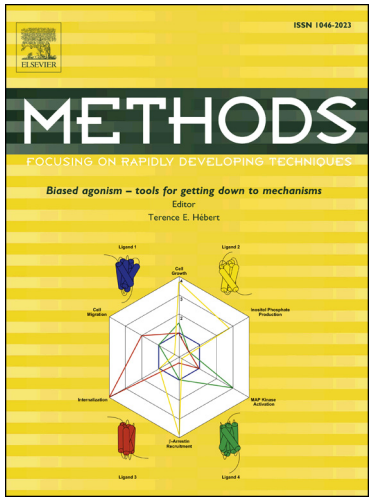
Accepted Manuscript



Discovering Disease Phenotypes from a Large Database of Inpatient Records: A Sepsis Study

Djordje Gligorijevic, Jelena Stojanovic, Zoran Obradovic

PII: S1046-2023(16)30232-8
DOI: http://dx.doi.org/10.1016/j.ymeth.2016.07.021
Reference: YMETH 4047

To appear in: *Methods*

Received Date: 2 May 2016
Revised Date: 26 July 2016
Accepted Date: 26 July 2016

Please cite this article as: D. Gligorijevic, J. Stojanovic, Z. Obradovic, Discovering Disease Phenotypes from a Large Database of Inpatient Records: A Sepsis Study, *Methods* (2016), doi: http://dx.doi.org/10.1016/j.ymeth.2016.07.021

Discovering Disease Phenotypes from a Large Database of Inpatient Records: A Sepsis Study

Djordje Gligorijevic[a,b], Jelena Stojanovic[a,b], Zoran Obradovic[a]

[a] *Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA 19122 USA*
[b] *Authors contributed equally*

Abstract

Data-driven phenotype discoveries on Electronic Health Records (EHR) data have recently drawn benefits across many aspects of clinical practice. In the method described in this paper, we map a very large EHR database containing more than a million inpatient cases into a low dimensional space where diseases with similar phenotypes have similar representation. This embedding allows for an effective segmentation of diseases into more homogeneous categories, an important task of discovering disease types for precision medicine. In particular, many diseases have heterogeneous nature. For instance, sepsis, a systemic and progressive inflammation, can be caused by many factors, and can have multiple manifestations on different human organs. Understanding such heterogeneity of the disease can help in addressing many important issues regarding sepsis, including early diagnosis and treatment, which is of huge importance as sepsis is one of the main causes of in-hospital deaths in the United States. This study analyzes state of the art embedding models that have had huge success in various fields, applying them to disease embedding from EHR databases. Particular interest is given to learning multi-type representation of heterogeneous diseases, which leads to more homogeneous groups. Our results show evidence that such representations have phenotypes of higher quality and also provide benefit when predicting mortality of inpatient visits.

Keywords: EHR databases, Phenotyping, Neural embedding models, Sepsis, Discovery of disease types

1. Introduction

Large-scale Electronic Health Records databases (EHRs) are an important source of detailed patient information that can potentially be used for more effective computational and statistical modeling aimed towards improved disease characterization and intervention [1, 2]. For example, benefits of big EHR analytics were evident in improving precision medicine by reducing uncertainty in decision-making and in design of preventive and therapeutic strategies [2, 3], in discovering novel relationships between human phenotypes and genotypes [4], and in improving overall healthcare by unearthing deeper medical insights. EHR modeling has been the focus of many studies aimed to improve healthcare [5, 6, 7, 8]. In clinical practice, these studies can allow medical practitioners to obtain novel insights in the patients' conditions and therapeutic processes, thus improving treatment and accelerating medical research. Such discoveries are especially important for infectious diseases such as influenza or Ebola that can spread rapidly and for complex diseases with fast progression such as sepsis that are insufficiently understood [9]. An emphasis has recently been placed on the effective mining of those big EHR databases in order to obtain actionable insights for improving healthcare, a concept often termed "data driven healthcare" [10, 11]. However, mining such data comes with challenges as it is often sparse, heterogeneous, noisy and biased due to different hospital and insurance company polices and non-standardized physician practices [12].

Large-scale efforts for generating and sharing phenotypes were established recently [13, 14]. The initial result of these initiatives is that many phenotypes are now shared via Electronic Medical Records and Genomics (eMERGE) Network [15] or the Observational Medical Outcomes Partnership (OMOP) [16]. However, many of EHR-derived phenotypes are based on supervised, rule-based or heuristic-based approaches and often require a consensus of medical experts, thus limiting their scalability [12]. Necessary human annotations require substantial time, effort, and expert knowledge to develop, and these limitations further complicate phenotyping approaches [3]. A potential method of mitigating this issue is using active learning approaches to compensate for the lack of labeled samples [17, 18]. However, this approach falls short when a large number of labels are necessary to model noisy EHR data. Nevertheless, the state-of-the-art is far from being optimal, as the labeling process can be tedious, and models require a large number of labels to achieve satisfactory performance on noisy EHR data [12]. To create

a scalable phenotyping environment applicable to large databases, the phenotyping process needs to minimize human supervision and should be more automated [12].

Recently, promising approaches, known as *computational or electronic phenotyping*, have been proposed for data-driven phenotyping [19]. Computational phenotyping refers to the process of mapping raw patient EHRs into meaningful medical concepts, which can be seen as a feature selection process [20, 21]. It is worth noting that this is typically done on tens of thousands of EHRs although there are tens of millions available [22]. Data-driven healthcare approaches differ in their general objectives and data used for their research. However, the main contribution of such a body of work can be seen as learning useful representations of human interactome [1], whether that be a phenotype network of human diseases [23, 24, 4], network of human genes and proteins [25, 26], temporal networks of hospital records [8, 27, 20], or general tensor representations for discovering latent dimensions of the concepts involved [6, 7, 3, 28, 19]. Most of these approaches can be unified by their aim to exploit available EHR data [1] in order to develop representations of medical concepts for their further utilization in precision medicine by improving the understanding of disease etiology.

Causes of health and wellness span multiple body systems and physiologic processes, thus the complexity of the phenotyping process is increased. This creates a nonlinear relationship among observed measurements, making the process of learning robust representations of human physiology challenging [19]. The discovery of disease types can benefit both the practice and science of medicine [29]. For physicians, having defined disease types of good quality can decrease uncertainty in diagnosing and monitoring patients' wellness resulting in improved treatment decisions. It can also aid in prognosis of, i.e. treatment outcome or expected cost of care [30]. For researchers in medical science, it can provide a novel lens allowing for more focused analysis. Furthermore, it is in the interest of many researchers to discover segments of diseases in order to better understand more homogeneous subsets [31, 32, 33]. Previously studied disease segmentation approaches often consisted of observing metabolic, genetic or proteogenomic interactions, thus differing from the purely EHR-based approach proposed in this study. Our

[1]Human interactome is defined as all interactions(connections) of diseases, genes, and proteins discovered on humans.

goal is to automatically detect such segments of diseases from large EHR databases by exploiting disease comorbidity information contained in patient discharge records.

To provide evidence of benefits from using the proposed disease multi modal embedding approach, we conducted a case study discovering segments for all sepsis related diagnoses. Sepsis is a potentially life-threatening complication of pathogen infection that triggers the systemic inflammatory response [34, 9]. Such systemic and progressive inflammation can lead to multiple organ dysfunction syndrome and even death [35]. It can occur due to many reasons (i.e. infection from bacteria, fungi, viruses, or other organisms on different organs, etc.) and it has a wide range of symptoms. Hence, sepsis is not a yet fully understood condition while treatments are still far from optimal; it is often diagnosed too late, which can result in a mortality rate as high as 30–50% in the case of septic shock [36, 34].

It is a disease that afflicts a large population [37] and was the largest cause of death in the state of California from 2003 to 2011 (Figure 1). Furthermore, sepsis is recognized as one of the main causes of in-hospital deaths in the United States [38], with more than 750,000 cases annually [39], and it contributes to 1 in every 2 to 3 deaths [40]. In addition to overwhelming presence of the sepsis, hospital costs of over $20 billion in 2011 in the United States [41] provide a huge motivation for research in fields of understanding, diagnosing and treating such condition, as the incidence of sepsis is rising [42]. Therefore, complicated coding techniques are applied by the physicians to discriminate between different sepsis cases while documenting patients' discharge records [43]. In this study, we aim to exploit such information recorded in a large EHR database in order to automatically build multi modal representations of sepsis diagnoses with the purpose of proposing a system for improving sepsis diagnostics and potentially aiding in early prediction of outcomes.

The proposed novel, multi modal neural embedding model [44, 45] is adapted for use in medical records for disease embeddings [30], following from the major success of such models in the field of Natural Language Processing (NLP) and other fields [46, 47, 48]. Unsupervised neural embeddings have shown promising disease modeling capabilities from EHR data [30], outperforming representatives of other state-of-the-art approaches on predicting hospital quality indicators such as length of stay, total charges and mortality. The goal of these models is to learn low dimensional distributed representations of diseases by utilizing context from inpatient diagnoses and learn

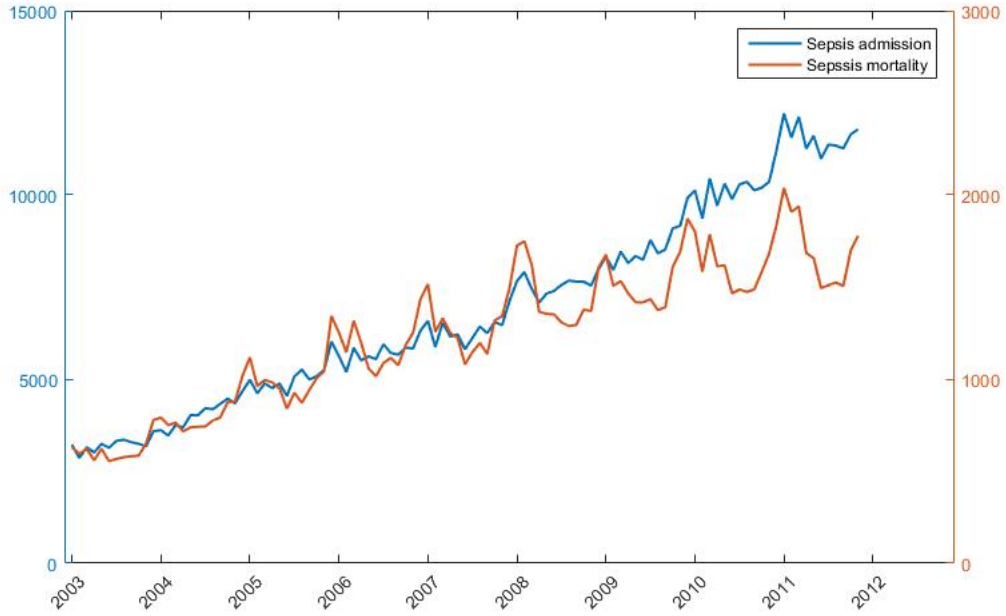


Figure 1: Admission (blue) and mortality (red) trends of sepsis diagnoses in California for the period 2003-2009

multiple type-specific embeddings for diseases of interest that would differ in the embedded space according to differences in contexts. The models for such tasks are described in Section 3. Such embeddings were shown to be able to capture disease-disease and disease-procedure relations, while also being very useful in further analyses in preventative and responsive medicine. This study further improves representational power of neural embeddings for learning distributed disease representations by allowing them to capture disease heterogeneities and automatically discover disease types. As discussed earlier, this is of great importance for highly heterogeneous diseases such as sepsis. Disease embedding approaches are described in Section 3.2, while novel type-specific approaches are described in Sections 3.2.2 and 3.2.3. Their benefits are evaluated and discussed in detail in Section 4 followed by conclusions in Section 5.

2. Large Electronic Health Records Database from California

The rapid growth in the development of healthcare information systems has led to an increased interest in utilizing the patient Electronic Health Records (EHR) in attempts to better understand human diseasome.

2.1. Healthcare Cost and Utilization Project (HCUP) data

For the purpose of this study, the State Inpatient Database (SID)[2], an archive that stores the inpatient discharge abstracts from a number of data organizations, is explored. 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, 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, is utilized. SID data provides discharge records for each inpatient, which may contain up to 25 diagnosis codes in an International Classification of Diseases coding schema that were applied during this particular admission of the patient. This coding schema[3] originates from the 9^{th} revision of the International Classification of Diseases (ICD9), a hierarchical coding scheme that is a part of standard diagnostic tools for epidemiology, health management, and clinical practice. Additionally, the SID database contains demographic information about each inpatient (e.g., age, birth year, sex, and race), as well as detailed 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 the ICD9 schema).

2.2. Sepsis inpatient discharge records dataset

We sample only discharge records containing one of the sepsis related codes from the SID CA database. Among the conditions considered we have included Systemic inflammatory response syndrome (SIRS), sepsis, and septicemia (names and ICD-9 codes given in Table 1). SIRS is defined as a clinical response to an insult, infection, or trauma that includes a systemic

[2]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

[3]`http://www.who.int/classifications/icd/en/`, accessed May 2016

Table 1: ICD-9 codes related to septic inpatients

Diagnosis code	Diagnosis name
995.90	Systemic inflammatory response syndrome, unspecified
995.91	Sepsis
995.92	Severe sepsis
995.93	Systemic inflammatory response syndrome due to noninfectious process without acute organ dysfunction
995.94	Systemic inflammatory response syndrome due to noninfectious process with acute organ dysfunction
785.52	Septic shock
038.9	Unspecified septicemia

inflammation as well as elevated or reduced temperature, rapid heart rate, rapid respiration, and elevated white blood cell count. Sepsis is additionally defined as SIRS due to infection without organ dysfunction, while severe sepsis is defined as SIRS due to infection with organ dysfunction. Please note that terms *septicemia* and *sepsis* are often used interchangeably, but are not considered synonyms in the ICD-9 coding. Septic shock is defined as a systematic disease associated with the presence of pathogenic microorganisms within the blood stream only. The selected sepsis targeted subset of the entire SID CA database constitutes $1,127,114$ discharge records, comprising 3.14% of total discharge records over the state of California from 2003 to 2011.

The process of coding sepsis in the EHR databases is tedious work, even under the most obvious circumstances, and requires proper application of the AHA Coding Clinic guidelines [49] and the Official Guidelines for Coding and Reporting for inpatient care [50], as well as well documented physician notes [43]. SIRS can be diagnosed with fairly easily, as there are strict physiological parameters that need to be satisfied. The EHR data records are represented by at least two codes, one for the underlying cause of infection (i.e., 038.xx, ...) and another for the sepsis subcategory (995.9x). Severe sepsis requires a minimum of three codes: a code for systemic infection (i.e., 038.xx, ...), the code 995.92 and the code for the associated organ failure. Septic shock is defined as severe sepsis with circulatory system failure, and in coding it only differs from severe sepsis in the second code where 995.92 is changed to 785.52. Finally, unspecified septicemia, code 038.9, is used when there is not

enough information in the doctors' notes and other diagnoses do not show a clear sign of the state of the patient's inflammation [43]. As can be seen from

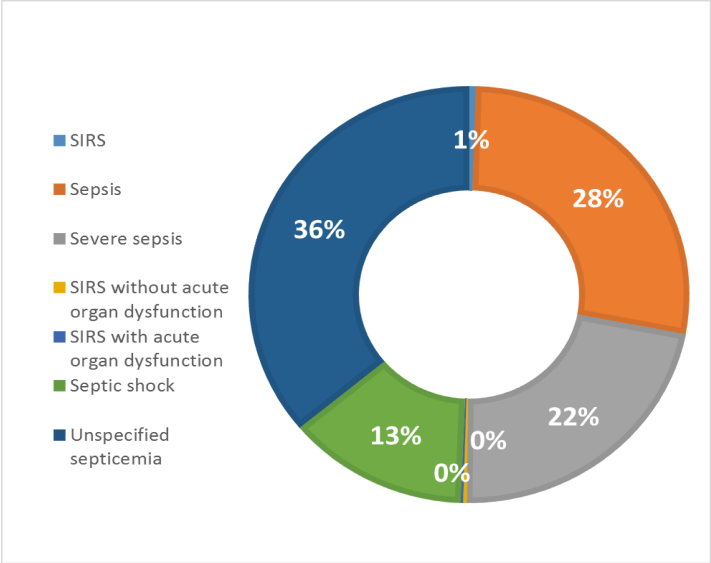


Figure 2: Prevalence of sepsis related diagnoses in SID California database

Figure 2, the SIRS conditions are the least prevalent, including virtually no cases where codes 995.90, 995.93 and 995.94 were used. On the other hand, the difficulty in properly diagnosing septic patients as described above is manifested, with the most dominant diagnosis being unspecified septicemia (0389) which was registered in around 36% of patient that were septic.

The discharge record containing a sepsis-related diagnosis is expected to have more than 2 diagnoses related to sepsis. Moreover, in the selected subset of the SID CA database, 16 diagnoses are observed on average per inpatient case. Thus, the context of one's inpatient stay includes other conditions observed in the record, which may provide additional insight in analyzing septic patient cases.

3. Methodology

We propose a new approach for the task of EHR phenotyping, motivated by the recent success of distributed language models [47, 51]. In NLP, distributed models were able to learn word representations in a low-dimensional

continuous vector space using a surrounding context of the word in a sentence, where in the resulting embedding space, semantically similar words are close to each other [47]. Previously, in the medical domain, such approaches have been applied to help understand physician notes or medical texts [52], while our goal is to apply them directly on the structured medical records (as described in Section 2.2), in order to learn meaningful low-rank disease representations. The advantage of such an approach is that diseases do not have to co-occur within same discharge record for the model to learn their connection, rather, their surrounding diseases, or disease context, has to be similar. Disease context is, as discussed before, governed by the proposed guidelines, and as such has a certain 'grammar' of diseases, which distributed language models can potentially exploit. However, there are aggravating factors when dealing with EHR records, i.e., inpatient diagnoses records vary in terms of both, type and physical system location, which increases heterogeneity of the data, but allows for discoveries of novel and interesting medical concepts. Such an approach would allow identifying similar diseases by trivial K-nearest-neighbor search in the new embedding space. Finding the nearest neighbor disease of a query disease will be referred to as *phenotyping* in this study, as neighboring diseases in the embedded space should have fairly similar traits. One shortcoming of such an approach is that each disease will be assigned a single vector, thus ignoring the heterogeneity present in the discharge records and resulting in a representation of lower quality.

In this paper, these issues are addressed by the applications of two state-of-the-art distributed language models [47] for learning disease representation, followed by two extensions aiming to learn multiple types [53] for selected diseases. We show that novel type-specific approaches are capable of learning more meaningful phenotypes, as well as aiding in patient mortality prediction.

3.1. Problem definition

We are given a set $\mathcal{P}$ of patient discharge records, where a patient's discharge record $p_i = (d_{i1}, \ldots, d_{iM_i}) \in \mathcal{P}$ is defined as a sequence of M_i diagnosed diseases $d_i \in \mathcal{D}$ at the end of the hospital stay. The objective is to find the D-dimensional real-valued representation $\mathbf{v}_d \in \mathcal{R}^D$ of each disease d such that diseases with similar phenotypes have similar representation.

3.2. Low-dimensional disease embeddings

Background. Neural language models take advantage of word order, and state the same assumption of n-gram language models that words closer 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, which act as the context. More formally, given a word sequence $(w_1, w_2, \ldots, w_T)$ in a training data, the objective of the model is to maximize the average log-likelihood,

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

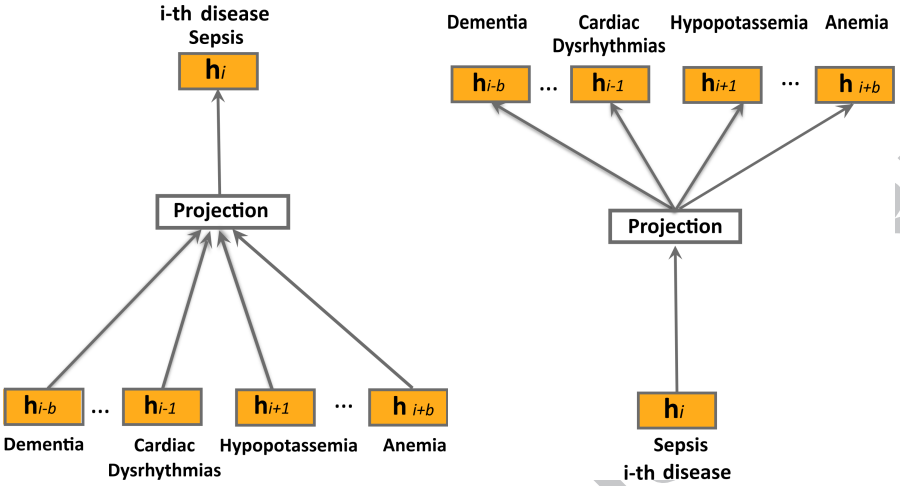
where w_t is the t-th word, and $w_{t-n+1} : w_{t-1}$ is a sequence of successive preceding words $(w_{t-n+1}, \ldots, w_{t-1})$ that act as the context to the word w_t. The probability distribution $\Pr(w_t | w_{t-n+1} : w_{t-1})$ is typically approximated using a neural network [54] trained to predict a word w_t by projecting the concatenation of vectors for context words $(w_{t-n+1}, \ldots, w_{t-1})$ into a latent representation with multiple non-linear hidden layers and the output softmax layer [54]. More recently, novel approaches have shown great improvements in representational power and training speed compared to the traditional neural embedding models [46]. Their representatives are discussed below.

3.2.1. Disease2vec disease representation

The method learns representations of diseases in a low-dimensional space using each patient discharge record as a "sentence" and the diseases within as "words", to borrow the terminology from the Natural Language Processing (NLP) domain. The diseases in each record are ordered by their importance with principal diseases coded at the beginning of the record. The disease2vec model has two architectures, differing in the independence assumption in the observed context.

CBOW disease2vec representation. In a continuous bag of words (CBOW) disease2vec approach disease representations are learned by maximizing the objective function $\mathcal{L}$ over the entire set $\mathcal{P}$ of records, as

$$\mathcal{L} = \sum_{p \in \mathcal{P}} \sum_{d_m \in p} \log \Pr(d_m | d_{m-b}, d_{m-1}, \ldots, d_{m+1}, d_{m+b}). \tag{2}$$



(a) CBOW disease2vec (b) SkipGram disease2vec

Figure 3: Graphical representations of disease2vec models

Probability $\Pr(d_m|d_{m-b} : d_{m+b})$ of observing a center disease d_m given its disease context $d_{m-b} : d_{m+b}$ is defined using the soft-max function,

$$\Pr(d_m|d_{m-b} : d_{m+b}) = \frac{\exp(\overline{\mathbf{v}}^\top \mathbf{v}'_{d_m})}{\sum_{d=1}^{D} \exp(\overline{\mathbf{v}}^\top \mathbf{v}'_d)}, \tag{3}$$

where $\mathbf{v}_d$ and $\mathbf{v}'_d$ are the input and output vector representations of D-dimensional disease d, and $2b$ is the length of the context for disease records. $\overline{\mathbf{v}}$ is obtained by averaging input vector representation of all diseases in observed context,

$$\overline{\mathbf{v}} = \frac{1}{T_c} \sum_{c=1}^{T_c} \mathbf{v}_{d_c} \tag{4}$$

As illustrated in Figure 3a and equation 3, CBOW disease2vec representation uses surrounding $T_c = 2b$ diseases $d_{m-b} : d_{m+b}$ to predict central disease d_m for each disease d_m in the discharge record. Thus, diseases that often co-occur and diseases with similar contexts (i.e., with similar neighboring diseases) will have similar representations.

SkipGram disease2vec representation. In SkipGram-based representation, central disease d_m is used to predict b diseases that occur before and b diseases that occur after it in the discharge record, as illustrated in Figure 3b and

equation 6. The SkipGram model introduces an additional assumption that neighboring diseases are independent of each other. Disease representations are learned by maximizing the objective function $\mathcal{L}$ over the entire set $\mathcal{P}$ of records, as

$$\mathcal{L} = \sum_{p \in \mathcal{P}} \sum_{d_m \in p} \sum_{-b \leq i \leq b, i \neq 0} \log \Pr(d_{m+i} | d_m). \tag{5}$$

The probability $\Pr(d_{m+i}|d_m)$ of observing a "neighboring" disease d_{m+i} given disease d_m is defined using the soft-max function,

$$\Pr(d_{m+i}|d_m) = \frac{\exp(\mathbf{v}_{d_m}^\top \mathbf{v}'_{d_{m+i}})}{\sum_{d=1}^{D} \exp(\mathbf{v}_{d_m}^\top \mathbf{v}'_d)}, \tag{6}$$

where $\mathbf{v}_d$ and $\mathbf{v}'_d$ are the input and output vector representations of disease d with dimensionality D, and $2b$ defines the length of the context for disease records.

3.2.2. Multi-type disease2vec disease representation

A major limitation of previously described models is that they assume a single vector representation for each disease. Such a disease representation is aimed to capture global trends in the discharge records, but it will not be able to represent the heterogeneity of each disease appropriately. For example, sepsis is a heterogeneous disease triggered by pneumonia, abdominal infection, kidney infection, bloodstream infection or other causes, and manifested on multiple organs, with different severity. Multi-type representations for such a complex disease can result in a more appropriate low-dimensional representation.

The multi-prototype approach for vector space models, which uses multiple representations to capture different senses and usages of a word is successfully used in the field of NLP [53] and a related approach is also applied to neural language models [44]. Here we also extend disease2vec models to a model using multiple types, which we call *t-CBOW* and *t-SkipGram*. In particular, we represent each discharge record by a sum of vectors of diagnoses found in that record. This global context representation dataset of inpatient visits is then clustered using K-means algorithm [53, 55] to obtain types of patient records that contained sepsis as a diagnosis. Finally, each sepsis occurrence in the discharge data is re-labeled to its associated cluster. Due to known heterogeneity of the discharge records data, sepsis types are obtained by clustering inpatient visit representation rather than observed

disease contexts as in [44]. New vectors of sepsis types are initialized as its global vector, and updated on the dataset such that the original sepsis disease spans a larger portion of the embedded space (via its types) thus capturing novel, previously undiscovered relationships.

This approach works globally for the entire dataset, in the form of a pipeline. However, it is possible to make disease2vec automatically model multiple types for each disease, specifically SkipGram, by locally discriminating contexts of each disease using either the MaxOut method or the K-means model and then deciding on the type vector update [45]. Such an approach is described in the following section.

3.2.3. Multi-sense SkipGram disease representation

This model, based on *multi-sense SkipGram* (MSSG) [45] (Figure 4), is capable of learning multiple types for each disease by locally discriminating contexts of each disease by either the MaxOut method or the K-means model. It performs multi-modal learning by clustering the embeddings of context around each disease. For each disease, clusters are maintained, and once the cluster is predicted the disease context representation for a disease type is updated. The difference between this and a multi-type *disease2vec* approach is that local contexts are clustered to decide the type of the disease and that the entire process is performed jointly by predicting the sense of the disease using the current parameter estimates. In the MSSG model, a global vector $\mathbf{v_g}(d)$ is assigned to each disease $d \in D$ and each type of the disease has a separate embedding $\mathbf{v_s}(d,k)$ $(k = 1, 2, \ldots, K)$, as well as a context cluster with center $\mu(d,k)$ $(k = 1, 2, \ldots, K)$. Clustering is performed in the following manner. First, for each disease d, a context vector is obtained by $\mathbf{v_{context}}(c_d) = \frac{1}{T_m} \sum_{c=1}^{T_c} \mathbf{v_g}(d_c)$, where c_d is context of disease d, and T_c is the size of the context. For context representation global vectors $\mathbf{v_g}$ are used rather than type-specific vectors to avoid additional computational complexity. Context representation $\mathbf{v_{context}}(c_d)$ is then used to predict the type of the disease d. In previous work [45], two approaches are discussed. Type of the disease s_k can be determined either by the MaxOut method:

$$s_k = \underbrace{argmax}_{k=1,2,\ldots,K} (\mathbf{v_s}(d,k)^\top \mathbf{v_{context}}(c_d)), \tag{7}$$

or by K-mean clustering:

$$s_k = \underbrace{argmax}_{k=1,2,\ldots,K} sim(\mu(d,k), \mathbf{v_{context}}(c_d)). \tag{8}$$

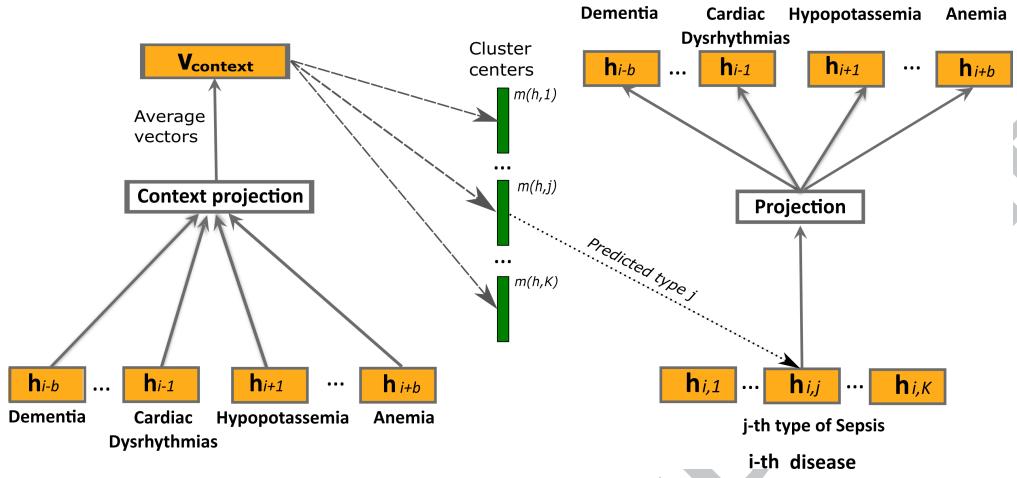


Figure 4: Graphical representations of the disease2vec MSSG model

Here the cluster center $\mu(d, k)$ is the average of all the context representation observed that belong to that cluster. For *sim*, cosine similarity is used in our experiments.

Finally, the objective function is obtained as in the SkipGram model (Eq. 5), with addition that the softmax function (Eq. 6) is conditioned on the cluster in which disease d belongs.

4. Experimental evaluations

In this section we describe experimental setups and the results obtained from such experiments. Mortality prediction results on sepsis-diagnosed patients using both type-specific and global embedding models are shown and an analysis of discovered types of sepsis related diagnoses is conducted.

All models were trained on $1,127,114$ sepsis diagnosed discharge records using a machine with 32GB of RAM memory and 4 cores. Diseases were mapped into $D = 200$ dimensional space. The value of parameter D was decided based on model complexity, and resulting model performance, where larger values marginally improved accuracy for mortality risk prediction of all models while making discovered types more overlapped and thus more difficult to interpret, while smaller values had worsen the accuracy of all models significantly. The context parameter b was varied in a set {2,4,16}, where 2 and 4 are determined with respect to coding patterns described in Section 2.2, and 16 was chosen to observe larger heterogeneous context as 16

was the average number of diagnoses in the dataset. We used 25 negative samples in each vector update for negative sampling, following a previously proposed approach for efficient learning [47]. The number of types K is considered in the range 1 to 15, where 15 is the number of reported different underlying infections causing sepsis according to potential causes listed in ICD-9 coding for 038.xx diagnoses[4]. The results reported in this section are obtained for $K = 5$ types based on the accuracy in mortality prediction.

4.1. Mortality prediction

In this section we evaluate the representational power of the discovered disease types. Feature vectors are learned in the embedded space for each disease and can be used for predictive tasks as such. Specifically, we used discovered sepsis types to predict patient survival probability, taking into consideration learned representations of diagnosed conditions and compared benefits of a type-specific approach versus predicting mortality based on global features of sepsis. The hypothesis evaluated in this experiment was that the multi-type sepsis vectors carry more information about mortality (some causes/effects can be more fatal than others) than the ones learned via global embedding models. We compared embeddings learned by four models from the family of type-specific embeddings (*t-CBOW*, *t-SkipGram*, *MSSG MaxOut* and *MSSG K-means*) to two global embedding models (learned by *CBOW* and *SkipGram*).

Features learned by those 6 embedding models were the input to the Logistic Regression algorithm used for mortality prediction (similar results were obtained by running SVM and neural network based classifiers). The model is trained on different subsets using 10% to 90% of data obtained as a balanced random sample and 10-fold validation for each sample size to remove any sampling bias. Learned models were then evaluated on the remaining EHR data. The results show stable performance (low variance of the obtained results from 10-fold validations) of both accuracy and F1 measure, as well as its components sensitivity and specificity, with respect to the entire range of training data sizes. The mentioned metrics are defined in terms of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) of binary classification results. Accuracy is computed

[4]http://www.icd9data.com/2013/Volume1/001-139/030-041/038/, acc. May 2016

Table 2: Accuracy, F-1 measure, Sensitivity and Specificity aggregated over 90 experiments for Logistic Regression model used on features learned by 6 embedding models: 4 type-specific and 2 global, for three values of hyperparameter b. The best results are bolded.

	Accuracy			F1-measure		
	b=2	b=4	b=16	b=2	b=4	b=16
t-CBOW	**77.2%**	**77.2%**	**76.1%**	**77.7%**	**77.9%**	**76.0%**
t-SkipGram	76.6%	76.9%	74.7%	77.1%	77.5%	74.3%
MSSG K-kmeans	67.9%	68.0%	69.3%	69.0%	69.2%	71.0%
MSSG MaxOut	67.9%	68.0%	69.1%	69.0%	69.2%	69.9%
CBOW	56.0%	56.1%	67.1%	58.3%	59.6%	67.8%
SkipGram	55.0%	55.4%	67.1%	57.0%	57.6%	69.8%
	Sensitivity			Specificity		
	b=2	b=4	b=16	b=2	b=4	b=16
t-CBOW	**79.4%**	**80.1%**	**78.2%**	**75.1%**	**74.4%**	**73.5%**
t-SkipGram	79.1%	79.5%	76.2%	74.0%	**74.4%**	71.6%
MSSG K-kmeans	71.3%	72.0%	73.9%	64.5%	63.9%	64.9%
MSSG MaxOut	71.3%	72.0%	72.7%	64.5%	63.9%	64.7%
CBOW	59.3%	59.4%	70.5%	52.4%	51.9%	62.0%
SkipGram	58.2%	58.2%	71.6%	51.5%	51.2%	62.0%

as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{9}$$

$$\text{F1 measure} = \frac{2TP}{2TP + FN + FP}, \tag{10}$$

and its components sensitivity and specificity as

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \qquad \text{and} \qquad \text{Specificity} = \frac{TN}{TN + FP}. \tag{11}$$

Therefore, Table 2 aggregates the evaluation results (accuracy, F-1 measure, sensitivity and specificity) of 6 models on 90 experiments from 10 validations on 9 different training-test sizes. Additionally, the influence of the context window size defined by parameter b on the overall predictive accuracy is examined, where b is chosen from a set {2,4,16}.

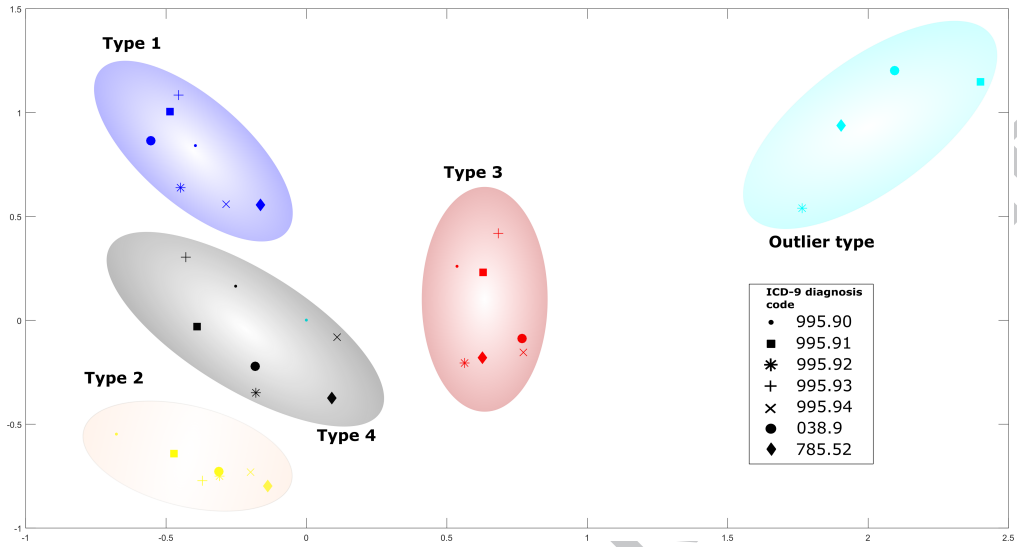


Figure 5: In the embedded space (here displayed 2D reduced space) each of seven sepsis-related ICD-9 diagnoses is partitioned to five types marked in different colors.

All type-specific embedding based sepsis mortality models were more accurate than the global models, where the best performing algorithm was the proposed *t-CBOW* model described in Section 3.2.2. The results were stable with respect to the parameter b when shorter context is used ($b = 2$ or 4), while larger context ($b = 15$) resulted in slightly decreased accuracy for *t-CBOW* and *t-SkipGram* models. The proposed multi-type approaches were more robust to the context window size. Larger context allows partial learning of broader concepts in a single type, which is why *CBOW* and *Skip-Gram* were more accurate with larger window size b. As the accuracy was fairly stable for multi-type models, but highly increased for global models for $b = 16$ (when compared to lower values of this parameter), phenotyping results are shown for the embedding with this parameter in Table 3 – 7. Finally, this experiment provides evidence that discovered sepsis segmentation is clinically relevant.

4.2. Disease types and their phenotypes

In this section we discuss sepsis disease phenotypes discovered by both global and type-specific embeddings. Here, phenotypes are defined as query disease's nearest neighbors in the embedded space. For the type-specific models, we discuss phenotypes found by the *t-CBOW* model, as it was the

best competing model for mortality prediction, and the *CBOW* model for the global embeddings, as there was no significant difference from the *SkipGram* model on the same task. Parameter b is fixed to be 16 in this section for both models, as results on mortality prediction gave the most balanced accuracy performance over all models examined.

We show the 5 embedded disease-types for each of the 7 sepsis diagnoses in Figure 5. The five discovered disease types emancipated cluster-like groupings in the new embedded space. Furthermore, we observe that all diagnoses in the same type share similar phenotype properties. Concrete findings will be discussed in more details below. Another interesting finding is the *outlier type* (upper right corner of Figure 5). The observed type we refer to as the *outlier type* has low prevalence, with less than a thousand cases in our dataset (or less than 1% of the discharge records). As such, it will be removed from further discussion, even though it forms the purest phenotype cluster, given that the main focus of this study are prevalent phenotypes, while analysis of outliers will be left for future work.

Additionally, we have observed that SIRS conditions also have much lower prevalence than sepsis (less than 1% as shown in Figure 2). Thus, the analysis reported in this paper is focused on segmenting four types for each of four sepsis diagnoses as shown in Figure 6. Analysis of disease type records shows that each of the remaining four discovered types of diseases occur in at least 10% of discharge records (Figure 6), and therefore, are well represented in the dataset.

In Table 3, we list five nearest non-sepsis diseases to the sepsis diagnosis 995.91 in the embedded space representation learned by the *global* embedding model. *Sepsis global* phenotype shows heterogeneous properties where most similar diagnoses are infections on different parts of organs, but also abortion or fracture related diagnosis, which are known as possible sepsis causes or effects [56, 57]. For each of the five most similar diseases in the *Sepsis global* phenotype, their rankings by the type-specific models are provided in columns *type 1- type 4* (for each of the types). Globally relevant diseases are not particularly close in the embedded space for most homogeneous types of sepsis, which can also be concluded from Figure 5. Note that ICD-9 codes provide disease coding on a very fine scale. For instance, the same condition can be present in multiple locations of an organ, and there are multiple codes for such a disease. Fine scale disease coding is the cause of low ranks of globally relevant diseases in type-specific phenotypes as other similar but type-specific conditions are ranked higher, demonstrating

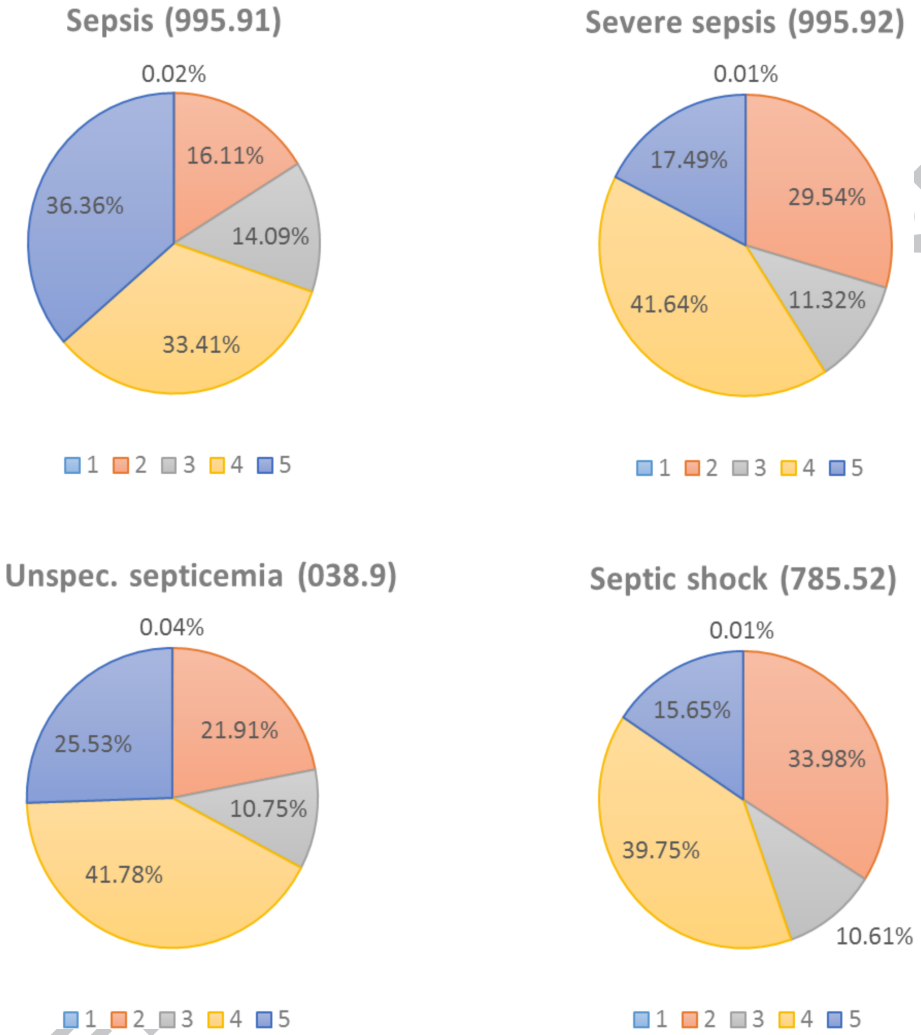


Figure 6: Fraction of disease *types 1-4* from Figure 5 in four sepsis-related ICD-9 diagnoses code groups (995.91, 995.21, 038.9 and 785.52)

the limitations of the global embeddings. Additionally, for each of the types, there is one (bolded) globally relevant disease that is higher ranked in that type than other diseases in the same type. For example, in sepsis disease *type 4*, which is represented in majority by the urinary related phenotype, the closest condition is the *urinary tract infection*, while the other conditions are at least three times lower ranked. In case of patients with sepsis and a urinary tract infection, physicians often use the term *urosepsis* [43] due to

Table 3: **Sepsis** (995.91 code) vector and its 5 nearest neighbors in the embedded global disease space vs 4 type-specific embedded disease space. For each of the types, there is one (bolded) globally relevant disease that is higher ranked in that type than other diseases in the same type.

Rank for (995.91) :	*global*	*type 1*	*type 2*	*type 3*	*type 4*
Closed fracture of lower and of forearm unspec.	1	5990	9118	604	4965
Acute upper respiratory infections of unspec. site	2	**637**	6746	1993	881
Urinary tract infection site not specified	3	5845	8185	**91**	**274**
Leukocytosis uspec.	4	4643	4649	1408	761
Legaly induced abortion with other spec. complications	5	9230	**350**	8770	2797

its prevalence, giving evidence of interpretability of obtained phenotypes.

Four discovered types of Sepsis (diagnosis 995.91) will be referred to as *Sepsis type 1* to *Sepsis type 4* and will be labeled as 995.91_1 to 995.91_4. For each of the types, five most similar diseases in the embedded space representation were listed in Table 4 as obtained based on the *t-CBOW* model for 995.91. The global rank for each of the listed diseases is also shown as obtained by the global *CBOW* embedding model.

As compared to the global phenotypes, sepsis type-specific phenotypes are more homogeneous. For example, sepsis in pregnant and postpartum women can develop as the result of many complications, such as miscarriages (spontaneous abortions) or induced abortions, prolonged or obstructed labor, ruptured membranes, cesarean sections, infection following a vaginal delivery, etc. [58, 59]. Some of these causes related to delivery (i.e., prolonged labor or ruptured membranes) are found in *Sepsis outlier type*, while causes related to abortions are found in *Sepsis type 1*. Both types have these causes ranked highly (they are close to sepsis vector in the embedded space) by the *t-CBOW* model, whereas the global ranking model assigns low ranks (e.g 1963, 2919, 8583), which shows the better representational ability of the proposed model over the global embedding.

Sepsis can cause a lot of damage in a person that is affected by this

Table 4: Four segments of **Sepsis** (995.91 code) and their 5 nearest neighbors in the embedded disease space

5 most related diagnoses in the embedded space to Sepsis (995.91)	Rank by types	Global rank
Sepsis type 1 ($995.91_{\mathbf{1}}$) [36.36%]		
Transient arthropathy shoulder region	1	1104
Tension headache	2	525
Unspec. abortion	3	786
Unspec. abortion complicated by damage to pelvic organs	4	2919
Paratyphoid fever A	5	566
Sepsis type 2 ($995.91_{\mathbf{2}}$) [33.41%]		
Variations in hair color	1	1410
Other persistent mental disorders	2	933
Paralysis agitans	3	913
Senile dementia uncomplicated	4	1525
Unspec. senile psychotic condition	5	4091
Sepsis type 3 ($995.91_{\mathbf{3}}$) [16.11%]		
Open skull fracture with cerebral laceration and contusion	1	3684
Nervous system complications from surg. implanted device	2	7486
Inclusion conjunctivitis	3	5704
Malignant neoplasm of other and unspec. testis	4	8215
Anemia of mother unspecified	5	8164
Sepsis type 4 ($995.91_{\mathbf{4}}$) [14.09%]		
Hypertensive chronic kidney disease (stage V)	1	7013
End stage renal disease	2	4992
Infection and inflammatory reaction due to oth. vascular device	3	1418
Complic. due to renal dialysis device implant and graft	4	7142
Hypertensive heart and chronic kidney disease	5	5254

disease and its treatment can also leave different consequences. The kidneys are often among the first organs to be affected by sepsis and published studies report that between 32% and 48% of acute kidney injury cases were caused by sepsis [60]. Therefore, it is not surprising that *Sepsis type 4* is very related to kidney diseases (not just for sepsis, but also for the other sepsis diseases shown in Tables 5 - 7).

Another category of sepsis consequences consists of mental and stress-related disorders, which are found in *Sepsis type 2*. It is reported that 17%

Table 5: Four segments of **Severe Sepsis** (995.92 code) and their 5 nearest neighbors in the embedded disease space

5 most related diagnoses in the embedded space to Severe Sepsis (995.92)	Rank in type	Global rank
Severe sepsis type 1 ($995.92_{\mathbf{1}}$) [17.49%]		
Hemicrania continua	1	1650
Chronic Eustachian salpingitis	2	1604
Other nongonococcal urethritis unspecified	3	3562
Other manifestations of yaws	4	9435
Acute pyelonephritis without lesion of renal medullary necrosis	5	466
Severe sepsis type 2 ($995.92_{\mathbf{2}}$) [41.64%]		
Chondrocalcinosis due to pyrophosphate crystals upper arm	1	9054
Meningitis in sarcoidosis	2	7550
Other persistent mental disorders due to conditions classified	3	933
Hyperosmolality and-or hypernatremia	4	6337
Paralysis agitans	5	913
Severe sepsis type 3 ($995.92_{\mathbf{3}}$) [29.54%]		
Burn involving 50-59 % of body surface w 3. degree burn 40-49%	1	8730
Letterer-siwe di. unspec. site extranodal and solid organ sites	2	8584
Pneumococcal peritonitis	3	9546
Defibrination syndrome	4	9352
Tuberculosis of intestines peritoneum and mes. glands tubercle bacilli	5	6584
Severe sepsis type 4 ($995.92_{\mathbf{4}}$) [11.32%]		
Hypertensive chronic kidney disease (V or end stage renal dis.)	1	7013
Nephrotic syndrome in diseases classified elsewhere	2	7728
End stage renal disease	3	4992
Diabetes with renal manifestations type II ...	4	5307
Other complications due to renal dialysis device implant and graft	5	7142

of elderly sepsis survivors developed dementia and around 40% experienced nervous system damage and could not walk without assistance in the years after [61]. It has also been reported that sepsis patients can develop large amounts of stress molecules [62], i.e. cortisol which is known to accumulate in human hair thus leading to color changes. Stress related conditions for sepsis survivors are becoming more evident as they reportedly experience stress disorders, including Post-Traumatic Stress Disorder (PTSD), as a result of prolonged treatments in Intensive Care Units (ICUs) [63]. Conditions

described above are highly ranked by *t-CBOW* in *Sepsis type 2*.

Sepsis type 3 covers diseases related to serious brain tissue injuries and nervous system complications from surgically implanted devices, which can both lead to septic inflammation [64] and reproductive system related causations of sepsis [65]. Since, *Sepsis type 1* covers a large fraction of impatient record cases (36.6%), it is expected that this phenotype is the most heterogeneous among all. Therefore, in addition to abortion cases, we observe other possible causes and effects of this disease.

Discovered phenotypes of global and type-specific embeddings of severe sepsis, septic shock, and septicemia diagnoses are presented in Tables 5, 6, 7, respectively. We observe that disease types show similar traits, as anticipated from Figure 5. The phenotypes discovered for the three diseases are consistent with the sepsis types: type 4 sepsis diseases are related to kidney and urinal tract problems, type 2 sepsis diseases are related to nervous system inflammations, while type 1 and type 3 sepsis diseases are related to external irritations such as burns, fractures and different inflammations. As expected, severe sepsis and septic shock phenotypes share 65% of the closest diseases, as they are considered the same condition, with septic shock being a severe sepsis with circulatory system failure.

5. Conclusions

Neural embedding models have shown great promise in many fields, but they have not been used yet in the field of electronic phenotyping. Hence, this paper studied low-dimensional models for disease type discovery from large EHR databases. Such low-dimensional embedding can be very useful not only for disease phenotyping but also for more accurate diagnostics. In this study, several approaches were proposed for addressing disease phenotyping challenges related to disease heterogeneity. As a case study, the proposed methodology is applied to phenotype characterization of sepsis, which is a highly heterogeneous disease and one of the main causes of death in the US hospitals. Conducted experiments provide evidence that the proposed approach can effectively discover informative phenotypes for sepsis. The discovered phenotypes for identified homogeneous groups were more relevant as compared to global vectors for the same diseases. Benefits were also evident for a mortality prediction task, where an increase in accuracy and prediction quality was observed when using multi-type disease embedding rather than single global embedding. In our experiments, we have compared

Table 6: Four segments of **Septic Shock** (785.52 code) and their 5 nearest neighbors in the embedded disease space

5 most related diagnoses in the embedded space to Septic Shock (785.52)	Rank in type	Global rank
Septic shock type 1 ($785.52_{\mathbf{1}}$) [15.65%]		
Chronic Eustachian salpingitis	1	1604
Other nongonococcal urethritis unspecified	2	3562
Cocaine dependence episodic	3	2459
Encounter for removal of intrauterine contraceptive device	4	7827
Inconclusive mammogram	5	3110
Septic shock type 2 ($785.52_{\mathbf{2}}$) [39.76%]		
Chondrocalcinosis due to pyrophosphate crystals upper arm	1	9054
Meningitis in sarcoidosis	2	7550
Hyperosmolality and-or hypernatremia	3	6337
Closed lateral dislocation of elbow	4	5474
Paralysis agitans	5	913
Septic shock type 3 ($785.52_{\mathbf{3}}$) [33.98%]		
Defibrination syndrome	1	9352
Pneumococcal peritonitis	2	9546
Letterer-siwe disease unspec. site extranodal and solid organ sites	3	8584
Burn involving 50-59 % of body surface w 3. degree burn 40-49%	4	8730
Acute and subacute necrosis of liver	5	9741
Septic shock type 4 ($785.52_{\mathbf{4}}$) [10.61%]		
Hypertensive chronic kidney disease (V or end stage renal dis.)	1	7013
End stage renal disease	2	4992
Other complications due to renal dialysis device implant and graft	3	7142
Nephrotic syndrome in diseases classified elsewhere	4	7728
Hypertensive heart and chronic kidney disease w. heart failure and chronic kidney disease stage V or end stage	5	8387

two approaches for disease type discovery, a global clustering approach and an automatic approach, where disease types are learned within the model itself. Although easier to use, an automatic approach failed to outperform global clustering (*t-models*). However, it was better than the original single vector approach. Discovering disease types has shown great promise as a future research direction in electronic phenotyping, and further efforts will be taken to further the understanding of the discovered disease types as

Table 7: Four segments of **Septicemia** (038.9 code) and their 5 nearest neighbors in the embedded disease space

5 most related diagnoses in the embedded space to Septicemia (038.9)	Rank in type	Global rank
Septicemia type 1 ($038.9_{\mathbf{1}}$) [25.53%]		
Basal cell carcinoma of scalp and skin of neck	1	2172
Inappropriate diet and eating habits	2	1703
Screening for other disorders of blood and blood-forming organs	3	9037
Impairment of auditory discrimination	4	778
Other arthropod infestation	5	451
Septicemia type 2 ($038.9_{\mathbf{2}}$) [41.79%]		
Loose body in joint other specified sites	1	1337
Other circadian rhythm sleep disorder	2	1901
Meningitis in sarcoidosis	3	7550
Other persistent mental disorders due to conditions classified elsewhere	4	933
Variations in hair color	5	1410
Septicemia type 3 ($038.9_{\mathbf{3}}$) [21.92%]		
Burn involving 50-59 % of body surface w 3. degree burn 40-49%	1	8730
Congenital anomalies of corneal size and shape	2	6577
Open fracture of mandible alveolar border of body	3	9357
Open skull fracture, cerebral laceration, contusion, loss of consciousness	4	3684
Subarachnoid hemorrhage, open intracranial wound, loss of consciousne	5	5619
Septicemia type 4 ($038.9_{\mathbf{4}}$) [10.75%]		
Hypertensive chronic kidney disease (V or end stage renal dis.)	1	7013
End stage renal disease	2	4992
Nephrotic syndrome in diseases classified elsewhere	3	7728
Hypertensive heart and chronic kidney disease w. heart failure and chronic kidney disease stage V or end stage	4	5254
Other ectopic pregnancy without intrauterine pregnancy	5	1951

well as to build effective models capable of jointly using existing medical knowledge and big data to discover disease embeddings of higher quality.

6. Acknowledgements

The authors gratefully acknowledge the support of the Defense Advanced Research Project Agency (DARPA) GRAPHS program under Air Force Research Laboratory (AFRL) prime contract no. FA9550-12-1-0406., National

Science Foundation BIGDATA grant 14476570 and Office of Naval Research Mathematics of Data Science Project N00014-15-1-2729. Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality provided data used in this study. The authors also thank Aleksandar Obradovic for proofreading and editing the language of the manuscript.

References

[1] J. C. Denny, Chapter 13: mining electronic health records in the genomics era, PLoS Comput. Biol. 8 (12) (2012) e1002823.

[2] P. B. Jensen, L. J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nat. Rev. Genet. 13 (6) (2012) 395–405.

[3] J. C. Ho, J. Ghosh, 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.

[4] K. Sun, J. P. Gonçalves, C. Larminie, N. Pržulj, Predicting disease associations via biological network analysis, BMC Bioinformatics 15 (1) (2014) 1.

[5] T. Xiang, D. Ray, T. Lohrenz, P. Dayan, P. R. Montague, Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought, PLoS Comput. Biol. 8 (2012) e1002841.

[6] J. Zhou, F. Wang, J. Hu, 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.

[7] J. C. Ho, J. Ghosh, S. R. Steinhubl, W. F. Stewart, J. C. Denny, B. A. Malin, J. Sun, Limestone: High-throughput candidate phenotype generation via tensor factorization, J. Biomed. Inform. 52 (2014) 199–211.

[8] D. Gligorijevic, J. Stojanovic, 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.

[9] I. Stojkovic, M. F. Ghalwash, X. H. Cao, Z. Obradovic, Effectiveness of Multiple Blood-Cleansing Interventions in Sepsis, Characterized in Rats, Sci. Rep. 6 (24719) (2016) 1–11.

[10] Data driven healthcare, Vol. 117(5):119, MIT Technology Review, 2014.

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

[12] G. Hripcsak, D. J. Albers, Next-generation phenotyping of electronic health records, J. Am. Med. Inform. Assoc. 20 (1) (2013) 117–121.

[13] K. M. Newton, P. L. Peissig, A. N. Kho, S. J. Bielinski, R. L. Berg, V. Choudhary, M. Basford, C. G. Chute, I. J. Kullo, R. Li, et al., Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the emerge network, J. Am. Med. Inform. Assoc. 20 (e1) (2013) e147–e154.

[14] A. N. Kho, J. A. Pacheco, P. L. Peissig, L. Rasmussen, K. M. Newton, N. Weston, P. K. Crane, J. Pathak, C. G. Chute, S. J. Bielinski, et al., Electronic medical records for genetic research: results of the emerge consortium, Sci. Transl. Med. 3 (79) (2011) 79re1–79re1.

[15] C. A. McCarty, R. L. Chisholm, C. G. Chute, et al., The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies, BMC Med. Genomics 4 (1) (2011) 13.

[16] J. M. Overhage, P. B. Ryan, C. G. Reich, A. G. Hartzema, P. E. Stang, Validation of a common data model for active safety surveillance research, J. Am. Med. Inform. Assoc. 19 (1) (2012) 54–60.

[17] Y. Chen, R. J. Carroll, E. R. M. Hinz, A. Shah, A. E. Eyler, J. C. Denny, H. Xu, Applying active learning to high-throughput phenotyping algorithms for electronic health records data, J. Am. Med. Inform. Assoc. 20 (e2) (2013) e253–e259.

[18] D. Dligach, T. A. Miller, G. K. Savova, Active learning for phenotyping tasks, in: Proceedings of the Workshop on NLP for Medicine and Biology associated with RANLP, Citeseer, 2013, pp. 1–8.

[19] Z. Che, D. Kale, W. Li, M. T. Bahadori, Y. Liu, Deep computational phenotyping, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 507–516.

[20] C. Liu, F. Wang, J. Hu, H. Xiong, Temporal phenotyping from longitudinal electronic health records: A graph based framework, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 705–714.

[21] J. Pathak, A. N. Kho, J. C. Denny, Electronic health records-driven phenotyping: challenges, recent advances, and perspectives, J. Am. Med. Inform. Assoc. 20 (e2) (2013) e206–e211.

[22] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals, Circ. 101 (23) (2000) e215–e220.

[23] C. Hidalgo, N. Blumm, A. Barabasi, N. Christakis, A dynamic network approach for the study of human phenotypes, PLoS Comput. Biol.

[24] D. Davis, N. Chawla, Exploring and exploiting disease interactions from multi-relational gene and phenotype networks, PLoS ONE.

[25] G. K.I., C. M.E., V. D., C. B., V. M., B. A.L., The human disease network, Proc. Natl. Acad. Sci. USA.

[26] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, A.-L. Barabási, Uncovering disease-disease relationships through the incomplete interactome, Science 347 (6224) (2015) 1257601.

[27] D. Gligorijevic, J. Stojanovic, Z. Obradovic, Uncertainty propagation in long-term structured regression on evolving networks, in: Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), 2016.

[28] P. Schulam, F. Wigley, S. Saria, Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery, in: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[29] S. Saria, A. Goldenberg, Subtyping: What it is and its role in precision medicine, IEEE Intell. Syst. 30 (4) (2015) 70–75.

[30] R. V. D. N. G. M. Stojanovic J., Gligorijevic Dj., O. Z., Modeling healthcare quality via compact representations of electronic health records, IEEE/ACM Trans. Comput. Biol. Bioinf.

[31] A. Bauer-Mehren, M. Bundschus, M. Rautschka, M. A. Mayer, F. Sanz, L. I. Furlong, Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases, PLoS ONE 6 (6) (2011) e20284.

[32] D. He, Z.-P. Liu, L. Chen, Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach, BMC Genomics 12 (1) (2011) 1.

[33] M. Kikuchi, S. Ogishima, T. Miyamoto, A. Miyashita, R. Kuwano, J. Nakaya, H. Tanaka, Identification of unstable network modules reveals disease modules associated with the progression of alzheimers disease, PloS one 8 (11) (2013) e76162.

[34] R. P. Dellinger, M. M. Levy, A. Rhodes, D. Annane, H. Gerlach, S. M. Opal, J. E. Sevransky, C. L. Sprung, I. S. Douglas, R. Jaeschke, et al., Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock, 2012, Intensive Care Med. 39 (2) (2013) 165–228.

[35] J. Russel, The current management of septic shock., Minerva Med. 99 (5) (2008) 431–458.

[36] S. W. Thiel, J. M. Rosini, W. Shannon, J. A. Doherty, S. T. Micek, M. H. Kollef, Early prediction of septic shock in hospitalized patients, J. Hosp. Med. 5 (1) (2010) 19–25.

[37] Anonymous, Focus on sepsis, Nat. Med. 18 (997).

[38] V. Liu, G. J. Escobar, J. D. Greene, J. Soule, A. Whippy, D. C. Angus, T. J. Iwashyna, Hospital deaths in patients with sepsis from 2 independent cohorts, JAMA 312 (1) (2014) 90–92.

[39] S. M. Zuev, S. F. Kingsmore, D. D. Gessler, Sepsis progression and outcome: a dynamical model, Theor. Biol. Med. Mod. 3 (1) (2006) 8.

[40] M. M. Levy, M. P. Fink, J. C. Marshall, E. Abraham, D. Angus, D. Cook, J. Cohen, S. M. Opal, J.-L. Vincent, G. Ramsay, et al., 2001 sccm/esicm/accp/ats/sis international sepsis definitions conference, Intensive Care Med. 29 (4) (2003) 530–538.

[41] C. M. Torio, R. M. Andrews, National inpatient hospital costs: the most expensive conditions by payer, 2011.

[42] G. S. Martin, Sepsis, severe sepsis and septic shock: changes in incidence, pathogens and outcomes, Expert Rev. Anti Infect. Ther. 10 (6) (2012) 701–706.

[43] L. A. Wiedemann, Coding sepsis and sirs, J. AHIMA 78 (4) (2007) 76–78.

[44] E. H. Huang, R. Socher, C. D. Manning, A. Y. Ng, Improving word representations via global context and multiple word prototypes, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, Association for Computational Linguistics, 2012, pp. 873–882.

[45] A. Neelakantan, J. Shankar, A. Passos, A. McCallum, Efficient non-parametric estimation of multiple embeddings per word in vector space, arXiv preprint arXiv:1504.06654.

[46] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.

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

[48] N. Djuric, H. Wu, V. Radosavljevic, M. Grbovic, N. Bhamidipati, Hierarchical neural language models for joint representation of streaming documents and their content, in: International World Wide Web Conference (WWW), 2015.

[49] A. H. Association, et al., Aha coding clinic for icd-9-cm, AHA, Chicago.

[50] 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.

[51] J. Turian, L. Ratinov, Y. Bengio, Word representations: a simple and general method for semi-supervised learning, in: Proceedings of the ACL, Association for Computational Linguistics, 2010, pp. 384–394.

[52] C. Wang, L. Cao, B. Zhou, Medical synonym extraction with concept space models, arXiv preprint arXiv:1506.00528.

[53] J. Reisinger, R. J. Mooney, Multi-prototype vector-space models of word meaning, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2010, pp. 109–117.

[54] Y. Bengio, R. Ducharme, P. Vincent, C. Janvin, A neural probabilistic language model, J. Mach. Learn. Res. 3 (2003) 1137–1155.

[55] I. S. Dhillon, D. S. Modha, Concept decompositions for large sparse text data using clustering, Mach. Learn. 42 (1-2) (2001) 143–175.

[56] M. Kylänpää-Bäck, et al., Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis, Crit. Care Med. 20 (1992) 864–874.

[57] H. B. Nguyen, E. P. Rivers, F. M. Abrahamian, G. J. Moran, E. Abraham, S. Trzeciak, D. T. Huang, T. Osborn, D. Stevens, D. A. Talan, et al., Severe sepsis and septic shock: review of the literature and emergency department management guidelines, Ann. Emerg. Med. 48 (1) (2006) 54–e1.

[58] E. R. Fernandez-Perez, S. Salman, S. Pendem, J. C. Farmer, Sepsis during pregnancy, Crit. Care Med. 33 (10) (2005) S286–S293.

[59] M. E. Bauer, B. T. Bateman, S. T. Bauer, A. M. Shanks, J. M. Mhyre, Maternal sepsis mortality and morbidity during hospitalization for delivery: temporal trends and independent associations for severe sepsis, Anesth. Analg. 117 (4) (2013) 944–950.

[60] S. S. Waikar, K. D. Liu, G. M. Chertow, Diagnosis, epidemiology and outcomes of acute kidney injury, Clin. J. Am. Soc. Nephrol. 3 (3) (2008) 844–861.

[61] T. Iwashyna, E. Ely, D. Smith, K. Langa, Long-term cognitive impairment and functional disability among survivors of severe sepsis, JAMA 304 (16) (2010) 1787–1794.

[62] M. Adib-Conquy, J.-M. Cavaillon, Stress molecules in sepsis and systemic inflammatory response syndrome, FEBS Lett. 581 (19) (2007) 3723–3733.

[63] G.-B. Wintermann, F. M. Brunkhorst, K. Petrowski, B. Strauss, F. Oehmichen, M. Pohl, J. Rosendahl, Stress disorders following prolonged critical illness in survivors of severe sepsis, Crit. Care Med. 43 (6) (2015) 1213–1222.

[64] R. Okapa, S. Rak, J. Wenda, W. Marczyński, P. Walczak, J. Macias, [septic complications after multilocal fractures and multiple traumatic injury]., Chirurgia narzadow ruchu i ortopedia polska 76 (4) (2010) 214–218.

[65] P. Sinha, M. Otify, Genital tract sepsis: early diagnosis, management and prevention, Obstet. Gynecol. 14 (2) (2012) 106–114.