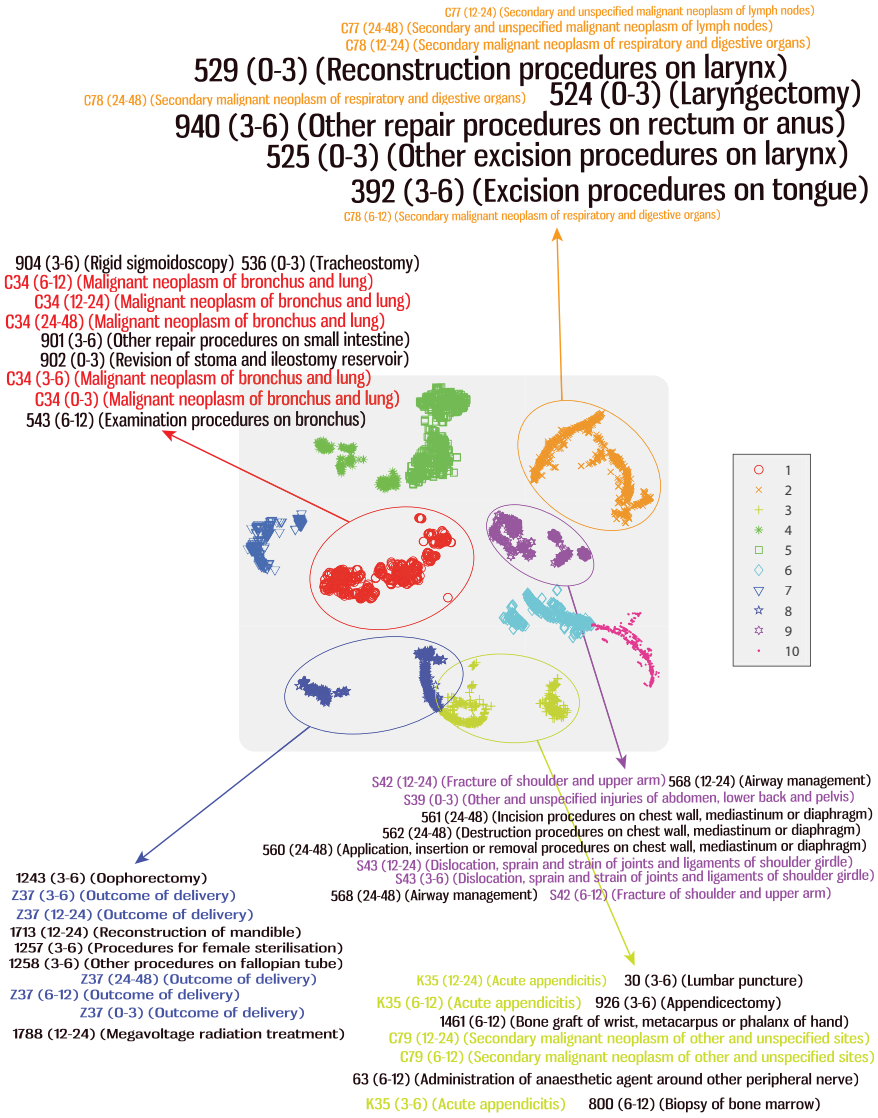

Figure 5: Disease and procedure embedding (model *DIAG+PROC*). Codes were first embedded into 200 dims using *e*NRBM, then projected onto 2D using t-SNE [23]. Color shows disease clusters discovered by k-means with 10 clusters. Font size indicates nearness to respective cluster centers.

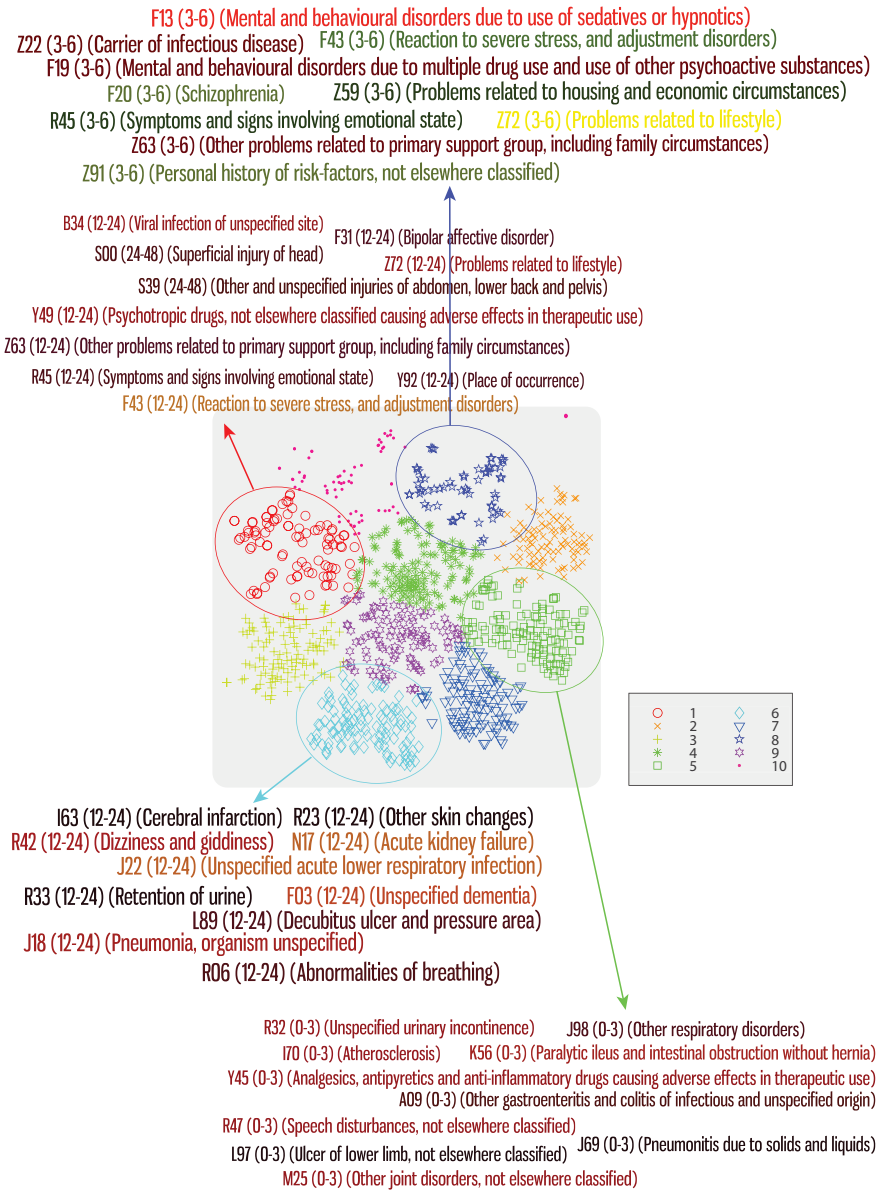

Figure 6: Disease embedding (model *DIAG*). Diseases were first embedded into 200 dims using RBM, then projected onto 2D using t-SNE [23]. Color shows disease clusters discovered by k-means with 10 clusters.

Rank: Moderate-risk	Rank: High-risk
1: *Z22* (3–6; 24–48)	1: *X61* (3–6); X62 (0–3)
Z29 (0–3; 3–6; 6–12)	*X64* (0–3; 3–6)
2: *R94* (all intervals)	*X65* (3–6)
3: *S61* (0–3; 3–6; 6–12)	2: *T50* (6–12; 12–24; 24–48)
S62 (0–3; 6–12)	*T51* (0–3; 3–6)
4: *F03* (0-3; 3–6; 6–12)	3: *T39* (all intervals)
F05 (3–6; 24–48)	4: *Z29* (0–3; 3–6; 6–12)
5: *E66* (all intervals)	*Z22* (3–6; 24–48)
	5: *S52* (0–3; 3–6; 6–12; 12–24)
	S51 (0–3)

Table 1: Top five feature groups corresponding to moderate-risk and high-risk suicide events, one per row, ranked by the weight in the corresponding logistic classifiers. Each group has top 5 discovered comorbidities coded in ICD-10 scheme, ranked by their mapping weight W_{ik}. Time periods for each comorbidity is described in the bracket, e.g., 3-6 means the comorbidity is recorded 3-6 months prior to the assessment point. See Tab. 2 for description of codes.

4.4. Risk groups

To identify which feature group was predictive of future risk, we used the posterior embedding of patients (see Eq. (5)) as inputs for two logistic regression classifiers, one for the moderate-risk class, the other for the high-risk class. Groups were ranked by their regression coefficients.

Table 1 presents top five feature groups corresponding to moderate-risk and high-risk classes (model DIAG). Moderate-risk groups consist of abnormality in function findings (ICD-10: *R94*), non-fatal hand injuries (ICD-10: *S6x*), mental disorders such as dementia (ICD-10: *F03*) and (ICD-10: *F05*), obesity (ICD-10: *F66*), and potential hazards related to communicable diseases (ICD-10: *Z2s*). High-risk groups involve self-harms (ICD-10: *X6s*) as the top risk, followed by poisoning (ICD-10: *T39*, *T5s*), hazards related to communicable diseases (ICD-10: *Z2s*), and finally hand injuries (ICD-10: *S5s*).

4.5. Risk stratification

We now report results on suicide risk stratification for a 3-month horizon. Fig. 7 shows the relative performance of the *e*NRBM (for representation learning) coupled with logistic regression classifiers (for classification), in comparison

E66: Obesity
F03: Unspecified dementia
F05: Delirium
R94: Abnormal functions
S51: Open wound of forearm
S52: Fracture of forearm
S61: Open wound of wrist and hand
S62: Fracture at wrist and hand level
T39: Poisoning by nonopioid analgesics
T50: Poisoning by diuretics
T51: Toxic effect of alcohol
X61: Intentional self-poisoning by psychotropic drugs
X62: Intentional self-poisoning by psychodysleptics
X64: Intentional self-poisoning by unspecified drugs
X65: Intentional self-poisoning by alcohol
Z22: Carrier of infectious disease
Z29: Need for other prophylactic measures

Table 2: Top ICD-10 codes contributing to suicide risk, as identified in Tab. 1.

with support vector machines (SVM) that ran on raw EMR data and on PCA-derived features. Using the full EMR-derived data leads to better results than those using the diagnoses alone, suggesting the capability in data fusion by the eNRBM.

Table 3 presents more detailed results. The F-scores achieved by eNRBM are 0.212 and 0.359 for moderate-risk and high-risk, respectively. The high-risk F-score is already three times better than the performance achieved by clinicians who admitted the risk assessment [18, 19]. The F-scores are also competitive with the results obtained by rival methods: SVM on raw features obtained F-score of 0.156 and 0.340; and SVM on PCA-derived features yielded 0.135 and 0.325 for moderate and high-risk, respectively. We ran a bootstrap simulation and found that (i) for moderate-risk, eNRBM is significantly better than SVM or RBM at $p = 0.05$; (ii) for high-risk, there is no statistical difference, largely due to the smaller number of high-risk cases.

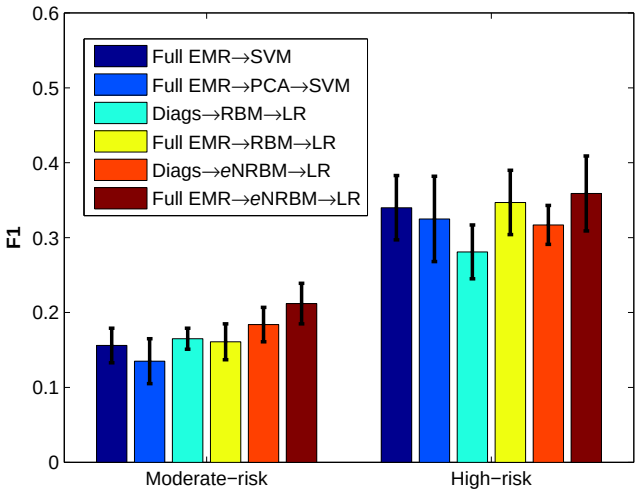

Figure 7: F-scores (F_1) for moderate and high-risk within 3 months. Arrows indicate the flow. *Diags* means using only diagnoses as input. Full EMR contains demographics, diagnoses, procedures, diagnosis related groups (DRG) and Elixhauser comorbidities [2].

5. Discussion

*5.1. e*NRBM *as a model of EMR*

The *e*NRBM belongs to, but differs radically from the rest of the latent variable family used in biomedical fields [24]. The family includes traditional methods such as factor analysis [25] and modern models such as latent Dirichlet allocation [26] and Indian buffet processes [27]. All of these existing models can be represented as directed graphical models whose inference is usually expensive. Importantly, while these methods are effective in analyzing latent factors or thematic structures, they are not typically designed for data representation on which further manipulations can be performed. The *e*NRBM, on the other hand, is undirected and permits fast inference and learning on massive high-dimensional data. The *e*NRBM offers multiple benefits: nonlinear; compact distributed representation; embedding medical objects into Euclidean space; and feature grouping. Importantly, the *e*NRBM can compute predictive

	Recall	Precision	F-measure
Full EMR→SVM			
Moderate-risk	0.251	0.114	0.156
High-risk	**0.455**	0.271	0.340
Full EMR→PCA→ SVM			
Moderate-risk	0.208	0.103	0.135
High-risk	0.433	0.268	0.325
Diags→RBM→LR			
Moderate-risk	0.234	0.127	0.165
High-risk	0.342	0.239	0.281
Full EMR→RBM→ LR			
Moderate-risk	0.226	0.125	0.161
High-risk	0.424	0.294	0.347
*Diags→*e*NRBM→LR*			
Moderate-risk	0.260	0.143	0.184
High-risk	0.384	0.271	0.317
*Full EMR→*e*NRBM→ LR*			
Moderate-risk	**0.310**	**0.161**	**0.212**
High-risk	0.445	**0.301**	**0.359**

Table 3: Performance of various classifiers with several input preprocessing techniques (PCA and *e*NRBM). Diags means we used only diagnoses as input. Full EMR contains demographics, diagnoses, procedures, diagnosis related groups (DRG) and Elixhauser comorbidities [2]. Bold numbers are highest in their category.

representations.

The feature grouping capability facilitates better understanding of feature interactions. This is critical in modern medicine where multimorbidity is the rule, not exception, especially among the elderly [28]. The illness trajectories and healthcare processes become increasingly interwoven [29], and it is crucial to automatically disentangle these dependencies.

The direct modeling of dependencies between clinical variables has been studied in Bayesian networks [30, 31]. The main difficulties are: designing acyclic structures, and slow inference in large networks. The *e*NRBM, on the other hand, requires no structure design, and is fast with only a single matrix operation.

Finally, we wish to emphasize that the RBM is a fully generative model of

EMRs with distribution $P(\boldsymbol{v})$. The RBM can simulate EMRs whose distribution follows $P(\boldsymbol{v})$. This offers a new solution for data sharing without compromising privacy. Details of the simulation are beyond the scope of this paper, but in general they are based on Monte Carlo simulation (see for example, [**?**]). For this paper, code and simulated data are available for download[4]. The data was sampled from a RBM which was learnt from the real data. Thus the simulated data reflects the true statistical properties of the real source.

5.1.1. Embedding medical objects

Medical objects and events are discrete in nature. This creates significant computational challenges for symbolic representation. First, the number of unique objects (e.g., diagnosis codes) is often very large, and the number of events grows in time. Second, rare objects (e.g., rare diseases) are not robust to quantify statistically. And third, relations such as nearness with continuously varying degrees are hard to specified to fine details.

This calls for an embedding of objects into low-dimensional spaces (e.g., see also [32] for similar arguments in linguistics). In other words, the representation of an object is *distributed.* Embedding promotes algebraic manipulations such as similarity computation and retrieval. It is also easy to assess the relatedness between objects of different kinds (e.g., a disease and a procedure), as we have seen in Fig. 5. Once objects have been embedded, an event can be considered as a set of objects observed in a period of time. The discussion can be extended to relations, for example, the parent-child relationship in the disease taxonomy: A parent is close to its children in the embedding space. This offers a novel way of exploiting existing medical knowledge bases.

[4] http://prada-research.net/∼truyen/code/eNRBM-jbi.zip

5.1.2. Risk group discovery

The *e*NRBM applied to mental health, as shown in Table 1, discovered risk factors that resemble those well-documented in the literature [19, 33]. For instance, psychiatric problems and prior attempts are well-recognized risk factors [34, 35]. Our method differs in that it is hypothesis-free and time-specific.

Comorbidities that appear remotely related to psychiatric issues were also discovered, for example infectious diseases [36, 37] and obesity [38, 39, 40]. While these findings are interesting to warrant a deeper analysis, a full clinical investigation is beyond the scope of this paper. Finally, the automatic grouping suggests a potential in automated phenotyping [4, 6].

5.2. Limitations

We recognize several limitations. First, a relation was defined if two ICD-10 codes shared the first character and the first digit, and the relation strength was always 1. This could be extended to be more flexible. For example, F20 and F31 share the parent F (Mental and behavioural disorders), so the relation strength can be thought as a half of that between F20 and F21. Determining the precise strength is a difficult problem itself. First, the *e*NRBM primarily ran on binary (or probability-like) observations. However, model can be easily extended to other data types such as counts (e.g., number of previous admissions) and continuous variables (e.g., lab test measurements) or a mixture of these [41, 42]. This suggests an interesting integration of multiple modalities, such as administrative data (this work), text (e.g., carer notes), and medical images [43]. Extension to unstructured clinical notes is not difficult: time-stamped notes can be aggregated into intervals just like other composite events (such as admissions), and known relations between concepts (e.g., using the UMLS or SNOMED-CT) can be naturally encoded into the *e*NRBM.

Second, some discovered groups may not be clinically relevant but a data

artifact. However, the structural relations can be modified without difficulty to encode known phenotypes and to prevent meaningless grouping.

Finally, the empirical study has been limited to EMRs from a single institution. The EMR is known for its quality issues [44]. However, EMRs are comprehensive and readily available, making them an attractive alternative to standard clinical data collection. In fact, the quality of the Charlson comorbidity index computed from EMR is comparable to that computed from the standard chart [45, 46]. The *e*NRBM is cohort-independent, and thus it is possible to run on multiple databases. Alternatively, *e*NRBM could be evaluated intensively using simulated data with controlled variations so that its behaviors and performance can be assessed. However, faithfully generating EMR data is a challenging research topic by itself (see, for example, a recent work by [**?**]).

5.3. Conclusion

We have proposed a novel model called EMR-driven nonnegative restricted Boltzmann machine (*e*NRBM) for EMR modeling. The *e*NRBM supports a variety of healthcare analytics tasks with minimal manual feature engineering. The model learns EMR representation by embedding features and trajectories into a low-dimensional space. Through nonnegativity and domain-specific structural constraints, intrinsic dimensionality can be estimated, meaningful grouping of medical objects can be discovered. The homogeneous representation leads to simple algebraic manipulations and easy use with existing classifiers. Experimental results on suicide risk stratification demonstrate that the proposed method is competitive in predictive performance. The model paves a pathway toward EMR-driven phenotyping.

Appendix A. Details on *e*NRBM

Appendix A.1. Model properties

To see how the nonnegativity constraints in the eNRBM let the grouping emerge, consider the activation probability of the hidden unit in Eq. (5):

$$\rho_k = P\left(h_k = 1 \mid \boldsymbol{v}\right) = \sigma\left(b_k + \sum_i W_{ik} v_i\right) \tag{A.1}$$

Suppose for the moment that $|b_k|$ is bounded from above. Then, the visible units must "compete" against each other to turn on the k-th hidden unit by making $\{b_k + \sum_i W_{ik} v_i\} \geq 0$, since $\{v_i\}$ are nonnegative. The result is that some elements of the k-th column vector $W_{\bullet k}$ are driven to zeros. The remaining elements will self-organized into the k-th group.

Since the bipartite structure of the eNRBM has no within-layer connections, the conditional distributions over visible and hidden units can be factorized as:

$$p\left(\boldsymbol{v} \mid \boldsymbol{h}\right) = \prod_{i=1}^{N} p\left(v_i \mid \boldsymbol{h}\right) \tag{A.2a}$$

$$p\left(\boldsymbol{h} \mid \boldsymbol{v}\right) = \prod_{k=1}^{K} p\left(h_k \mid \boldsymbol{v}\right) \tag{A.2b}$$

Thus inference can be efficiently performed by layer-wise sampling. Model density can be estimated as

$$P(\boldsymbol{v}) = \frac{1}{S} \sum_{s=1}^{S} P\left(\boldsymbol{v} \mid \boldsymbol{h}^{(s)}\right) \tag{A.3}$$

using S random samples $\left\{\boldsymbol{h}^{(s)}\right\}$ for $s = 1, 2, ..., S$.

Appendix A.2. Model estimation

Learning in the eNRBM was carried out by maximizing the data log-likelihood $\log P(\boldsymbol{v})$ subject to several constraints:

- *Nonnegativity*: $W_{ik} \geq 0$ for all i, k. For simplicity, we used the barrier function $B(W_{ik}) = W_{ik}^2$ if $W_{ik} < 0$ and 0 otherwise.
- *Bounding*: $|a_i|, |b_k| \leq c$. This could be realized by adding a penalty term to the data likelihood $\sum_i a_i^2 + \sum_k b_k^2$
- *Structural smoothness*: similar features should share similar weights, as encoded in the regularizer $\Omega(\boldsymbol{W})$ in Eq. (6).

Finally, the augmented log-likelihood is

$$L(\boldsymbol{W}) = \log P(\boldsymbol{v}) - \frac{\alpha}{2} B(W_{ik}) - \frac{\beta}{2} \left(\sum_i a_i^2 + \sum_k b_k^2 \right) - \frac{\lambda}{2} \Omega(\boldsymbol{W}) \qquad \text{(A.4)}$$

where $\alpha, \beta, \gamma > 0$ are tunable hyperparameters.

The structural smoothness can be rewritten as

$$\Omega(\boldsymbol{W}) \quad = \quad \sum_k \boldsymbol{W}_{\bullet k}^\top \boldsymbol{L} \boldsymbol{W}_{\bullet k}$$

where $L_{ii} = \sum_{j \neq i} \gamma_{ij}$; $L_{ij} = -\gamma_{ij}$. The matrix $\boldsymbol{L}$ is known as the Laplacian of the graph whose edge weight is γ_{ij}.

Finally, the parameter update rule becomes:

$$\begin{aligned}
a_i &\leftarrow a_i + \eta \left(v_i - \langle v_i \rangle_P - \beta a_i \right) \\
b_k &\leftarrow b_k + \eta \left(\rho_k - \langle h_k \rangle_P - \beta b_k \right) \\
W_{ik} &\leftarrow W_{ik} + \eta \left(v_i \rho_k - \langle v_i h_k \rangle_P - \alpha \lceil W_{ik} \rceil^- - \lambda \boldsymbol{L} W_{ik} \right)
\end{aligned}$$

where $\lceil W_{nk} \rceil^-$ denotes the negative part of the weight. The "contrastive divergence" procedure [10] was used to approximate expectations with respect to the model distribution $P(\boldsymbol{v}, \boldsymbol{h})$. The Markov chain started from the observation $\boldsymbol{v}$, runs for one step, then the pair $(\boldsymbol{v}, \boldsymbol{h})$ was collected to approximate P.

References

[1] 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.

[2] A. Elixhauser, C. Steiner, D. R. Harris, R. M. Coffey, Comorbidity measures for use with administrative data, Medical care 36 (1) (1998) 8–27.

[3] W. M. Tierney, B. Y. Takesue, D. L. Vargo, Using electronic medical records to predict mortality in primary care patients with heart disease, Journal of general internal medicine 11 (2) (1996) 83–91.

[4] G. Hripcsak, D. J. Albers, Next-generation phenotyping of electronic health records, Journal of the American Medical Informatics Association 20 (1) (2013) 117–121.

[5] D. He, S. C. Mathews, A. N. Kalloo, S. Hutfless, Mining high-dimensional administrative claims data to predict early hospital readmissions, Journal of the American Medical Informatics Association 21 (2) (2014) 272–279.

[6] T. A. Lasko, J. C. Denny, M. A. Levy, Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data, PloS one 8 (6) (2013) e66341.

[7] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (8) (2013) 1798–1828.

[8] G. Hinton, R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.

[9] Y. Freund, D. Haussler, Unsupervised learning of distributions on binary vectors using two layer networks, Advances in Neural Information Processing Systems (1993) 912–919.

[10] G. Hinton, Training products of experts by minimizing contrastive divergence, Neural Computation 14 (2002) 1771–1800.

[11] M. Stacey, C. McGregor, Temporal abstraction in intelligent clinical data analysis: A survey, Artificial Intelligence in Medicine 39 (1) (2007) 1–24.

[12] T. Nguyen, T. Tran, D. Phung, S. Venkatesh, Learning Parts-based Representations with Nonnegative Restricted Boltzmann Machine , in: Proc. of 5th Asian Conference on Machine Learning (ACML), Canberra, Australia, 2013.

[13] M. Prince, V. Patel, S. Saxena, M. Maj, J. Maselko, M. R. Phillips, A. Rahman, No health without mental health, The lancet 370 (9590) (2007) 859–877.

[14] M. K. Nock, J. G. Green, I. Hwang, K. A. McLaughlin, N. A. Sampson, A. M. Zaslavsky, R. C. Kessler, Prevalence, correlates, and treatment of lifetime suicidal behavior among adolescentsresults from the national comorbidity survey replication adolescent supplementlifetime suicidal behavior among adolescents, JAMA psychiatry 70 (3) (2013) 300–310.

[15] G. Borges, M. K. Nock, J. M. H. Abad, I. Hwang, N. A. Sampson, J. Alonso, L. H. Andrade, M. C. Angermeyer, A. Beautrais, E. Bromet, et al., Twelve month prevalence of and risk factors for suicide attempts in the WHO

World Mental Health Surveys, The Journal of clinical psychiatry 71 (12) (2010) 1617.

[16] M. Large, C. Ryan, O. Nielssen, The validity and utility of risk assessment for inpatient suicide, Australasian Psychiatry 19 (6) (2011) 507–512.

[17] C. Ryan, O. Nielssen, M. Paton, M. Large, Clinical decisions in psychiatry should not be based on risk assessment, Australasian Psychiatry 18 (5) (2010) 398–403.

[18] T. Tran, W. Luo, D. Phung, R. Harvey, M. Berk, R. L. Kennedy, S. Venkatesh, Risk stratification using data from electronic medical records better predicts suicide risks than clinician assessments, BMC psychiatry 14 (1) (2014) 76.

[19] T. Tran, D. Phung, W. Luo, S. Venkatesh, Stabilized sparse ordinal regression for medical risk stratification, Knowledge and Information Systems (2014) 1–28.

[20] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2) (2005) 301–320.

[21] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (2011) 27:1–27:27.

[22] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, A. McCallum, Optimizing semantic coherence in topic models, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 262–272.

[23] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research 9 (2579-2605) (2008) 85.

[24] S. Rabe-Hesketh, A. Skrondal, Classical latent variable models for medical research, Statistical methods in medical research 17 (1) (2008) 5–32.

[25] F. J. Floyd, K. F. Widaman, Factor analysis in the development and refinement of clinical assessment instruments., Psychological assessment 7 (3) (1995) 286.

[26] D. Blei, A. Ng, M. Jordan, Latent dirichlet allocation, The Journal of Machine Learning Research 3 (2003) 993–1022.

[27] F. J. Ruiz, I. Valera, C. Blanco, F. Perez-Cruz, Bayesian nonparametric comorbidity analysis of psychiatric disorders, arXiv preprint arXiv:1401.7620.

[28] A. Prados-Torres, B. Poblador-Plou, A. Calderón-Larrañaga, L. A. Gimeno-Feliu, F. González-Rubio, A. Poncel-Falcó, A. Sicras-Mainar, J. T. Alcalá-Nalvaiz, Multimorbidity patterns in primary care: interactions among chronic diseases using factor analysis, PloS one 7 (2) (2012) e32190.

[29] J. M. Corbin, A. Strauss, A nursing model for chronic illness management based upon the trajectory framework, Research and Theory for Nursing Practice 5 (3) (1991) 155–174.

[30] P. J. Lucas, L. C. van der Gaag, A. Abu-Hanna, Bayesian networks in biomedicine and health-care, Artificial Intelligence in medicine 30 (3) (2004) 201–214.

[31] F. Stella, Y. Amer, Continuous time Bayesian network classifiers, Journal of biomedical informatics 45 (6) (2012) 1108–1119.

[32] P. D. Turney, P. Pantel, et al., From frequency to meaning: Vector space models of semantics, Journal of artificial intelligence research 37 (1) (2010) 141–188.

[33] G. K. Brown, A. T. Beck, R. A. Steer, J. R. Grisham, Risk factors for suicide in psychiatric outpatients: a 20-year prospective study., Journal of consulting and clinical psychology 68 (3) (2000) 371.

[34] X. Gonda, M. Pompili, G. Serafini, F. Montebovi, S. Campi, P. Dome, T. Duleba, P. Girardi, Z. Rihmer, Suicidal behavior in bipolar disorder: epidemiology, characteristics and major risk factors, Journal of affective disorders 143 (1) (2012) 16–26.

[35] C. Martin-Fumadó, G. Hurtado-Ruíz, Clinical and epidemiological aspects of suicide in patients with schizophrenia, Actas Esp Psiquiatr 40 (6) (2012) 333–45.

[36] S. C. Segerstrom, G. E. Miller, Psychological stress and the human immune system: a meta-analytic study of 30 years of inquiry., Psychological bulletin 130 (4) (2004) 601.

[37] J. P. Godbout, R. Glaser, Stress-induced immune dysregulation: implications for wound healing, infectious disease and cancer, Journal of Neuroimmune Pharmacology 1 (4) (2006) 421–427.

[38] K. M. Carpenter, D. S. Hasin, D. B. Allison, M. S. Faith, Relationships between obesity and DSM-IV major depressive disorder, suicide ideation, and suicide attempts: results from a general population study, American Journal of Public Health 90 (2) (2000) 251.

[39] C. U. Onyike, R. M. Crum, H. B. Lee, C. G. Lyketsos, W. W. Eaton, Is obesity associated with major depression? results from the third national

health and nutrition examination survey, American journal of epidemiology 158 (12) (2003) 1139–1147.

[40] A. J. Stunkard, M. S. Faith, K. C. Allison, Depression and obesity, Biological psychiatry 54 (3) (2003) 330–337.

[41] T. Tran, D. Phung, S. Venkatesh, Mixed-variate restricted Boltzmann machines, in: Proc. of 3rd Asian Conference on Machine Learning (ACML), Taoyuan, Taiwan, 2011.

[42] T. Nguyen, T. Tran, D. Phung, S. Venkatesh, Latent patient profile modelling and applications with mixed-variate restricted Boltzmann machine, in: Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Gold Coast, Queensland, Australia, 2013.

[43] R. D. Hjelm, V. D. Calhoun, R. Salakhutdinov, E. A. Allen, T. Adali, S. M. Plis, Restricted boltzmann machines for neuroimaging: An application in identifying intrinsic networks, NeuroImage 96 (2014) 245–260.

[44] L. I. Iezzoni, Assessing quality using administrative data, Annals of internal medicine 127 (8_Part_2) (1997) 666–674.

[45] H. Quan, G. A. Parsons, W. A. Ghali, Validity of information on comorbidity derived from ICD-9-CCM administrative data, Medical care 40 (8) (2002) 675–685.

[46] M. Nuttall, J. van der Meulen, M. Emberton, Charlson scores based on ICD-10 administrative data were valid in assessing comorbidity in patients undergoing urological cancer surgery, Journal of clinical epidemiology 59 (3) (2006) 265–273.