SN Computer Science (2021) 2:276
https://doi.org/10.1007/s42979-021-00656-y

ORIGINAL RESEARCH

Medical Provider Embeddings for Healthcare Fraud Detection

Justin M. Johnson[1] · **Taghi M. Khoshgoftaar**[1]

Received: 27 December 2020 / Accepted: 19 April 2021

Abstract
Advances in data mining and machine learning continue to transform the healthcare industry and provide value to medical professionals and patients. In this study, we address the problem of encoding medical provider types and present four techniques for learning dense, semantic embeddings that capture provider specialty similarities. The first two methods (GloVe and Med-W2V) use pre-trained word embeddings to convert provider specialty descriptions to phrase embeddings. Next, HcpsVec and RxVec embeddings are constructed from publicly available big data using specialty-procedure and specialty-drug occurrence matrices, respectively. We evaluate the learned provider type embeddings on two real-world medicare fraud classification problems using logistic regression (LR), random forest (RF), gradient boosted tree (GBT), and multilayer perceptron (MLP) learners. Through repetition, statistical analysis, and feature importance measures, we confirm that semantic embeddings for provider types significantly improve fraud classification results. Finally, t-SNE visualizations are used to show that the learned provider type embeddings capture meaningful specialty characteristics and provider type similarities. Our primary contributions are two novel methods for encoding medical specialties using procedure-level statistics and the evaluation of four encoding techniques on two large-scale healthcare fraud classification tasks. Since all data sources are publicly available, these encoding techniques can be readily adopted and applied in future machine learning applications in the healthcare industry.

Keywords Machine learning · Healthcare · Semantic embeddings · Big data · Fraud detection

Introduction

Advancements in data mining and machine learning have the potential to transform the healthcare industry. Scalable data mining and machine learning techniques are able to process large volumes of structured and unstructured data from electronic health records (EHR) and historical claims data, far exceeding the capabilities of any human analysis [35]. From this analysis, valuable insights can be extracted to guide medical professionals on operational decisions and patient care. At the operational level, machine learning can be used to assist facilities with scheduling, forecasting expenses [63], information retrieval [27], and reducing patient wait times [57]. Machine learning for patient care can assist medical providers by detecting and diagnosing illnesses [28], suggesting personalized treatment plans [70], monitoring physiological signals [34], and forecasting epidemic trends [66]. Fueled by big data [32] and a variety of supervised and unsupervised learning methods [25], these applications have the potential to improve healthcare efficiency, lower patient costs, improve patient satisfaction, and increase life expectancy rates.

Machine learning techniques for automating the detection of fraudulent activity is another healthcare application that has the potential to drastically lower patient costs and improve the quality of patient care. In this study, we explore new techniques for detecting fraud within the United States (U.S.) Medicare program using historical claims data. The U.S. Medicare program provides affordable health insurance to individuals 65 years and older, and other individuals with permanent disabilities [68]. There are currently more than 62.2 million U.S. citizens enrolled in Medicare [10], and

This article is part of the topical collection "Artificial Intelligence for HealthCare" guest-edited by Lydia Bouzar-Benlabiod, Stuart H. Rubin and Edwige Pissaloux.

✉ Justin M. Johnson
jjohn273@fau.edu

Taghi M. Khoshgoftaar
khoshgof@fau.edu

[1] College of Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA

Published online: 15 May 2021

2019 expenditures exceeded $796 billion [14]. The Federal Bureau of Investigation estimates that fraud accounts for 3–10% of all billings [51], i.e. $23–$79 billion per year in the Medicare program. Examples of fraudulent activities include billing for appointments that a patient did not keep, billing for more complex services than those that were performed, or billing for services that were not provided at all. In an effort to reduce fraud, the Centers for Medicare and Medicaid Services (CMS) makes Medicare data sets publicly available for analysis [11].

This study uses two publicly available data sets from the CMS: (1) the 2012–2016 Medicare Provider Utilization and Payment Data: Physician and Other Supplier Public Use File (Part B) data sets [12], and (2) the 2013–2017 Medicare Provider Utilization and Payment Data: Part D Prescriber (Part D) data sets [13]. These big data sets contain millions of records that summarize the utilization and payments for medical procedures, services, and prescription drugs provided to Medicare beneficiaries. Each record describes a specific medical provider and their billing activity for a given year, e.g. the provider specialty, the number of times a procedure was performed or a drug was prescribed, the number of patients seen, and the average amount billed for a service. Fraudulent and non-fraudulent class labels are mapped to the Part B and Part D data claims data using the List of Excluded Individuals and Entities (LEIE) [53]. The LEIE is maintained by the Office of Inspector General (OIG), and it lists providers that are prohibited from participating in Federal healthcare programs. By mapping real-world fraud labels to the Part B and Part D claims data sets, we can employ supervised learning algorithms to predict fraudulent versus non-fraudulent providers from claims activity.

We improve upon existing Medicare fraud detection rates by employing semantic embeddings for the medical provider type (specialty) feature contained within the Part B and Part D data sets. The provider type attribute describes a provider's medical specialty, e.g. Internal Medicine, Anesthesiology, and Cardiology, and is also referred to as the specialty or specialty group throughout this paper. This is a categorical variable that contains 123 and 103 unique values in the Part B and Part D data sets, respectively. In related Medicare fraud classification studies, the provider type has either been excluded from the feature set or represented as a one-hot vector. A one-hot encoding is a method for converting a categorical variable to a binary vector with a length equal to the total number of categories, where the ith entry corresponding to the ith category is 1 and all other entries are 0 [71]. Since the provider type variables have high cardinality (> 100), their resulting one-hot representations become large and sparse, both increasing model complexity and degrading predictive performance through the curse of dimensionality [16]. Furthermore, the one-hot vectors are not able to capture the relationships that exist between similar provider types in practice. We address this by exploring four semantic embedding techniques for the provider type variable, i.e. dense numeric representations that capture meaningful characteristics of each provider type.

The concept of learning low-rank semantic representations for medical specialty groups is largely inspired by word embeddings. Semantic and distributed representations for words have transformed natural language processing and improved upon the state-of-the-art in a variety of domains, e.g. sentiment analysis [48], machine translation [72], question and answering [60], and named entity recognition [22]. By encoding semantic meaning into each word's representation using dense continuous vectors, machine learning algorithms are able to detect similarities between words, similarities that go undetected when using one-hot vectors. Mikolov et al. [50] proposed two Word2Vec models for efficiently learning high-quality representations of words. Pennington et al. [55] later proposed global vectors for word representation (GloVe), a method for generating word embeddings from a global word-word co-occurrence matrix. Both Word2Vec and GloVe embeddings are word-level representations, i.e. they do not take into account the order of words. Embeddings from language models (ELMo) uses a bidirectional long short-term memory (LSTM) algorithm to encode context-aware word embeddings [56]. Similarly, bidirectional encoder representations from transformers (BERT) achieves contextualized word embeddings by pre-training masked language models with a multi-layer bidirectional transformer encoder [24]. In this paper, we use pre-trained word embeddings to capture provider type similarities and improve Medicare fraud prediction performance. Additionally, we propose two techniques that encode provider type variables using historical Part B and Part D claims data.

We consider four techniques for encoding semantic meaning into medical specialty embeddings, i.e. entity embeddings [29]. The first approach (GloVe) converts the provider type variable to its equivalent word embedding by replacing the tokens in the provider type textual description with their respective GloVe word embeddings. Most provider types are short phrases, e.g. 2–4 words, so we combine them to a single vector by taking their unweighted average [4]. In a similar manner, the second approach (Med-W2V) converts provider type variables to word embeddings using publicly available word embeddings that have been pre-trained on clinical notes from PubMed and Pubmed Central Open Access (PMC OA) [59]. Since word embeddings have already been shown to capture syntactic and semantic meaning for words through the distributional hypothesis [62], we are merely transferring that knowledge to our Medicare fraud classification problem as phrase embeddings. Next, we construct HcpcsVec provider type embeddings from Medicare Part B claims data using a specialty-procedure occurrence matrix created from the Healthcare Common

Procedure Coding System (HCPCS) codes reported within each specialty group. Finally, we construct RxVec provider type embeddings from Medicare Part D claims data using a specialty-drug occurrence matrix that is derived from the drugs prescribed by each specialty group. Generally, medical providers within a particular specialty group provide a subset of services and prescriptions that are relevant to their practice, and similar provider types are expected to provide similar services and prescriptions. Therefore, we expect both HcpcsVec and RxVec to encode semantic meaning for specialty groups such that similar provider types share similarities in one or more dimensions of embedding space. The pre-trained GloVe and Med-W2V embeddings have a dimensionality of 300 and 200, respectively. The HcpcsVec and RxVec methods produce numeric representations for specialties with a dimensionality of 7527 and 3545, respectively. To provide a fair comparison over a range of embedding sizes, we use principal component analysis (PCA) [71] to further reduce the dimensionality of the embeddings and compare results using embedding sizes between 32 and 128.

Quantitative and qualitative experimental results are presented to demonstrate the efficacy of semantic embeddings for medical specialty groups. First, we compare sparse one-hot representations to the four semantic embeddings using the two Medicare Part B and Medicare Part D fraud classification problems. Logistic regression (LR), gradient boosted tree (GBT), random forest (RF), and multilayer perceptron (MLP) learners are trained and evaluated using each embedding technique and data set. Performance is measured using the area under the receiver operating characteristic curve (AUC) [58], and 30 repetitions of each experiment are used to perform statistical analysis with Tukey's honestly significant different (HSD) test [67]. AUC results averaged across all experiments show that the HcpcsVec, RxVec, and Med-W2V methods significantly outperform one-hot embeddings on the Medicare fraud classification task. Furthermore, we show that these new embedding techniques outperform previous works [38]. Next, we use the GBT feature importance measure to show that the GBT learner assigns a higher weight to the provider type attribute when it is represented with one of the four semantic embeddings. Finally, we use t-SNE visualizations [49] to qualitatively evaluate the HcpcsVec and RxVec embeddings and illustrate how similar provider types are encoded with similar numeric representations.

This work expands upon our previous works on semantic embeddings for medical providers [41] by incorporating the Medicare Part D data set, RxVec embeddings, an analysis of feature importance, and t-SNE visualizations. Our primary contributions are the evaluation of semantic embeddings for Medicare fraud prediction and the HcpcsVec and RxVec methods for constructing specialty type embeddings from publicly available big data. Having been constructed from publicly available Medicare data with up to 122 million data points, we believe these high-quality embeddings for medical specialty groups can be extended to other healthcare applications that depend on medical providers and their specialties.

The remainder of this paper is structured as follows. In the next section, we review previous work related to Medicare fraud prediction and embedding methods for medical concept extraction. In the subsequent sections, we describe data sets used, embedding techniques, and the experiment design. Before the last section, classification, feature importance, and t-SNE results are presented. The final section concludes with a summary of our results and areas for future work.

Related Work

Medicare Fraud Prediction

Medicare fraud prediction has received a lot of attention in recent years. In many of these studies, however, the provider type attribute is not used. Bauder and Khoshgoftaar [5] estimate Medicare payments using five regression models. Actual payment amounts are compared to estimated payments, and the deviations are used to flag potentially fraudulent providers. Ko et al. [45] use a linear regression model to analyze the variability of service utilization and payments in 2012 CMS Part B data. Both studies use only a subset of medical specialties and model each specialty separately. Branting et al. [8] use graph-based features and a decision tree learner to predict Medicare fraud using the 2012–2014 CMS Part B and Part D data sets. Advanced features are constructed from behavioral similarity between providers and risk propagation through geospatial collocation, but the provider type predictor is not included.

Several studies have explored the relationship between provider specialties and the procedures that they bill for. In [6], Bauder et al. use a Naive Bayes learner to predict Medicare provider specialties from HCPCS procedure occurrences. Results show that 7 of 82 provider types scored very highly (F1-score > 0.90), and 18 provider types scored reasonably ($0.5 <$ F1-score < 0.90). This suggests that the majority of the provider types have overlapping procedure activity that prevents accurate provider type prediction. Herland et al. [30] expand on this by incorporating 2014 CMS Part B data and real-world fraud labels defined by the LEIE data set. A Naive Bayes learner is used to predict a provider's specialty from their respective HCPCS frequencies, and misclassified provider types are assigned a fraudulent class label. Herland et al. discover that grouping similar provider types improves overall classification performance. Chandola et al. [15] use healthcare claims and fraudulent

provider labels provided by the Texas OIG exclusion database to detect anomalies and bad actors. The authors model providers as documents and use latent Dirichlet allocation to identify 20 hidden topics from a provider-diagnosis matrix. Chandola et al. show how some topics are dominated by diagnoses belonging to the same area of medicine, e.g. Oncology and Ophthalmology, and suggest that the topic distributions can be used as features for downstream learning. Experimental results showed that the inclusion of the provider type attribute increases AUC score from 0.716 to 0.814. These related works stress the importance of the provider type feature in predicting Medicare fraud, and allude that more semantic representations for provider types can improve the performance of Medicare fraud classification.

Another subset of Medicare fraud studies encode medicare provider types using one-hot vectors. Herland et al. [31] explore fraud prediction using three 2012–2015 CMS Medicare data sets, i.e. Part B, Part D, and the Durable Medical Equipment, Prosthetics, Orthotics and Supplies (DMEPOS). Fraudulent providers are identified using the LEIE data set and the feature set includes provider activity statistics, e.g. the average amount billed, average amount paid, and average number of beneficiaries treated. Part B, Part D, and DMEPOS claims data are used independently to perform cross-validation with LR, RF, and GBT learners. The LR learner performed best on the Part B data set with a maximum AUC score of 0.805. In another study [38], we reuse the CMS Part B data set from Herland et al. to evaluate deep neural networks and various techniques for addressing class imbalance. Data-level and algorithm-level methods are used to treat class imbalance, and results show that balancing training data with random over-sampling (ROS) and random under-sampling (RUS) maximizes performance with an average AUC of 0.8506. Johnson and Khoshgoftaar [39] explored the effects of data sampling further on Part B and Part D data, and found that that the hybrid ROS–RUS method also maximizes performance on the Part D data set. We believe that the one-hot encoding of provider types in these previous works is insufficient, and that meaningful relations between provider types are lost in equidistant one-hot vectors. To address this information loss, this paper explores multiple encoding techniques that capture meaningful relations between similar provider types.

Two related works evaluate procedure code embeddings for predicting healthcare fraud. Fursov et al. [26] present a procedure code embedding technique for identifying billing errors. A total of 2205 unique treatments are encoded using a private data set. Johnson and Khoshgoftaar [40] expand on this by encoding 7527 HCPCS procedure codes by training Word2Vec models on publicly available claims data. Results show that embeddings for procedure codes consistently outperform one-hot encodings. In this study, we aim to obtain similar results by employing provider type embeddings.

In a previous study [41], we explored embedding techniques for medical provider types using the Medicare Part B data set. This introduced the HcpcsVec embedding technique, where provider type embeddings are inferred from specialty-HCPCS occurrences and compared to word embeddings on the Part B fraud classification task. Classification results showed that the HcpcsVec embedding technique outperforms one-hot, GloVe, and Med-W2V embeddings based on the AUC metric. In this paper, we expand on these results by introducing the RxVec embedding technique, a second data set for validation (Part D), an analysis of feature importance, and t-SNE visualizations of embeddings.

Medical Concept Embeddings

Several research groups have extended word embedding techniques to the biomedical domain and learned word representations for specific clinical concepts, including disease codes, prescription drugs, and medical procedures. Two recent survey papers summarize these methods in great detail. Khattak et al. [44] survey methods for training word embeddings on clinical text and cover a range of topics, e.g. word representation, clinical text corpora, pre-trained clinical embeddings, applications, and their limitations. Embedding techniques covered include Word2Vec, GloVe, FastText, ELMo, and several variations of BERT, and the authors have suggested using t-SNE for visualizing word embeddings to verify similar words form clusters in feature space. Kalyan and Sangeetha [43] provide a second survey, focusing on embedding techniques for clinical natural language processing objectives. Medical corpora for training word embeddings are categorized as EHR data, medical social media data, online medical knowledge resources, and scientific literature. The Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) data set [69], and subsequent versions MIMIC-II and MIMIC-III, are introduced as the largest publicly available EHR data sets, containing patient demographics, vital signs, laboratory tests, and medications. Kalyan and Sangeetha also present solutions to common embedding challenges, e.g. small corpus sizes, multi-sense embeddings, domain adaptation, and out-of-vocab words.

De Vine et al. [23] produce medical concept embeddings from free text in clinical records and medical journal abstracts. Sequences of medical concepts are first created by mapping the free text to concepts defined in the Unified Medical Language System (UMLS) Metathesaurus using MetaMap [3]. Skip-gram models are trained on the sequences of medical concepts using a range of embedding and window sizes. Semantic similarity results are evaluated against six baseline methods using two data sets of similarity pairs that were produced by expert medical judges. The skip-gram embeddings perform best overall according to the

Pearson correlation measure and on average, larger embedding and window sizes improve performance.

Choi et al. [19] learn low-dimensional embeddings for clinical concepts from multiple sources using the Word2Vec model and compare their results to the medical concept embeddings from medical journals (MCEMJ) provided by De Vine et al. [23]. Medical concept embeddings from medical claims (MCEMC) are trained on claims data that spans over four million patients between 2005 and 2013. Medical concept embeddings from clinical narratives (MCECN) are trained on concept co-occurrence matrices from 20 million clinical notes across 19 years of data. The MCEMJ embeddings score the highest overall on a medical concept similarity measure and the MCEMC embeddings score the highest overall on a medical relatedness measure. Results show that clinical embeddings learned on different sources of data capture different semantics, suggesting that different embedding sources may perform better on specific downstream tasks. We explore this phenomenon in this study by comparing pre-trained GloVe embeddings to embeddings that have been pre-trained on medical corpora. In another work, Choi et al. [18] propose Med2Vec for learning embeddings for medical codes and patient visits using procedure code co-occurrences and sequential patient data. Med2Vec outperforms Skip-Gram, GloVe, one-hot encoding, and stacked autoencoder embeddings on the task of predicting future medical codes. Yuqi et al. [64] evaluate the impact of word-level and context-sensitive word representations on the task of clinical concept extraction. The authors compare Word2Vec, GloVe, fastText [7], ELMo, and BERT on the i2b2 and SemEval data sets. Both open-domain and MIMIC-III are used for pre-training word embeddings, and results show that pre-training on clinical corpora generally performs better. Huang et al. [33] propose ClinicalBERT for generating and evaluating representations of clinical notes. ClinicalBERT is trained on clinical notes from a patient's intensive care notes and discharge summary, and is evaluated on a 30-day hospital readmission prediction task. The proposed model outperforms a bag-of-words model and a bidirectional LSTM network that is trained using Word2Vec embeddings trained on the MIMIC-III clinical data set.

Several works supplement medical concept embeddings with ontologies, or relational knowledge graphs, e.g. the Clinical Classifications Software (CSS) categorization scheme [21]. Choi et al. [17] proposed the GRaph-based Attention Model (GRAM), which learns representations for diagnosis codes from a combination of CSS ontological ancestors via a neural attention weighting mechanism. GRAM is compared to five baseline methods using two next-visit diagnosis prediction tasks and a heart failure prediction task, and results suggest GRAM outperforms baseline methods on low-frequency diseases and small data sets. Ma et al. [47] propose the knowledge-based attention model (KAME) for predicting patients' future health conditions. Unlike GRAM, KAME exploits medical knowledge throughout the whole prediction process, i.e. code representations, visit embeddings, and down-stream predictions. Ma et al. show that KAME outperforms GRAM and three other baseline methods on three diagnosis prediction tasks. Song et al. [65] extend the GRAM model with Medical Concept Embeddings with Multiple Ontological REpresentations (MMORE). MMORE, which outperforms GRAM, combats the inconsistencies between EHR data and medical ontologies by allowing multiple representations of hierarchical ontology concepts to be learned. In each of these works, ontology relationships help align the learned embeddings with expert medical knowledge and improve the quality of low-frequency concept embeddings.

In summary, the related work shows that semantic embeddings for medical concepts improve performance on downstream classification tasks. Many of these related works focus on learning embeddings for diagnosis codes and medications for the purpose of evaluating semantic similarity and predicting future diagnosis. In this study, we focus specifically on constructing semantic embeddings for medical specialty groups. We employ word embedding techniques from natural language processing using open-domain and clinical-domain pre-trained embeddings, and we present two new methods for constructing provider type embeddings from publicly available Medicare data. Given the success of word embeddings for representing Medicare provider types, we plan to explore more advanced language model embeddings (e.g. ELMo, BERT, and ClinicalBERT) and supplementing embeddings with ontologies in future works.

Data Sets

This study uses two publicly available Medicare data sets: (1) Medicare Provider Utilization and Payment Data: Physician and Other Supplier (Part B), and (2) Medicare Provider Utilization and Payment Data: Part D Prescriber (Part D). The Part B and Part D claims data sets have been made available by the Centers for Medicare and Medicaid Services (CMS), and can be downloaded in a tab delimited format on the CMS website [1]. Data sets are released on an annual basis, and each yearly data set summarizes the utilization and payments for procedures, services, and prescription drugs provided to Medicare beneficiaries by medical professionals for that year. Providers are identified within each data set by their National Provider Identifier (NPI), a unique 10-digit identification number for healthcare providers [52].

The Part B data sets used in this study cover years 2012–2016 and contain approximately 47 million records. The Part B claims data set describes the services and procedures that healthcare professionals provide to Medicare's

Fee-For-Service beneficiaries. Records within the data set contain various provider-level attributes, e.g. NPI, first and last name, gender, credentials, and provider type. More importantly, records contain specific claims details that describe a provider's activity within Medicare. Examples of claims data include the procedure performed, the average charge submitted to Medicare, the average amount paid by Medicare, and the total number of beneficiaries that received the service. There are 7527 unique procedures codes, and they are encoded using the Healthcare Common Procedures Coding System (HCPCS) [9]. For example, HCPCS codes 99219 and 88346 are used to bill for hospital observation care and antibody evaluation, respectively. Part B data is aggregated by: (1) provider NPI, (2) HCPCS code, and (3) place of service. A summary of the Part B features used in this study are listed in Table 1.

We use the Medicare Part D data sets from CMS for years 2013–2017, i.e. approximately 122 million instances in total. The Part D data sets provide information on prescription drugs prescribed to Medicare beneficiaries by medical professionals. Many of the Part D provider-level attributes are the same as those in the Part B data set, e.g. NPI, name, and provider type. The numeric attributes of the Part D data set describe the provider's activity relative to a specific prescription drug for a given year. Examples of prescription attributes include the drug name, the number of beneficiaries receiving the medication, the quantity being prescribed, and the cost paid for all claims. A description of the Part D attributes used in this study are listed in Table 2.

The feature of interest in this study is the Provider_type attribute, a categorical variable that describes the provider's medical specialty. The Part B and Part D data sets used in this study contain 123 and 103 unique provider type values, respectively. Examples of provider types include Internal Medicine, Nurse Practitioner, and Urology. In related works, the provider type attribute has been encoded using one-hot encodings. This subjects predictive models to relatively large, sparse, and uninformative feature vectors. We address these concerns in this study by constructing dense, semantic embeddings for provider type specialty groups using the Part B and Part D claims data sets.

The exclusion column in the Part B and Part D data sets are real-world fraud labels derived from the LEIE [53]. These fraud labels are mapped to the Part B and Part D data sets by joining each data set with the LEIE data set on the NPI column. The LEIE is maintained by the Office of Inspector General and it lists providers that are prohibited from the Medicare program. In addition to the provider's NPI, the LEIE also includes the reason for exclusion and a reinstatement date, if applicable. Following previous works [31], we use a subset of exclusion types representative of fraud to label Part B and Part D providers as fraudulent and

Table 1 Description of Part B data

Feature	Description
NPI	Unique provider identification number
Provider_type	Medical provider's specialty (or practice)
Nppes_provider_gender	Provider's gender
HCPCS	Code for medical service furnished by provider
Line_srvc_cnt	Number of procedures the provider performed
Bene_unique_cnt	Number of beneficiaries receiving the service
Bene_day_srvc_cnt	Number of beneficiary/per day services
Avg_submitted_chrg_amt	Avg. of charges that provider submitted
Avg_medicare_payment_amt	Avg. payment made to provider per claim
Exclusion	Fraud labels from the LEIE data set

Table 2 Description of Part D data

Feature	Description
NPI	Unique provider identification number
Provider_type	Medical provider's specialty (or practice)
Drug_name	Brand name of drug prescribed
Bene_count	Number of distinct beneficiaries receiving drug
Total_claim_count	Number of Medicare Part D claims, including refills
Total_30_day_fill_count	Number of standardized 30-day fills
Total_day_supply	Number of day's supply
Total_drug_cost	Cost paid for all associated claims
Exclusion	Fraud labels from the LEIE data set

non-fraudulent. This additional Exclusion column creates a labelled data set for training predictive models to classify fraudulent providers, allowing us to evaluate our embeddings on two fraud classification tasks.

Methods

Runtime Environment

All experiments are carried out on a high-performance computing environment running Scientific Linux 7.4 (Nitrogen) [46]. Jobs are dispatched onto CPU nodes with 20 Intel(R) Xeon(R) CPU E5-2660 v3 2.60GHz processors and 128GB of RAM. The scikit-learn package [54] (version 0.21.1) is used for pre-processing data and training the LR, RF, GBT, and t-SNE models. Neural networks are implemented using the Keras [20] open-source deep learning library written in Python with its default backend, i.e. TensorFlow [2]. The specific library implementations used in this study are the default configurations of Keras 2.4.0 and TensorFlow 2.3.0.

Embedding Techniques for Medical Specialties

We employ four provider type embedding techniques that were inspired by advances in natural language processing, and we compare results to those achieved with traditional one-hot vectors. One-hot vector encodings of provider types are sparse, atomic, and uninformative representations that fail to capture any form of provider similarity. Consider, for example, the provider types Certified Clinical Nurse Specialist, Certified Nurse Midwife, and Orthopedic Surgeon. With one-hot encoding, all specialties are equidistant in vector space and no pair of specialties is geometrically closer than another. We begin with two word embedding techniques, GloVe and Med-W2V, encoding each specialty type with a phrase embedding using pre-trained word embeddings. Next, we leverage the large-scale Medicare data made available by CMS to construct two sets of embeddings from the historical claims made within each specialty group.

Provider type attributes are nothing more than short phrases, e.g. 1–4 word textual descriptions. Therefore, we can capture the semantic relationships between provider types by replacing their textual descriptions with their word embeddings. Our first word embedding technique uses GloVe embeddings to represent each provider type using an unweighted average of word embeddings. For example, we encode Orthopedic Surgeon by adding the GloVe embedding for orthopedic and the GloVe embedding for surgeon, and then dividing by two. We chose to use unweighted averages because they are relatively simple and effective for short phrases [4]. Similarly, we repeat this process using domain-specific word embeddings. We refer to these embeddings as Med-W2V embeddings in this study, because they were induced from large biomedical corpora using a Word2Vec model [59]. The pre-trained GloVe and Med-W2V embeddings have dimensionality of 300 and 200, respectively. For both word embedding techniques, we use the *P+PCA* algorithm [61] to further reduce the dimensionality of the word embeddings to lengths of 32, 64, and 128 based on preliminary results.

The HcpcsVec embedding technique constructs dense provider type representations from a specialty-HCPCS occurrence matrix using the Medicare Part B data set. This is achieved by aggregating HCPCS occurrences over all providers within a single specialty group. First, we group the Part B data on (NPI, provider type) and sum the total number of times a given HCPCS was performed using the Line_srvc_cnt attribute. Since there are 7527 unique HCPCS values, and most providers only use a small subset of procedures, this produces a sparse HCPCS occurrence vector for each (NPI, provider type) pair. Next, we normalize each row of the occurrence matrix by dividing by the maximum occurrence in each row. We apply this normalization step because preliminary analysis shows that some providers are significantly more active than others, but the relative HCPCS frequencies are assumed to be similar for providers within the same specialty group [6]. This normalization step also reduces the magnitude of our final embeddings to the desired interval of [0, 1]. Finally, we group the normalized HCPCS occurrence matrix on the provider type and take the average for each HCPCS column. The resulting specialty-HCPCS occurrence matrix has a size of 123×7527, and each row summarizes the average procedure activity for a particular specialty. We summarize this process visually using a small subset of Medicare data in Fig. 1. When constructing HcpcsVec embeddings, the service column in Fig. 1 corresponds to the HCPCS attribute included in the Part B data set. We use PCA to reduce the dimensionality of the occurrence vectors to lengths of 32, 64, and 123. Note that we were not able to achieve an embedding size of 128 when using HcpcsVec because there are only 123 provider types.

The RxVec embedding technique produces dense embeddings for provider types from a specialty-drug occurrence matrix using the Medicare Part D data set. There are 3545 unique drug names in total, and providers within a given specialty generally prescribe only a subset of drugs related to their primary practice. Hence, the drugs prescribed by a specialty group can be used to characterize that specialty group and capture relationships between similar provider types. Much like the HcpcsVec embeddings, the RxVec embedding technique first groups the Part D data on (NPI, provider type), and then sums the total number of times a given drug was prescribed using the Total_claim_count feature. Each provider-drug occurrence vector is normalized by dividing by the maximum occurrence in each row. Next,

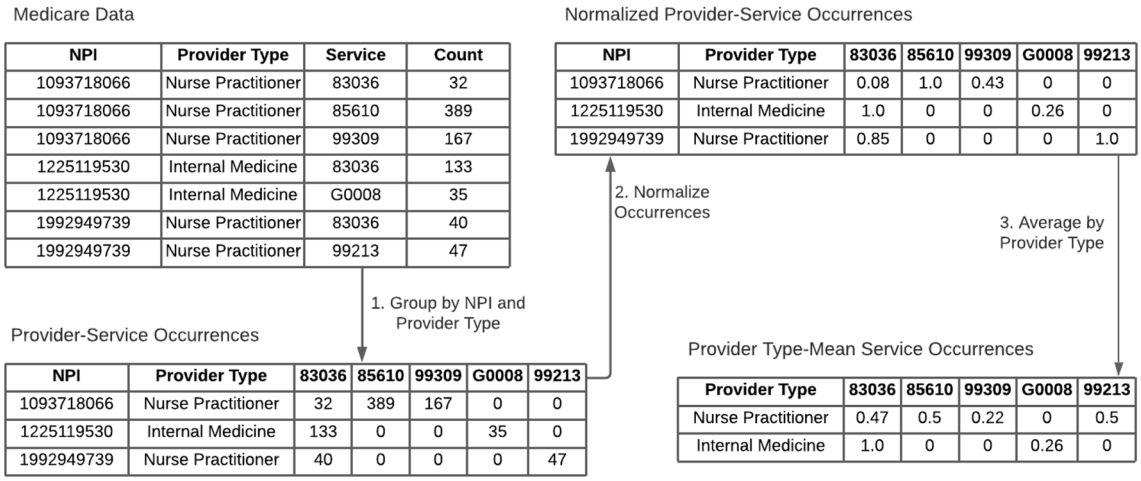

Fig. 1 HcpcsVec and RxVec embedding techniques

a specialty-drug occurrence matrix is created by aggregating the provider-drug occurrence matrix on the provider type attribute and taking the average number of prescription claim counts for each drug name. This produces a specialty-drug occurrence matrix of size 103×3545, where each row describes a specific specialty group using the average number of prescription claims for each drug. Similar to HcpcsVec, PCA is used to reduce the dimensionality of the occurrence vectors to lengths of 32, 64, and 103, where the maximum dimension size is again limited by the total number of unique provider types. The process of constructing RxVec embeddings for medical provider types is achieved through the steps outlined in Fig. 1, where the Service column corresponds to the prescription drug name attribute in the Part D data set.

Performance Evaluation

Embedding techniques are first evaluated on the Medicare Part B and Medicare Part D fraud classification tasks. Hyperparameters for the LR, RF, and GBT learner are selected by maximizing performance on the training partition using fivefold cross-validation. For the LR learner, we set the maximum number of iterations to 200. For the RF and GBT ensembles, we restrict trees to a maximum depth of eight, and for the GBT learner we use the exponential loss. The fourth learner is a MLP neural network with two hidden layers and 32 neurons per layer. The MLP is implemented using the Keras deep learning library [20], and following our previous work [36], we implement the MLP using two hidden layers, ReLU activations, batch normalization, a dropout rate of 0.5, batch sizes of 128, and the Adam optimizer with a learning rate of 0.001. Unless stated otherwise here, the remaining hyperparameters are left at their default values per scikit-learn v0.21.1 and Keras v2.4.0.

Each learner is trained on the Part B and Part D data sets using the one-hot, GloVe, and Med-W2V embeddings. HcpcsVec and RxVec are also evaluated on their respective data sets, Part B and Part D. To allow for comparison with related works [39], the HCPCS and Drug_name attributes are not used as predictors for training predictive models. A train-test split of 80–20% is used to score each classifier, and test results are reported using the AUC metric. We use this metric to evaluate results because the threshold-agnostic score is well suited for class-imbalanced data and previous works have reported AUC results for comparison [37, 42]. Each experiment is repeated 30 times, and a significance level of $\alpha = 0.05$ is used to report the margin of error (MOE) and Tukey's HSD results. Tukey's HSD test is a multiple comparison procedure that determines which method means are statistically different from each other by identifying differences that are greater than the expected standard error. Methods are assigned to alphabetic groups based on the statistical difference of AUC means, e.g. group a is significantly different from group b. In other words, HSD results allow us to assert with 95% confidence that differences in AUC scores did not occur by chance.

Next, the GBT learner's feature importance score is used to evaluate the provider specialty type's feature importance. The GBT feature importance uses the depth of decision nodes for specific features and the impurity measure to estimate the predictive power of each predictor variable [54]. Each feature is scored over the interval [0, 1] and then averaged over the ensemble of decision trees to produce the final result. By comparing feature importance scores across the Part B and Part D data sets using different embedding techniques, we can interpret how the importance of the provider type variable varies as we change its representation. For the GloVe and Med-W2V embeddings, we compute feature importance

scores using the embedding size of 128. For the HcpcsVec and RxVec embeddings, we compute feature importance scores using embedding sizes of 123 and 103, respectively. Since each embedding type will yield between 103 and 128 different feature importance scores, one for each dimension, we compute the cumulative feature importance by taking the sum over all dimensions. We do not compute feature importance scores for embedding sizes of 32 and 64 because the one-hot vectors have dimensionality of 103–123, and the differences in dimensionality may yield misleading results.

Finally, we explore provider type similarities in embedding space by plotting HcpcsVec and RxVec results using t-SNE visualizations. If our embedding techniques are effective at capturing specialty similarities, then similar specialties will have similar numeric representations and will form clusters. Possessing these semantic qualities will support the use of HcpcsVec and RxVec embeddings in related healthcare applications, such as those discussed in "Related Work".

Results and Discussion

Distributed representations for medical specialties are evaluated on two large-scale fraud classification tasks using publicly available Medicare data from the CMS. Classification results with embedding techniques are compared to traditional one-hot embeddings using the LR, RF, GBT, and MLP learners. Next, we evaluate how the embedding type affects predictive models by comparing each embedding technique's feature importance using the GBT learner. Finally, we visualize HcpcsVec and RxVec embeddings using t-SNE plots to confirm that the learned embeddings capture the desired semantic meanings.

Medicare Fraud Classification Results

Table 3 lists the mean AUC score and margin of error (MOE) for each embedding technique evaluated on the Medicare Part B data set. To establish a baseline, we first compare model performance without the provider type attribute to model performance with one-hot encoded provider types. All four models perform significantly better when the provider type predictor is included in the model. We see the largest increases in performance using the LR and MLP learners, with an average AUC gain of 0.063 each. The RF and GBT learners improve by 0.013 and 0.027, respectively. When using one-hot encoded vectors, the MLP model performs best with an average AUC score of 0.852.

Next, we compare the performance of two word embedding techniques, GloVe and Med-W2V. The LR model performs approximately the same with GloVe and Med-W2V embeddings, and it consistently performs better with larger embeddings of 64 and 128. For the RF, GBT, and MLP learners, no single embedding size or word embedding technique consistently outperforms another, and all differences in average AUC scores are minimal. The RF and GBT learners both significantly outperform their one-hot encoded baselines on average by as much as 0.024 and 0.020, respectively. The LR and MLP learners perform worse with word embeddings, and they see a decrease in AUC of 0.003 and 0.001, respectively. Overall, GloVe and Med-W2V perform approximately the same for each learner, and the GBT learner with the GloVe-64 embedding maximizes performance with an average AUC of 0.870.

The HcpcsVec embedding scored the highest AUC score overall for the LR, RF and GBT learners. The RF and GBT learners saw a maximum AUC gain of 0.026 and 0.023 compared to one-hot encodings, respectively. The LR learner performed significantly better with HcpcsVec-123, but the performance gain was marginal. The MLP learner performs

Table 3 Medicare Part B embedding performance (30 runs)

Embedding method	Embedding size	Mean AUC ± 95% MOE			
		LR	RF	GBT	MLP
None	0	0.750 ±7.6e−4	0.817 ±4.2e−4	0.823 ±1.1e−3	0.788 ±6.4e−3
One-hot	123	0.813 ±1.8e−4	0.830 ±3.7e−4	0.850 ±1.1e−3	**0.852** ±1.2e−3
GloVe	32	0.794 ±2.5e−4	0.854 ±9.8e−4	0.867 ±9.8e−4	0.847 ±2.1e−3
	64	0.810 ±2.0e−4	0.851 ±7.0e−4	0.870 ±6.9e−4	0.847 ±3.6e−3
	128	0.811 ±2.2e−4	0.851 ±7.3e−4	0.866 ±8.3e−4	0.843 ±2.2e−3
Med-Word2Vec	32	0.801 ±3.4e−4	0.853 ±4.0e−4	0.868 ±8.6e−4	0.848 ±2.0e−3
	64	0.810 ±2.8e−4	0.853 ±2.9e−4	0.869 ±8.9e−4	0.851 ±9.7e−4
	128	0.811 ±1.4e−4	0.850±3.4e−4	0.863 ±7.7e−4	0.844 ±3.0e−3
HcpcsVec	32	0.786 ±2.0e−4	**0.856** ±2.7e−4	**0.873** ±6.3e−4	0.849 ±4.1e−3
	64	0.810 ±3.6e−4	0.855 ±3.3e−4	0.864 ±6.1e−4	0.849 ±1.2e−3
	123	**0.814** ±2.8e−4	0.852 ±3.4e−4	0.865 ±9.3e−4	0.846 ±1.3e−3

best with the HcpcsVec-32 embedding, and while the MLP's highest AUC score is recorded using one-hot encoding, the difference in mean AUC scores is not statistically significant. The GBT model trained with the HcpcsVec-32 encoding scored the highest average AUC (0.873) on the Medicare fraud classification task and outperforms the best-known score of 0.851 from [36]. We believe that the HcpcsVec performs best overall on this Medicare task because it uses historical procedure-level data to capture provider type relationships instead of auxiliary natural language modeling tasks.

Table 4 lists the mean AUC score and MOE for each embedding technique evaluated on the Medicare Part D data set. Similar to the Part B classification results, we see that all four models perform significantly better when the provider type predictor is included in the model. The MLP learner has the largest increase in performance with one-hot encodings, increasing from 0.690 to 0.818 when adding the provider type feature with a one-hot encoding. The GBT learner performs second best when using the one-hot encoded provider types with an average AUC of 0.794, and the RF learner performs the worst with an average AUC of 0.756.

For the two word embedding techniques, GloVe and Med-W2V, the MLP learner performs best with a maximum AUC score of 0.830 when using Med-W2V with an embedding size of 64 and outperforms previous works [36]. The GBT model also obtains its best performance when using the Med-W2V embeddings, while the RF learner obtains its best performance with the GloVe embedding. Similar to Part B results, the LR learner performs worse with word embeddings than it does with one-hot encoding, and decreases significantly from an average AUC of 0.784 down to 0.768. For the remaining three learners, there is very little difference in AUC scores across the different embedding sizes.

Finally, we compare the RxVec embeddings across the four learners using the Part D data set. The RF, GBT, and MLP learners trained on RxVec embeddings perform significantly better than when trained with one-hot encodings. Compared to one-hot encoding results, the RF and GBT learners see the greatest increase in performance, increasing average AUC scores by 0.030 and 0.016, respectively. For the LR learner, the RxVec embeddings consistently outperform GloVe and Med-W2V, but do not perform better than the one-hot encodings. Unlike Part B results, the word embedding techniques GloVe and Med-W2V perform best on the Part D classification task. Nevertheless, all three semantic embedding techniques, including RxVec, significantly outperform one-hot encodings on the Part D fraud classification problem.

Table 5 summarizes the Part B and Part D classification results by taking the average AUC result from both data sets over all embedding sizes and computing Tukey's HSD groups. Methods with non-overlapping HSD groups have significantly different means with a confidence of 95%. We have grouped the two methods proposed in this study, HcpcsVec and RxVec, as they are both derived from publicly available claims data following a similar procedure. All three semantic embedding techniques belong to HSD group *a* and perform approximately the same with average AUC scores between 0.818 and 0.821. The one-hot encoding results have an average AUC of 0.812 and belong to HSD group *b*. Based on these results, the embeddings derived from historical claims data (HcpcsVec/RxVec) perform significantly better than results obtained with one-hot encodings

Table 5 Summary of fraud classification results

Embedding method	Mean AUC	HSD group
HcpcsVec/RxVec	0.821	a
Med-W2V	0.820	a
GloVe	0.818	ab
One-hot	0.812	b
None	0.755	c

Table 4 Medicare Part D embedding performance (30 runs)

Embedding method	Embedding size	Mean AUC ± 95% MOE			
		LR	RF	GBT	MLP
None	0	0.723 ±1.4e−4	0.720 ±6.8e−4	0.728 ±8.8e−4	0.690 ±9.8e−3
One-hot	123	**0.784** ±6.6e−4	0.756 ±8.3e−4	0.794 ±6.5e−4	0.818 ±1.6e−3
GloVe	32	0.763 ±6.0e−4	**0.787** ±5.7e−4	0.808 ±7.5e−4	0.820 ±2.2e−3
	64	0.768 ±5.5e−4	0.785 ±1.7e−4	0.807 ±7.3e−4	0.822 ±1.7e−3
	128	0.763 ±6.6e−4	0.784 ±2.1e−4	0.803 ±7.6e−4	0.821 ±2.1e−3
Med-Word2Vec	32	0.764 ±8.8e−4	0.786 ±1.1e−4	**0.812** ±7.1e−4	0.828 ±1.1e−3
	64	0.765 ±9.9e−4	0.785 ±1.3e−4	0.808 ±7.2e−4	**0.830** ±1.3e−3
	128	0.765 ±4.4e−4	0.784 ±1.1e−4	0.808 ±7.3e−4	0.825 ±1.7e−3
RxVec	32	0.772 ±5.5e−4	0.785 ±1.6e−4	0.806 ±6.9e−4	0.826 ±1.6e−3
	64	0.774 ±1.1e−3	0.786 ±2.1e−4	0.809 ±6.6e−4	0.825 ±2.1e−3
	123	0.772 ±5.8e−4	0.786 ±1.6e−4	0.810 ±5.4e−4	0.825 ±1.7e−3

on the Medicare fraud classification task. GloVe and one-hot embeddings both belong to HSD group *b* and perform similarly. Results obtained with no provider specialty type attributes belong to HSD group *c* with an average AUC of 0.755, performing the worst overall. In summary, the medical specialty attribute is a significant factor of classification performance, and semantic embedding techniques produce higher classification AUC results overall.

To illustrate the effect that specialty embeddings have on our predictive models, we use the GBT learner's built-in feature importance measure to compare the importance of provider type attributes across each embedding technique. Feature importance results for the Part B and Part D data set are illustrated in Fig. 2. For the Part D data set, the one-hot encoding technique yields a provider type feature importance of 0.199. When using semantic embeddings for provider types, the provider type feature importance increases significantly to 0.280–0.283, obtaining the highest importance when using the RxVec embedding. For the Part B data set, the one-hot encoding technique yields a provider type feature importance of 0.293. When we switch to the semantic embeddings, feature importance scores increase to 0.376–0.378, where the Med-W2V method obtains the highest feature importance overall. For both data sets, the GBT learner gives higher weight to the provider type feature set when we encode it using dense, semantic embeddings. This suggests that the semantic embeddings have encoded valuable characteristics about each provider specialty group that are conducive to training predictive models. Furthermore, these results align with the average AUC scores obtained in Tables 3 and 4, where the GBT learner trained with semantic embeddings consistently outperforms the GBT learner trained with one-hot encodings.

Visualizing Semantic Embeddings

To be effective in a variety of healthcare applications, the HcpcsVec and RxVec embeddings must encode meaningful characteristics from each provider specialty type. We evaluate embeddings in this section by plotting embeddings with t-SNE visualizations and comparing their similarities in feature space.

HcpcsVec embeddings with 123 dimensions are illustrated in Fig. 3, where we have annotated a sample of provider specialties for comparison. It is immediately clear that related provider types have similar embeddings, as they cluster accordingly in the t-SNE visualization. For example, in the lower-left quadrant we observe several specialty types related to the anesthesiology practice, i.e. CRNA, Certified Registered Nurse Anesthetist, and Anesthesiology Assistant. Cardiac surgery and thoracic surgery types, both of which entail chest and heart surgical procedures, are clustered closely together at the top of the figure. We observe five surgical specialty types clustered together in the upper right quadrant, two psychiatric specialty types in the upper left, and several cardiology specialty types in the lower right. This subset of results suggests that the HcpcsVec embeddings have succeeded in capturing the desired semantics.

RxVec embeddings with 103 dimensions are illustrated in Fig. 4. Similar to the HcpcsVec analysis, the t-SNE plot of RxVec embeddings illustrates that related provider types have created clusters in embedding space. In the upper-left corner there is a cluster of dental provider types, and in the lower left quadrant there is a cluster of surgical provider types. Four psychiatric and behavior related provider types have similar embeddings (top right), and three cardiology provider types have created a cluster in the lower-right quadrant. We also observe four child birth and pregnancy-related specialty types clustered near the center of the plot, i.e. midwife, obstetrics/gynecology, certified nurse midwife, and skilled nursing facility.

The t-SNE visualizations in this section confirm that both sets of embedding techniques presented in this study have successfully captured meaningful qualities from each provider type with a dense numeric representation. This is expected, as each embedding technique constructs its numeric representation for provider types from historical claims data, and each provider within a specialty group generally provides similar procedures and prescribes similar drugs. Given these qualities, we feel that these embeddings can be further applied to a wide range of data mining and machine learning applications in the healthcare industry that would benefit from provider specialty type features.

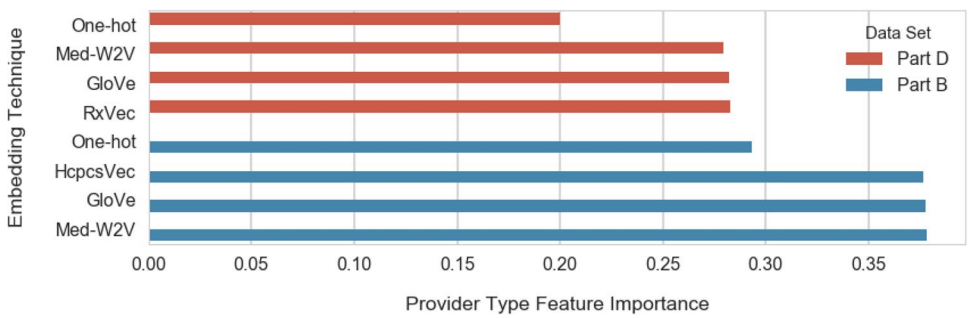

Fig. 2 Medical specialty feature importance using GBT learner

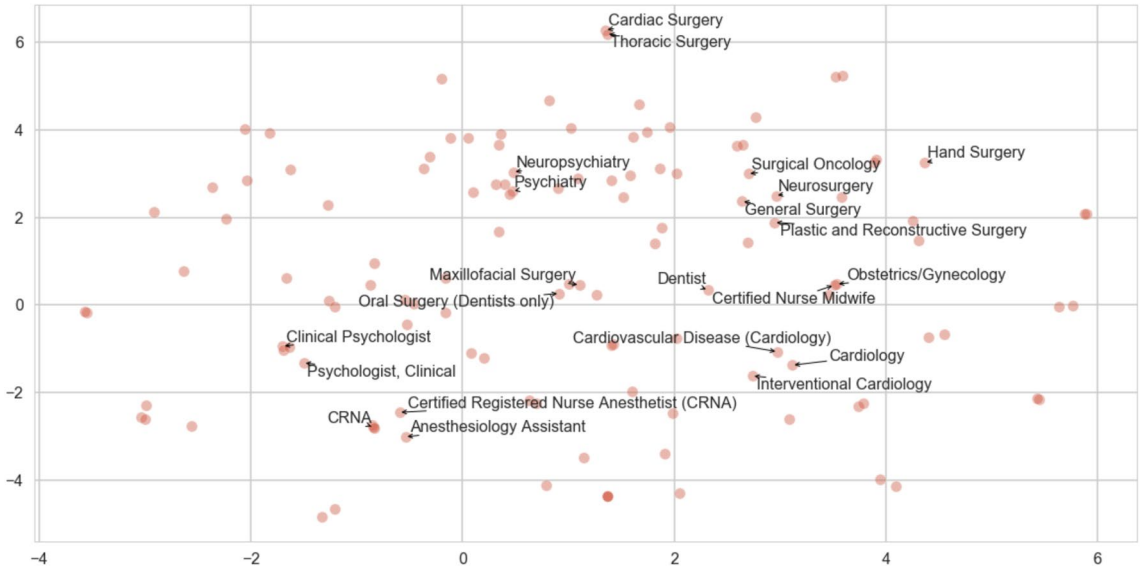

Fig. 3 t-SNE visualization of HcpcsVec embeddings

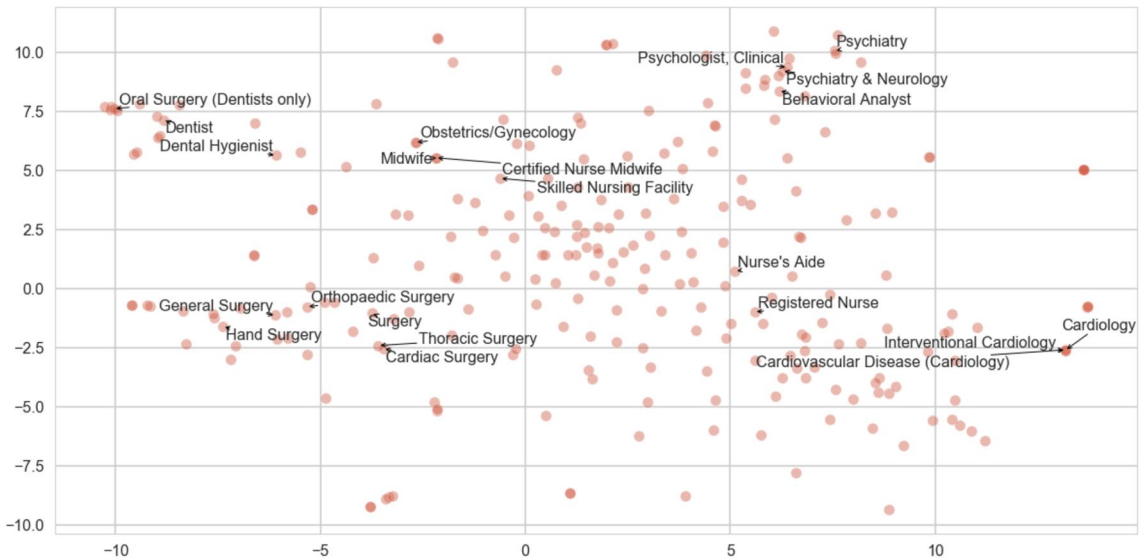

Fig. 4 t-SNE visualization of RxVec embeddings

Conclusion

This study presented four techniques for constructing low-rank semantic representations for medical specialty groups. Medical provider types were converted to phrase embeddings using pre-trained global word embeddings (GloVe) and pre-trained medical word embeddings (Med-W2V). From the Medicare Part B claims data set, HcpcsVec embeddings for provider types were generated from specialty-HCPCS occurrence matrices using 47 million data points. Similarly, RxVec embeddings for provider types were generated from specialty-drug occurrence matrices using 122 million data points from the Medicare Part D data set. Embeddings were evaluated on two large-scale Medicare fraud classification data sets. The RF and GBT learners perform best overall on the Medicare Part B data set using HcpcsVec embeddings, yielding average AUC scores of 0.856 and 0.857, respectively. On the Medicare Part D data set, the GBT and MLP learners obtain the best performance using Med-W2V embeddings, with respective average AUC scores of 0.812 and 0.830.

While RxVec embeddings did not perform best on the Medicare Part D data set, the RF, GBT, and MLP learners all performed significantly better when trained with RxVec embeddings versus one-hot embeddings. Feature importance results show that predictive models trained with semantic embeddings weight the provider type attributes more than when trained with one-hot encodings. Finally, t-SNE visualizations were provided to show how the HcpcsVec and RxVec specialty embeddings capture the inherent relationships between similar provider types.

Having been constructed from publicly available data, all four embedding techniques presented in this study lend themselves to future works in healthcare applications that utilize medical provider types and specialty groups. In addition to evaluating these embeddings on alternative medical benchmarks, future works will explore alternative encoding and dimension reduction techniques.

Acknowledgements The authors would like to thank the reviewers in the Data Mining and Machine Learning Laboratory at Florida Atlantic University.

Author Contributions JMJ performed the literature review, executed the experiment design, and drafted the manuscript. TMK worked with JMJ to develop the article's framework and focus. All authors have read and approved the final manuscript.

Funding Not applicable.

Data Availability All data analysed during this study are referenced in this published article.

Code Availability All software packages used during this study are open-source and are referenced in this published article.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Medicare Provider Utilization and Payment Data. Centers for Medicare & Medicaid Services. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/index 2020, Accessed 15 Feb 2020.
2. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado G.S, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X. TensorFlow: large-scale machine learning on heterogeneous systems. http://tensorflow.org/ 2015, Accessed 15 Feb 2020.
3. Aronson A, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc JAMIA. 2010;17:229–36. https://doi.org/10.1136/jamia.2009.002733.
4. Arora S, Liang Y, Ma T. A simple but tough-to-beat baseline for sentence embeddings. In: ICLR; 2017.
5. Bauder RA, Khoshgoftaar TM. A novel method for fraudulent medicare claims detection from expected payment deviations (application paper). In: 2016 IEEE 17th international conference on information reuse and integration (IRI); 2016. p. 11–19. https://doi.org/10.1109/IRI.2016.11.
6. Bauder RA, Khoshgoftaar TM, Richter A, Herland M. Predicting medical provider specialties to detect anomalous insurance claims. In: 2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI); 2016. p. 784–790. https://doi.org/10.1109/ICTAI.2016.0123.
7. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist. 2017;5:135–46. https://doi.org/10.1162/tacl_a_00051.
8. Branting L.K, Reeder F, Gold J, Champney T. Graph analytics for healthcare fraud risk estimation. In: 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM); 2016. p. 845–851. https://doi.org/10.1109/ASONAM.2016.7752336.
9. Centers For Medicare & Medicaid Services: Hcpcs general information. https://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/index.html 2018, Accessed 15 Feb 2020.
10. Centers for Medicare & Medicaid Services: medicare enrollment dashboard. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Dashboard/Medicare-Enrollment/Enrollment%20Dashboard.html 2019, Accessed 15 Feb 2020.
11. Centers For Medicare & Medicaid Services: medicare provider utilization and payment data. https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/medicare-provider-charge-data 2019, Accessed 15 Feb 2020.
12. Centers For Medicare & Medicaid Services: medicare provider utilization and payment data: physician and other supplier. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier 2020, Accessed 15 Feb 2020.
13. Centers For Medicare & Medicaid Services: medicare provider utilization and payment data: part d prescriber. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Part-D-Prescriber 2020, Accessed 15 Feb 2020.
14. Centers For Medicare & Medicaid Services: trustees report & trust funds. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/ReportsTrustFunds/index.html 2020, Accessed 15 Feb 2020.
15. Chandola V, Sukumar SR, Schryver JC. Knowledge discovery from massive healthcare claims data. In: KDD; 2013.
16. Chen L. Curse of dimensionality. Boston: Springer; 2009. p. 545–6. https://doi.org/10.1007/978-0-387-39940-9_133.
17. Choi E, Bahadori M.T, Song L, Stewart W.F, Sun J. Gram: graph-based attention model for healthcare representation learning. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, KDD '17, p. 787–795. Association for Computing Machinery, New York, NY, USA; 2017. https://doi.org/10.1145/3097983.3098126.
18. Choi E, Bahadori T, Searles E, Coffey C, Thompson M, Bost J, Tejedor-Sojo J, Sun J. Multi-layer representation learning for medical concepts. In: 22nd ACM SIGKDD international conference; 2016. p. 1495–1504. https://doi.org/10.1145/2939672.2939823.
19. Choi Y, Chiu CYI, Sontag DA. Learning low-dimensional representations of medical concepts. AMIA Summits Transl Sci Proc. 2016;2016:41–50.
20. Chollet F, et al. Keras. https://keras.io (2015), Accessed 15 Feb 2020.

21. Cost H.C.H., (HCUP), U.P. Clinical classifications software (ccs) for icd-9-cm. www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp 2017, Accessed 15 Feb 2020.
22. Das A, Ganguly D, Garain U. Named entity recognition with word embeddings and wikipedia categories for a low-resource language. ACM Trans Asian Lowresour Lang Inf Process. 2017. https://doi.org/10.1145/3015467.
23. De Vine L, Zuccon G, Koopman B, Sitbon L, Bruza P. Medical semantic similarity with a neural language model. In: Proceedings of the 23rd ACM international conference on conference on information and knowledge management, CIKM '14, p. 1819–1822. Association for Computing Machinery, New York, NY, USA; 2014. https://doi.org/10.1145/2661829.2661974.
24. Devlin J, Chang M.W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT; 2019.
25. Ferdous M, Debnath J, Chakraborty N.R. Machine learning algorithms in healthcare: a literature survey. In: 2020 11th International conference on computing, communication and networking technologies (ICCCNT); 2020. p. 1–6. https://doi.org/10.1109/ICCCNT49239.2020.9225642.
26. Fursov I, Zaytsev A, Khasyanov R, Spindler M, Burnaev E. Sequence embeddings help to identify fraudulent cases in healthcare insurance. ArXiv abs/1910.03072. 2019.
27. Gudivada A, Tabrizi N. A literature review on machine learning based medical information retrieval systems. In: 2018 IEEE symposium series on computational intelligence (SSCI); 2018. p. 250–257. https://doi.org/10.1109/SSCI.2018.8628846.
28. Hafiz AM, Bhat GM. A survey of deep learning techniques for medical diagnosis. In: Tuba M, Akashe S, Joshi A, editors. Information and communication technology for sustainable development. Singapore: Springer; 2020. p. 161–70.
29. Hancock JT, Khoshgoftaar TM. Survey on categorical data for neural networks. J Big Data. 2020;7(1):28. https://doi.org/10.1186/s40537-020-00305-w.
30. Herland M, Bauder RA, Khoshgoftaar TM. Medical provider specialty predictions for the detection of anomalous medicare insurance claims. In: 2017 IEEE international conference on information reuse and integration (IRI); 2017. p. 579–588. https://doi.org/10.1109/IRI.2017.29.
31. Herland M, Khoshgoftaar TM, Bauder RA. Big data fraud detection using multiple medicare data sources. J Big Data. 2018;5(1):29. https://doi.org/10.1186/s40537-018-0138-3.
32. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. J Big Data. 2014;1(1):2. https://doi.org/10.1186/2196-1115-1-2.
33. Huang K, Altosaar J, Ranganath R. Clinicalbert: modeling clinical notes and predicting hospital readmission. ArXiv abs/1904.05342. 2019.
34. Jeyaraj PR, Nadar ERS. Smart-monitor: patient monitoring system for IoT-based healthcare system using deep learning. IETE J Res. 2019. https://doi.org/10.1080/03772063.2019.1649215.
35. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2(4):230–43. https://doi.org/10.1136/svn-2017-000101.
36. Johnson JM, Khoshgoftaar TM. Deep learning and data sampling with imbalanced big data. In: 2019 IEEE 20th international conference on information reuse and integration for data science (IRI); 2019. p. 175–183.
37. Johnson JM, Khoshgoftaar TM. Deep learning and thresholding with class-imbalanced big data. In: 2019 18th IEEE international conference on machine learning and applications (ICMLA); 2019. p. 755–762. https://doi.org/10.1109/ICMLA.2019.00134.
38. Johnson JM, Khoshgoftaar TM. Medicare fraud detection using neural networks. J Big Data. 2019;6(1):63. https://doi.org/10.1186/s40537-019-0225-0.
39. Johnson JM, Khoshgoftaar TM. The effects of data sampling with deep learning and highly imbalanced big data. Inf Syst Front. 2020;22(5):1113–31. https://doi.org/10.1007/s10796-020-10022-7.
40. Johnson JM, Khoshgoftaar TM. Hcpcs2vec: healthcare procedure embeddings for medicare fraud prediction. In: 2020 IEEE 6th international conference on collaboration and internet computing (CIC); 2020.
41. Johnson JM, Khoshgoftaar TM. Semantic embeddings for medical providers and fraud detection. In: 2020 IEEE 21st international conference on information reuse and integration for data science (IRI); 2020. p. 224–230. https://doi.org/10.1109/IRI49571.2020.00039.
42. Johnson JM, Khoshgoftaar TM. Thresholding strategies for deep learning with highly imbalanced big data. Singapore: Springer; 2021. p. 199–227. https://doi.org/10.1007/978-981-15-6759-9_9.
43. Kalyan KS, Sangeetha S. Secnlp: a survey of embeddings in clinical natural language processing. J Biomed Inform. 2020;101:103323. https://doi.org/10.1016/j.jbi.2019.103323.
44. Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F. A survey of word embeddings for clinical text. J Biomed Inform X. 2019;4:100057. https://doi.org/10.1016/j.yjbinx.2019.100057. http://www.sciencedirect.com/science/article/pii/S2590177X19300563.
45. Ko J, Chalfin H, Trock B, Feng Z, Humphreys E, Park SW, Carter B, Frick DK, Han M. Variability in medicare utilization and payment among urologists. Urology. 2015. https://doi.org/10.1016/j.urology.2014.11.054.
46. Linux S. About. https://www.scientificlinux.org/about/ (2014), Accessed 15 Jan 2020.
47. Ma F, You Q, Xiao H, Chitta R, Zhou J, Gao J. Kame: knowledge-based attention model for diagnosis prediction in healthcare. In: Proceedings of the 27th ACM international conference on information and knowledge management, CIKM '18, p. 743–752. Association for Computing Machinery, New York, NY, USA; 2018. https://doi.org/10.1145/3269206.3271701.
48. Maas A, Daly R.E, Pham P.T, Huang D, Ng A.Y, Potts C. Learning word vectors for sentiment analysis. In: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies; 2011. p. 142–150.Accessed 15 Feb 2020.
49. Maaten LVD, Hinton G. Visualizing data using t-sne. J Mach Learn Res. 2008;9:2579–605.
50. Mikolov T, Chen K, Corrado GS, Dean J. Efficient estimation of word representations in vector space. CoRR abs/1301.3781. 2013.
51. Morris L. Combating fraud in health care: an essential component of any cost containment strategy. Health Aff. 2009;28:1351–6. https://doi.org/10.1377/hlthaff.28.5.1351.
52. National Plan & Provider Enumeration System: Nppes npi registry. https://npiregistry.cms.hhs.gov/registry/ 2020, Accessed 15 Feb 2020.
53. Office of Inspector General: Leie downloadable databases. https://oig.hhs.gov/exclusions/exclusions_list.asp (2019).
54. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
55. Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. In: Empirical methods in natural language processing (EMNLP); 2014. p. 1532–1543.
56. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. In:

Proceedings of the 2018 conference of the North american chapter of the association for computational linguistics: human language technologies, vol. 1 (long papers), p. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana; 2018. https://doi.org/10.18653/v1/N18-1202.
57. Pianykh OS, Guitron S, Parke D, Zhang C, Pandharipande P, Brink J, Rosenthal D. Improving healthcare operations management with machine learning. Nat Mach Intell. 2020;2(5):266–73. https://doi.org/10.1038/s42256-020-0176-3.
58. Provost F, Fawcett T. Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In: Proceedings of the third international conference on knowledge discovery and data mining; 1999. p. 43–48.
59. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. Distributional semantics resources for biomedical text processing. In: Proceedings of languages in biology and medicine; 2013.
60. Rajpurkar P, Zhang J, Lopyrev K, Liang P. Squad: 100, 000+ questions for machine comprehension of text. In: EMNLP; 2016.
61. Raunak V, Gupta V, Metze F. Effective dimensionality reduction for word embeddings. In: Proceedings of the 4th workshop on representation learning for NLP (RepL4NLP-2019), p. 235–243. Association for Computational Linguistics, Florence, Italy; 2019. https://doi.org/10.18653/v1/W19-4328.
62. Sahlgren M. The distributional hypothesis. Ital J Linguist. 2008;20:33–54.
63. Shailaja K, Seetharamulu B, Jabbar M. Machine learning in healthcare: a review. In: 2018 Second international conference on electronics, communication and aerospace technology (ICECA), IEEE; 2018. p. 910–914.
64. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. J Am Med Inform Assoc JAMIA. 2019. https://doi.org/10.1093/jamia/ocz096.
65. Song L, Cheong C.W, Yin K, Cheung W.K, Fung B.C.M, Poon J. Medical concept embedding with multiple ontological representations. In: Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI-19, p. 4613–4619. International Joint Conferences on Artificial Intelligence Organization; 2019. https://doi.org/10.24963/ijcai.2019/641.
66. Sun J, Chen X, Zhang Z, Lai S, Zhao B, Liu H, Wang S, Huan W, Zhao R, Ng MTA, Zheng Y. Forecasting the long-term trend of covid-19 epidemic using a dynamic model. Sci Rep. 2020;10(1):21122. https://doi.org/10.1038/s41598-020-78084-w.
67. Tukey JW. Comparing individual means in the analysis of variance. Biometrics. 1949;5(2):99–114.
68. U.S. Government, U.S. Centers for Medicare & Medicaid Services: the official U.S. government site for medicare. https://www.medicare.gov/. Accessed 15 Feb 2020.
69. Villarroel M, Reisner A, Clifford G, Lehman LW, Moody G, Heldt T, Kyaw T, Moody B, Mark R. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. Crit Care Med. 2011;39:952–60. https://doi.org/10.1097/CCM.0b013e31820a92c6.
70. Wang M, Zhang Q, Lam S, Cai J, Yang R. A review on application of deep learning algorithms in external beam radiotherapy automated treatment planning. Front Oncol. 2020;10:2177. https://doi.org/10.3389/fonc.2020.580919.
71. Witten IH, Frank E, Hall MA, Pal CJ. Data mining, fourth edition: practical machine learning tools and techniques. 4th ed. San Francisco: Morgan Kaufmann Publishers Inc.; 2016.
72. Zou WY, Socher R, Cer D, Manning CD. Bilingual word embeddings for phrase-based machine translation. In: Proceedings of the 2013 conference on empirical methods in natural language processing; 2013. p. 1393–1398.