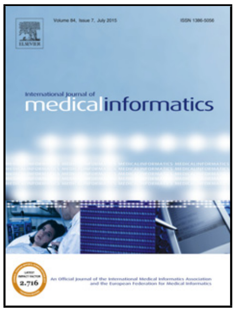

Automated ICD coding via unsupervised knowledge integration (UNITE)

Aaron Sonabend W, Winston Cai, Yuri Ahuja, Ashwin Ananthakrishnan, Zongqi Xia, Sheng Yu, Chuan Hong

PII: S1386-5056(19)31302-4

DOI: https://doi.org/10.1016/j.ijmedinf.2020.104135

Reference: IJB 104135

To appear in: *International Journal of Medical Informatics*

Received Date: 19 November 2019

Revised Date: 14 February 2020

Accepted Date: 26 March 2020

Automated ICD Coding via Unsupervised Knowledge Integration (UNITE)

Aaron Sonabend W.[1∗], Winston Cai[2∗], Yuri Ahuja[1], Ashwin Ananthakrishnan[3], Zongqi Xia[4], Sheng Yu[5,6,7*], Chuan Hong[8,*]

[1]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA

[2]Bronx Science, New York City, NY, USA

[3]Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School

[4]Department of Neurology and Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA

[5]Center for Statistical Science, Tsinghua University, Beijing, China

[6]Department of Industrial Engineering, Tsinghua University, Beijing, China

[7]Institute for Data Science, Tsinghua University, Beijing, China

[8]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

∗Sonabend and Cai contributed equally

*Yu and Hong contributed equally

Highlights

- Unsupervised ICD coding without requiring human labor
- Analyzing clinical narrative notes via semantic relevance assessment
- Stable performance and high portability across EMRs in different institutions.

ABSTRACT

Objective: Accurate coding is critical for medical billing and electronic medical record (EMR)-based research. Recent research has been focused on developing supervised methods to automatically assign International Classification of Diseases (ICD) codes from clinical notes. However, supervised approaches rely on ICD code data stored in the hospital EMR system and is subject to bias rising from the practice and coding behavior. Consequently, portability of trained supervised algorithms to external EMR systems may suffer.

Method: We developed an unsupervised knowledge integration (UNITE) algorithm to automatically assign ICD codes for a specific disease by analyzing clinical narrative notes via semantic relevance assessment. The algorithm was validated using coded ICD data for 6 diseases from Partners HealthCare (PHS) Biobank and Medical Information Mart for Intensive Care (MIMIC-III). We compared the performance of UNITE against penalized logistic regression (LR), topic modeling, and neural network models within each EMR system. We additionally evaluated the portability of UNITE by training at PHS Biobank and validating at MIMIC-III, and vice versa.

Results: UNITE achieved an averaged AUC of 0.91 at PHS and 0.92 at MIMIC over 6 diseases, comparable to LR and MLP. It had substantially better performance than topic models. In regards to portability, the performance of UNITE was consistent across different EMR systems, superior to LR, topic models and neural network models.

Conclusion: UNITE accurately assigns ICD code in EMR without requiring human labor, and has major advantages over commonly used machine learning approaches. In addition, the UNITE attained stable performance and high portability across EMRs in different institutions.

Keywords: automated ICD assignment; electronic medical records; portability; semantic embedding; knowledge integration; unsupervised learning

INTRODUCTION

A key task in mining Electronic Medical Records (EMR) is to properly assign International Classification Diseases (ICD) codes. While ICD codes are important for making clinical and billing decisions, manual assignment of proper ICD codes to a patient encounter is time-consuming, error-prone and expensive. In current practice, after reviewing the diagnosis entered by physicians and other information pertaining to a visit, medical coders manually assign the most appropriate ICD codes according to coding guidelines. However, miscoding and unbundling errors (i.e., charges that should be bundled together were entered separately) frequently occur during the manual coding given the imprecision due to abbreviations, synonyms, and the existing hierarchy of ICD codes [1] [2]. The financial cost for improving coding quality and for reducing coding errors is estimated to be $25 billion per year in the US [3] [4].

Prior methods to automatically assign ICD codes to clinical documents via supervised machine learning approaches during the last two decades included logistic regression, K-nearest neighbors, naïve Bayes, support vector machines, Bayesian ridge regression and deep learning, all showing promising results [5] [6] [7] [8]. In general, these supervised approaches all rely on ICD code data stored in the EMR. Although supervised learning algorithms are usually more accurate than unsupervised learning by leveraging existing ICD codes within the same healthcare system, they are subject to bias arising from the practice and coding behavior of the specific healthcare system. Consequently, supervised algorithms trained in one system often have poor portability to external healthcare systems [9].

Existing unsupervised learning methods, including rule-based approaches using mentions of disease and topic modeling for medical note classification, have their limitations. Rule-based approaches usually have low specificity, whereas topic modeling approaches [10] have low accuracy because the topic can be too general for specific diseases [11] and the optimal number of topics is generally unknown. Interestingly, Kavaluru 2013 et al. [12] reported an unsupervised approach to extract ICD codes from medical notes by using a combination of named entity recognition (NER), knowledge-based graph mining, and extractive text summarization. However, this approach essentially maps concept unique identifier (CUI) to ICD codes, while a large amount of CUIs were not utilized, which leads to loss in accuracy.

Embedding approaches such as *word2vec* and *GloVE* that use distributed representations of words have become increasingly popular [13] [14]. Different from conventional dictionary-based approaches, by exploiting semantic similarities between words, these approaches produce dense vector representation of words that usually lead to improved accuracy in machine learning tasks. More recently, Beam et al. (2018) constructed a comprehensive set of embeddings for medical concepts, referred to as cui2vec, by combining large sources of multimodal healthcare data [15].

In this paper, we aimed to develop an "unsupervised knowledge integration" (UNITE) algorithm to predict ICD codes by analyzing clinical narrative notes via semantic relevance assessment, without requiring coded ICD data for training. We hypothesize and validate that the proposed algorithm will retain much of the accuracy of the supervised methods and show greater accuracy than the existing unsupervised methods when testing within the same EMR system, while performing better in portability than the supervised methods across different EMR systems.

METHODS

A flow-chart summary of UNITE is shown in Figure 1. For a target disease $\mathcal{D}$, UNITE predicts the ICD code that should be assigned per visit for any specific disease $\mathcal{D}$ in three key steps: knowledge extraction via online sources, EMR and knowledge source text analysis via natural language processing (NLP) and embeddings, and diagnosis assessment by summarizing notes.

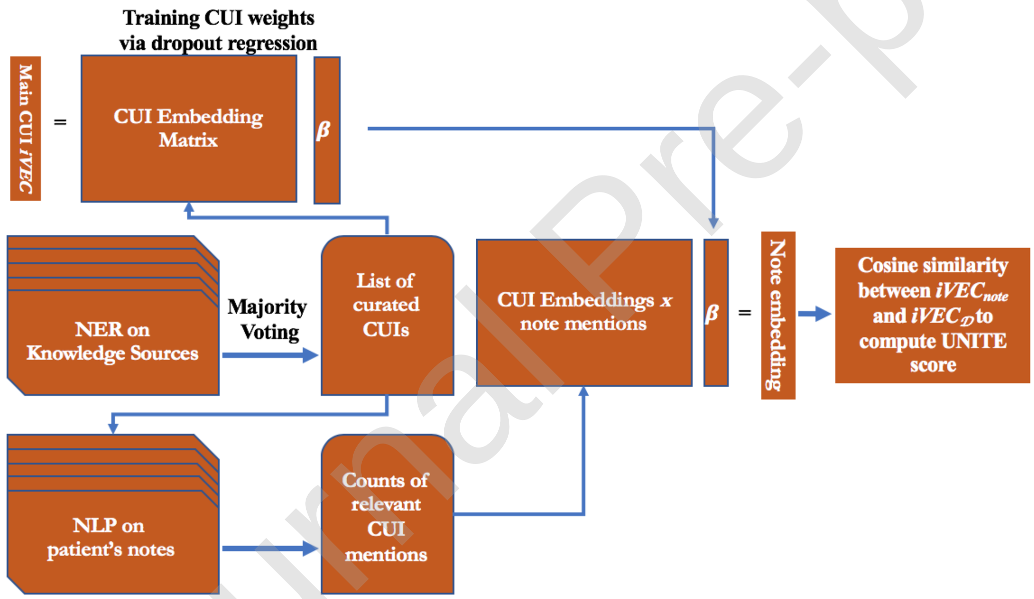

Figure 1. Workflow of UNITE in predicting the presence of ICD for one specific disease. NLP: natural language processing; NER: named entity recognition; CUI: concept unique identifier. β is the regression coefficient which serves as the weights for importance of the CUIs.

Step I. Knowledge Integration with Online Sources

Clinical terms relevant to $\mathcal{D}$ were extracted via NER as in Yu et al. (2017) [16], from five online knowledge sources (KS) on $\mathcal{D}$, including Wikipedia, Medscape eMedicine, Merck Manuals

Professional Edition, MedlinePlus Medical Encyclopedia, and Mayo Clinic Diseases and Conditions. The identified terms were mapped to their corresponding Concept Unique Identifiers (CUIs) listed in the Unified Medical Language System (UMLS) to assemble a CUI list for $\mathcal{D}$. CUIs that appeared in at least two of the five online sources were retained in candidate CUI set C for predictive modeling for the target disease. More details can be found in Appendix A of the Supplementary Materials.

Step II. EMR and Knowledge Source Text Analysis

Semantic vectors (VECs) were trained from PHS biobank-linked EMR for the concepts in the EMR corpus using the *cui2vec* method [17]. Specifically, notes from EMR were parsed using the NLP software NILE [18] to extract positive mentions of concepts in the EMR corpus. A CUI-CUI pointwise mutual information (PMI) matrix was then constructed with cooccurrence data using a 30-day window:

$$PMI\ (w,c) = log\frac{p(w,c)}{p(w)*p(c)}, \qquad (1)$$

where $p(w,c)$ is the number of times CUI w and CUI c occur in the same context window divided by the total number of CUI-context pairs (i.e., pairs of CUI w and other CUIs in the context) whereas $p(w)$, $p(c)$ were individual CUI frequencies. We further constructed the shifted PMI matrix – a sparse matrix defined by

$$SPPMI\ (w,c) = max(PMI(w,c) - \log(k)\,, 0), \qquad (2)$$

where k is the number of negative samples in the original *word2vec* paper ($k = 5$ to 20 works for smaller dataset, while $k = 2$ to 5 works for larger dataset). We finally conducted singular value decomposition (SVD) to factorize the SPPMI matrix, and used the resulting left-singular vectors as CUI VECs. Together with published semantic VECs pre-trained from PubMed Central articles and the Stanford EMR [19], we represented each CUI using an integrated VEC (iVEC) by concatenating the three L2-normalized CUI VECs (PHS biobank EMR, PubMed Central and Stanford EMR with 1000, 500 and 300 dimensions respectively) to create large-dimension embeddings for the CUIs that contain semantic information from several sources.

To determine the relevance of a CUI to a specific disease, the iVEC of the main CUI for disease $\mathcal{D}$ (the CUI associated to the disease in question) was regressed against the iVECs of all CUIs selected in Step I as features, applying penalized linear regression and randomly replacing 50% of the entries with the column mean (dropout training) [20],

$$iVEC_{main} = \beta' iVEC_{all} + \epsilon, \ \ (3)$$

where β is the regression coefficient serving as importance weights, and ϵ is a random error term. As an example, for the main CUI of rheumatoid arthritis (C0003873), the iVEC of C0003873 was regressed against the iVEC of all the CUIs related to rheumatoid arthritis selected through the KS (including C0003873 itself, the dropout guaranteed that some of the weights for CUIs other than C0003873 would be different that zero).

For the j^{th} CUI, we defined the CUI importance W_j as the product of the regression coefficient and the logarithm of the term frequency of $\mathrm{CUI_j}$ in the KS articles:

$$W_j = \beta_j \times \log\left(\text{Frequency of } \mathrm{CUI_j} + 1\right). \ \ (4)$$

Step III. Diagnosis Assessment by Summarizing Notes

A note level iVEC was calculated as the average of iVECs of candidate CUIs weighted by the product of the term frequency-inverse document frequency (TFIDF) of the CUIs in the note and W:

$$iVEC_{note} = \frac{\sum_{c \in C} W_c \cdot TFIDF_{c,note} \cdot iVEC_c}{\sum_{c \in C} W_c \cdot TFIDF_{c,note}}, \quad (5)$$

where W_c was the importance of CUI C, and $IDF_{c,note}$ was obtained from the EMR corpus.

A reference iVEC for $\mathcal{D}$ was created in a similar fashion by treating the KS articles combined as a single note:

$$iVEC_{\mathcal{D}} = \frac{\sum_{c \in C} W_c \cdot TFIDF_{c,\mathcal{D}} \cdot iVEC_c}{\sum_{c \in C} W_c \cdot TFIDF_{c,\mathcal{D}}}, \quad (6)$$

where $IDF_{c,\mathcal{D}}$ was based on the five KS articles. UNITE scored the likelihood of having an ICD code for $\mathcal{D}$ as the cosine similarity between $iVEC_{note}$ and $iVEC_{\mathcal{D}}$.

Performance Assessment of UNITE using EMR data

Partners HealthCare Biobank (PHS) contained EMR data anchored by two large tertiary care hospitals: Brigham and Women’s Hospital and Massachusetts General Hospital in Boston [21].

The PHS EMR consisted of both structured data (e.g., ICD code) and unstructured clinical notes for both inpatients and outpatients. Notes might not explicitly list the patients' diagnoses like in discharge summaries. However, each visit had its own ICD code in the structured data that served as the gold standard, and we aimed to predict the ICD code that should be assigned per visit. Because the total number of notes was extraordinarily large, we randomly sampled 1,590 patients from PHS Biobank, with 193,677 clinical notes starting from June 3rd, 1987 to March 4th, 2016. The choice of our sample size was based on the total number of concepts included for each disease. Across the 6 diseases, the average number of CUIs was 702.

Medical Information Mart for Intensive Care III (MIMIC III) was a large, publicly available database that contained de-identified health-related data from over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center in Boston between 2001 and 2012 [22]. The MIMIC consists of inpatient data; thus we used the discharge summaries, which explicitly listed the patients' diagnoses. We tested 52,691 discharge summary notes from the 46,520 available patients. NLP was run on the notes to obtain counts of disease-relevant CUIs and discharge diagnoses.

To evaluate the performance of the UNITE, we used both PHS and MIMIC to construct datamarts for conducting automated ICD coding of six diseases, including rheumatoid arthritis (RA), coronary artery disease (CAD), lung cancer (LC), multiple sclerosis (MS), ulcerative colitis (UC), and Crohn's disease (CD). The prevalence of an ICD code at visit levels for RA, CAD, LC, MS, UC, and CD were estimated at 3.0%, 7.7%, 0.4%, 0.3%, 0.6% and 0.5% at PHS, and 1.3%, 31.0%, 0.2%, 0.6%, 0.5% and 0.7% at MIMIC, respectively.

Methods for Comparison

Three baseline models were considered for comparison. (1) Penalized Logistic Regression (LR). Lasso LR was trained with the presence of ICD of a specific disease as outcome (which was readily available in the EMR) and the logarithm of the count of each CUI mention in the note as features. Cross validation was used for tuning the regularization parameter. (2) Multilayer perceptron (MLP). We trained a dense neural network with two hidden layers, each with 128 units, and regularized with 40% and 30% dropout rates, respectively. (3) Topic Models. We ran three types of topic models to cluster patient's notes into two groups (topics), including latent Dirichlet allocation (LDA) model via variational expectation maximization (VEM), LDA with Gibbs sampling for estimating the posterior distributions, and correlated LDA with VEM. The estimated probability of the topic that had the most cases was used for prediction. It is worth noting that the features used in the comparison models were based on the KS selection of CUIs.

To understand the effect of the various components of UNITE, we additionally tested supervised models using UNITE's features (the iVECs) as the input. (1) The *UNITE-LR* and *UNITE-MLP* use as input features the note vector derived from UNITE, that is, the note level vector $iVEC_{note}$. (2) The *mCUI-LR* and *mCUI-MLP* use as input features the mean CUI embedding. These were based on the methods for representing clinical text by Catling et al (2018) [23], where the mean CUI embedding is defined as the sum of the iVECs weighted by TFIDF. (3) The *TFIDF-LR* and *TFIDF-MLP* use as input features the CUI counts in the notes, normalized by inverse document frequency.

Furthermore, we implemented several state-of-the-art methods to the MIMIC data, including a recurrent neural network (NN), convolutional NN, gated recurrent unit NN and a long short-term memory NN model. The models and architectures chosen are based on Huang et al. (2019) [24], we focus on results from the best performing models in the mentioned paper. The architectures of these networks are shown in Table S1 of the Supplementary Materials. More details of the methods of comparison are shown in Appendix B of the Supplementary Materials.

Overview of analyses

We compared the performances of UNITE and the comparison models in predicting ICD for six diseases, including RA, CAD, LC, MS, UC, and CD, using data from both PHS and MIMIC-III EMRs. We split the dataset from each EMR into training, validation and test sets with 60%, 20% and 20% of the data respectively. To test the portability of the methods, all methods were trained separately on each of the two sources and tested in held-out sets of both sources. Prediction performance of each algorithm was reported based on the following metrics evaluated using a validation set with labels: the area under the receiver operating characteristic curve (AUC), positive predictive value (PPV), sensitivity (SE), F-score, and negative predictive value (NPV), where the thresholds for binary classifier were chosen to maximize the F-scores for classification of cases. As the deep NN models require computational resources that were not available at PHS cluster, we could not train models in their EMR system. However, we assessed the portability of these methods by training the models using the MIMIC-III dataset and test on both MIMIC-III and PHS.

RESULTS

Relative CUI Importance

UNITE allows for an interpretable evaluation of the results. Figure 2 shows the relative CUI importance for each disease assessed by UNITE using a word cloud representation. The magnitude of the font size is proportional to the absolute value of the CUI importance W. The results were largely consistent with clinical knowledge. Take CAD for example, the three most important CUIs besides the main CUIs were myocardial infarction, acute coronary syndrome and angina, which were all highly indicative of CAD. In the cases of UC and CD, which are important differential diagnosis to each other and should have large negative coefficients [25], we see that they both got a large positive weight in predicting themselves, and a large negative weight in predicting the other.

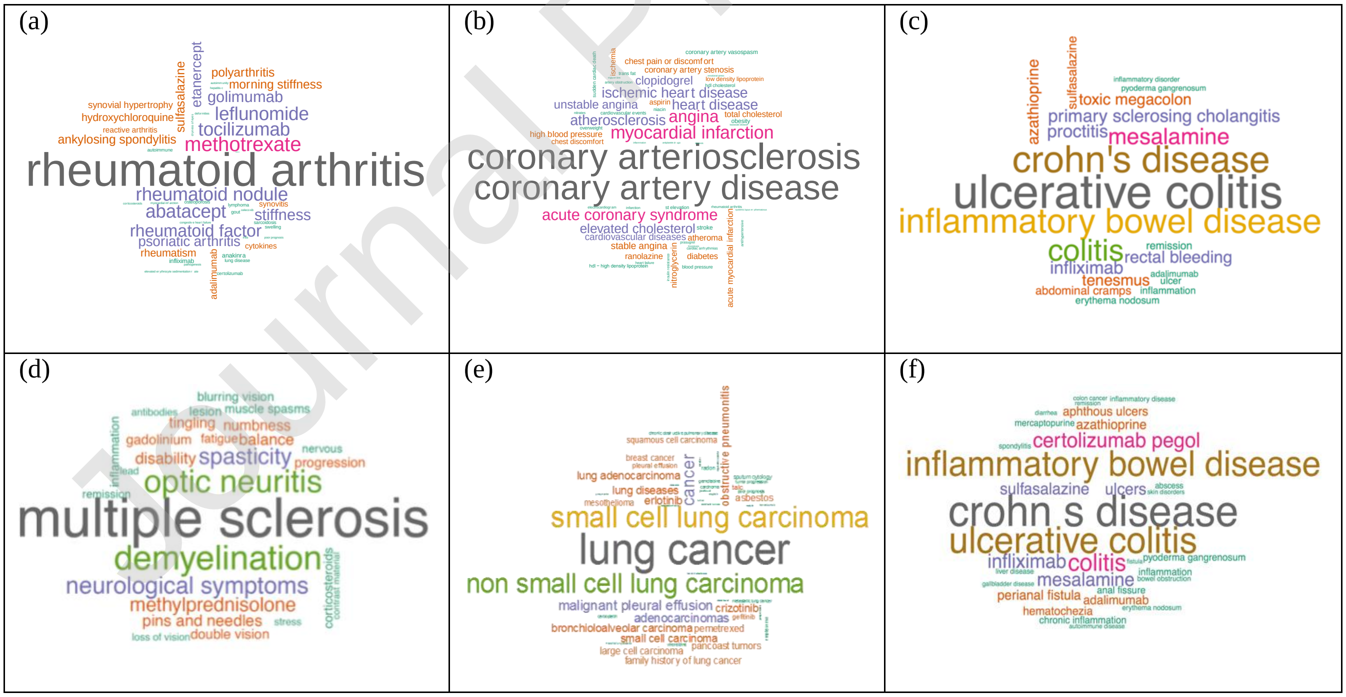

Figure 2. Relative CUI importance for RA, CAD, UC, MS, LC, CD using a word cloud representation with the magnitude of the font size proportional to the CUI importance.

Below, we show performance metrics for methods trained and tested in the same EMR system and across EMR systems. Predictive probability thresholds were chosen to maximize the F-scores for classification of cases.

Validation within the same EMR system

The two left panels in Figure 3 summarize the AUC of UNITE and baseline methods within the same EMR system. Across all diseases, UNITE predicted ICD codes with overall comparable accuracy when compared to baseline models LR and MLP, and performed substantially better than the unsupervised counterparts: the three topic modeling approaches. Among the supervised learning methods with UNITE features as input, we observe that UNITE-MLP and mCUI-MLP have a slight increase in performance over the regular MLP, presumably due to the advantage of introducing embeddings.

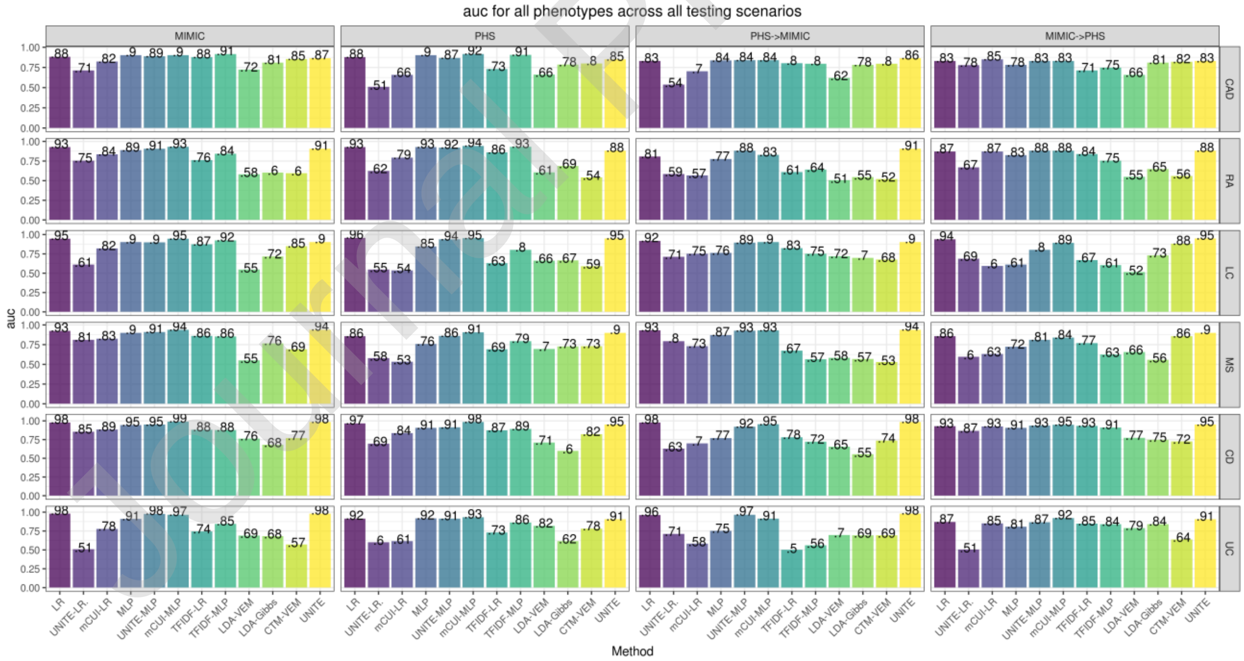

Figure 3. AUC for penalized logistic regression (LR), penalized logistic regression with UNITE (UNITE-LR), penalized logistic regression with the mean CUI (mCUI-LR), penalized logistic regression with the TFIDF (TFIDF-LR), multiple layer perceptron (MLP), multiple layer

perceptron with UNITE features (UNITE-MLP), multiple layer perceptron with the mean CUI (mCUI-MLP), multiple layer perceptron with TFIDF (TFIDF-MLP), topic modeling based on LDA with 2 topics fit with VEM (LDA_VEM) and Gibbs (LDA_Gibbs), CTM with 2 topics fit with VEM (CTM_VEM) and UNITE. Each row is a different disease, columns show performance for methods trained and tested on either the same hospital system (first two panels), or trained and tested in different hospital systems (last two panels).

Similarly, UNITE-predicted binary classification (yes versus no) had an accuracy comparable to the supervised approach (LR and MLP). Across the six diseases, UNITE classified all diseases with a higher F-score than LDA. As shown in Figure 4, using MS at MIMIC as an example, F-scores were 0.76 vs. 0.06, 0.07 and 0.07 for UNITE and the three topic models, respectively. The results for the other three metrics including PPV, sensitivity and NPV are displayed in Figures S1-S3 and Table S2 with confidence intervals estimated from 200 bootstraps in the Supplementary Materials.

The results of deep NN are summarized in Table S3 of the Supplementary Materials. Deep NN show above average performance when testing within EMR system. Particularly in the case of the gated recurrent unit NN (GRU). For example, the AUCs for RNN (ranging from 0.53 to 0.82) were systematically lower than those of UNITE (ranging from 0.87 to 0.98), which is expected due to the long sequences in the clinical notes. CNN had higher AUCs for CAD and LC (0.96 and 0.99, respectively) compared with UNITE, but lower AUCs for the remaining diseases. The GRU model performs relatively better among the state-to-art methods, and better than UNITE except for UC and RA. Finally, UNITE is better or as good as LSTM in RA, CD and UC.

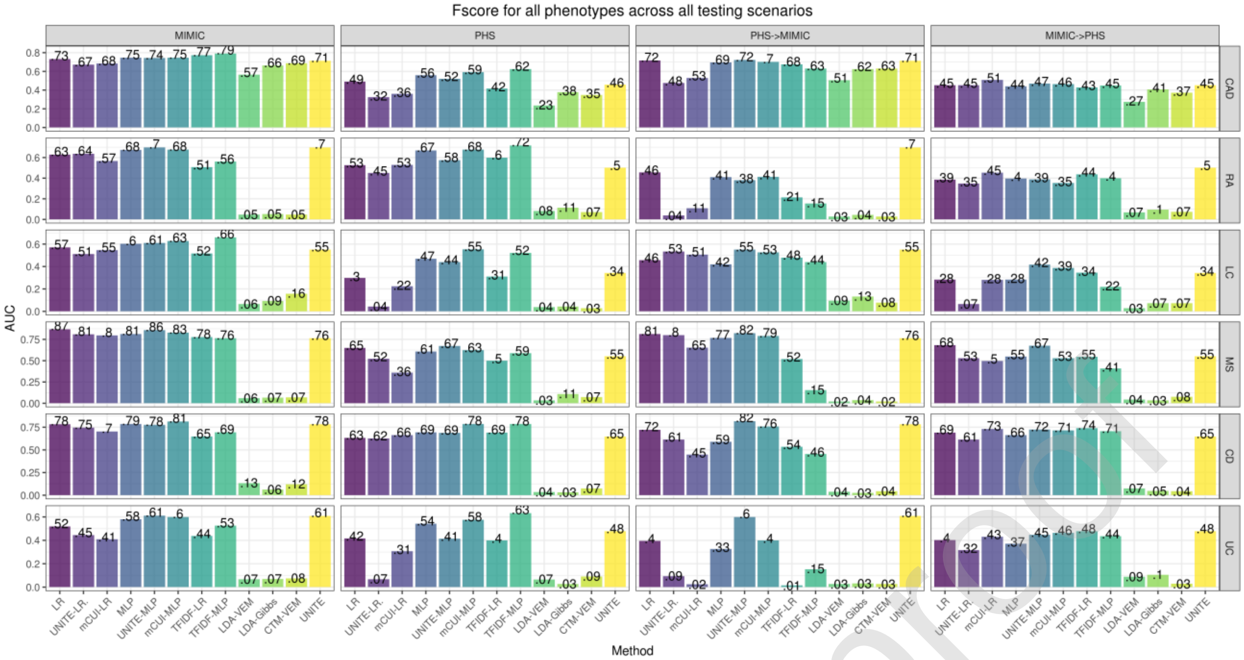

Figure 4. F-score for penalized logistic regression (LR), penalized logistic regression with UNITE (UNITE-LR), penalized logistic regression with the mean CUI (mCUI-LR), penalized logistic regression with the TFIDF (TFIDF-LR), multiple layer perceptron (MLP), multiple layer perceptron with UNITE features (UNITE-MLP), multiple layer perceptron with the mean CUI (mCUI-MLP), multiple layer perceptron with TFIDF (TFIDF-MLP), topic modeling based on LDA with 2 topics fit with VEM (LDA_VEM) and Gibbs (LDA_Gibbs), CTM with 2 topics fit with VEM (CTM_VEM) and UNITE.

Validation across different EMR systems

The two right columns in Figure 3 summarize the AUCs of cross-EMR performance, following training at one EMR system and validating at a different EMR system. UNITE consistently showed the best portability by demonstrating the same high AUCs when validating at different EMR systems as when validating within the same EMR. In addition, the comparisons between UNITE and the two supervised algorithms (LR and MLP) in regards to the portability loss are summarized in Table S4. Using RA as an example, UNITE had a cross-EMR AUC of 0.91 when training at PHS and validating at MIMIC, same to training and validating both within MIMIC. On the other hand, LR had an AUC of 0.81 when training at PHS and validating at MIMIC, substantially lower than the AUC the 0.93 when training and validating within MIMIC. We found similar findings for the other diseases with UNITE performing the best in portability across different EMR systems.

Among the supervised methods with UNITE features as input, it is particularly interesting that UNITE-MLP has better transportability performance than the other two MLPs, specially from PHS to MIMIC, for example in UC and RA the difference in AUC is .98 vs. 97 and .91 vs .88 respectively. This highlights the advantage of incorporating the weights learned from the drop-out self-regression, a crucial step in UNITE. However, as opposed to regular UNITE, this is still a supervised method and performance does decrease when transported to a different EMR system. For example, the AUCs of MLP, UNITE-MLP and mCUI-MLP were 0.9, 0.89 and 0.9, respectively when training and validating at MIMIC, and were all 0.84 when training at PHS and validating at MIMIC.

Deep NN have the most significant drop when transferring the models across EMR systems. For example, GRU drops from AUC of 0.97 to 0.57 and from 0.91 to 0.73 for predicting CAD and RA ICD codes respectively. Overall phenotyping performance is consistently low across diseases and models.

Computing Efficiency

We illustrate the speed with training the CAD model on PHS data. The NLP cost was omitted as it was a common factor for all methods. All computation was performed on a cluster that had 50 GB of RAM and a 2.1 GHz Intel(R) Xeon(R) processor. Training UNITE's weights took under 10 seconds, lasso logistic regression training with 10-fold cross validation took around 55 seconds for the low dimensional features (CUI using counts, on average 701.5) and 12 minutes for high dimensional features (using embeddings of 1800 dimensions), MLP took 2.3 minutes, LDA models on average took 20 seconds, and CTM took 2 minutes. Deep NN methods were trained using 8 Tesla V100, 8 Tesla M40 and 16 Tesla K80 GPUs, and took an average of 10 hours.

DISCUSSION

The utility of accurately and efficiently assigning ICD codes based on clinical notes is not limited to reducing coding burden. Many EMR research efforts involving clinical diseases require time-consuming chart review. A score of semantic relevance can prioritize review of notes most relevant to the target disease by reducing the time reviewing uninformative notes. The sum of the predicted probabilities for a particular ICD from clinical notes of a given patient can potentially improve the phenotyping accuracy [26] and replace conventional ICD code counts in phenotyping models [25]. Additionally, as shown in Figure 2, the method provides insight into what CUIs are relevant for each disease, allowing for an interpretable evaluation of the notes.

A clear advantage of UNITE is its higher portability across different institutions compared to other supervised (and unsupervised) methods. The ICD data that are available in a given EMR system for algorithm training are often subject to bias arising from the practice and coding behavior of the specific healthcare system. Thus, the portability of trained supervised algorithms to external healthcare systems is often poor. Using RA as an example, the LR algorithm trained at PHS had an AUC of 0.93 when validating at PHS, but only 0.81 when validating at MIMIC. In contrast, without using any existing ICD code, UNITE is reasonably portable across PHS and MIMIC EMRs. The algorithm trained at PHS attained similar accuracy as the algorithm trained at MIMIC when validating at MIMIC, and likewise for validating MIMIC-trained algorithm at PHS. In addition to taking advantage of the recent development of unsupervised learning of semantic vectors, the utilization of existing knowledge sources was another key factor that contributed to the portability. Unlike the deep learning models that try to exploit information from every word in the note, UNITE's knowledge search allows it to focus on only hundreds of features, which reduces the chance of overfitting. Another advantage of UNITE is its computing efficiency. On average, UNITE takes half the time of MLP and CTM.

In this paper, we chose drop-out regularization for estimating the regression coefficient β in Step II of UNITE. This was motivated by the fact that many CUIs were clinically similar or related, and thus they should have similar iVECs. The dropout training helped alleviate the collinearity issue by randomly replacing entries with the mean, and assigns similar iVECs with similar weights. For example, as Figure 2 shows, *angina*, *myocardial infarction*, and *acute coronary syndrome*, which were highly related, all had relatively equal weights from the CAD regression. This happened to all similar iVECs across diseases.

In this study, AUC and F-score were two major metrics we used to evaluate the performances of methods of comparison. Under some scenarios such as rare event, F-score or PPV will be a better choice than AUC for comparing the performances. However, the applications of the UNITE method are for both clinical use and research use. For the former, F-score might be a more relevant metric since ultimately a decision needs to be made about whether to code a disease or not, while for the latter, when directly utilizing the continuous probability score instead of a binary prediction, such as in Sinnott et al. 2018 [26], the AUC would be more informative.

To investigate the reason of false predictions of UNITE method, we performed an error analysis on predicted UC ICD codes for individual visit at PHS. We first investigated the confident false positive predictions. We randomly sampled 7 visits out of the 48 visits which got predicted probabilities of UC ICD code greater than 0.97 but were not actually assigned UC ICD, and conducted chart review. All of the visits had at least one mention of the main term *ulcerative colitis* or its variations in the medical notes, and 37% have more than one mention of the main term. However, those mentions were related to the patient’s medical history of UC, while the current visits were for other purposes, such as different diseases or annual exam. We then investigated the confident false negative predictions. We randomly sampled 5 visits from the 9 visits with predicted probabilities of UC ICD code less than 0.03 but were actually assigned UC ICD, and conducted chart review. We identified 2 visits with miscoded ICD of UC. Specifically, a visit of patient who had history of irritable bowel syndrome but no mention of UC was coded as UC; the other visit was regarding use of hormone replacement therapy but was coded as UC. We further reviewed numerous notes during the same period for the same patient, but found nothing relevant to UC. We therefore concluded that these 2 visits were potentially miscoded as UC. This example demonstrated the advantage of UNITE method in correcting the coding error in ICD labels.

One limitation of the proposed approach is labels were used to choose the threshold for the binary classification. Alternatively, we applied an unsupervised thresholding approach without using any label. Specifically, we set the threshold as the median UNITE score among notes that have at least one mention of the disease CUI. In rare diseases such as UC and CD the median is often zero, in which case we use a higher quantile. This way, the binary classification for whether a note corresponds to a disease ICD code was assigned without the need of any labels. To evaluate the performance of this unsupervised thresholding procedure, we compared the results of binary metrics obtained by both supervised thresholding approach using labels and unsupervised thresholding approach. As shown in Table S5 of the Supplementary Materials, the metrics estimated using unsupervised thresholding approach are comparable to that of the supervised approach, suggesting the use of the unsupervised thresholding approach as an alternative way.

The main advantage of using deep NN models is their ability to exploit the sequential nature of text. That is, these models have good a memory system to deal with long sequences, which makes them very sensitive to how the training text is structured. As this text structure changes between EMR systems, transportability is affected. This can be shown in the performance of the MIMIC-III trained models when tested on PHS. As expected, there is a clear drop in performance, especially in AUC. We believe this drop is due to two things which relate to how text is structured within vs. between systems: 1) many drugs and terms relevant to the diseases of interest don't overlap EMR systems such as: *gastroenteric* which is relevant to UC and CD, or *cardiazol* a drug formerly used as a respiratory and circulatory stimulant. Additionally, there are abbreviations such as *sbp170*, *sbp89* that are used in one but not both EMR systems. 2) There is a high number of misspellings in both EMR systems which are more consistent within that between systems. For example, we found there are some medications that are consistently misspelled in either EMR system such as *zyflo* instead of *zyflow*. 3) Finally, as Huang et. al. (2019) [24], we trained the word embeddings using the MIMIC data. This is a reasonable approach as these word embeddings capture better the medical context of the terms, and more specifically how they are used in the MIMIC hospitals. However, this affects the portability since there is a distributional change in the language used in Partner’s system notes. There are many transfer learning methods that can help alleviate the decrease in performance for these models, however one would need to customize the process in each target EMR system and would need a large dataset to retrain these models. This sheds light on the practical aspect of unsupervised, easily-trained methods like UNITE, that can handle bag-of-words data and can be ported. Additionally, UNITE is not highly sensitive to language structure change between EMR systems as it is unsupervised and relies only on CUI and not word embeddings.

Findings from this study inform several future directions. Firstly, in this study, we only trained CUI VECs from PHS biobank EMR. Leveraging information from different EMR data types will help further improve the performance of the CUI VECs and will be a future topic of interest. A potential strategy is to integrate coded data (such as ICD, laboratory values, and procedure codes) with narrative notes to create an even larger and better embedding that could potentially enhance the quality of CUI VECs. In our current approach, we determined the CUI importance by regression using the iVEC of the main CUI as the regressand. In future studies, we will test the performance by directly regressing the iVEC of the target ICD code instead of the main CUI. Secondly, we have shown that UNITE could provide accurate ICD assignment for six representative diseases in the absence of coded ICD data across EMRs. In the future, we will assess its performance across a broader range of diseases, particularly for those whose CUIs have relatively poor accuracy in predicting the ICD (such as bipolar disorder and schizophrenia). In these situations, a combination of CUIs and a small amount of ICD codes in calculating the iVEC could potentially perform better than CUIs alone. Finnaly, UNITE only assigns ICD codes for one disease at a time. There is a need to assign multiple ICD codes to a patient encounter. Recent efforts to assign multiple diagnosis codes have room for improvement [27]. We plan to further develop UNITE to enable multiple ICD code assignment.

CONCLUSION

We introduced a novel unsupervised method (UNITE) for accurate and efficient ICD assignment from clinical notes without using existing ICD data for training. More importantly, UNITE shows superior portability across different EMR systems. The proposed method has the potential to greatly reduce the human burden of ICD coding and promote patient safety by reducing medical errors in coding and providing interpretable results. Finally, UNITE can potentially improve the accuracy of clinical phenotyping by ranking the informativeness of each note with respect to a medical condition.

Authors' contributions

AS, WC and CH conceived the study design. CH extracted data for analyses. AS and WC conducted the data analyses. ZX, AA and YA provided clinical guidance. SY and CH provided statistical guidance. AS and WC wrote the manuscript and all authors contributed to the writing. All authors approve the final manuscript.

Statement on conflicts of interest

All authors have declared that they have no financial or non-financial interests that may be relevant to the submitted work; no other relationships or activities that could appear to have influenced the submitted work.

Summary Points

- Accurate ICD coding is critical for medical billing and electronic medical record-based research.
- Supervised coding algorithms trained in one system often have poor portability to external healthcare systems.
- Unsupervised coding algorithms such as topic modeling approaches have low accuracy because the topic can be too general for specific diseases.
- As the text structure changes between EMR systems, transportability of deep NN using the raw text as input is affected.
- UNITE accurately assigns ICD code in EMR without requiring human labor, and has major advantages over commonly used machine learning approaches.
- A clear advantage of UNITE is it attained stable performance and high portability across EMRs in different institutions.

REFERENCE

[1] K. O'malley, K. Cook, M. Price, K. Wildes, J. Hurdle and C. Ashton, "Measuring diagnoses: ICD code accuracy," Health Services Research, vol. 40, no. 5p2, pp. 1620-1639, 2005.

[2] J. E. Sheppard, L. C. Weidner, S. Zakai, S. Fountain-Polley and J. Williams, "Ambiguous abbreviations: an audit of abbreviations in paediatric note keeping," Arch. disease childhood, vol. 93, p. 204–206, 2008.

[3] D. Lang, "Consultant report-natural language processing in the health care industry," Cincinnati Children’s Hospital Medical Center, vol. Winter, no. 6, 2007.

[4] R. Farkas and G. Szarvas, "Automatic construction of rule-based icd-9-cm coding systems," BMC bioinformatics 9, vol. 9, no. 10, 2008.

[5] L. L. and C. B., "Automatic assignment of ICD9 codes to discharge summaries," Technical report, University of Massachusetts at Amherst, Amherst, MA., 1995.

[6] B. Ribeiro-Neto, A. Laender and L. De Lima, "An experimental study in automatically categorizing medical documents," Journal of the American society for Information science and Technology, vol. 52, no. 5, pp. 391-401, 2001.

[7] J. Medori and C. Fairon, "Machine learning and features selection for semi-automatic ICD-9-CM encoding," In Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, vol. Association for Computational Linguistics, pp. 84-89, 2010.

[8] S. Pakhomov, J. Buntrock and C. Chute, "Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques," Journal of the American Medical Informatics Association, vol. 13, no. 5, pp. 516-525, 2006.

[9] R. Carroll, W. Thompson, A. Eyler, A. Mandelin, T. Cai, R. Zink, J. Pacheco, C. Boomershine, T. Lasko, H. Xu and E. Karlson, "Portability of an algorithm to identify rheumatoid arthritis in electronic health records," Journal of the American Medical Informatics Association, vol. 19, no. e1, pp. e162-e169, 2012.

[10] R. Farkas and G. Szarvas, "Automatic construction of rule-based ICD-9-CM coding systems," BMC bioinformatics , vol. 9, no. 3, p. S10, 2008.

[11] J. Boyd-Graber, D. Mimno and D. Newman, "Care and feeding of topic models: Problems, diagnostics, and improvements," 2014, p. 225255.

[12] R. Kavuluru, S. Han and D. Harris, "Unsupervised extraction of diagnosis codes from EMRs using knowledge-based and extractive text summarization techniques," Canadian conference on artificial intelligence, pp. 77-88, 2013.

[13] Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin, "A neural probabilistic language model," Journal of machine learning research, vol. 3, no. Feb, pp. 1137-1155, 2003.

[14] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint , p. arXiv:1301.3781., 2013.

[15] A. Beam and I. Kohane, "Translating artificial intelligence into clinical care," JAMA, vol. 316, no. 22, pp. 2368-2369., 2016.

[16] S. Yu, A. Chakrabortty, K. P. Liao, T. Cai, A. N. Ananthakrishnan, V. S. Gainer and e. al., "Surrogate-assisted feature extraction for high-throughput phenotyping," Journal of the American Medical Informatics Association, vol. 24, no. e1, pp. e143-e149, 2017.

[17] A. Beam, B. Kompa, I. Fried, N. Palmer, X. Shi, T. Cai and I. Kohane, "Clinical concept embeddings learned from massive sources of medical data," arXiv preprint, p. arXiv:1804.01486., 2018.

[18] S. Yu and T. Cai, "A short introduction to NILE," arXiv, vol. preprint arXiv:1311.6063., 2013.

[19] S. Finlayson, P. LePendu and N. Shah, "Building the graph of medicine from millions of clinical narratives," Scientific data, vol. 1, p. 140032., 2014.

[20] W. Ning, S. Chan, A. Beam, M. Yu, A. Geva, K. Liao, M. Mullen, K. Mandl, I. Kohane, T. Cai and S. Yu, "Feature extraction for phenotyping from semantic and knowledge resources," Journal of Biomedical Informatics, vol. 91, p. 103122, 2019.

[21] E. Karlson, N. Boutin, A. Hoffnagle and N. Allen, "Building the partners healthcare biobank at partners personalized medicine: informed consent, return of research results, recruitment lessons and operational considerations," Journal of Personalized Medicine, vol. 6, no. 1, p. 2, 2016.

[22] A. Johnson, T. Pollard, L. Shen, H. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Celi and R. Mark, "MIMIC-III, a freely accessible critical care database," Scientific data, vol. 3, p. 160035, 2016.

[23] F. Catling, G. P. Spithourakis and S. & Riedel, "Towards automated clinical coding," International journal of medical informatics, vol. 120, pp. 50-61, 2018.

[24] J. O. C. &. S. L. W. Huang, "An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes," Computer Methods and Programs in Biomedicine, vol. 17, pp. 141-153, 2019.

[25] S. Yu, K. Liao, S. Shaw, V. Gainer, S. Churchill, P. Szolovits, S. Murphy, I. Kohane and T. Cai, "Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources," Journal of the American Medical Informatics Association, vol. 22, no. 5, pp. 993-1000, 2015.

[26] J. Sinnott, F. Cai, S. Yu, B. Hejblum, C. Hong, I. Kohane and K. Liao, "PheProb: probabilistic phenotyping using diagnosis codes to improve power for genetic association studies," Journal of the American Medical Informatics Association, 2018.

[27] T. Baumel, J. Nassour-Kassis, R. Cohen, M. Elhadad and N. Elhadad, "Multi-label classification of patient notes: case study on ICD code assignment," in In Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence., 2018.