Modeling Healthcare Quality via Compact Representations of Electronic Health Records

Jelena Stojanovic, Djordje Gligorijevic, Vladan Radosavljevic, Nemanja Djuric, Mihajlo Grbovic, and Zoran Obradovic

Abstract—Increased availability of Electronic Health Record (EHR) data provides unique opportunities for improving quality of health services. In this study we couple EHRs with the advanced machine learning tools to predict three important parameters of healthcare quality. More specifically, we describe how to learn low-dimensional vector representations of patient conditions and clinical procedures in an unsupervised manner, and generate feature vectors of hospitalized patients useful for predicting their length of stay, total incurred charges, and mortality rates. In order to learn vector representations we propose to employ state-of-the-art language models specifically designed for modeling co-occurrence of diseases and applied clinical procedures. The proposed model is trained on a large-scale EHR database comprising more than 35 million hospitalizations in California over a period of 9 years. We compared the proposed approach to several alternatives and evaluated their effectiveness by measuring accuracy of regression and classification models used for three predictive tasks considered in this study. Our model outperformed the baseline models on all tasks, indicating a strong potential of the proposed approach for advancing quality of the healthcare system.

Index Terms—Electronic Health Records; healthcare quality; embedding models; neural language models.

1 Introduction

Inpatient Quality Indicators (IQIs) were developed as a set of measures that provide a perspective on quality of patient care in hospitals[1]. These indicators include inpatient mortality for certain procedures and medical conditions [1], length of stay [2], and total charges of an inpatient stay[2], and can be considered as important metrics for evaluating quality of care [3]. These measures can be used to help hospitals identify potential problem areas that might need further studies and provide the opportunity to assess quality of care inside hospitals using administrative data found in typical discharge records. On the other hand, transparency of these indicators may help potential users of hospital care choose a hospital that will fit their needs and their financial constraints. This aspect is becoming an increasingly important issue as healthcare users are reportedly declaring personal bankruptcies during hospitalizations either due to high hospital care prices, or due to inpatient staying too long in a hospital when this might not be necessary [4], [5], [6], [7].

Unsurprisingly, one of the important metrics that the patients are worried about is how high their final hospital bill will be. However, computing this value upfront is not a trivial task, as pricing of health care services vary significantly among different providers even for the most common procedures. Each provider takes into account many parameters before charging a patient, and the process is different for different players in the industry. For example, Medicare takes more than one hundred parameters to determine a hospitalization reimbursement[3]. For these reasons, many economists, employers and health plans are advocating for providing the price quote of health care services as a way to encourage consumers to choose low-cost, high-quality providers and to promote competition based on the value of care[4].

Length of stay (LoS) is another important metric for assessing the quality of health care, also useful for planning scheduling capacity within a hospital. For instance, the United Kingdom's Department of Health treats LoS as a key performance indicator and uses it both to monitor hospital quality and to manage patients' expectations [8]. The length of time patients spend in hospital beds is known to be a good measure of utilization for a number of hospital resources, including staffing and equipment. As a result, the department publishes average LoS on the National Health Service (NHS) website[5] as a hospital operations parameter to help patients make more informed choices on which hospital to visit. Through such increased transparency pressure is put on hospitals to improve patient care, which involves providing more cost efficient and standardized services often reflected in duration of the service [2]. Thus, gaining a better understanding of LoS provides an opportunity to reduce the time patients stay in hospitals without affecting the quality of service[6], which is in the financial and personal interests of hospitals and patients. Additionally, early and accurate knowledge of LoS can aid hospital administrators in management of bed occupancy. This is a crucial problem faced by hospitals, which are pressured to shorten the LoS, potentially increasing risk of patient complications after discharge. Medicare was among the first insurance companies to consider predicting

- *Authors associated with the Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA, 19122.*
E-mail: {jelena.stojanovic, gligorijevic, vladan, nemanja, mihajlo, zoran.obradovic}@temple.edu

Manuscript received October, 2015; revised May, 2016s.

1. http://www.qualityindicators.ahrq.gov/Default.aspx, acc. October, 2015
2. http://www.cha.com/Documents/Publications/2012_Charge_Report.aspx, accessed October, 2015
3. http://www.beckershospitalreview.com/finance/100-things-to-know-about-medicare-reimbursement.html, acc. Oct 2015
4. http://www.commonwealthfund.org/publications/newsletters/quality-matters/2012/april-may/in-focus, accessed October 2015
5. NHS Choices. [http://www.nhs.uk], accessed October 2015
6. http://www.institute.nhs.uk/quality_and_service_improvement_tools/quality_and_service_improvement_tools/length_of_stay.html, acc. Oct 2015

length of hospital stay for each inpatient and using it for diagnosis of related groups [3]. The acceptance of length of stay as an indicator of resource utilization has caused a surge of interest across the healthcare industry in the predictability of LoS.

Increased penetration of information technologies in hospital systems has enabled collections of vast amounts of data in a form of large-scale Electronic Health Records (EHRs), which became an important source of detailed patient information within hospitals [9]. EHR data presents an unique opportunity for data–driven progress in early and accurate diagnostics and therapy, allowing medical staff to improve patient's care by learning from previous encounters [9], [10].

In recent years an increasing emphasis is given to the effective mining of clinical data in order to obtain actionable insights for improving healthcare delivery, a concept often termed "data-driven healthcare" [11], [12]. Data–driven health care practitioners have been addressing various problems aimed to improve healthcare quality [13], [14], [15], [10], [16]. The overall objective is to build a stable framework for modeling different aspects of the healthcare systems, and to provide significant insights to healthcare institutions and patients alike. Some particularly important and impactful applications are aimed towards predictive modeling of health outcomes in terms of diseases, procedures, mortality, and other measures that may have a huge impact on quality of patient treatment. The models are used to improve detection of high-risk groups of patients, or detect important effects not taken into consideration in prior medical treatments.

However, the modeling process is very challenging, as healthcare observational data are often sparse, heterogeneous, and/or incomplete due to different hospital and insurance policies, further aggravated by non-standardized physician practices [17]. The existing data mining tools are not fully capable of addressing the important task of healthcare modeling [18], and, in order to make use of multifaceted, noisy healthcare data sources, development of novel efficient and effective machine learning approaches is required.

In this study we address this important problem, and propose a novel approach that makes use of the latest advances in the representation learning for the task of predicting inpatient length of stay, pricing, and survival rates, with the objective of modeling the quality of healthcare services. In the following section we present the proposed approach. Section 3 describes large scale EHR database used in empirical analysis. The analysis and experimental results are described in detail in Section 4. Finally, we conclude our study and discuss drawbacks of the current approach and provide suggestions for future work in Section 5.

2 The proposed approach

In this section we present a novel approach for learning low-dimensional, distributed representations of patient EHRs. As a first step, we describe how to apply state-of-the-art, unsupervised neural language models for learning embeddings of *diseases* and applied clinical *procedures* from the EHR data of individual patients. Then, the obtained embeddings are employed to find useful *inpatient* feature vectors, used to train predictive models of the healthcare quality indicators in a supervised manner. The entire pipeline of the proposed methodology is illustrated in Figure 2 and each step is presented in more details in the following sections.

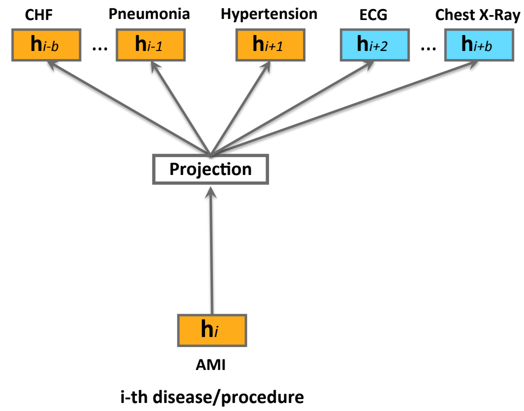

Fig. 1: Graphical representations of the *disease+procedure2vec* model. The model uses central disease/procedure h_i to predict b diseases/procedures (colored yellow and blue, respectively) that come before and b that come after it in the discharge record.

2.1 Low-dimensional embedding models

Assume we are given a set $\mathcal{R}$ of N hospital inpatient discharge records (representing a single hospital visit) and sets $\mathcal{D}$ of possible diseases and $\mathcal{P}$ procedures. Then, a discharge record $r_i = [(d_{i1}, \ldots, d_{iD_i}), (p_{i1}, \ldots, p_{iP_i})] \in \mathcal{R}, i = 1, \ldots, N$, of the i^{th} patient is defined as a sequence of diseases $d_i \in \mathcal{D}$ and procedures $p_i \in \mathcal{P}$ at the end of a hospital stay. Here, D_i is the number of diagnosed diseases and P_i is the number of applied procedures in the sequence, so that $D_i + P_i = H_i$ and that record is represented as $r_i = (h_{i1}, \ldots, h_{iH_i}) \in \mathcal{R}$, where h_{il} can be a disease or a procedure in the sequence. Then, using the set $\mathcal{R}$, the objective is to find M-dimensional real-valued representations $\mathbf{v}_d \in \mathbb{R}^M$ for every disease d and $\mathbf{v}_p \in \mathbb{R}^M$ for every procedure p, such that similar diseases and procedures lie nearby in the joint M-dimensional vector space and to use them to build a patient vector representation $x_i \in \mathbb{R}^M$ for training predictive models of the healthcare quality indicators.

Before discussing applications to specific healthcare related prediction problems, it is intuitive to introduce neural language models as applied to NLP. These methods take advantage of word order, and assume that closer words in the word sequence are statistically more dependent. Typically, a neural language model learns the probability distribution of the next word given a fixed number of preceding words that act as the context. More formally, given a sequence of words $(w_1, w_2, \ldots, w_T)$ from the training data, the objective of the model is to maximize the average log-likelihood function,

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \log \mathbb{P}(w_t | w_{t-b+1} : w_{t-1}), \tag{1}$$

where w_t is the t^{th} word, and $w_{t-b+1} : w_{t-1}$ is a sequence of b successive preceding words that act as the context to the word w_t. A typical approach to approximate the probability distribution $\mathbb{P}(w_t | w_{t-b+1} : w_{t-1})$ is to use a neural network model architecture [19]. The neural network is trained by projecting the vectors for context words $(w_{t-b+1}, \ldots, w_{t-1})$ into a latent representation with multiple non-linear hidden layers and the output softmax layer comprising W nodes, where W is the vocabulary size (in our task equal to the number of diseases and procedures $|\mathcal{D}| + |\mathcal{P}|$), while attempting to predict word w_t with high probability.

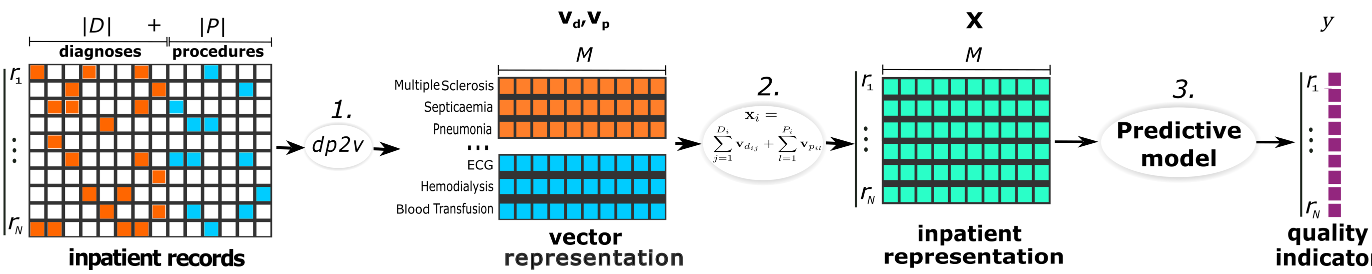

Fig. 2: Pipeline of the proposed approach: 1) Use the proposed embedding methodology to learn compact vector representation of diseases and procedure $\mathbf{v}_d, \mathbf{v}_d \in \mathbb{R}^M$ using raw EHR data $\in \mathbb{R}^{|D|+|P|}$; 2) generate inpatient representation $\boldsymbol{X}$ from the learned embeddings $\mathbf{v}_d$ and $\mathbf{v}_d$; 3) train regression and classification models to predict important indicators of healthcare quality y (LoS, TOTCHG and mortality for certain procedures and medical conditions of an inpatient).

When working with large-scale data, the vocabulary size W can easily reach millions. In those cases, training of the neural network becomes a challenging task, as updates of word vectors become computationally expensive. For that reason, recent approaches [20] propose log-linear models which aim to reduce the computational complexity. The use of hierarchical softmax [21] or negative sampling [20] is shown to be effective in substantially speeding up the training process.

2.2 disease+procedure2vec method

In this section we propose *disease+procedure2vec* (dp2v) approach for learning diseases and procedures representations (step 1 in Figure 2) that extend models of the recently proposed *word2vec* algorithm [20]. The key insight is that we can represent the patients' lists of diseases and procedures from EHRs as sequences of tokens, and view each sequence as a sample from some unknown language. Following this reasoning, the language model learns representations of diseases and procedures in a low-dimensional space using each patient discharge record as a "sentence" and the diseases and procedures within the record as "words", to borrow the terminology from the NLP domain. Low-dimensional representations for diseases and procedures are learned by maximizing the objective function $\mathcal{L}$ over the entire set $\mathcal{R}$ of records as follows,

$$\mathcal{L} = \sum_{r \in \mathcal{R}} \sum_{h_i \in r} \sum_{-b \le m \le b, m \ne 0} \log \mathbb{P}(h_{i+m}|h_i). \tag{2}$$

Probability $\mathbb{P}(h_{i+m}|h_i)$ of observing some "neighboring" disease/procedure h_{i+m} given the current disease/procedure h_i is defined using the soft-max function as

$$\mathbb{P}(h_{i+m}|h_i) = \frac{\exp(\mathbf{v}_{h_i}^{\top} \mathbf{v}'_{h_{i+m}})}{\sum_{h=1}^{H} \exp(\mathbf{v}_{h_i}^{\top} \mathbf{v}'_{h})}, \tag{3}$$

where $\mathbf{v}_h$ and $\mathbf{v}'_h$ are the input and output M-dimensional vector representations of disease/procedure h and hyper-parameter b represents the length of the context for disease records. Note that h can represents either d or p, with $H = |\mathcal{D}| + |\mathcal{P}|$.

As illustrated in Figure 1 and equation (3), *disease+procedure2vec* uses central disease/procedure h_i to predict b diseases/procedures that come before and b diseases/procedures that come after it in the discharge record, an architecture known as the SkipGram. As a result, diseases and procedures that often co-occur and have similar contexts (i.e., with similar neighboring diseases and procedures) will have similar representations as learned by our model. Additionally, we have considered a continuous bag of words architecture (CBOW), that uses context diseases and procedures to predict a central disease or procedure, however, the SkipGram architecture was consistently more accurate than the CBOW (as shown in Figure 3) and as such was the one used in *disease+procedures2vec* model.

TABLE 1: Number of inpatient stays and number of diagnoses and procedure codes used for different healthcare providers

Provider	N	$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{P}\rvert$	$\lvert\mathcal{D}\rvert+\lvert\mathcal{P}\rvert$
Medicare	11,300,025	11,636	3,649	15,285
Medicaid	9,134,840	12,237	3,668	15,905
Private insurance	12,344,355	12,458	3,737	16,195
Self-pay	1,247,209	10,640	3,230	13,870

The *disease+prodedure2vec* model was optimized using stochastic gradient ascent, suitable for large-scale problems. However, computation of gradients is proportional to the number of unique disease and procedures in the datasets, which may be computationally expensive in practical tasks. As an alternative, we used negative sampling approach [20], which significantly reduces the computational complexity.

2.2.1 Patient visit representation

Having learned the disease and procedure vectors, we aim to exploit them for the purpose of predicting total charges, length of stay, and mortality. For this purpose, we generate a data set $\mathcal{M} = \{(\mathbf{x}_i, y_i), i = 1, ..., N\}$, where for each record r_i the value of $y_i \in \mathcal{Y}$ represents one of the target variables: LoS, total charges (TOTCHG), or binary mortality indicator, and $\mathbf{x}_i \in \mathbb{R}^M$ is a patient's feature vector calculated by summing vectors of diseases and procedures that appear in that record [22] (step 2 in Figure 2),

$$\mathbf{x}_i = \sum_{j=1}^{D_i} \mathbf{v}_{d_{ij}} + \sum_{l=1}^{P_i} \mathbf{v}_{p_{il}}. \tag{4}$$

Once the data set $\mathcal{M}$ is generated, the learning task is to find a prediction function $f : \mathbb{R}^M \to \mathcal{Y}$, which maps each patient visit into one of the three variables of interest depending on the task (step 3 in Figure 2). When predicting LoS and TOTCHG this results in a regression problem, while for mortality prediction the problem can be viewed as a classification task.

2.2.2 The analysis of model parameters

In Figure 3 results obtained by varying vector dimension and window size for both CBOW and SkipGram models are shown for the task of predicting total charges. The SkipGram model was

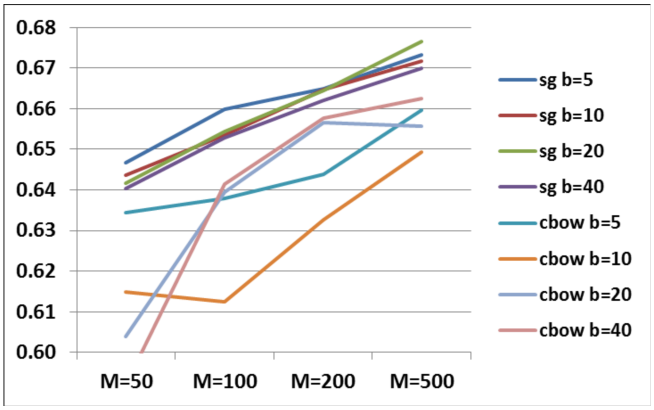

Fig. 3: R^2 results obtained by varying vector dimension ($\mathcal{M}$) and window size (b) for SkipGram (sg) and CBOW ($cbow$) models for the task of predicting total charges.

consistently more accurate than the CBOW model, thus we opted to use this model in *disease+procedures2vec* approach. Varying parameter b did not introduce much variation in the results for SkipGram, thus we chose to set context neighborhood size to $b = 40$, such that model captures larger context and most of the diseases and procedures in that record. From Figure 3 we can see that increasing parameter M improves the accuracy, however dimensionality is increased, leading to a more complex model that is more difficult to train. Dimensionality of the embedding space was set to $M = 200$, the parameter M was chosen in such a manner as to avoid larger dimensionality of the learned model while obtaining good predictive accuracy. Finally, we used 25 negative samples in each vector update for negative sampling. Similarly to the approach presented in [20], the most frequent diseases and procedures were sub-sampled during the training phase.

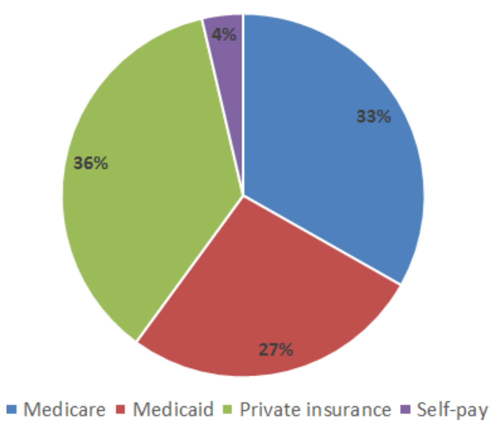

Fig. 4: Distribution of California inpatient hospital admissions by the primary payer (for a 2003-2011 period)

3 EHR DISCHARGE DATABASE

For the purpose of this study we explored the State Inpatient Database (SID)[7], an archive that stores the inpatient discharge abstracts from a number of data organizations. The data is provided by the Agency for Healthcare Research and Quality and is included in the Healthcare Cost and Utilization Project (HCUP). In particular, we used the SID California database, which contains 35,844,800 inpatient discharge records over a period of 9 years (from January 2003 to December 2011) in 474 different hospitals. SID data provides discharge records for each inpatient that may contain up to 25 diagnosis codes and up to 15 procedure codes in ICD9 coding schema that were applied during this particular admission of the patient. This coding schema[8] originates from the 9$^{\text{th}}$ revision of the International Classification of Diseases (ICD9), a hierarchical coding scheme which is a part of standard diagnostic tools for epidemiology, health management, and clinical purposes.

The disease coding process of EHR databases is tedious work, even under the most obvious circumstances. It requires proper application of the AHA Coding Clinic guidelines [23] and the Official Guidelines for Coding and Reporting for inpatient care [24], and documented physician notes are mandatory for precise coding [25]. Thus diagnoses found in the EHR records are ordered by their importance to the patient's reason of admission and hospital stay while respecting given guidelines of diagnoses coding. As such, EHR data possess a 'grammar' of diagnoses and procedures codes, where contexts of different diseases and procedures in discharge records may provide significant additional information for the prediction of hospital quality indicators.

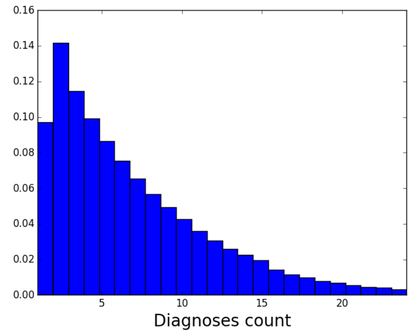

Fig. 5: Histogram of diagnoses counts for 35 million hospitalizations in California (on average 6.78 diagnoses were given per patient hospitalization)

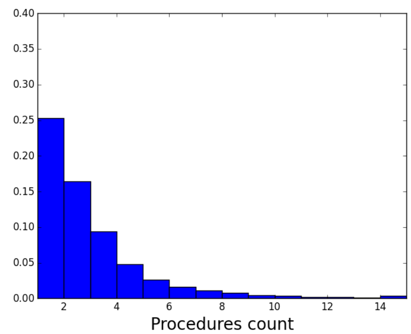

Fig. 6: Histogram of procedures counts for 35 million hospitalizations in California (on average 1.61 procedures were administered per patient hospitalization)

7. HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2005-2009. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp

8. http://www.who.int/classifications/icd/en/, accessed September 2015

TABLE 2: Association of procedures to two high-mortality diseases discovered by measuring cosine distance on features obtained using dp2v embedding model

Neighbors of *respiratory failure*	Neighbors of *congestive heart failure*
Insertion of endotracheal tube	Insertion of implantable heart assist system
Tracheostomy toilette	Implantation of cardiac resynchronization defibrillator total system (CRT-D)
Other lavage of bronchus and trachea	Implantation of cardiac resynchronization defibrillator pulse generator (CRT-D)
Bronchoscopy through artificial stoma	Insertion of percutaneous external heart assist device
Other oxygen enrichment	Heart transplantation
Other repair and plastic operations on trachea	Excision destruction or exclusion of left atrial appendage (LAA)
Fiber-optic bronchoscopy	Aquapheresis
Infusion of vasopressor agent	Automatic implantable cardioverter-defibrillator (AICD) check
Replacement of tracheostomy tube	Noninvasive programmed electrical stimulation (NIPS)
Replacement of gastrostomy tube	Removal of lead(s) [electrode] without replacement
Complete glossectomy	Endovascular removal of obstruction from head and neck vessel(s)
Other intubation of respiratory tract	Replacement of automatic cardioverter-defibrillator lead(s) only

Additionally, the SID database contains information about a hospital stay, including length of stay, total charges, type of payment, insurance type, discharge month, and survival information. In total, the SID California database covers 13,004 unique disease codes (out of around 14,000 present in ICD9 schema), and 3,830 procedure codes (out of around 4,000 present in ICD9 schema).

In Figure 4 we plot the distribution of inpatient admissions by primary payer (i.e., type of insurance). Histograms of diagnoses and procedures counts per visit are shown in Figures 5 and 6, respectively. Additionally, we show the number of records N, unique diseases $|\mathcal{D}|$, and procedures $|\mathcal{P}|$ for four types of health insurance in Table 1. To address different practices of health insurance providers, we built non-overlapping cohorts for each of four insurance groups and trained separate embedding models for each of them. The experimental setup and results are presented in the following section.

4 Empirical evaluation

In this section we first explore the embedding space learned using the proposed method, validating that the vector representations are meaningful and insightful. Then, we discuss linear predictive models used in the experiments, and describe baseline approaches for low-dimensional embedding. Lastly, we discuss experimental setup, give evaluation metrics, and present the obtained results.

4.1 Exploring associations in the embedding space

The dp2v model maps each disease and procedure into a common low-dimensional space, and in this section we provide evidence that such learned mappings are indeed medically relevant. In particular, we explored the embedding space by retrieving the nearest procedures to diseases found in the SID California database. This is done by choosing most similar procedures for a query disease via calculating cosine similarity of their vectors.

As examples of learned associations between diseases and procedures we selected to find nearest procedures for *respiratory failure* and *congestive heart failure* (CHF), two conditions that exhibit high mortality among patients. We retrieved 12 nearest procedures for each query disease, and show the results in Table 2. We can see that for the respiratory failure the method retrieved several procedures that serve to aid in breathing of the patient, such as insertion of endotracheal tube, tracheostomy toilette, repair and plastic operations on trachea, replacement of tracheostomy and gastrostomy tube, intubation of respiratory tract, and oxygen enrichment. We also see procedures that are commonly applied prior to bronchus examination and for bronchus cleaning, such as bronchoscopy for throat, trachea examination, and lavage of bronchus and trachea.

For the *congestive hearth failure* disease discovered associated procedures also confirm that dp2v embeddings are medically relevant. Several procedures in the top 12 list include different implants aimed to assist the heart (e.g., CRT, AICD) or electro method performed to stimulate heart pumping (e.g., NIPS). Other procedures include heart transplantation, aquapheresis (which treats fluid overflow that can be caused by CHF), or endovascular removal of blood clots that can be caused by a heart attack. The results validate the quality of the learned representations, where medically relevant diseases and procedures were found to be nearby in the embedding space.

4.2 Predictive models

Several penalized linear models for regression and classification tasks are used in our experiments, as suggested in the relevant literature [26], [27]. In particular, for regression problems we apply linear regression,

$$y_i = f(\mathbf{w}, \mathbf{x}_i) = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2), \tag{5}$$

where ε is a zero-mean Gaussian noise with variance σ^2. On the other hand, for the classification problem we use the logistic regression model,

$$y_i = f(\mathbf{w}, \mathbf{x}_i) = I\Big(\frac{1}{1 + \exp\big(-(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)\big)} > 0.5\Big). \tag{6}$$

Vector $\mathbf{w}$ is an unknown set of weights for both prediction models, and $I(\cdot)$ is an indicator function equal to 1 if the argument is true and 0 otherwise.

In addition, for both models we explored a number of regularization approaches, ranging from ℓ_1 Lasso to overlapping group Lasso penalizations. We summarized the training objectives of five penalized linear models in Table 3, where ℓ_1 indicates Lasso norm and ℓ_q is norm of the non-overlapping groups, $\mathbf{w}_i$ and $\mathbf{w}_{G_i}$ indicate a single dimension of the weight vector and a group of dimensions defined by the index set G_i, respectively. For the sparse group Lasso, the index sets G_i do not overlap (i.e., $G_i \cap G_j = \emptyset, \forall i \neq j$), which is not the case for the overlapping group Lasso. The index sets G_i for group Lasso models were defined in groups of ten consecutive features, indexed from 1 to 10, 11 to 20, and so on until $M - 9$ to M (smaller groups showed better performance). For the overlapping group Lasso the index sets were defined as 1 to 20, 11 to 30, and so on.

TABLE 3: Overview of linear models used in this study

Penalty	Optimization problem	Model name	Abbreviation	Description
Lasso	$\min_{\mathbf{w}} f(\mathbf{w}, \mathbf{x}) + \lambda \Vert\mathbf{w}\Vert_1$	LeastR LogisticR	LR logR	Least squares loss Logistic loss
Group Lasso	$\min_{\mathbf{w}} f(\mathbf{w}, \mathbf{x}) + \lambda \Vert\mathbf{w}\Vert_{q,1}$	glLeastR glLogisticR	glLR glLogR	Least Squares Loss Logistic Loss
Fused Lasso	$\min_{\mathbf{w}} f(\mathbf{w}, \mathbf{x}) + \lambda_1 \Vert\mathbf{w}\Vert_1 + \lambda_2 \sum_{i=1}^{M-1} \vert\mathbf{w}_i - \mathbf{w}_{i+1}\vert$	fusedLeastR fusedLogisticR	fLR fLogR	Least Squares Loss Logistic Loss
Sparse group Lasso	$\min_{\mathbf{w}} f(\mathbf{w}, \mathbf{x}) + \lambda \Vert\mathbf{w}\Vert_1 + \sum_{i=1}^{g} \lambda_{G_i} \Vert\mathbf{w}_{G_i}\Vert_2$	sgLeastR sgLogisticR	sgLR sgLogR	Least Squares Loss Logistic Loss
Overlapping group Lasso	$\min_{\mathbf{w}} f(\mathbf{w}, \mathbf{x}) + \lambda \Vert\mathbf{w}\Vert_1 + \sum_{i=1}^{g} \lambda_{G_i} \Vert\mathbf{w}_{G_i}\Vert_2$	overlapping LeastR overlapping LogisticR	olLR olLogR	Least Squares Loss Logistic Loss

All λ parameters were set to be equal and chosen from range $[0.01, 0.1]$, determined through cross-validation. In the conducted experiments, an implementation from the efficient SLEP[9] package [28] is used for training the models.

4.3 Low-dimensional embedding baselines

As the objective of our work is to find meaningful representations of diagnoses and procedures in a low-dimensional space, we compare the proposed embedding approach to a number of state-of-the-art alternatives. More specifically, we considered Latent Dirichlet Allocation (LDA) [29], as a representative of topic learning models, as well as spectral clustering [30] and modularity [31] approaches used for low-dimensional representations of nodes in an undirected graph representing co-occurrence of diagnoses and procedures. In addition, we examined binary encoding in the original $\mathbb{R}^{|\mathcal{D}|+|\mathcal{P}|}$ space and applied PCA on such sparse representation. In the following sections we briefly describe the baseline embedding methods.

4.3.1 Binary coding with dimensionality reduction (dPCA)

A high-dimensional representation of EHR records is obtained by creating a binary vector of $|\mathcal{D}| + |\mathcal{P}|$ entries corresponding to the total number of unique diagnoses and procedures found in the SID California database (the values of $|\mathcal{D}|$ and $|\mathcal{P}|$ can be found in Table 1). Each entry in the extended representation is either 0 or 1 depending whether that particular diagnoses or procedure occurred in that discharge record. As the dimensionality of this problem is large, we apply PCA [32] to reduce dimensionality of the problem to M dimensions (in our experiments we set the dimensionality of the embedding space to $M = 200$ for all methods).

4.3.2 Spectral clustering (Spec)

If we consider an undirected network $\mathcal{G}$ of co-occurrences of diagnoses and procedures in hospital discharge data, we can use advanced tools to learn node representation in $\mathbb{R}^M$ space using the information from the graph. The spectral clustering method generates a representation in $\mathbb{R}^M$ space from the first M eigenvectors of $\mathbf{L}$, a normalized graph Laplacian of graph $\mathcal{G}$ [30]. The Laplacian is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D} = \text{diag}(d_1, d_2, ..., d_N, p_1, p_2, ..., p_N)$ and $\mathbf{A}$ is the adjacency matrix of $\mathcal{G}$. The normalized Laplacian $\mathbf{L}$ is then defined as

$$\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}. \tag{7}$$

Then, we find the first M eigenvectors of the normalized Laplacian and treat them as latent dimensions of nodes from the graph $\mathcal{G}$, thus inferring low-dimensional representations for both procedures and diagnoses.

4.3.3 Modularity (Mod)

This method generates a representation in $\mathbb{R}^M$ space from the top M eigenvectors of $\mathbf{B}$, the modularity matrix of $\mathcal{G}$. For two nodes i and j in the graph $\mathcal{G}$ with degrees d_i and d_j, respectively, the expected number of edges between these two nodes in a uniform random graph model is $\frac{d_i d_j}{2m}$, where m represents the total number of edges in the graphs. Modularity matrix $\mathbf{B}$ measures the deviation of adjacency matrix $\mathbf{A}$ from a uniform random graph with the same degree distribution,

$$\mathbf{B} = \mathbf{A} - \frac{1}{2m}\mathbf{d}\mathbf{d}^{\top}. \tag{8}$$

While in many real graphs the adjacency matrix $\mathbf{A}$ is typically very sparse, the modularity matrix $\mathbf{B}$ is typically dense. The matrix $\mathbf{B}$ is then decomposed using SVD method and the obtained eigenvectors of $\mathbf{B}$ encode information in $\mathbb{R}^M$ space about modular partitions of the graph $\mathcal{G}$ [31], which are used to represent the nodes in a lower-dimensional space.

4.3.4 Latent Dirichlet Allocation (LDA)

LDA is a popular latent topic model [29], shown to obtain a state-of-the-art performance in a number of tasks both within and outside of the domain of the natural language processing. Assuming a fixed number of topics that generated the data, the model learns a topic distribution over the diseases and procedures, effectively embedding them in the topic space. Then, the found topical representations can be used as feature vectors in the classification and regression models.

4.4 Evaluation metrics

For evaluation of the proposed regression methods we use a goodness-of-fit metric R^2 defined as follows,

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \mu)^2}, \tag{9}$$

where y_i and $\hat{y}_i$ are true and predicted values of the target variable for the record r_i, respectively, and μ is the mean value for all records in the set $\mathcal{R}$.

For evaluation of patient survival analysis we use an accuracy measure defined as follows,

$$accuracy = \frac{tp + tn}{tp + fp + tn + fn}, \tag{10}$$

9. http://www.yelab.net/software/SLEP/, accessed October 2015

TABLE 4: Average total charges, length of stay in days, and survival rate for four datasets from SID California database

Provider	TOTCHG	LoS	Survival rate
Medicare	$50,878.02	5.94	0.96
Medicaid	$30,264.11	4.51	0.99
Private insurance	$29,412.26	3.71	0.99
Self–pay	$31,824.64	3.97	0.98

where tp and tn denote true positives and true negatives, respectively (i.e., correctly classified cases), while fp and fn denote false positive and false negative test examples, respectively (i.e., mistakenly classified cases).

4.5 Results

In this section we provide experimental results of three predictive tasks on four insurance data sets. Different representations of diagnoses and procedures were trained for each insurance data set, and learned using five competing approaches. In particular, four datasets were created for each of the insurance categories. From the first month of the observation period we sampled 100,000 records for training and testing predictive models, while the remaining data was used for learning the embedding models. From the 100,000 sampled examples, 80% were randomly chosen for regression and classification training, while 20% were used for testing. In addition, as hospitals currently report mean values for TOTCHG, LoS, and survival rate, shown in Table 4, we also use these values as a naïve baseline. We further comment on their performance in the following sections.

4.5.1 Prediction of total charges (TOTCHG)

In this section we address the problem of predicting total charges for a patient per hospital visit. As discussed previously, there are more than 100 factors that may influence hospital charges, making the estimation of the exact value a non-trivial problem. For example, Table 4 suggests that Medicare patients are charged almost twice as much as the other three groups of patients (which are similar with respect to average charges). As Medicare patients are people of age, we can assume that they are diagnosed with more conditions and have more procedures performed compared to the other three insurance groups.

We first used the mean TOTCHG computed on the training data as a trivial baseline predictor and measured its accuracy on the test data for each provider. We observed that this trivial predictor underperformed and obtained $R^2 < 0$. The result indicates that the information provided by hospitals is of little value for an individual patient, and in the following we explore more involved approaches for this predictive task, where as an input we take into account diagnosed diseases for a specific patient and a list of procedures that might be applied.

In Table 5, we show the results in terms of R^2 measure obtained by five regression models for four insurance categories, making use of a 200-dimensional representations obtained by various embedding methods. We observe that the proposed dp2v model outperformed the baseline approaches in all 20 experiments (for all five regression models and for all four insurance categories). The R^2 improvement of using the proposed embedding over the best performing alternative is on average around 20%. The obtained results strongly suggest that the most useful representation for predicting total charges is learned using dp2v model. We also see that the LR regression model outperformed alternatives in this application, and that the most difficult task was to estimate costs for patients on Medicaid insurance.

TABLE 5: R^2 results obtained for predicting total charges by five regression models for four insurance categories

	LR	glLR	fLR	sgLR	olLR
Medicare					
dp2v	***0.6454***	**0.6388**	**0.5846**	**0.3641**	**0.4204**
Spec	0.5584	0.5274	0.3487	≤ 0	0.02218
Mod	0.5635	0.5235	0.3628	≤ 0	≤ 0
LDA	0.2022	0.2040	0.1955	0.2141	0.2008
dPCA	0.5059	0.4805	0.3300	≤ 0	0.0005
Medicaid					
dp2v	***0.5850***	**0.5805**	**0.5646**	**0.4550**	**0.4550**
Spec	0.5155	0.5138	0.4423	0.1892	0.2836
Mod	0.5163	0.5092	0.4490	0.0945	0.1769
LDA	0.2052	0.2046	0.1974	0.1630	0.1511
dPCA	0.4112	0.4118	0.3094	0.0601	0.1166
Private insurance					
dp2v	***0.6553***	**0.6434**	**0.5930**	**0.2903**	**0.3773**
Spec	0.5744	0.5539	0.4401	0.1038	0.1801
Mod	0.5757	0.5516	0.4111	0.0196	0.0374
LDA	0.1936	0.1932	0.1692	0.1610	0.1516
dPCA	0.5688	0.5438	0.4967	0.0768	0.1875
Self-pay					
dp2v	***0.6093***	**0.5954**	**0.5575**	**0.3281**	**0.3375**
Spec	0.5246	0.4989	0.4100	0.0686	0.1491
Mod	0.4756	0.4672	0.3680	0.0194	0.0879
LDA	0.0939	0.0945	0.0864	0.0787	0.0455
dPCA	0.6048	0.5706	0.4390	0.1057	0.1689

4.5.2 Prediction of length of stay (LoS)

The length of stay is one of the most important indicators of quality of a hospital system, and is an important parameter considered when choosing a hospital. Therefore, providing LoS estimation for a specific visit is a very important task. Many hospitals are handling these predictions by reporting the mean length of stay. Similarly to the total charges, our experiments indicate that such a summary statistic is not informative for individual patients ($R^2 < 0$).

In this study we consider a patient that is diagnosed with several diseases, and we account for procedures suggested for this patient in order to estimate the patient's length of stay. The results of five regression models learned on latent features projected by the competing models are shown at Table 6. We observe that the proposed dp2v model was the best choice in 18 out of 20 experiments, obtaining average accuracy improvements up to 34% for Medicare, 19% for Medicaid, and 20% for self-pay patients over the best performing alternative. Interestingly, for private insurances the proposed model did not provide improvement for all predictive models. Nevertheless, the model that performed the best on this dataset used features learned by the dp2v embedding method. We can conclude that the proposed embedding approach provides the best features for prediction of length of stay among the considered models overall.

TABLE 6: R^2 results obtained for predicting LoS by five regression models for four insurance categories

	LR	glLR	fLR	sgLR	olLR
Medicare					
dp2v	***0.4356***	**0.4260**	**0.3872**	**0.2687**	**0.3411**
Spec	0.4092	0.3989	0.2840	0.0598	0.0935
Mod	0.4136	0.3955	0.2569	≤ 0	≤ 0
LDA	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
dPCA	0.3337	0.3149	0.2538	≤ 0	0.0005
Medicaid					
dp2v	***0.3220***	**0.3178**	**0.3089**	**0.1876**	**0.1964**
Spec	0.2691	0.2571	0.1906	0.0392	0.0818
Mod	0.2910	0.2641	0.1813	0.0093	0.0259
LDA	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
dPCA	0.2715	0.2575	0.1703	0.0253	0.0423
Private insurance					
dp2v	**0.3657**	**0.3599**	***0.3874***	0.0493	0.1230
Spec	0.3463	0.3507	0.2528	0.0155	0.0321
Mod	0.3508	0.3574	0.2404	≤ 0	0.0125
LDA	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
dPCA	0.2893	0.3448	0.2342	**0.0702**	**0.1254**
Self-pay					
dp2v	***0.2402***	**0.2383**	**0.2137**	**0.0766**	**0.0945**
Spec	0.1402	0.1279	0.0813	≤ 0	0.0026
Mod	0.1459	0.1290	0.0743	≤ 0	≤ 0
LDA	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0
dPCA	0.0876	0.0774	0.0432	≤ 0	0.0015

TABLE 7: Mortality prediction accuracy by five classification models for four insurance categories

	logR	glLogR	fLogR	sgLogR	olLogR
Medicare					
dp2v	***0.6256***	**0.6131**	**0.5385**	**0.5433**	**0.5332**
Spec	0.4923	0.4923	0.4923	0.4923	0.4923
Mod	0.4923	0.4923	0.4923	0.4923	0.4923
LDA	0.4928	0.4928	0.4928	0.4928	0.4928
dPCA	0.4825	0.4825	0.4825	0.4825	0.4825
Medicaid					
dp2v	***0.8289***	**0.8273**	**0.7796**	**0.7566**	**0.7928**
Spec	0.5066	0.5066	0.5066	0.5066	0.5066
Mod	0.5066	0.5066	0.5066	0.5066	0.5066
LDA	0.5164	0.5164	0.5164	0.5164	0.5164
dPCA	0.5000	0.5000	0.5000	0.5000	0.5000
Private insurance					
dp2v	***0.8714***	**0.8643**	**0.7405**	**0.7619**	**0.7524**
Spec	0.5167	0.5167	0.5167	0.5167	0.5167
Mod	0.5167	0.5167	0.5167	0.5167	0.5167
LDA	0.4881	0.4881	0.4881	0.4881	0.4881
dPCA	0.5769	0.5769	0.5769	0.5769	0.5769
Self-pay					
dp2v	***0.8435***	**0.8252**	**0.6125**	**0.6357**	**0.5391**
Spec	0.4951	0.4951	0.4951	0.4951	0.4951
Mod	0.4951	0.4951	0.4951	0.4951	0.4951
LDA	0.4792	0.4792	0.4792	0.4792	0.4792
dPCA	0.4764	0.4764	0.4764	0.4764	0.4764

4.5.3 *Prediction of inpatient survival*

Lastly, we turn our attention to estimating patients mortality, which we use as an ultimate quality indicator of hospital care considered in this study [33]. More specifically, the prediction task was to estimate patient's survival probability, taking into consideration diagnosed conditions and conducted procedures.

From Table 4, we observe that data sets for this prediction task are highly imbalanced. Therefore, in order to make a fair comparison we drew a balanced sample for each of the insurance categories and learned classification models on such data. From Table 7 we observe that survival for the Medicare group was the most difficult to predict, and that for the private insurance group classification models perform the best when compared to other insurance categories. Nevertheless, mirroring the result from the previous experiments, we can see that the features learned by the dp2v method resulted in the highest accuracy, outperforming the competing approaches by a significant margin.

5 Conclusion

In this paper we proposed a novel unsupervised approach for learning representations of inpatients, diseases and procedures from large hospitalization records database, building upon the latest advances in neural embedding language models. We compared our approach to four competitive baselines on three different predictive tasks, where we applied five regression and classification models. Experiments on predicting important inpatient quality indicator values for a potential patient stay were conducted on a large-scale inpatient EHR database, with four cohorts defined according to insurance categories. Benefits of using the proposed embedding approach versus the alternatives were shown of a majority of conducted experiments, demonstrating the power of the proposed approach and its potential for modeling healthcare quality. However, the methodology still possesses drawbacks in terms of modeling diseases and procedures embeddings. For example, currently the model does not account for the concept of primary diagnosis and secondary diagnoses, heterogeneity of a disease is not captured well by the given approach and multiple visits of same patients, including readmission, are not included in the modeling process. Modeling longitudinal effects and addressing disease heterogeneity will be the focus of our future work.

Acknowledgments

This research was supported in part by DARPA grant FA9550-12-1-0406 negotiated by AFOSR, NSF BIGDATA grant 14476570 and ONR grant N00014-15-1-2729. Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality, provided data used in this study.

References

[1] J. B. Dimick, H. G. Welch, and J. D. Birkmeyer, "Surgical mortality as an indicator of hospital quality: the problem with small sample size," *Jama*, vol. 292, no. 7, pp. 847–851, 2004.

[2] P. P. Goodney, T. A. Stukel, F. L. Lucas, E. V. Finlayson, and J. D. Birkmeyer, "Hospital volume, length of stay, and readmission rates in high-risk surgery," *Annals of surgery*, vol. 238, no. 2, p. 161, 2003.

[3] T. A. Marciniak, E. F. Ellerbeck, M. J. Radford, T. F. Kresowik, J. A. Gold, H. M. Krumholz, C. I. Kiefe, R. M. Allman, R. A. Vogel, and S. F. Jencks, "Improving the quality of care for medicare patients with acute myocardial infarction: results from the cooperative cardiovascular project," *Jama*, vol. 279, no. 17, pp. 1351–1357, 1998.

[4] D. L. Barlett and J. B. Steele, *Critical condition: how health care in America became big business–and bad medicine*. Broadway, 2006.

[5] T. Gross and M. J. Notowidigdo, "Health insurance and the consumer bankruptcy decision: Evidence from expansions of medicaid," *Journal of Public Economics*, vol. 95, no. 7, pp. 767–778, 2011.

[6] D. U. Himmelstein, E. Warren, D. Thorne, and S. J. Woolhandler, "Illness and injury as contributors to bankruptcy," *Available at SSRN 664565*, 2005.

[7] C. Zhan and M. R. Miller, "Excess length of stay, charges, and mortality attributable to medical injuries during hospitalization," *Jama*, vol. 290, no. 14, pp. 1868–1874, 2003.

[8] E. M. Carter and H. W. Potts, "Predicting length of stay from an electronic patient record system: a primary total knee replacement example," *BMC medical informatics and decision making*, vol. 14, no. 1, p. 26, 2014.

[9] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, no. 6, pp. 395–405, 2012.

[10] J. C. Ho, J. Ghosh, and J. Sun, "Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 115–124.

[11] *Data driven healthcare*. MIT Technology Review, 2014, vol. 117(5):119.

[12] L. B. Madsen, *Data-Driven Healthcare: How Analytics and BI are Transforming the Industry*. Wiley, 2014.

[13] T. Xiang, D. Ray, T. Lohrenz, P. Dayan, and P. R. Montague, "Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought," *PLoS Comput. Biol*, vol. 8, p. e1002841, 2012.

[14] J. Zhou, F. Wang, J. Hu, and J. Ye, "From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 135–144.

[15] J. C. Ho, J. Ghosh, S. R. Steinhubl, W. F. Stewart, J. C. Denny, B. A. Malin, and J. Sun, "Limestone: High-throughput candidate phenotype generation via tensor factorization," *Journal of biomedical informatics*, vol. 52, pp. 199–211, 2014.

[16] D. Gligorijevic, J. Stojanovic, and Z. Obradovic, "Improving confidence while predicting trends in temporal disease networks," in *4th Workshop on Data Mining for Medicine and Healthcare, 2015 SIAM International Conference on Data Mining*, 2015.

[17] G. Hripcsak and D. J. Albers, "Next-generation phenotyping of electronic health records," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 117–121, 2013.

[18] P. Chowriappa, S. Dua, and Y. Todorov, "Introduction to machine learning in healthcare informatics," in *Machine Learning in Healthcare Informatics*. Springer, 2014, pp. 1–23.

[19] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Mar. 2003.

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013, pp. 3111–3119.

[21] F. Morin and Y. Bengio, "Hierarchical probabilistic neural network language model," in *Proceedings of the international workshop on artificial intelligence and statistics*, 2005, pp. 246–252.

[22] M. Grbovic, N. Djuric, V. Radosavljevic, F. Silvestri, and N. Bhamidipati, "Context- and content-aware embeddings for query rewriting in sponsored search," in *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2015.

[23] A. H. Association *et al.*, "Aha coding clinic for icd-9-cm," *AHA, Chicago*, 2008.

[24] C. for Medicare, M. Services *et al.*, "Icd-9-cm official guidelines for coding and reporting," *Baltimore, CMS and NCHS, 2008Centers for Medicare and Medicaid Services (CMS), the National Center for Health Statistics (NCHS), Baltimore CMS and NCHS*, 2011.

[25] L. A. Wiedemann, "Coding sepsis and sirs," *Journal of AHIMA*, vol. 78, no. 4, pp. 76–78, 2007.

[26] W. M. Tierney, J. F. Fitzgerald, M. E. Miller, M. K. James, and C. J. McDonald, "Predicting inpatient costs with admitting clinical data," *Medical care*, pp. 1–14, 1995.

[27] J. L. Moran, P. J. Solomon, A. R. Peisach, and J. Martin, "New models for old questions: generalized linear models for cost prediction," *Journal of evaluation in clinical practice*, vol. 13, no. 3, pp. 381–389, 2007.

[28] J. Liu, S. Ji, and J. Ye, *SLEP: Sparse Learning with Efficient Projections*, Arizona State University, 2009. [Online]. Available: http://www.public.asu.edu/~jye02/Software/SLEP

[29] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[30] L. Tang and H. Liu, "Leveraging social media networks for classification," *Data Mining and Knowledge Discovery*, vol. 23, no. 3, pp. 447–478, 2011.

[31] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[32] L. J. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality reduction: A comparative review," 2009.

[33] D. G. Pfister, D. M. Rubin, E. B. Elkin, U. S. Neill, E. Duck, M. Radzyner, and P. B. Bach, "Risk adjusting survival outcomes in hospitals that treat patients with cancer without information on cancer stage," *JAMA oncology*, vol. 1, no. 9, pp. 1303–1310, 2015.

Jelena Stojanovic was born in Belgrade, Serbia, in 1989. She received the B.S. and M.S. degree in Information Systems and Technologies from the University of Belgrade, Belgrade, Serbia, in 2012. and 2013, respectively. She is currently working toward the Ph.D. degree in Computer and Information Sciences at Temple University, Philadelphia, PA. Her research interests include data mining, machine learning and modeling of partially observed data.

Djordje Gligorijevic was born in Kragujevac, Serbia, in 1990. He received the B.S. degree in Information Systems and Technologies from the University of Belgrade, Belgrade, Serbia, in 2013. He is currently working toward the Ph.D. degree in Computer and Information Sciences at Temple University, Philadelphia, PA. His research interests include spatiotemporal data mining, machine learning, predictive uncertainty modeling and healthcare data analysis.

Vladan Radosavljevic received his B.A. and M.S. degree in electrical engineering from the University of Belgrade in 2003. He obtained a Ph.D. degree in computer and information sciences in 2011 at Temple University, Philadelphia, PA. He is currently working as an applied scientist at Uber Advanced Technologies Center in the machine learning group. His research interests include machine learning, data mining, geospatial modeling, and computational advertising.

Nemanja Djuric received his B.A. and M.S. degree in electrical engineering from the Faculty of Technical Sciences, University of Novi Sad in 2009. He obtained a Ph.D. degree in computer and information sciences in 2014 at Temple University, Philadelphia, PA. He is currently working as an applied scientist at Uber Advanced Technologies Center in the machine learning group. His research interests include machine learning, data mining, large-scale learning, and computational advertising.

Mihajlo Grbovic received his B.A. and M.S. degree in electrical engineering from the Faculty of Technical Sciences, University of Novi Sad in 2007. He obtained a Ph.D. degree in computer and information sciences in 2012 at Temple University, Philadelphia, PA. He is currently working as a senior research manager at Yahoo!. His research interests include machine learning, data mining, data-driven fault detection, and computational advertising.

Zoran Obradovic is an elected member of the Academia Europaea (the Academy of Europe) and a Foreign Academician at the Serbian Academy of Sciences and Arts, a L.H. Carnell Professor of Data Analytics at Temple University, Professor in the Department of Computer and Information Sciences with a secondary appointment in Statistics, and is the Director of the Data Analytics and Biomedical Informatics Center. He is the executive editor at the journal on Statistical Analysis and Data Mining, which is the official publication of the American Statistical Association and is an editorial board member at eleven journals. He is the chair at the SIAM Activity Group on Data Mining and Analytics and was co-chair for 2013 and 2014 SIAM International Conference on Data Mining and was the program or track chair at many data mining and biomedical informatics conferences. His work is published in more than 300 articles and is cited more than 15,000 times (H-index 48). For more details see http://www.dabi.temple.edu/∼zoran/