Learning Low-Dimensional Representations of Medical Concepts

Youngduck Choi[1], **Chill Yi-I Chiu MS**[1], **David Sontag PhD**[1]
[1]**New York University, New York, NY**

Abstract

We show how to learn low-dimensional representations (embeddings) of a wide range of concepts in medicine, including diseases (e.g., ICD9 codes), medications, procedures, and laboratory tests. We expect that these embeddings will be useful across medical informatics for tasks such as cohort selection and patient summarization. These embeddings are learned using a technique called neural language modeling from the natural language processing community. However, rather than learning the embeddings solely from text, we show how to learn the embeddings from claims data, which is widely available both to providers and to payers. We also show that with a simple algorithmic adjustment, it is possible to learn medical concept embeddings in a privacy preserving manner from co-occurrence counts derived from clinical narratives. Finally, we establish a methodological framework, arising from standard medical ontologies such as UMLS, NDF-RT, and CCS, to further investigate the embeddings and precisely characterize their quantitative properties.

1 Introduction

Ontologies of medical concepts such as the Unified Medical Language System (UMLS) or the International Classification of Diseases (ICD-9, ICD-10) are widely used for epidemiology, health management, and clinical purposes. For example, ICD9 codes are widely used by physicians and coders to record patient symptoms and diagnoses for the purpose of submitting claims to payers for reimbursement and for clinical research. The UMLS Metathesaurus is widely used as a means of standardizing the documentation of clinical concepts found in a patient's electronic health record. LOINC (Logical Observation Identifiers Names and Codes) is one of the standards used in the U.S. for identifying medical laboratory observations. NDC (National Drug Code) is a universal product identifier for human drugs in the US.

There are over 3,100,000 concepts in the UMLS, close to 70,000 ICD-10-CM diagnosis codes, 70,000 ICD-10-PCS procedure codes, 70,000 LOINC codes, and over 360,000 NDC codes. Because of the large number of concepts, hierarchical structure found in these and related ontologies help make the tasks of identifying specific concepts and finding related concepts significantly easier. However, hierarchies are limited by their top-down structure, and although the UMLS has a large number of relationships between concepts, mapping between concepts (for example, finding the LOINC codes corresponding to the lab tests used to diagnose or measure progression of a disease with a specific ICD9 code) remains challenging.

In this paper, we show how to learn low-dimensional representations (also called embeddings) of medical concepts, putting all ICD9 diagnosis and procedure codes, LOINC laboratory codes, and NDC drug codes in a common space. We show in Figure 1 what such an embedding might look like (in practice we learn embeddings with over 100 dimensions). The closer two concepts are to each other in the embedded space, the more similar their meaning. Since medications and diagnosis codes are embedded in the same space, by searching for the nearest neighbors of a specific ICD9 code one may be able to find the medications that are relevant to the corresponding disease, such as for treatment or preventative purposes.

If high-quality embeddings of medical concepts can be learned, they could be of widespread use across medical informatics. For example, in clinical information retrieval, a query involving a specific UMLS concept could be expanded to include nearby concepts, e.g. the 10 nearest neighbors in the embedded space. Patient similarity metrics, rather than looking at the distance between patients in terms of their ICD9 codes, could use a Wasserstein distance between the codes where the weight between any pair of codes is derived from the Euclidean distance in the embedded space. Electronic phenotyping, which typically involves the laborious process of specifying inclusion criteria, could be made much faster using a tool which allows the lookup of related LOINC, ICD9 and NDC codes given a query code. Machine learning to predict disease onset or progression, or hospital readmission, may need significantly less data by using a feature vector derived by pooling the embeddings for the codes in a patient's medical history, which would have a dimension in the hundreds instead of the tens of thousands [1, 2].

Learning distributed representations or word embeddings [3, 4] has proven particularly useful in various natural language processing tasks ranging from simple language modeling and part-of-speech tagging to information

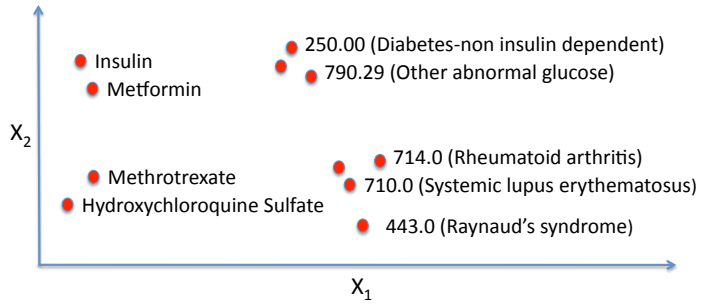

Figure 1: Illustration a low-dimensional representation (in this case, 2 dimensions) of medical concepts. Similar concepts are close to each other in Euclidean space.

extraction. Although many models have been proposed for learning distributed representations, the most popular is the skip-gram model of Mikolov *et al.* [5], implemented in the *word2vec* system. The key idea is that words with similar contexts should have similar meanings. For example, if we see the two sentences "the patient *complained* of flu-like symptoms" and "the patient *reported* flu-like symptoms", we might infer that "complained" means the same thing as "reported". As a result, these two words should be close in the representation space.

With the success of using distributed representations in the natural language processing domain, there has been a surge of interest in learning representations of concepts in medical domains. For example, Minarro-Gimenez *et al.* [6] learn embeddings from unstructured medical copora crawled from PubMed, Merck Manuals, Medscape and Wikipedia. The corpora are processed by removing the punctuations and forming medically related multi-words terms. De Vine *et al.* [7] learn embeddings of concepts by first extracting UMLS concepts from two sets of free text, clinical patient records and medical journal abstracts, then learning the embeddings using documents obtained by concatenating all of the extracted concepts.

In this paper, we take this line of work further by showing how to learn medical concept embeddings from health care claims. Specifically, we show how to use a claims dataset consisting of the ICD9 diagnosis codes, CPT procedure codes, medication and laboratory records of over 4 million patients longitudinally for 2-4 years per patient. Similar data is widely available both to providers such as health systems and to payers such as health insurance companies or the Center for Medicare and Medicaid services. We show that with simple algorithmic adjustments, it is possible to use the *word2vec* algorithm to learn embeddings on this type of longitudinal non-textual data. We also demonstrate how to learn medical concept embeddings in a privacy preserving manner from co-occurrence counts derived from clinical narratives, learning embeddings of UMLS concepts using the publicly available "graph of medicine" published by Finlayson *et al.* [8]. Finally, we create several benchmark tasks from standard medical resources such as the UMLS, the National Drug File Reference Terminology (NDF-RT), and the Agency for Healthcare Research and Quality's clinical classification software (CCS), and use these to evaluate the embeddings and characterize their properties. We find that the embeddings derived from the claims dataset are substantially better than those learned by De Vine *et al.* [7] on these benchmarks. Both our embeddings and open-source code to reproduce the benchmark results are available at `http://clinicalml.org`.

2 Methods

2.1 Background

Neural probabilistic language models are widely used in natural language processing to learn distributed representations or embeddings of words. Our learning algorithms are based on recent work on log-bilinear language models [9], and in particular makes use of the skip-gram architecture and training strategy implemented in *word2vec* [5, 10].

Let V denote the set of all concepts. We associate every concept $v \in V$ with a *concept embedding* $\Phi_v \in \mathbb{R}^d$ and a *context embedding* $\tilde{\Phi}_v \in \mathbb{R}^d$, where d is the dimension of the embedding. The context embeddings are used within the learning algorithm, but are then typically discarded, whereas the concept embeddings Φ_v are the final output. Given the context j (e.g., a neighboring concept), the log-bilinear skip-gram model defines a distribution over concepts given by the softmax function, i.e.

$$\Pr(i \mid j) = \frac{\exp\left(\Phi_i \cdot \tilde{\Phi}_j\right)}{\sum_{k \in V} \exp\left(\Phi_k \cdot \tilde{\Phi}_j\right)}. \quad (1)$$

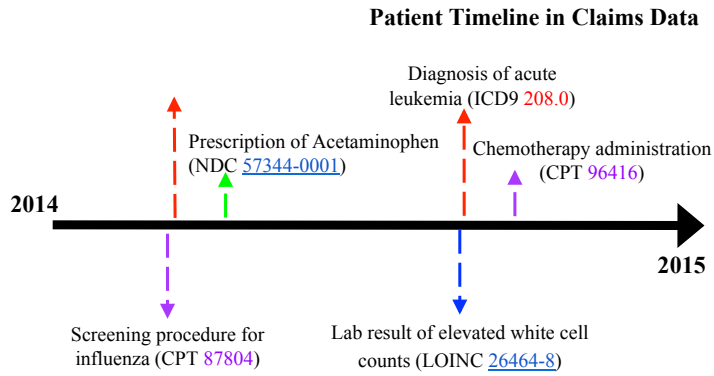

Figure 2: Illustration of the data used to learn embeddings of medical concepts, for a single patient.

Given an observed pair (i, j), the *loss* of the prediction is typically measured using log loss, $-\log \Pr(i|j)$. Note that this is 0 if i is perfectly predicted given j, and non-negative otherwise. In neural probabilistic language models, embeddings are viewed as parameters of the model, and given a set of concepts and their corresponding contexts, learning seeks the embeddings that minimize the sum of the losses.

2.2 Medical Concept Embeddings

In this section, we describe the three medical concept embeddings that are considered in this work. The three types of medical concept embedding are respectively learned from medical journals, medical claims, and clinical narratives. The one from medical journals, the most direct application of neural language modeling to medical concepts, has been recently published [7], and is the baseline to which we compare our new embeddings. We then introduce two new medical concept embeddings learned from (a) medical claims and (b) a graph of medicine constructed from co-occurence counts across clinical narratives [8].

2.2.1 Baseline: Medical Concept Embeddings from Medical Journals (MCEMJ)

We first introduce the embeddings by De Vine *et al.* [7], which were shared with us by the authors, as it will provide a valuable reference to how the embeddings we introduce differ from the previous works. As mentioned before, this type of embedding exhibits the most direct form of applying the neural language modeling techniques to medical concepts.

In this case, the main source of data is from the OHSUMED dataset, which consists of $348,566$ medical journal abstracts in the TREC 2000 Filtering Track. The authors converted the free text to UMLS concept unique identifiers (CUIs) using MetaMap v11.2, a state-of-the-art concept identification system. After obtaining the CUI sequences, they learned the embeddings using *word2vec*'s skip-gram model with hierarchical softmax, corresponding to predicting the nearby CUIs from each other in the given context of the medical journals. They experimented with different dimensions $d = 100, 200, 400$ and various window sizes $r = 2, 5, 10$. The embeddings we used have $d = 200$ and $r = 5$ as it yielded the best result in the authors' evaluations. We refer to this embedding as MCEMJ in the Results section.

2.2.2 Medical Concept Embeddings from Medical Claims (MCEMC)

The second set of embeddings are learned from a private medical claims dataset obtained from a health insurance company. The de-identified dataset consists of roughly four million persons' health claims data from 2005 to 2013. The data of each individual is in a structured format which contains information including diagnose codes (ICD9), medical visits, lab test results (LOINC), and drug usage (NDC).

Figure 2 gives an illustration of what the claims data for a single patient looks like. Using this type of temporal data creates new obstacles not found in natural language processing. Most algorithms for learning word embeddings consider a small context of 5-20 words to the left or right of each word. Once one goes beyond language to other types of data, the notion of context is less clear. Even if one takes time as the dimension of ordering, there is a newly added complexity of multiple concepts co-occurring at a single time. In the claims data, there tend to be many duplicate codes that affect the distributional pattern in an uninformative way. Furthermore, there can be multiple events that can happen in a short period of time, and the precise ordering of the events within this time window may be unimportant.

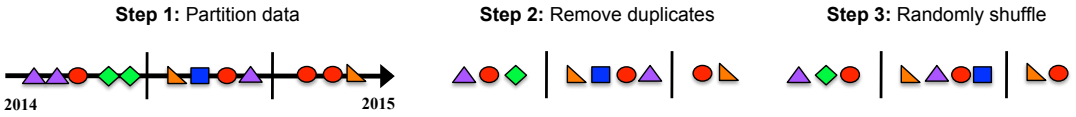

Figure 3: Our simple modified algorithm takes as input temporal data such as that shown in Figure 2 (just one temporal record is shown, but typically there would be many, e.g. one per patient). It first partitions the data into time intervals of size T (here, 1/3 of a year). Then it removes duplicate concepts, and finally shuffles the concepts so that they are in a random order. Each partition is then treated as a single sentence, and stochastic gradient descent of a bilinear skip-gram model is performed using *word2vec* [10].

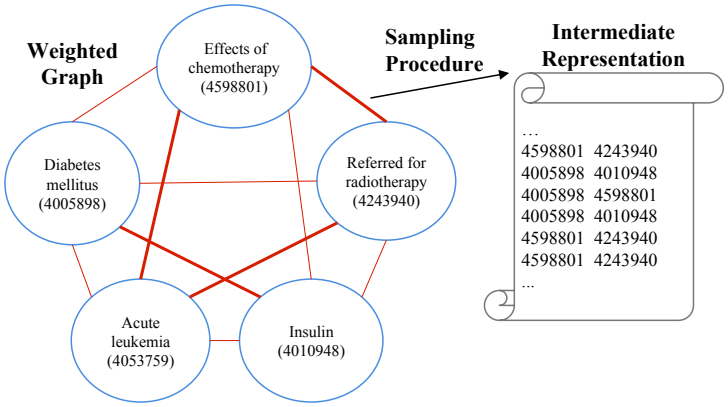

Figure 4: Shown on the left is the input, a weighted graph where the nodes correspond to each concept. We sample edges with probability proportional to the edge weight, with replacement. This results in the intermediate representation shown on the right. Each line corresponds to two prediction tasks, where the goal is to predict one of the concepts given the other. The learning algorithm iterates through the intermediate representation and performs gradient updates to the embeddings to minimize the loss on the corresponding prediction problems.

Our solution is to use partitioning and random-shuffling. Given a group of patient records P, the data for the i-th patient is denoted as p_i (e.g., as shown in Figure 2). We assume that associated with each concept occurence is a timestamp. Let T denote a time interval used for partitioning. First, we use these timestamps to partition the data into intervals of size T, as illustrated in Figure 3 (left), resulting in a set of partitions $\mathcal{E}_i$. Next, we remove duplicate mentions of a concept within each partition $e_i \in \mathcal{E}_i$ (Figure 3 middle), and then randomly shuffle the concepts within e_i, resulting in a sequence of concepts e_i' (Figure 3 right). Finally, we treat e_i' as a single sentence to be given to *word2vec* [10], which will perform $O(|e_i|)$ stochastic gradient descent (SGD) steps on the corresponding bilinear skip-gram model.

We refer to the embedding learned using this algorithm as MCEMC in the Results section. The simple algorithmic modifications that we introduce to the technique of neural language modeling were crucial for embedding medical concepts through this particular type of data; without these modifications and applying the techniques directly, we were not able to obtain sensible embeddings.

2.2.3 Medical Concept Embeddings from Clinical Narratives (MCECN)

The data used for learning the third set of embeddings is the publicly available electronic health record data released by [8]. The data consists of co-occurrence matrices of 30 million terms mapped to 1 million medical concepts from the UMLS, calculated from the raw text of 20 million clinical notes spanning 19 years of data from Stanford Hospital and Clinics. The co-occurrence matrices are constructed by using the pairwise mentions of terms and concepts in 1, 7, 30, 90, 180, 365, and ∞-day bins. The frequency counts are aggregated in two ways: (1) **per bin** counts co-occurrence of concepts at most once for each temporal bin of a patient (2) **per patient** counts co-occurrence of concepts in the same temporal bin at most once for each patient. In particular, in the ∞-day setting, the result of per bin and per patient will be the same because the complete record of patient is treated as a bin. We limit our attention to the **per bin** co-occurrence matrices to learn our embeddings.

For this clinical narrative data, we used two strategies to learn the embedding. The first, described in Figure 4, is to sample edges from the graph proportionally to the (non-negative) edge weights corresponding to the co-occurrence counts, and then to present these word pairs as examples to *word2vec*, whose skip-gram model can be shown to only depend on pairs of nearby words. The second strategy makes use of Levy and Goldberg's observation [11] that *word2vec* is implicitly factorizing the *shifted positive pointwise mutual information* (SPPMI) matrix of words and contexts (e.g., neighboring words). This approach simply performs a singular value decomposition of the SPPMI matrix derived from the weighted graph to obtain the corresponding embeddings, and as

such is a deterministic alternative to *word2vec*'s stochastic learning algorithm. We refer to the two embeddings as MCECN-SGD and MCECN-SVD in the Results section.

3 Results

In this section, we investigate various properties of the three introduced medical concept embeddings. Even from our preliminary studies, we qualitatively observed that there were significant differences among the embeddings. For instance, in one set of embeddings, diagnosis codes for various kinds of cancer formed a cluster, whereas in another set, a cluster around lung cancer would include drugs and treatments related to lung cancer. The challenge then was to provide a concrete, quantitative methodology, with which we could precisely characterize each set of embeddings.

To that end, we first identify two abstract notions of *medical relatedness* and *medical conceptual similarity* for embeddings. Then, leveraging existing medical resources such as AHRQ's clinical classification software (CCS), UMLS, and the National Drug File - Reference Terminology (NDF-RT), we define surrogate measures that reasonably compute the degrees to which a specific embedding space exhibit the abstract properties.

Experimental Setup: To perform comparisons between the embeddings we must map the concepts into the same vocabulary space. The UMLS provides a unique mapping from ICD9 diagnosis codes to UroMLS concept unique identifiers (CUIs). To map NDC drug codes to pharmacological substance CUIs (used in Section 3.2), we first map each NDC code to the corresponding clinical drug CUIs. Then, to map the clinical drug CUIs to pharmacological substance CUIs we use the UMLS relations specifying each drug's ingredients. Note that the mapping from NDC to clinical drug CUI is many-to-one and the mapping from clinical drug CUIs to pharmacological substance CUIs is many-to-many. Therefore, the mapping from NDC to pharmacological substance CUIs is many-to-many. Our method of merging the NDC embeddings is to take the mean of the embeddings of all the NDCs associated with the same CUI (we tried various alternatives, and this gave the best results for the baseline).

For each of the experiments discussed below, the vocabulary used corresponds to the intersection of the vocabularies available for each of the embeddings that are being compared. This ensures that the best possible nearest neighbor set is the same for all embeddings, providing a fair comparison. For the precise set of concepts used in each table, we refer the reader to the benchmark code available at `http://clinicalml.org`.

3.1 The Conceptual Similarity Property

We systematically evaluate whether or not a particular set of medical concept embeddings cluster the conceptually similar medical concepts as neighbors by leveraging the UMLS. We consider the following six medical concept types from the UMLS: pharmacologic substance, disease or syndrome, neoplastic process, clinical drug, finding, and injury or poisoning.

Concretely, we define the Medical Conceptual Similarity Measure (MCSM) of a set of concepts V with respect to a conceptual type set T induced by the UMLS (e.g., neoplastic process), parameterized by a size of the neighborhood k, as:

$$\text{MCSM}_{\text{UMLS}}(V, T, k) \quad = \quad \frac{1}{V(T)} \sum_{v \in V(T)} \sum_{i=1}^{k} \frac{\mathbb{1}_T(v(i))}{\log_2(i+1)},$$

where $V(T) \subset V$ is the set of concepts of type T, $v(i)$ denotes the ith closest neighbor of the chosen medical concept v, and $\mathbb{1}_T$ is an indicator function which is 1 if concept $v(i)$ is of type T, and 0 otherwise. For example, we denote the Medical Conceptual Similarity Measure of MCEMJ with respect to a conceptual type of "Neoplastic Process" from the UMLS with $k = 40$ neighbors by $\text{MCSM}_{\text{UMLS}}(\text{MCEMJ}, \text{Neoplastic Process}, 40)$. The scoring function adopts the widely used measure called Discounted Cumulative Gain (DCG) from the information retrieval literature. Note that as $k \to \infty$, the measure loses its meaning, and if k is too small it introduces too much variance to the measure. $|V|$ in our case is roughly $20,000$, and we use $k = 40$ for the experiments.

Table 1 shows how the measure is computed. As the medical concept under consideration, 4003436, has a UMLS type of neoplastic process, the top neighbors (the decimals show the computed inner product) that have the same type will contribute to the measure, modulo the index discount factors. Averaging over all the medical concepts in the vocabulary of the corresponding type, we obtain the $\text{MCSM}_{\text{UMLS}}$ of a given embedded space. Intuitively, we say that an embedded space exhibits the medical conceptual similarity property if the $\text{MCSM}_{\text{UMLS}}$ measure is high in comparison to those of other spaces.

With the $\text{MCSM}_{\text{UMLS}}$ measure defined, we now compute the measure with respect to six different medical concept types for the MCEMJ and MCECN embeddings. Table 2 shows that in 5 out of 6 cases the MCEMJ

Table 1: Display of a sub-computation for $\text{MCSM}_{\text{UMLS}}(\text{MCECN}, \text{Neoplastic Process}, 8)$. The sub-computation concerns the neighborhood of the medical concept 4003436 (Carcinoma, non-small-cell lung). The medical concept type annotations are shown in the square brackets. The numerical values represent the cosine distance of the corresponding medical concept from the query 4003436.

Neighbors of CUI 4003436 (Carcinoma, non-small-cell lung) ['Neoplastic Process']
4069419 (small cell carcinoma of lung, C0149925, ['Neoplastic Process']) : 0.956
4394316 (carcinoma of lung, C0684249, ['Neoplastic Process']) : 0.934
4125384 (malignant neoplasm of lung, C0242379, ['Neoplastic Process']) : 0.929
4070138 (adenocarcinoma of lung (disorder), C0152013, ['Neoplastic Process']) : 0.925
4555365 (tarceva, C1135136, ['Organic Chemical', 'Pharmacologic Substance']) : 0.918
4069342 (lung mass, C0149726, ['Finding']) : 0.914
4542086 (alimta, C1101816, ['Organic Chemical', 'Pharmacologic Substance']) : 0.903
4148168 (non-small cell lung cancer metastatic, C0278987, ['Neoplastic Process']) : 0.900

Table 2: Medical conceptual similarity property comparison of MCEMJ and MCECN-SGD through $\text{MCSM}_{\text{UMLS}}$. We display the evaluations of six different medical concept types with their standard deviations. Overall, we observe that MCEMJ has a stronger medical conceptual similarity property in comparison to MCECN-SGD.

	$\text{MCSM}_{\text{UMLS}}(\text{MCEMJ}[7], -, 40)$	$\text{MCSM}_{\text{UMLS}}(\text{MCECN-SGD}, -, 40)$
Pharmacologic Substance	$\mathbf{6.74 \pm 3.21}$	2.95 ± 2.15
Disease or Syndrome	$\mathbf{5.41 \pm 2.48}$	4.28 ± 1.60
Neoplastic Process	$\mathbf{6.74 \pm 3.47}$	4.54 ± 0.11
Clinical Drug	$\mathbf{1.01 \pm 0.12}$	0.12 ± 0.18
Finding	$\mathbf{2.85 \pm 1.90}$	2.15 ± 1.35
Injury or Poisoning	2.67 ± 2.40	$\mathbf{2.92 \pm 2.80}$

has a higher $\text{MCSM}_{\text{UMLS}}$ than MCECN-SGD has. In the case of injury or poisoning, MCEMJ achieves a higher score, but the difference is not statistically significant. Notably, the medicine related concept types, such as pharmacologic substance and clinical drug, exhibit the biggest differences. Overall, we can see that the set of MCEMJ embeddings exhibit a medical conceptual similarity property, when compared with the set of MCECN-SGD embeddings.

We have chosen MCECN-SGD instead of MCECN-SVD to minimize the algorithmic differences between two selected embeddings, as MCEMJ also arises from a SGD method. Given that the algorithmic differences are minimal, the core difference in the two embeddings lies in the type of distributional pattern of medical concepts that the learning algorithm uses. For the MCEMJ case, the embeddings arise from the linguistic distributional pattern. In contrast, for the MCECN-SGD case, the embeddings arise from the temporal distributional pattern, as we replace a series of extracted medical concepts from a particular medical journal with a series of medical concepts of a particular patient's clinical narratives through time. A possible explanation can be that conceptually similar concepts will inherently have similar linguistic structures around themselves (i.e. "Breast cancer is treated with..." and "Lung cancer is treated with ..."). Investigating the above claim would be an interesting direction for future work.

3.2 The Medical Relatedness Property

In this section, we systematically evaluate whether or not a particular medical concept embedding space has medically related concepts as neighbors by leveraging the NDF-RT and the hierarchical ICD9 groupings from the CCS. We refer to this particular property as the Medical Relatedness Property. The NDF-RT provides sets of relations that exist between drugs and diseases. For instance, there exists a relation called May-Treat that provides a list of drug-disease pairs, where a drug and a disease form a pair if the drug may be used to treat the disease. We study whether the nearest neighbors of a disease include drugs used to prevent or treat the disease. The hierarchical ICD9 grouping given by the CCS collapses diagnosis codes into clinically meaningful categories that are used to identify populations for specific studies or to perform reporting. We study whether the nearest

Table 3: Display of a sub-computation for $\text{MRM}_{\text{NDF-RT}}(\text{MCECN}, \text{May-Treat}, 8, -)$. In contrast to Table 1, we now have the medical concepts 4555365 (tarceva) and 4542086 (alimta) highlighted, as they are medications that are used to treat lung cancers, which the NDF-RT May-Treat relation encodes. The use of NDF-RT allows us to quantitatively and quickly evaluate the Medical Relatedness Property on a large number of test cases.

Neighbors of CUI 4003436 (Carcinoma, non-small-cell lung) ['Neoplastic Process']
4069419 (small cell carcinoma of lung, C0149925, ['Neoplastic Process']) : 0.956
4394316 (carcinoma of lung, C0684249, ['Neoplastic Process']) : 0.934
4125384 (malignant neoplasm of lung, C0242379, ['Neoplastic Process']) : 0.929
4070138 (adenocarcinoma of lung (disorder), C0152013, ['Neoplastic Process']) : 0.925
4555365 (tarceva, C1135136, ['Organic Chemical', 'Pharmacologic Substance']) : 0.918
4069342 (lung mass, C0149726, ['Finding']) : 0.914
4542086 (alimta, C1101816, ['Organic Chemical', 'Pharmacologic Substance']) : 0.903
4148168 (non-small cell lung cancer metastatic, C0278987, ['Neoplastic Process']) : 0.900

neighbors of a disease include other diseases that are related to it, quantified by being in the same category as it in the CCS. For the analysis with NDF-RT, we extend the previous neighbor analysis further by introducing the notion of analogical reasoning which has been a topic of great interest in the word embedding literature [5].

Concretely, for NDF-RT we define the Medical Relatedness Measure (MRM) of a set of concepts V with respect to a medical relation R, parameterized by a size of the neighborhood k and choice of a seed pair s, as:

$$\text{MRM}_{\text{NDF-RT}}(V, R, k, s) \quad = \quad \frac{1}{|V^*|} \sum_{v \in V^*} 1_R\Big(\bigcup_{i=1}^{k} (v - s)(i) \Big),$$

where $V^* \subset V$ denotes the set of concepts for which NDF-RT specifies at least one pharmacological substance with the given relation, and 1_R is the indicator function which returns 1 if *any* of the medical concepts in the top-k neighborhood of the selected medical concept is an element with the given relation R, and 0 otherwise. The notation $v - s$ means to first subtract the vector corresponding to a seed pair s (e.g., $s = \Phi_{\text{Acute leukemia}} - \Phi_{\text{Revlimid}}$) from Φ_v prior to searching for its nearest neighbors. In particular, the $s = 0$ case is precisely the typical neighborhood structure for v (no analogical reasoning).

We introduce analogical reasoning because, interestingly, we observed qualitatively that relations such as:

$$\Phi_{\text{Lung Cancer}} - \Phi_{\text{Tarceva}} + \Phi_{\text{Revlimid}} \quad \approx \quad \Phi_{\text{Acute leukemia}},$$

hold for the learned embeddings (Tarceva and Revlimid are medications for lung cancer and acute leukemia, respectively). The algebra signifies vector addition and subtraction. Because of this approximate equality, the concept 'Acute leukemia' is likely to be the nearest neighbor of the resulting vector. For analogical reasoning, picking different seeds has a great influence on the result. Therefore, in our experiments we record the result of using all possible drug-disease pairs as the seed, and report the mean and max hit rate.

Table 3 shows how the defined neighbor measure is computed. In contrast to the previous analysis, we do not have the medical concepts of a neoplastic process type highlighted; rather, the medical concepts 'tarceva' and 'alimta' are highlighted, as they separately form elements of the May-Treat relation with the medical concept 'carcinoma, non-small-cell lung'. It is important to note that the contribution of this subcomputation to the overall measure is precisely adding 1 to the numerator, as we only consider whether or not there is an element in the neighborhood that is related, as opposed to counting how many there are. We found this characterization to be more informative in our experiments.

Using the CCS hierarchical grouping of ICD9 codes, we generate two sets of disease groups: the most fine-grained groups (leaf nodes) and a coarse-grained grouping (cut off at the second level). When querying a disease, if another disease in the same group appears, we consider it as a hit. We define the Medical Relatedness Measure with respect to a granularity of disease grouping G induced by the CCS hierarchy and neighborhood size k as:

$$\text{MRM}_{\text{CCS}}(V, G, k) \quad = \quad \frac{1}{|V(G)|} \sum_{v \in V(G)} \sum_{i=1}^{k} \frac{1_G(v(i))}{\log_2(i+1)},$$

where $V(G) \subset V$ denotes all ICD9 codes in the vocabulary and 1_G considers whether the ith nearest neighbor $v(i)$ is in the same group as v according to G. Notice that the computational structure is similar to $\text{MCSM}_{\text{UMLS}}$,

Table 4: The Medical Relatedness Property comparison of various embeddings through $\text{MRM}_{\text{NDF-RT}}$. The results are of the form (neighbors/avg-seed/max-seed).

	$\text{MRM}_{\text{NDF-RT}}(-, \text{May Treat}, 40, -)$	$\text{MRM}_{\text{NDF-RT}}(-, \text{May Prevent}, 40, -)$
$\text{MCEMJ}_{\text{r=5,d=200}}$ [7]	12.59% / 31.56% / 53.92%	18.12% / **35.20%** / 55.88%
$\text{MCEMC}_{\text{month,ns20}}$	10.93% / 28.67% / 57.01%	5.88 % / 29.45% / **57.35%**
$\text{MCEMC}_{\text{month,hs}}$	19.24% / **37.68%** / **60.57 %**	8.82 % / 30.20% / **57.35%**
$\text{MCECN-SGD}_{\text{1Bil,7d,ns20}}$	36.81% / 33.94% / 57.48%	27.94% / 30.42% / 45.59%
$\text{MCECN-SGD}_{\text{10Bil,7d,ns20}}$	38.72% / 34.90% / 57.95%	32.95 % / 31.99% / 48.53%
$\text{MCECN-SVD}_{\text{7d,ns10}}$	**52.26%** / 35.70 % / 53.21%	**39.71%** / 32.32% / 50.00%

but this is now a measure of the Medical Relatedness Property. This highlights how the measures we introduce could be further extended with reasonably chosen relations from any medical ontology.

Table 4 contains various evaluations for the Medical Relatedness Property using the May-Treat and May-Prevent relations from NDF-RT applied to various embeddings. Unlike the previous analyses, we not only distinguish the data sources, but also show the algorithmic differences as well. First, we see that the MCECN embeddings capture medical relatedness through the neighborhood structure significantly more than the other embeddings. It can be seen that the SVD method from Levy and Goldberg preserves these particular types of semantics more effectively than the SGD method of learning. Additionally, running SGD more epochs achieves a higher measure, so this can likely be attributed to optimization rather than the slight differences in objective. In the case of the MCEMC embeddings, we see that using a hierarchical soft-max to efficiently compute the normalization in Eq. 1 preserves the semantics more so than negative sampling [10].

We observe that the results depend on the type of relation under consideration. In particular, $\text{MCEMC}_{\text{month,hs}}$ and MCEMJ reverse in their comparative positions with respect to the NDF-RT neighbor analysis when going from the May-Treat to the May-Prevent relation. Importantly, this analysis reveals that the use case for which you consider the embeddings matters. For instance, for analogical reasoning, $\text{MCEMC}_{\text{month,hs}}$ has higher performance than MCECN, capturing both relations better, but this order is swapped when you consider just the neighborhood structure. Analogical reasoning significantly improves the $\text{MRM}_{\text{NDF-RT}}$ measure in nearly all cases. This suggests that these embeddings may be very well suited for use within machine learning or patient similarity metrics, where the medical relatedness property is likely important.

Table 5: The Medical Relatedness Property comparison of various embeddings through MRM_{CCS}.

	$\text{MRM}_{\text{CCS}}(-, \text{Fine-grained}, 40)$	$\text{MRM}_{\text{CCS}}(-, \text{Coarse-grained}, 40)$
$\text{MCEMJ}_{\text{r=5,d=200}}$ [7]	0.2293	0.2490
$\text{MCEMC}_{\text{month,ns20}}$	0.4127	0.4422
$\text{MCEMC}_{\text{month,hs}}$	**0.4536**	**0.4804**
$\text{MCECN-SGD}_{\text{1Bil,7d,ns20}}$	0.2966	0.3319
$\text{MCECN-SGD}_{\text{10Bil,7d,ns20}}$	0.3087	0.3420
$\text{MCECN-SVD}_{\text{7d,ns10}}$	0.3461	0.3776

Table 5 compares the various embeddings using the Medical Relatedness Property evaluated using the fine-grained and coarse-grained groupings of ICD9 codes according to the CCS hierarchy. In this case, the MCEMC embedding achieves the highest measures across the board, which is consistent with the good qualitative results that we show in the next section. Furthermore, we see that overall the newly introduced embeddings exhibit more of the Medical Relatedness Property than the MCEMJ embeddings from [7], which is consistent with what we saw in the NDF-RT experiments.

All in all, we have introduced the concept of the Medical Relatedness Property and provided a concrete metric for evaluating it using standard medical resources such as NDF-RT and CCS. Although the results regarding the analogical reasoning should be further be investigated, we have shown through various empirical results that the two sets of embeddings that this work introduces exhibit the Medical Relatedness Property through their neighborhood structures more strongly than the previous work.

Table 6: The neighborhood of the diagnosis code 710.0 in the MCEMC. We display the top 5 neighbors for each type of code, filtering duplicates.

	Nearest Neighbors of ICD9 710.0 (Systemic lupus erythematosus) in MCEMC
	Diagnoses (ICD9)
1	695.4 (Lupus erythematosus)
2	710.9 (Unspecified diffuse connective tissue disease)
3	710.2 (Sicca syndrome)
4	795.79 (Other and unspecified nonspecific immunological findings)
5	443.0 (Raynaud's syndrome)
	Laboratory tests (LOINC)
1	4498-2 (Complement C4 in Serum or Plasma)
2	4485-9 (Complement C3 in Serum or Plasma)
3	5130-0 (DNA Double Strand Ab) in Serum)
4	14030-1 (Smith Extractable Nuclear Ab+Ribonucleoprotein Extractable Nuclear Ab in Serum)
5	11090-8 (Smith Extractable Nuclear Ab in Serum)
	Drugs (NDC)
1	00378037301 (Hydroxychloroquine Sulfate 200mg)
2	00024156210 (Plaquenil 200mg)
3	51927105700 (Fluocinolone Acetonide Miscell Powder)
4	00062331300 (All-flex Contraceptive Diaphragm Arcing Spring Ortho All-flex 80mm)
5	00054412925 (Cyclophosphamide 25mg)

3.3 Qualitative evaluation

We already gave in Table 3 one example of the neighborhood structure for the newly introduced MCECN embedding, learned using clinical narratives. In Table 7 we show the nearest neighbors of various genetic mutation concepts. In Table 6 we display an example using the newly introduced MCEMC embedding, learned using health insurance claims. The neighborhood structures themselves may provide sources of insights to practitioners searching for undiscovered medical relations.

4 Discussion

This work demonstrates that learning distributed representations or embeddings, originally used in the natural language processing community, has new applicability in biomedical informatics. Furthermore, we provide a precise characterization of the new embeddings in that they exhibit the Medical Relatedness Properties in their neighborhood structures when contrasted with the embeddings learned in previous work.

We introduced two algorithms for learning distributed representations from temporal data. The first algorithm takes as input the raw data, and as such leads to a flexible framework that is more easily extended. The second algorithm takes as input a weighted graph, where the weights are derived from the co-occurence counts of concepts within fixed time intervals. Notably, this allows us to learn embeddings from *privacy-preserving* data which is aggregated across patients prior to being presented to the learning algorithm. Additionally, it may be more broadly applicable to other tasks where embeddings are to be learned from a weighted graph, and where the weights simply denote the strength of the interaction between the corresponding pair of concepts.

There are several directions for future work. For example, whereas we used a fixed time interval to partition the data (e.g., using weighted graphs computed from co-occurences in a single day, month, or year), we could instead attempt to learn simultaneously on all time intervals, which may lead to higher quality embeddings by asking them to be predictive for all of these distinct prediction tasks (the context vectors could be allowed to vary). Similarly, the algorithm illustrated in Figure 3 only attempts to predict *within* a time interval, but it would be interesting to consider attempting to predict *across* time intervals. We could also introduce additional parameters to the model to capture the temporal ordering of the data, so that when learning the embeddings we can take advantage of ordering information, e.g. whether a disease is first diagnosed one month before a lab test was performed, or vice-versa.

Table 7: A few neighborhood examples from MCECN illustrating genotypic-phenotypic relations.

	(cd52, C2733653)		**(bcl1, C2599665)**
1	(cd52 protein, human, C0376272)	1	(cyclins, C0079183)
2	(mycosis fungoides/sezary syndrome nos, C0862196)	2	(proliferating cell nuclear antigen, C0072108)
3	(t-cell receptor, C0034790)	3	(lymphoplasmacytic lymphoma, C2700641)
4	(lymphoma, t-cell, cutaneous, C0079773)	4	(paired box 5 protein, C0167636)
5	(pralatrexate, C1721300)	5	(cyclin d1, C0174680)
	(jak2 mutation, C2827348)		**(kras mutation, C2747837)**
1	(refractory anemia with ringed sideroblasts, C1264195)	1	(mesothelioma, C0025500)
2	(large platelets (finding), C1148412)	2	(cdx2 protein, human, C1505661)
3	(anagrelide, C0051809)	3	(cdx2 antigen, C1829706)
4	(hypercellular bone marrow, C1334068)	4	(pleural mass, C1709576)
5	(myeloid metaplasia, C0027013)	5	(braf protein, human, C1259929)

Our experimental results illustrated that, not surprisingly, embeddings learned on different types of data capture different semantics. Future work should consider multi-modal learning algorithms that incorporate data from various sources to learn high-quality distributed representations, such as using both medical claims and clinical narratives for the same patient, and using prior knowledge gleaned from medical journals.

Acknowledgments

The authors gratefully acknowledge support by Independence Blue Cross.

References

[1] Choi E, Bahadori MT, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. CoRR. 2015;abs/1511.05942.

[2] Krompaß D, Esteban C, Tresp V, Sedlmayr M, Ganslandt T. Exploiting Latent Embeddings of Nominal Clinical Data for Predicting Hospital Readmission. KI. 2015;29(2):153–159.

[3] Bengio Y, Ducharme R, Vincent P, Janvin C. A Neural Probabilistic Language Model. J Mach Learn Res. 2003 Mar;3:1137–1155.

[4] Collobert R, Weston J. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In: Proceedings of the 25th International Conference on Machine Learning. ICML '08. New York, NY, USA: ACM; 2008. p. 160–167.

[5] Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. CoRR. 2013;abs/1301.3781.

[6] Miñarro-Giménez JA, Marín-Alonso O, Samwald M. Exploring the Application of Deep Learning Techniques on Medical Text Corpora. In: e-Health - For Continuity of Care - Proceedings of MIE2014, the 25th European Medical Informatics Conference; 2014. p. 584–588.

[7] De Vine L, Zuccon G, Koopman B, Sitbon L, Bruza P. Medical Semantic Similarity with a Neural Language Model. In: Proceedings of CIKM '14. New York, NY, USA: ACM; 2014. p. 1819–1822.

[8] Finlayson S, LePendu P, Shah N. Data from: Building the graph of medicine from millions of clinical narratives. Dryad Digital Repository; 2014.

[9] Mnih A, Hinton G. Three new graphical models for statistical language modelling. In: Proceedings of the 24th international conference on Machine learning. ACM; 2007. p. 641–648.

[10] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and their Compositionality. In: Advances in Neural Information Processing Systems 26; 2013. p. 3111–3119.

[11] Levy O, Goldberg Y. Neural word embedding as implicit matrix factorization. In: Advances in Neural Information Processing Systems; 2014. p. 2177–2185.