Original Paper

Strategies to Address the Lack of Labeled Data for Supervised Machine Learning Training With Electronic Health Records: Case Study for the Extraction of Symptoms From Clinical Notes

Marie Humbert-Droz[1], PhD; Pritam Mukherjee[1], PhD; Olivier Gevaert[1,2], PhD

[1]Stanford Center for Biomedical Informatics Research, Department of Medicine, Stanford University, Stanford, CA, United States

[2]Department of Biomedical Data Science, Stanford University, Stanford, CA, United States

Corresponding Author:
Olivier Gevaert, PhD
Stanford Center for Biomedical Informatics Research
Department of Medicine
Stanford University
Medical School Office Building
Stanford, CA, 94305
United States
Phone: 1 650 721 2378
Email: olivier.gevaert@stanford.edu

Abstract

Background: Automated extraction of symptoms from clinical notes is a challenging task owing to the multidimensional nature of symptom description. The availability of labeled training data is extremely limited owing to the nature of the data containing protected health information. Natural language processing and machine learning to process clinical text for such a task have great potential. However, supervised machine learning requires a great amount of labeled data to train a model, which is at the origin of the main bottleneck in model development.

Objective: The aim of this study is to address the lack of labeled data by proposing 2 alternatives to manual labeling for the generation of training labels for supervised machine learning with English clinical text. We aim to demonstrate that using lower-quality labels for training leads to good classification results.

Methods: We addressed the lack of labels with 2 strategies. The first approach took advantage of the structured part of electronic health records and used diagnosis codes (International Classification of Disease–10th revision) to derive training labels. The second approach used weak supervision and data programming principles to derive training labels. We propose to apply the developed framework to the extraction of symptom information from outpatient visit progress notes of patients with cardiovascular diseases.

Results: We used >500,000 notes for training our classification model with International Classification of Disease–10th revision codes as labels and >800,000 notes for training using labels derived from weak supervision. We show that the dependence between prevalence and recall becomes flat provided a sufficiently large training set is used (>500,000 documents). We further demonstrate that using weak labels for training rather than the electronic health record codes derived from the patient encounter leads to an overall improved recall score (10% improvement, on average). Finally, the external validation of our models shows excellent predictive performance and transferability, with an overall increase of 20% in the recall score.

Conclusions: This work demonstrates the power of using a weak labeling pipeline to annotate and extract symptom mentions in clinical text, with the prospects to facilitate symptom information integration for a downstream clinical task such as clinical decision support.

(JMIR Med Inform 2022;10(3):e32903) doi: 10.2196/32903

KEYWORDS

clinical text mining; weak supervision; text classification; symptom extraction; EHR; machine learning; natural language processing

Introduction

Background

Unstructured text from electronic health records (EHR) contains myriads of information that is not encoded in the structured part of EHRs, such as symptoms experienced by the patient. Structuring and managing symptom information is a major challenge for research owing to their complex and multidimensional nature. Extracting symptom information from clinical text is critical; for example, for phenotypic classification, clinical diagnosis, or clinical decision support [1-3]. More specifically, symptoms are crucial to the assessment and monitoring of the general state of the patient [1,4] and are critical indicators of quality of life for chronically ill patients [5,6]. Their evolution through time can be a string indicator of the patient's clinical status change. Finally, in the context of pandemic prevention, symptoms are used for syndromic surveillance [7,8] and patient characterization [9,10].

Using natural language processing (NLP) and machine learning to process and use clinical text for such applications has great potential [11-14]. Unfortunately, machine learning, and more specifically supervised machine learning, requires a great amount of labeled data to train a model, which is at the origin of the main bottleneck of model development [15]. Manually labeling data sets is extremely costly and time-consuming as multiple experts need to manually review and annotate several hundreds of clinical notes [13,16]. Moreover, the development of such a resource presents unique challenges as the text contains personal information, and access to such data is usually restricted.

Throughout the past years, shared resources such as Informatics for Integrating Biology and the Bedside (i2b2) have generated deidentified and annotated data sets for the development of NLP systems for specific tasks. Such resources remain limited, as most of the annotated data sets contain only hundreds to a few thousands of notes. Moreover, these data sets come from a limited number of institutions, making the development of an NLP system with such data unlikely to generalize to other institutions or other tasks.

To develop NLP systems and models that are transferable between multiple institutions and free of overfitting, a large amount of data needs to be available for training. To do so, alternatives to supervised machine learning have been explored, such as distant supervision, which seeks to include information from existing knowledge bases [17] or active learning, which involves human experts in the machine learning process [18-20]. One method in particular, weak supervision, is attracting increasing attention for the automatic generation of lower-quality labels for unlabeled data sets [21-25].

Objective

To address the lack of labeled data, we propose 2 alternatives to manual labeling for the generation of training labels for supervised machine learning with clinical text. The first approach takes advantage of the structured part of EHRs and uses diagnosis codes to derive training labels. The second approach uses weak supervision and data programming principles to derive training labels. We propose to apply the developed framework to the extraction of symptom information from outpatient visit progress notes of patients with cardiovascular diseases.

Extracting symptoms from clinical narratives is not a straightforward task as symptoms are often expressed in an abstract manner. A straightforward way of deriving labels from EHR would be to take advantage of their coded part and use the International Classification of Disease–10th revision–Clinical Modification (referred to as ICD-10, henceforth) codes. This approach has challenges, as demonstrated in multiple studies [2,9,26-30]. This is especially true if the target information is symptoms, as the corresponding ICD-10 chapter is typically used when a sign or symptom cannot be associated with a definitive diagnosis. Thus, their occurrence in EHR is very scarce and expected to be incomplete. Despite issues related to inaccuracy in ICD-10 coding, we propose to use such codes to label our training set, with the assumption that with sufficient training data, the poor quality of the labels will be balanced out. Although inaccurate and possibly biased, the use of ICD-10 data is considered standard in many classification studies involving clinical text [15,31-41]. Moreover, we propose to complement the use of ICD-10 codes with a weak supervision approach to derive labels. Weak supervision has gained a great amount of traction in the past years [21-25] as a response to the increased need for training data for machine learning. We used the Snorkel library [42] to combine a large number of clinical reports with noisy labeling functions and unsupervised generative modeling techniques to generate labels for our models. Finally, we test the models on external cohorts as a way to assess the bias and test the generalizability of the models.

We successfully demonstrate that by using a large number of notes for training, we can train a classification model able to recognize specific classes of symptoms using low-quality labels. The resulting model is independent of the prevalence of positive instances and is transferable to a different institution. We show that training our model on such pseudolabels results in a good predictive performance when tested on a data set containing gold labels.

Methods

Cohort Description

Our data set consisted of 20,009,822 notes from January 1, 2000, to December 31, 2016, for 134,000 patients with cardiovascular diseases from Stanford Health Care (SHC), collected retrospectively in accordance with the approved institutional review board protocol (IRB-50033) guidelines. Progress notes from outpatient office visits were selected. As the ICD-10 codes for symptoms were chosen for initial labels, encounters without R codes were discarded. Finally, short notes (ie, <350 characters) were also discarded. The final cohort contained 545,468 notes for 93,277 patients (Figure 1).

For prototyping purposes and to evaluate the effect of the training set size on the performance, subsets of the full cohort were created, leading to the following three data set sizes: I (patients: 717/93,277, 0.77%), II (patients: 5611/93,277, 6.02%), and III (patients: 93,277/93,277, 100%). Patients were split into training, validation, and test sets using a 60:20:20 ratio. Table 1 provides a more detailed description of the data sets.

ICD-10 codes describing symptoms and signs involving the circulatory and respiratory systems were used to label the notes for the text classification task. The symptoms considered were only coded at the highest level of the ICD-10 hierarchy. The prevalence of the R codes was low, between 2% and 10% of positive instances (see Table S1 in Multimedia Appendix 1 for details).

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram for Stanford Health Care–electronic health record symptom extraction. Our full cohort consisted of 20 million notes and 134,000 patients. We selected progress notes from outpatient visits from encounters with International Classification of Disease–10th revision (ICD-10) codes from the chapter R. Notes <350 characters were discarded, yielding 545,468 notes for 93,277 patients.

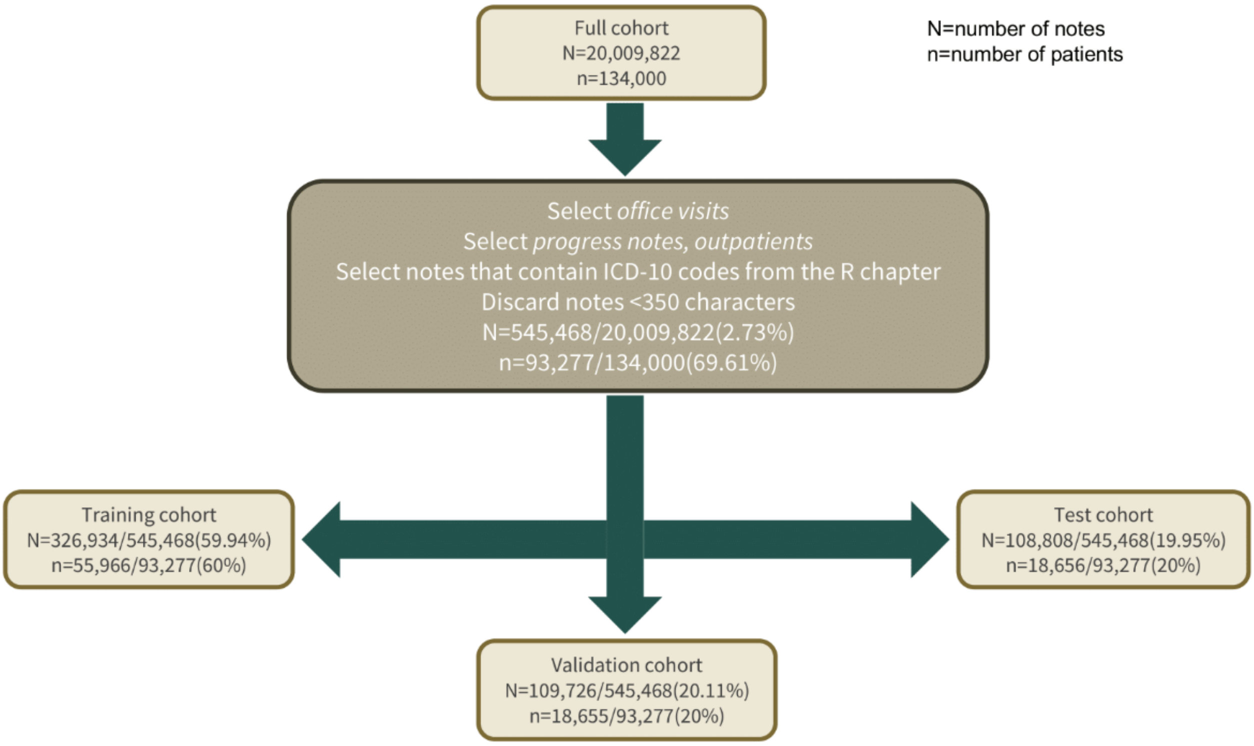

Full cohort
N=20,009,822
n=134,000
N=number of notes
n=number of patients
Select *office visits*
Select *progress notes, outpatients*
Select notes that contain ICD-10 codes from the R chapter
Discard notes <350 characters
N=545,468/20,009,822(2.73%)
n=93,277/134,000(69.61%)
Training cohort
N=326,934/545,468(59.94%)
n=55,966/93,277(60%)
Test cohort
N=108,808/545,468(19.95%)
n=18,656/93,277(20%)
Validation cohort
N=109,726/545,468(20.11%)
n=18,655/93,277(20%)

Table 1. Patient and note distribution for each data set considered in this study.

Data set	I[a] (N=717)	II[a] (N=5611)	III (N=93,277)	IV[b] (93,277)	V[c] (N=75.692)
Train set, n (%)	430 (59.9)	3360 (59.88)	55,966 (59.99)	55,966 (59.99)	38,381 (50.71)
Validation set, n (%)	143 (19.9)	1123 (20.01)	18,655 (19.99)	18,655 (19.99)	18,655 (24.65)
Test set, n (%)	144 (20.1)	1128 (20.10)	18,656 (20)	18,656 (20)	18,656 (24.65)
Age (years), mean (SD)	60 (23)	58 (23)	59 (23)	59 (23)	53 (23)
Gender, n (%)					
Men	306 (42.7)	2381 (42.43)	51,876 (55.61)	51,876 (55.61)	43,765 (57.82)
Women	410 (57.2)	3229 (57.55)	41,396 (44.38)	41,396 (44.38)	31,925 (42.18)
Unknown	1 (0.1)	1 (0.02)	5 (0.005)	5 (0.005)	2 (0.003)
Total notes, n (%)	4245 (100)	34,368 (100)	545,468 (100)	871,753 (100)	544,907 (100)
Train set	2480 (58.42)	20,500 (59.65)	326,934 (59.94)	653,219 (74.93)	326,373 (59.89)
Validation set	704 (16.58)	6698 (19.49)	109,726 (20.12)	109,726 (12.59)	109,726 (20.14)
Test set	794 (18.70)	6494 (18.89)	108,808 (19.95)	108,808 (12.48)	108,808 (19.97)

[a] Data sets I and II are subsets of data set III.

[b] Data set IV represents the hybrid data set of labeled and unlabeled notes considered for the weak supervision experiment.

[c] Data set V contains the set of unlabeled notes from IV.

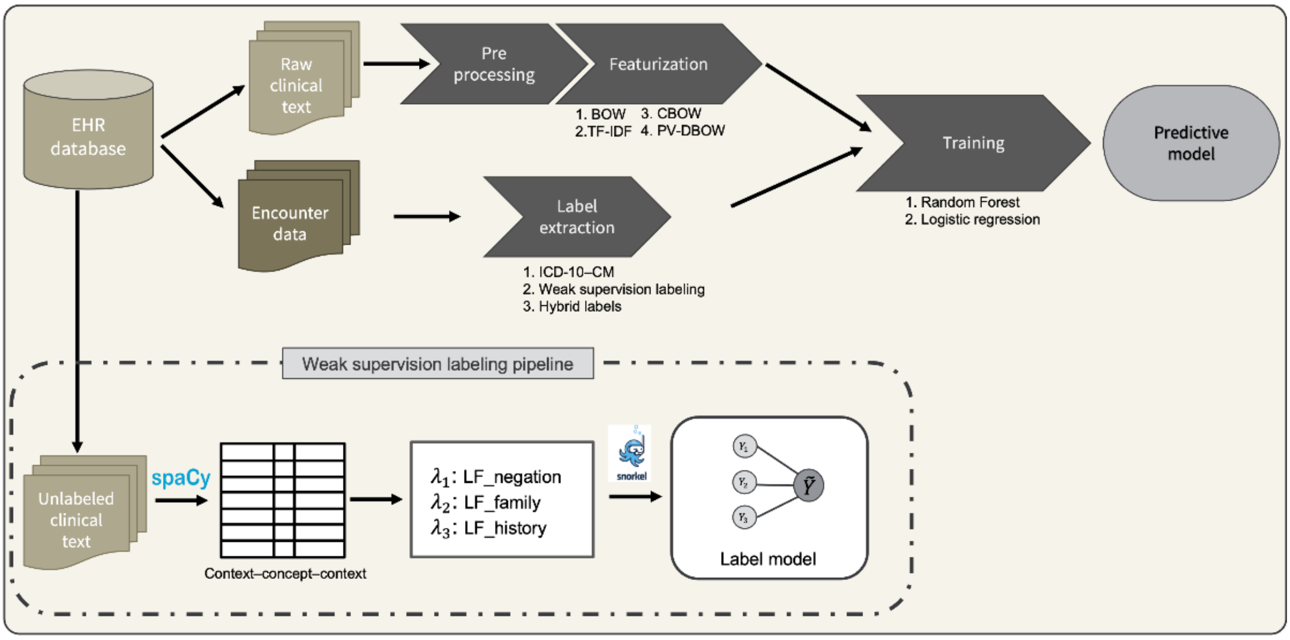
Pipeline

We defined our task of extracting symptom information from clinical notes as a multi-class classification problem. Machine learning algorithms were trained to classify whether each input note contained a specific class of symptoms.

The proposed pipeline used a subset of the ICD-10 chapter containing symptoms, signs, and abnormal clinical and laboratory findings. The codes in this chapter are typically used when a sign or symptom cannot be associated with a definitive diagnosis. As their occurrence in EHR is expected to be incomplete, we assumed that the presence of a code is associated with the observation of the symptom, but the absence of a code cannot be associated with the absence of the symptom in question.

The full pipeline developed for this study is depicted in Figure 2. We obtained the raw clinical text and encounter data from the SHC database. The raw text was first preprocessed for standardization purposes. Then, the text was transformed into a numerical format (ie, featurization) so that it can be used as input features for our model training. Then, ICD-10 codes were extracted from the structured encounter data to use as labels. A multi-class classification model was then trained to predict the presence of symptoms in the text. Next, we propose a weak supervision labeling pipeline as an additional method for extracting labels for the downstream prediction task. For that additional part, notes that were initially discarded because of the lack of symptom codes in the encounter data were processed using an entity recognition model with the spaCy library [43] and labeled using a labeling model generated using the Snorkel package [42].

Figure 2. End-to-end pipeline developed for extracting pseudolabels out of an electronic health record (EHR) database and training a text classifier for recognition of presence or absence of symptoms. The approach leverages the structured part of EHR (International Classification of Disease–10th revision–Clinical Modification [ICD-10–CM] codes) and weak supervision to generate labeled training corpus. Three types of labels are used for the training: ICD-10–CM codes; noisy labels obtained by a weak supervision pipeline; and hybrid labels, containing both ICD-10–CM codes and noisy labels. Two machine learning algorithms are considered: random forest and logistic regression. Four featurization methods are considered: bag-of-words (BOW), term frequency–inverse document frequency (TF-IDF), continuous BOW (CBOW), and paragraph vector–distributed BOW (PV-DBOW). LF: labeling function.



Preprocessing

To facilitate machine learning techniques, the clinical notes were standardized in the following manner: special characters and numbers were removed; the text was transformed into lower case only; frequent words (eg, the, as, and thus) often denoted as stop words were removed, except negative attributes such as no or not; next, each note was standardized using the Porter stemming algorithm; and finally, the text was tokenized into individual words. Sectioning of the notes was not performed; thus, the entire note was included in the featurization step.

Featurization

In this report, we evaluated the following approaches for featurization of the clinical notes. The first method, bag-of-words (BOW), is a simple yet effective method to represent text data for machine learning and acts as a baseline. In this method, the frequency of each word is counted, yielding a vector representing the document. As each word represents a dimension of the document vector, the size of the latter is proportional to the size of the vocabulary used. As words are represented by their document frequency, the resulting document vector does not contain any syntactic or contextual information.

Next, we used term frequency–inverse document frequency (TF-IDF), a weighting scheme, in addition to BOW whereby word frequencies from BOW are weighted according to their IDF. This reweighting of the frequencies dampens the effect of extremely frequent or rare words.

Next, we used the continuous BOW (CBOW; also referred to as *word2vec*) algorithm [44]. CBOW is an algorithm that generates word vectors based on a prediction task via a neural network. The output of such a network is an embedding matrix that is used to encode each word into a specific vector. The embedding matrix used in this project was trained on biomedical text (PubMed and Medical Information Mart for Intensive Care–III [MIMIC-III]) by Zhang et al [45]. Word vectors were generated using these pretrained embeddings and then averaged to yield a single document vector representing the entire note. As a result, the document embedding vector was of dimension 200.

Finally, the paragraph vector–distributed BOW (PV-DBOW; also referred to as *doc2vec*) [46], an extension of CBOW to paragraphs, was used to add some syntactic knowledge in the encoding of each document. The vector size for the document was 300 and was independent of the corpus size.

Weak Labeling

To address the problem of a lack of labels for EHR-based supervised learning, a weak supervision pipeline using the Snorkel package [42] was implemented. Weak supervision allows us to create a set of noisy labels for an unlabeled data set. The noisy labels are generated using a set of *labeling functions*, namely, a set of heuristic rules.

For this project, we implemented labeling functions based on pattern recognition applied to a 20 token–context window (10 tokens before and 10 tokens after the target term) to determine the negation, temporality, and experiencer of the target symptom. We used the publicly available *clinical event recognizer* base terminology [47] to match our context window with negative expressions, historical expressions, and family mentions. If a mention is matched within the context window of a given term, it is labeled accordingly: *absent* if negative expression is matched, *history* if historical expression is matched, and *family* if family mention is matched. Target symptoms that were positive, experienced by the patient, and not part of the past medical history were labeled positive. Occurrences deviating from this pattern were labeled negative.

Symptom recognition was performed using a ScispaCy [48] pipeline trained to recognize biomedical entities. The process of extracting the presence or absence of symptoms belonging to the R00-R09 categories was implemented as follows: the full clinical note is processed with spaCy [43] using the entity recognition model from the ScispaCy library, trained on BioCreative V Chemical Disease Relation corpus, a corpus of 1500 PubMed articles annotated for chemicals, disease, and chemical–disease interactions [49] (en_ner_bc5cdr_md [48]). As we were classifying the notes using only the 3 characters categories of the ICD-10 codes, each entity that was tagged needed to be associated to its corresponding category. For that purpose, we normalized them to the concept unique identifiers from the unified medical language system with the highest similarity score. This allowed us to group each entity to their corresponding ICD-10 category (see Table S2 in Multimedia Appendix 1 for a list of concept unique identifiers). Then, the labeling functions defined earlier were used to generate noisy labels, which can finally be used to train a machine learning model.

Modeling

The input features were used to predict a set of symptoms related to abnormalities in the circulatory and respiratory systems (ICD-10 codes R00-R09). The problem was approached as a text classification task using a subset of the ICD-10-R codes for the class labels. The classes are not mutually exclusive; therefore, a *one-versus-all* classification was chosen. We compared two classification algorithms for this task, namely random forests [50] and logistic regression [51]. We only report the results obtained with 100 estimators for the random forest and the limited-memory Broyden–Fletcher–GoldfarbShanno solver. The detailed parameters used for each model are provided in the Multimedia Appendix 1.

Performance Evaluation

We used the following classification metrics to evaluate each model: recall, F1 score, and average precision score. We also computed the receiver operating characteristic (ROC) curves and precision-recall curves. Owing to the class imbalance, we gave more importance to the precision-recall curve. For example, in the case of *hemorrhage from respiratory passages* class of symptoms (R04), the positive instances represent only approximately 1% of the data points. We also considered computation time and memory requirements as important metrics to determine the best classification model. Given the size of our data set, an efficient implementation was of paramount importance for the success of our predictive model.

External Validation

To assess the impact of training the model on low-quality labels, the models were tested on an external data set developed for symptom extraction by Steinkamp et al [52]. Their work provides an open-source annotated data set for symptom extraction. The notes were 1008 deidentified discharge summaries from the i2b2 2009 Medication Challenge [53]. The set of notes was annotated by 4 independent annotators for all symptom mentions, whether *positive*, *negative,* or *uncertain*. To benchmark our study, we chose three classes of symptoms that were both present in our study and in the annotated data set of Steinkamp et al [52], namely cough (R05), abnormalities of breathing (R06), and pain in throat and chest (R07). As the annotations were performed at the *mention* level but our study was performed at the *note* level, a majority voting algorithm was chosen to assess the note-level polarity of the symptom mention to generate note-level labels. On the basis of the SHC experiments, only models showing the best promise in terms of predictive performance were chosen for this step. More specifically, models trained with the logistic regression algorithm using TF-IDF and PV-DBOW features were chosen for the external validation.

Results

Logistic Regression Performs Better Than Random Forest for Predicting the Presence of Symptoms in Outpatient Progress Notes

Outpatient progress notes collected from January 1, 2000, to December 31, 2016, from the SHC EHR database were used to train a text classifier to extract symptoms related to abnormalities in the circulatory and respiratory systems (Figure 1). Two machine learning algorithms were considered, namely random forest and logistic regression. The models were first built on a subset of the cohort for prototyping purposes (Table 1: data set I). Random forest showed poor predictive performance, with no or few positive instances predicted (Figure 3). Without exception, logistic regression outperformed random forest for all the considered data set sizes (Figure 4). Use of TF-IDF features to predict the presence of symptoms in the notes led to the best overall performance (Figure 5).

Figure 3. Histogram of predicted probabilities for the presence of the cough symptom (R05) in the outpatient progress note for data set I, with a comparison between probabilities predicted by logistic regression (LR) and random forest (RF) for term frequency–inverse document frequency (TF-IDF) and paragraph vector–distributed bag-of-words (PV-DBOW) feature extraction methods.

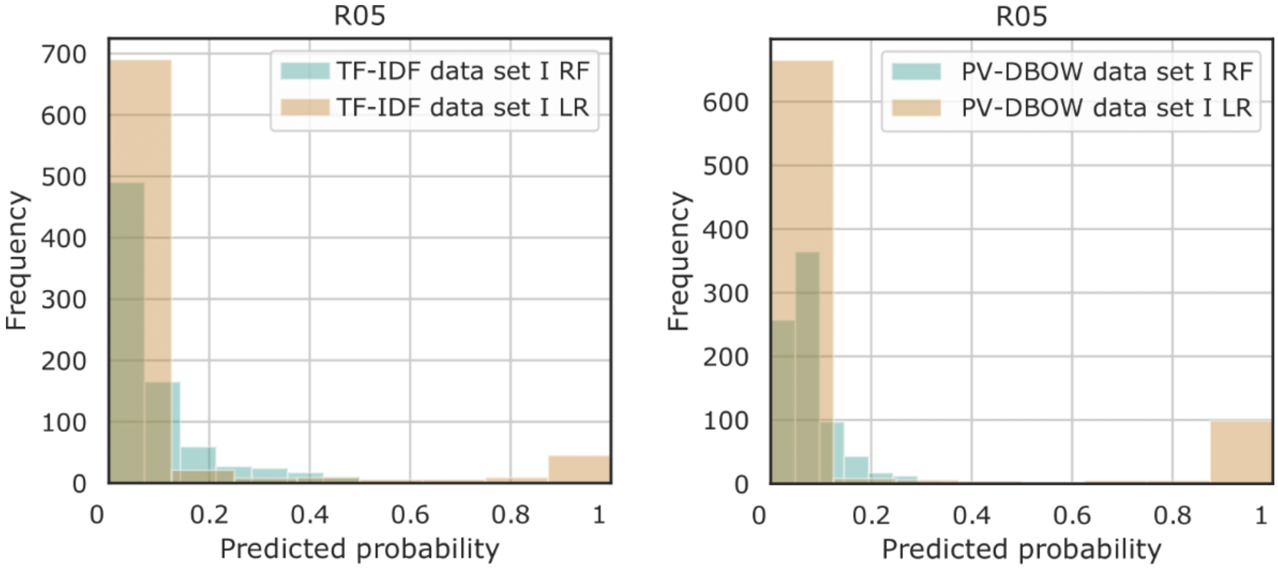


Figure 4. Summary of performance metrics averaged over all codes for all four considered feature extraction methods (bag-of-words [BOW], term frequency–inverse document frequency [TF-IDF], continuous BOW [CBOW], and paragraph vector–distributed BOW [PV-DBOW]). AUROC: area under the receiver operating characteristic curve; LR: logistic regression model; RF: random forest model.

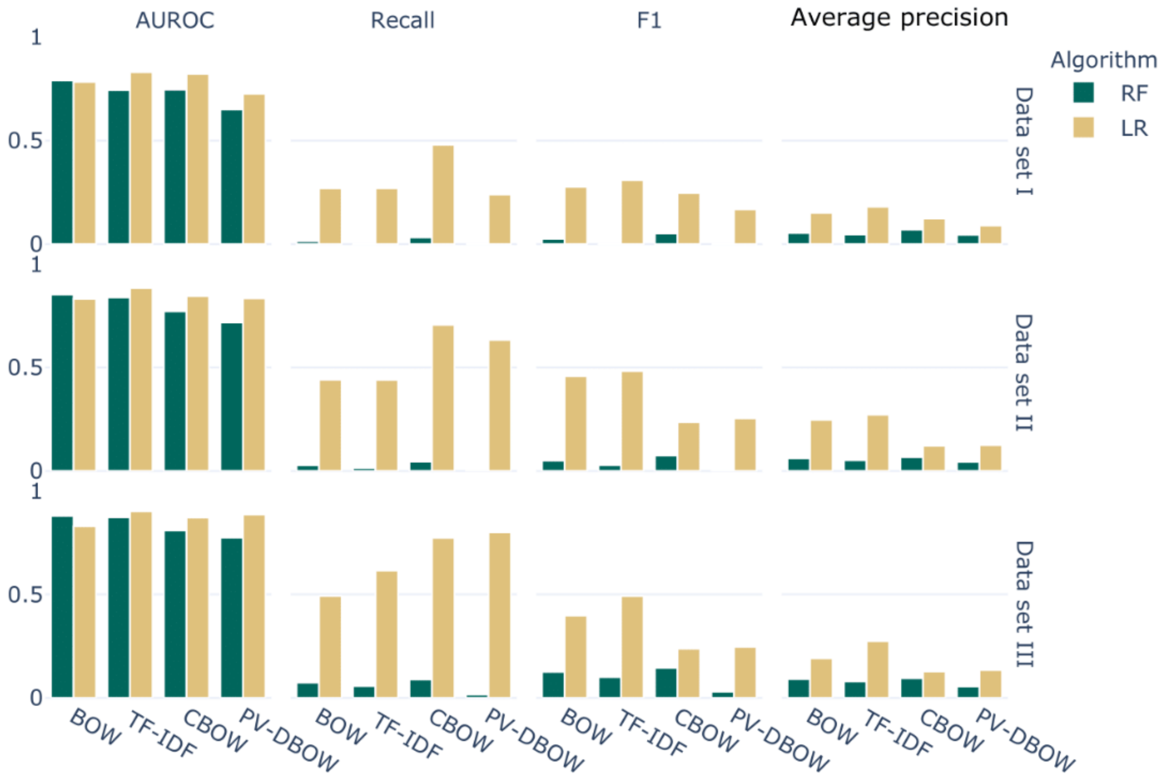
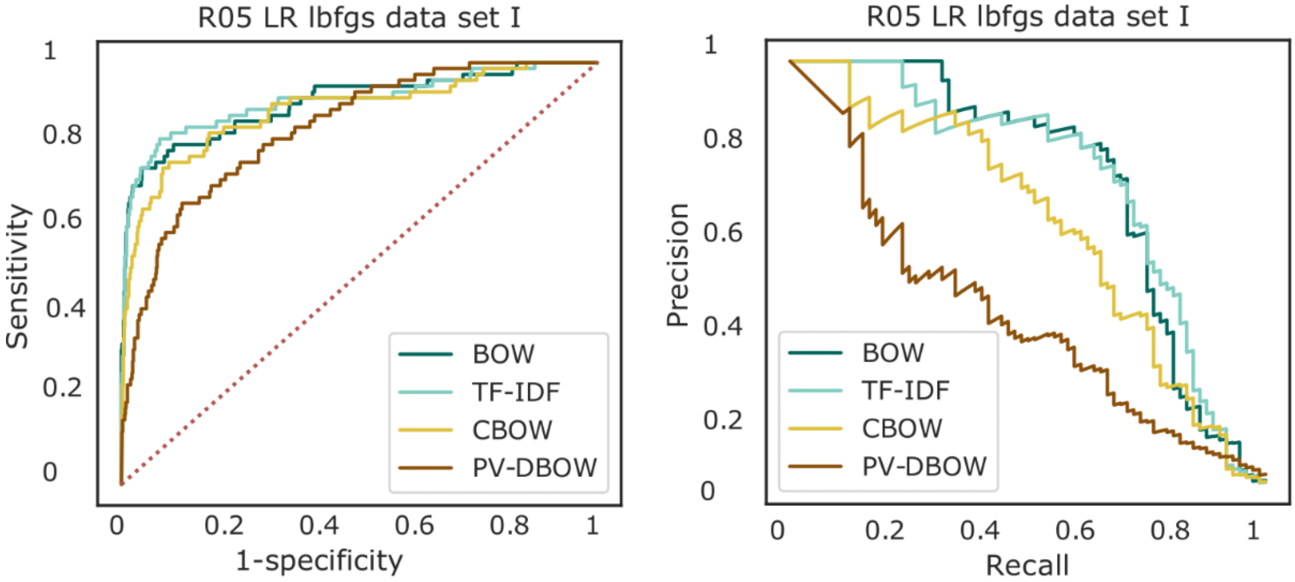


Figure 5. Receiver operating characteristic and precision-recall curves for the prediction on the test set (data set I described in Table 1) of presence of cough (R05) symptoms from outpatient progress notes using logistic regression (LR) with 4 feature extraction methods. BOW: bag-of-words; CBOW: continuous BOW; lbfgs: limited-memory Broyden–Fletcher–GoldfarbShanno solver; PV-BOW: paragraph vector–distributed BOW; TF-IDF: term frequency–inverse document frequency.



Embedding-Based Methods Perform Better With Increasing Data Set Size

To demonstrate that increasing the size of the training set significantly improves the performance of deep learning–based embedding methods, the classification task was performed on 3 different data set sizes, ranging from 0.75% (700/93,277) of patients to 100% (93,277/93,277) of patients (Table 1).

For all codes, the performance (area under the ROC [AUROC] curves and area under the precision-recall curves) of PV-DBOW features with logistic regression drastically improved with the size of the training set. For TF-IDF features also, there was a slight improvement, but it was less pronounced (Figure 6). More importantly, we observed that when increasing the size of the training set, the low prevalence of the symptoms does not affect the performance of embedding-based features (CBOW and PV-DBOW; Figure 7). Next, although the performance obtained with TF-IDF features was high, the computational performance was drastically affected by the increasing size of the training set. It takes 2 minutes and 1.6 GB of memory to train the model with PV-DBOW features, whereas the model with TF-IDF features requires 2.3 GB of memory and takes almost 3 hours (Table 2).

Figure 6. Comparison of receiver operating characteristic (left column) and precision-recall (right column) curves for the prediction of presence of cough (R05), abnormality of breathing (R06), and pain in throat and chest (R07) classes of symptoms from outpatient progress notes using logistic regression (LR) with the limited-memory Broyden–Fletcher–GoldfarbShanno (lbfgs) solver on data set I, data set II and data set III with term frequency–inverse document frequency (TF-IDF) and paragraph vector–distributed bag-of-words (PV-DBOW) features.

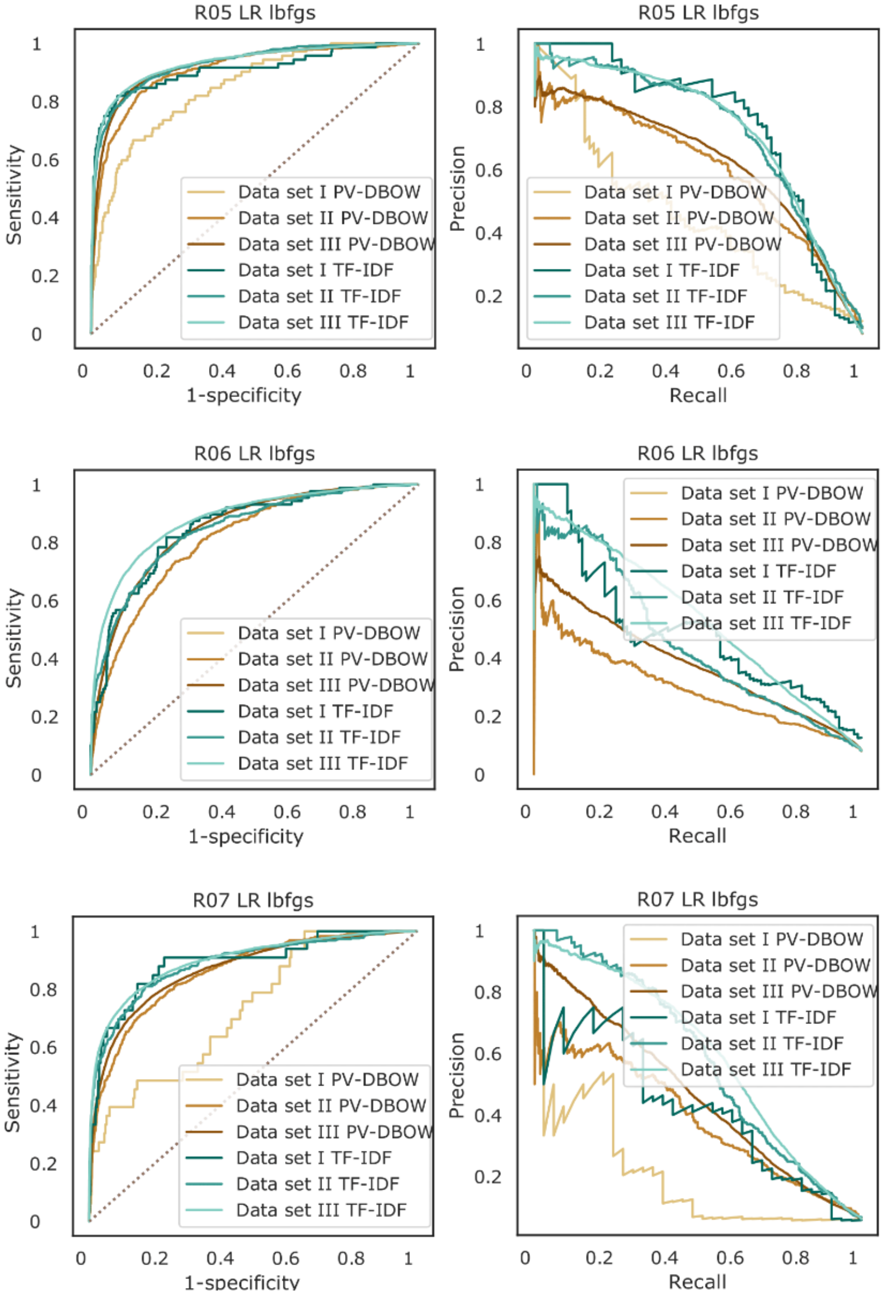


Figure 7. Recall scores as a function of the symptom prevalence in 3 considered data sets for all the features. BOW: bag-of-words; CBOW: continuous BOW; PV-BOW: paragraph vector–distributed BOW; TF-IDF: term frequency–inverse document frequency.

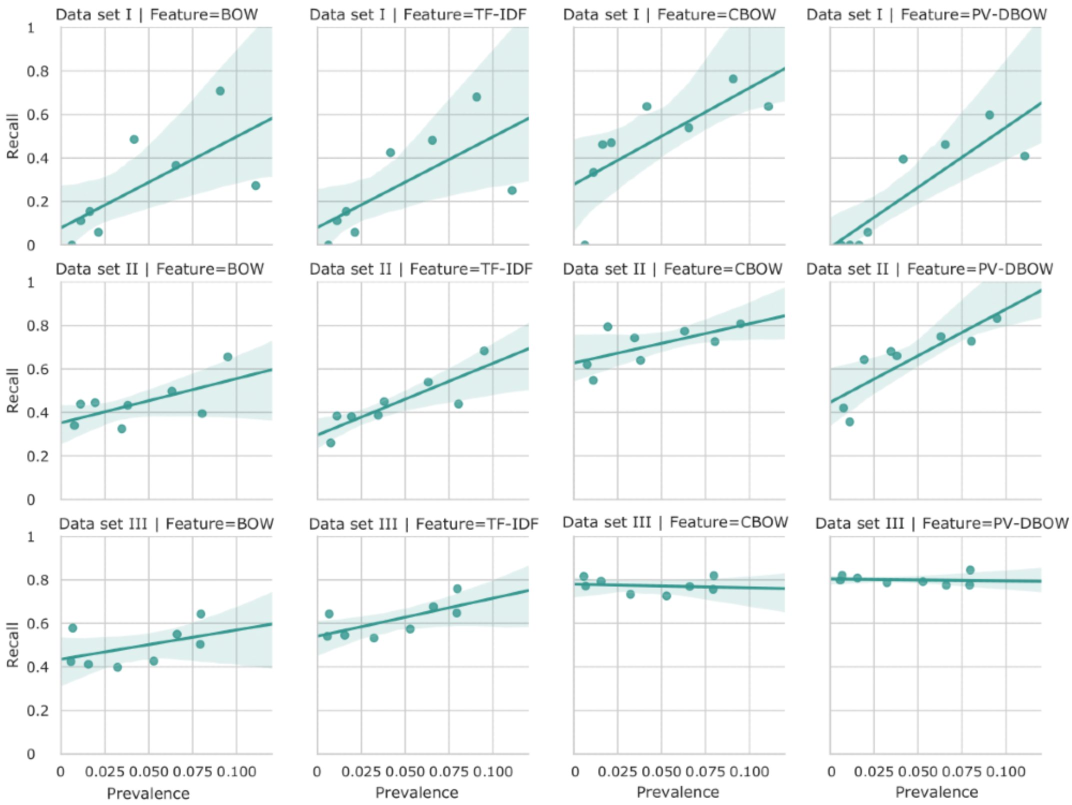


Table 2. Computational resources used for each classifier by feature type for data sets II and III.

Feature type and data set	Random forest		Logistic regression	
	Memory, MB	Run time, hours:minutes:seconds	Memory, MB	Run time, hours:minutes:seconds
BOW[a]				
II	310	00:04:10	340	00:21:35
III	3500	07:22:02	3400	23:17:20[b]
TF-IDF[c]				
II	310	00:04:15	270	00:03:04
III	3400	06:37:04	2300	02:47:30
CBOW[d]				
II	193	00:03:02	180	00:01:17
III	1700	01:21:11	1700	00:16:36
PV-DBOW[e]				
II	170	00:03:35	89	00:00:34
III	1100	01:41:18	1600	00:02:13

[a] BOW: bag-of-words.

[b] No convergence after 100,000 iterations.

[c] TF-IDF: term frequency–inverse document frequency.

[d] CBOW: continuous BOW.

[e] PV-DBOW: paragraph vector–distributed BOW.

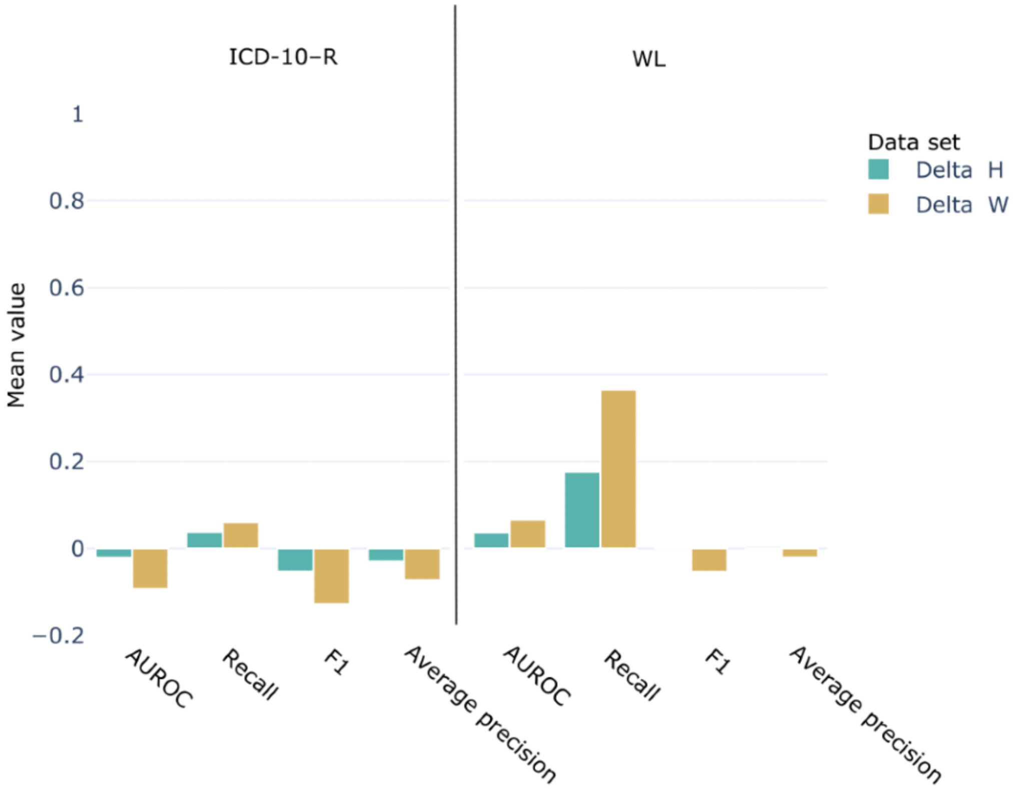
Enriching the Training Set With Weak Labels Enhances the Performance Further

The original cohort contained many notes that do not contain ICD-10 codes from the R chapter, leading to a substantial reduction in the number of notes available to train our model. Indeed, an additional 1,290,170 notes from the patients included in our cohort did not contain any ICD-10 code for symptoms.

To use these notes, they were processed using a weak supervision approach to determine the presence or absence of symptoms belonging to the R00-R09 categories. Then, the weakly labeled notes were added to data set D for training the classifier (ie, data set IV). For comparison, we also trained a model using only the weakly labeled notes (ie, data set V). Then, the 2 models were tested on test set III with ICD-10 codes for labels. The weak labeling model was also applied to the test set to extract weak labels for testing. Given the poor scaling performance of TF-IDF features compared with that of PV-DBOW, this experiment was performed solely with the PV-DBOW features.

Figure 8 shows the difference in performance between the enriched data set (IV) and the baseline data set (III). Overall, the recall score improved by 3.8%. However, the AUROC score was reduced by 2.1%. This decrease in the AUROC score can be attributed to the number of false-positive predictions. As the model was trained on mixed labels (ICD-10 and weak labels) but tested on ICD-10 codes, such increase in predictions flagged as false positives was expected. However, treating the weak labels as *true* labels for the test set led to an increase in recall score by 17.7% and an increase in AUROC score by 3.7%.

Figure 8. Performance metrics differential for the weak labeling experiment. Delta H represents the score difference between the hybrid data set IV and the baseline data set III (score [IV]–score [III]). Delta W represents the score difference between the weakly labeled data set V and the baseline data set III (score[V]–score [III]) The left panel shows the score calculated using International Classification of Disease–10th revision–R (ICD-10–R) codes for labels and the right panel shows the score calculated treating the weak labels (WL) as true labels in the test set. AUROC: area under the receiver operating characteristic curve.



Use of only weakly labeled notes for training (data set V) and testing on ICD-10 labels led to a 6% increase in recall score and a 9.3% decrease in the AUROC score. Finally, using the weak labels as *true* labels for the test set, the weakly labeled notes performed 36.6% (recall) and 6.6% (AUROC) better than the baseline data set.

Embedding-Based Features Perform Better Than TF-IDF Features on an External Validation Set

We selected a set of 56.65% (571/1008) notes from the i2b2 2009 challenge annotated for symptom extraction [52] containing mentions of symptoms of cough (R05), abnormalities of breathing (R06), and pain in throat and chest (R07). The logistic regression models trained on data set III using TF-IDF and PV-DBOW features were used to predict the presence of the 3 classes of symptoms.

Overall, the model trained with PV-DBOW features performed well when used to predict symptoms from the i2b2 notes. Figure 9 shows the difference in scores between the i2b2 data set and the baseline data set III trained using TF-IDF and PV-DBOW features for the set of 3 selected classes of symptoms. For R06 and R07, PV-DBOW, recall, and AUROC scores were within the range of the scores obtained when tested on the SHC notes.

However, the F1 and average precision scores were >40 points better on the i2b2 notes. On the other hand, the model trained with TF-IDF features performed poorly. The recall and AUROC scores were 20 to 30 points lower than when tested on the SHC notes. The F1 score was similar to that obtained with the SHC notes. However, the average precision was almost 30 points higher than that of the SHC notes (Figure 9). For both PV-DBOW and TF-IDF features, the performance of the symptom *cough* decreased when tested on the i2b2 set compared with the SHC notes.

Finally, the models trained with the hybrid labels and weak labels using the PV-DBOW features were also tested on the i2b2 notes. For both models, the recall and AUROC scores were within the range of those obtained with the SHC notes. However, the F1 and average precision scores were approximately 50 points higher than when tested with the SHC notes, reinforcing the conclusion that even though the models were trained on pseudolabels, they still perform well when tested on gold labels (Figure 10). Typically, recall performed better when hybrid or weak labels were used for training than when ICD-10 codes were used. Similar to the use of ICD-10 codes as labels, the performance for R05 decreased for the i2b2 notes.

Figure 9. Performance metrics differential for the external validation set. The score has been calculated as the difference between the score obtained on the external validation set and the baseline data set III (score [Informatics for Integrating Biology and the Bedside]–score [Stanford Health Care]). Term frequency–inverse document frequency (TF-IDF) represents the logistic regression model trained with TF-IDF features. Paragraph vector–distributed bag-of-words (PV-DBOW) represents the logistic regression model trained with PV-DBOW features. International Classification of Disease–10th revision–R codes have been used as reference labels to compute the metrics. AUROC: area under the receiver operating characteristic curve.

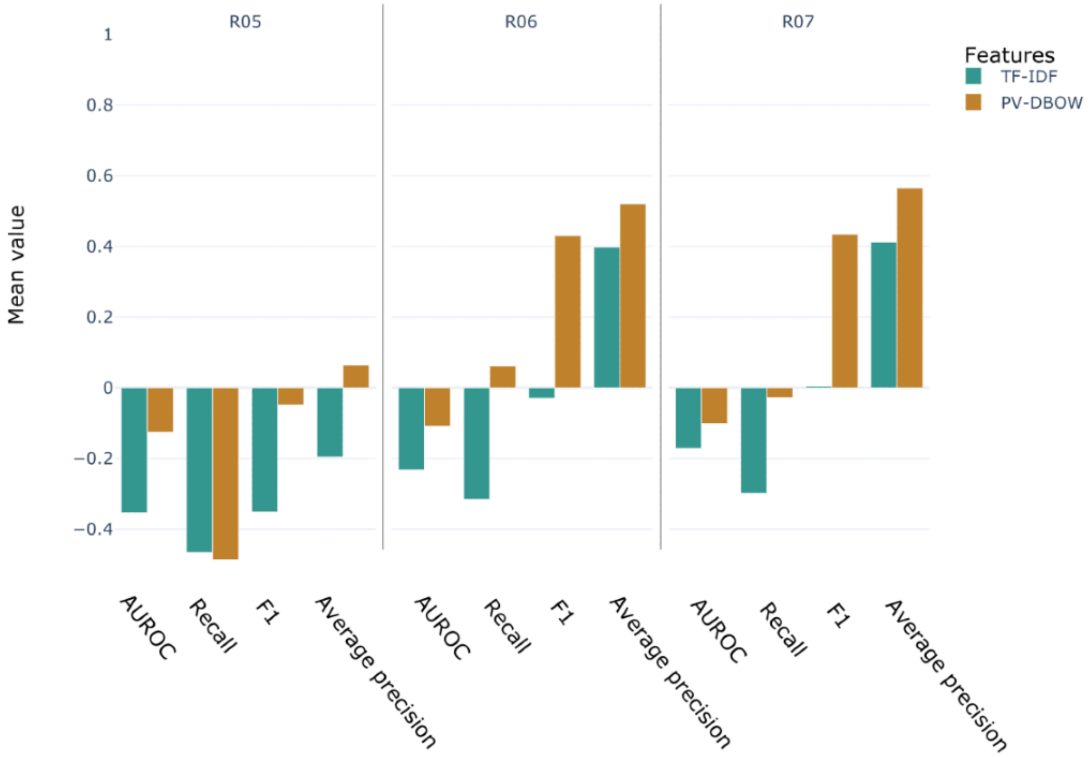
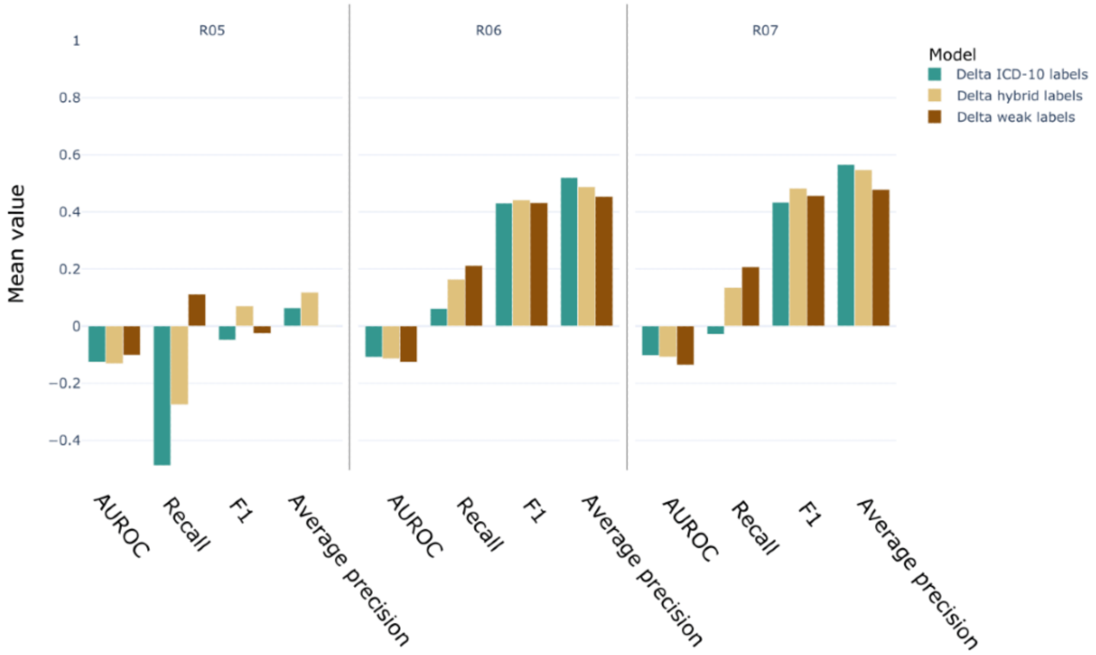


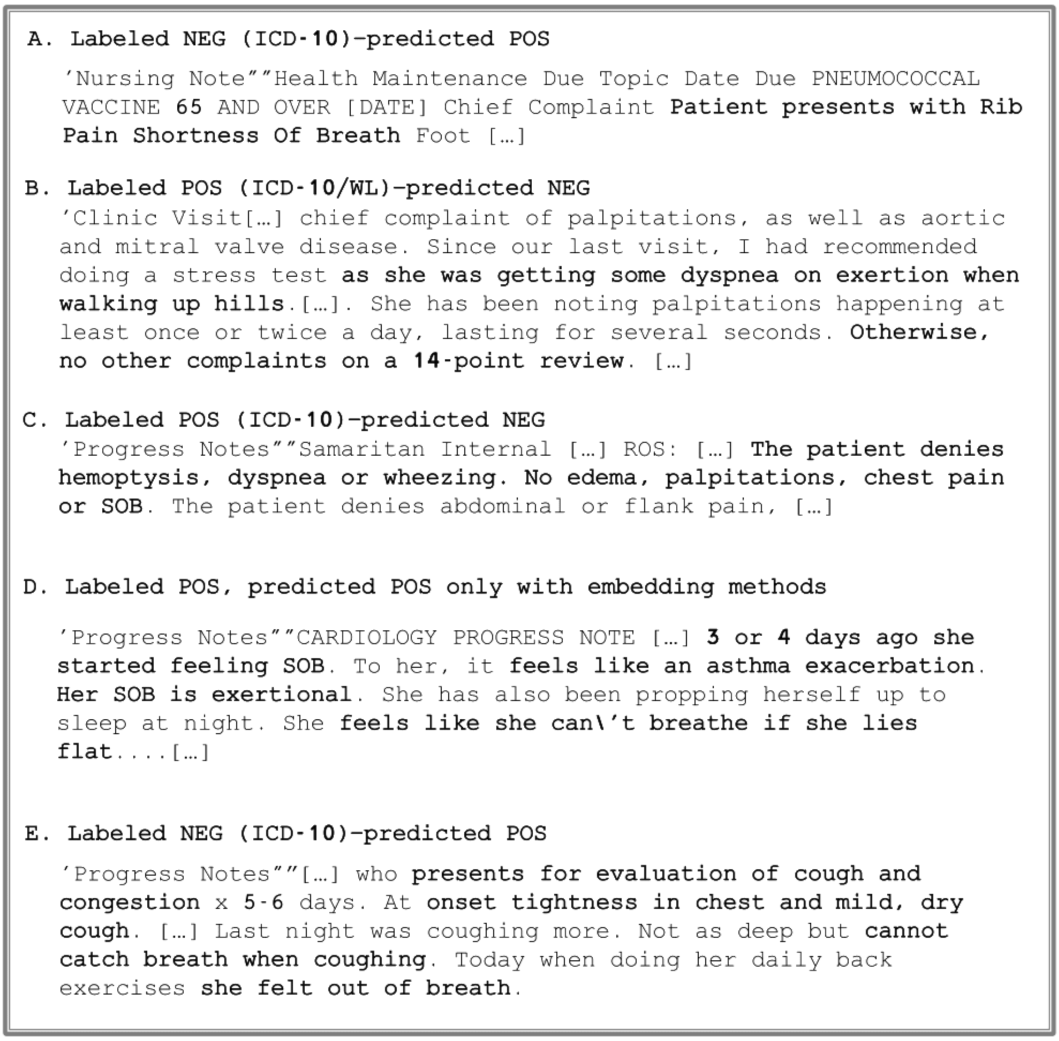
Figure 10. Performance metrics differential for the external validation set. The validation was performed for three models using paragraph vector–distributed bag-of-words features only, trained using different labels: International Classification of Disease–10th revision–R, the weak labels, and the hybrid labels. The score differences are computed relative to the baseline data set III (score [Informatics for Integrating Biology and the Bedside]–score [Stanford Health Care]). AUROC: area under the receiver operating characteristic curve.



Analysis of Misclassified Cases

To illustrate that despite the low quality of training labels used, the classification models were able to correctly classify notes, we show a few examples of the presence of abnormality of breathing symptoms in Figure 11. Snippets (A) and (E) show examples where the predictions were flagged as false positive but turned out to be true-positive cases. Snippets (B) and (C) show 2 examples that were flagged as false negative; however, when reading the note, the symptom was clearly absent (historical for (B) and negated for (C). Finally, snippet (D) shows an example that was correctly predicted only when embedding features were used.

Figure 11. Snippet examples of mislabeled notes for R06 class of symptoms. ICD-10: International Classification of Disease–10th revision; NEG: negative; POS: positive; WL: weak labels.



A. Labeled NEG (ICD-10)-predicted POS

'Nursing Note""Health Maintenance Due Topic Date Due PNEUMOCOCCAL VACCINE 65 AND OVER [DATE] Chief Complaint **Patient presents with Rib Pain Shortness Of Breath** Foot [...]

B. Labeled POS (ICD-10/WL)-predicted NEG

'Clinic Visit[...] chief complaint of palpitations, as well as aortic and mitral valve disease. Since our last visit, I had recommended doing a stress test **as she was getting some dyspnea on exertion when walking up hills**.[...]. She has been noting palpitations happening at least once or twice a day, lasting for several seconds. **Otherwise, no other complaints on a 14-point review**. [...]

C. Labeled POS (ICD-10)-predicted NEG

'Progress Notes""Samaritan Internal [...] ROS: [...] **The patient denies hemoptysis, dyspnea or wheezing. No edema, palpitations, chest pain or SOB**. The patient denies abdominal or flank pain, [...]

D. Labeled POS, predicted POS only with embedding methods

'Progress Notes""CARDIOLOGY PROGRESS NOTE [...] **3 or 4 days ago she started feeling SOB**. To her, it **feels like an asthma exacerbation**. **Her SOB is exertional**. She has also been propping herself up to sleep at night. She **feels like she can\'t breathe if she lies flat**....[...]

E. Labeled NEG (ICD-10)-predicted POS

'Progress Notes""[...] who **presents for evaluation of cough and congestion** x **5-6** days. At **onset tightness in chest and mild, dry cough**. [...] Last night was coughing more. Not as deep but **cannot catch breath when coughing**. Today when doing her daily back exercises **she felt out of breath**.

Discussion

Principal Findings

We trained *one-versus-all* multi-label classification models using four featurization methods, namely BOW, TF-IDF, CBOW, and PV-DBOW, to predict the presence of signs and symptoms related to abnormalities in the circulatory and respiratory systems. The challenging lack of labels for training such models was addressed using 2 label extraction strategies. First, we extracted labels based on a subset of ICD-10 codes from EHR encounter data. This approach yielded good predictive performance, as evidenced by external validation. Relying on the coded part of EHR to extract training labels leaves a large part of progress notes untouched, as ICD-10 codes for symptoms are rarely used. The second approach we used was a method to extract training labels by leveraging clinical named entity recognition and a weak supervision pipeline. This approach not only allowed us to make use of a much larger set of notes for training but also significantly improved the predictive performance, both on an SHC test set and an external validation set.

Although TF-IDF features yielded the best performance overall (Figure 4), the size of the feature vector is the size of the corpus, leading rapidly to intractable size and computational inefficiency when the corpus size increased (Table 2), whereas embedding methods such as CBOW and PV-DBOW led to a fixed feature vector length, independent of the training corpus size. The main computing cost in such an approach lies in the pretraining of the embedding vectors, which must be performed only once. Training a classifier on any data set size led only to a minor increase in computational cost, making this approach more desirable.

Unfortunately, the results on a small training set were not satisfactory as these types of models are known to be extremely data hungry. The performance is expected to be more reasonable with larger data set sizes. We observed this in our experiments; when the training set size was increased, the performance also increased significantly. For example, the most notable performance improvement was observed for the recall, which increased from 0.25 to 0.8 for PV-DBOW features (Figure 4). This is important because when predicting the presence or absence of symptoms, minimizing the false-negative rate is desirable. Moreover, owing to the nature of our training labels, the absence of an ICD-10 code does not mean the absence of the symptom, whereas the presence of the code more likely signifies the presence of the symptom. Moreover, the effect of the low prevalence of some codes on the performance became negligible with increasing data set size and the use of PV-DBOW features, suggesting that the use of a resampling method is not necessary if training on larger data sets (Figure 6).

Next, enriching the largest data set with unlabeled notes using a weak supervision approach for labeling yielded an overall gain in performance. This result not only suggests that more is better but also points to the conclusion that the use of ICD-10 codes as labels to extract the presence of symptoms from clinical notes can be improved by using weak labeling pipelines to label previously unlabeled notes. Indeed, external validation of our models showed a large increase in performance of the PV-DBOW features. We attribute this gain to the quality of labels in the external validation data set, resulting in a drop in false-positive predictions. This experiment also suggests that although the quality of the labels used to train the models was not optimal, the model was still able to learn enough to reliably predict the presence of symptoms. On the other hand, the poor performance of the TF-IDF features suggests that the high performance observed on the SHC notes might be owing to overfitting of the features rather than a good predictive power. However, the increase in average precision suggests that the false-positive rate is reduced owing to the higher quality of the labels. Although TF-IDF seems to work well within one context, it is likely to fail when testing at other sites.

It is worth noting that the performance for cough symptoms (R05) decreased significantly when tested on our external validation data set. The causes for such a drop have not been investigated, but Figure 10 offers some hints about a labeling issue. Indeed, the recall score performed poorly when using the model trained with ICD-10 codes as labels but increased when using the weak labels as ground truth for training.

The automatic classification of clinical text into specific ICD codes is a common task, and various state-of-the-art models have been developed over the years. Although our objective is different, it is worth comparing our classification results with some of the available work. Moons et al [54] recently compared multiple state-of-the-art models for ICD coding of clinical records, using public data sets encoded with both ICD-9 (MIMIC-III [55]) and ICD-10 (CodiEsp [56]). They reported micro- and macro-F1, micro-AUROC, and Precision@5 for multiple subsets of MIMIC-III and CodiEsp using multiple deep learning architectures. As they did not report recall or the prevalence of each class, a direct comparison with our work is difficult. However, it is worth noting that the best-performing model on the MIMIC-III data set (using ICD-9 codes) yields a macro-F1 of 64.85. Their best-performing model on CodiEsp (using ICD-10 codes) yields a macro-F1 of 11.03. Our macro-F1 of 24.66 falls in between these values, suggesting that our performance lies within the range of some of the best-performing deep learning models available.

We note that although we are using a data set containing gold standard annotations, a direct comparison with previous results from Steinkamp et al [52] is not possible. Both experiments are fundamentally different. Our objective was to lay out strategies to generate training labels for a symptom classification task and demonstrate that if sufficient training data are provided, such strategies will yield good predictive performance. We did not aim to extract all symptoms from the notes or create new named entity recognition models. The use of the external data set, labeled by Steinkamp et al [52], was meant to show that (1) our models, although trained on SHC data, perform well on another institution's data and, (2) considering that our models were trained on pseudolabels, they performed well on a test set containing gold labels.

Recent work has also seen the rise in transformers for NLP tasks. Although these methods are gaining popularity, the adaptation of such language model to the clinical use case is not straightforward. First, transformer models usually have a relatively short fixed maximum input length (eg, 412 tokens for bidirectional encoder representations from transformers [BERT]–based models). Clinical notes in general, and progress notes in particular, tend to be much longer than that (eg, in our case, the note length is closer to a couple of thousands of tokens). Moreover, transformer-based models trained on open domain text are not suitable for clinical text and must be fine-tuned to maximize performance. Although some BERT adaptations for the clinical domain have been released recently (eg, ClinicalBERT [57], BioBERT [58], or BlueBERT [59]), these publicly available models might not be suitable for the task at hand. Reasons why BERT-based models might not be suitable include attention dilution and the use of subword tokenization rather than word-level tokenization [60]. Finally, finding the best embedding method for note classification was outside the scope of our study. For these reasons, we did not include transformers in our comparison.

Conclusions

In this study, we introduced 2 methods to extract labels from EHR data sets for the training of a classifier for clinical notes. Multiple featurization methods were investigated, showing that PV-DBOW is clearly superior in terms of transferability and scaling. Although the use of ICD-10 codes present in the encounter data is a simple way of extracting training labels, the poor accuracy of the coding leads to less accurate models. Using a weak labeling pipeline to extract such labels yields improved performance and allows for the use of more notes as we are not relying on the presence of codes. Both approaches have been validated with an external set of notes containing gold labels, which showed the superiority of the weak labeling approach. Using ICD-10 codes for initial labels, we grouped a wide variety of signs and symptoms under the same label, learning classes of symptoms rather than specific symptoms. For example, R06 (abnormalities of breathing) covers a variety of breathing abnormalities; for example, dyspnea, wheezing, or hyperventilation. Such granularity in the symptoms is beyond the scope of this study and thus has not been investigated. However, the good performance of the weak labeling pipeline suggests that such an approach to generate more granular labels (eg, to distinguish between wheezing and shortness of breath in the R06 category) could be used. Moreover, the nature of the *one-versus-all* approach allows us to add a new category without having to retrain our model on all labels. Finally, the good performance and computational efficiency of the PV-DBOW features with logistic regression model would make such an expansion of the model computationally cheap.

Data and Code Availability

Protected Health Information restrictions apply to the availability of the Stanford Health Care clinical data set presented here, which were used under institutional review board approval for use only in this study and thus are not publicly available. The code can be made available upon request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary tables.
[DOCX File , 43 KB-Multimedia Appendix 1]

References

1. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc 2019 Apr 01;26(4):364-379 [FREE Full text] [doi: 10.1093/jamia/ocy173] [Medline: 30726935]
2. Forbush TB, Gundlapalli AV, Palmer MN, Shen S, South BR, Divita G, et al. "Sitting on pins and needles": characterization of symptom descriptions in clinical notes". AMIA Jt Summits Transl Sci Proc 2013;2013:67-71 [FREE Full text] [Medline: 24303238]
3. Adnan K, Akbar R, Khor S, Ali A. Role and challenges of unstructured big data in healthcare. In: Data Management, Analytics and Innovation. Singapore: Springer; 2020:301-323.
4. Koleck TA, Tatonetti NP, Bakken S, Mitha S, Henderson MM, George M, et al. Identifying symptom information in clinical notes using natural language processing. Nurs Res 2021;70(3):173-183. [doi: 10.1097/NNR.0000000000000488] [Medline: 33196504]
5. Luo X, Gandhi P, Storey S, Zhang Z, Han Z, Huang K. A computational framework to analyze the associations between symptoms and cancer patient attributes post chemotherapy using EHR data. IEEE J Biomed Health Inform 2021 Nov;25(11):4098-4109. [doi: 10.1109/jbhi.2021.3117238]
6. Forsyth AW, Barzilay R, Hughes KS, Lui D, Lorenz KA, Enzinger A, et al. Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. J Pain Symptom Manag 2018 Jun;55(6):1492-1499 [FREE Full text] [doi: 10.1016/j.jpainsymman.2018.02.016] [Medline: 29496537]
7. Mandl KD, Overhage JM, Wagner MM, Lober WB, Sebastiani P, Mostashari F, et al. Implementing syndromic surveillance: a practical guide informed by the early experience. J Am Med Inform Assoc 2004;11(2):141-150 [FREE Full text] [doi: 10.1197/jamia.M1356] [Medline: 14633933]
8. Katz R, May L, Baker J, Test E. Redefining syndromic surveillance. J Epidemiol Glob Health 2011 Dec;1(1):21-31 [FREE Full text] [doi: 10.1016/j.jegh.2011.06.003] [Medline: 23856373]
9. Crabb BT, Lyons A, Bale M, Martin V, Berger B, Mann S, et al. Comparison of international classification of diseases and related health problems, tenth revision codes with electronic medical records among patients with symptoms of coronavirus disease 2019. JAMA Netw Open 2020 Aug 03;3(8):e2017703 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.17703] [Medline: 32797176]
10. Wagner T, Shweta F, Murugadoss K, Awasthi S, Venkatakrishnan AJ, Bade S, et al. Augmented curation of clinical notes from a massive EHR system reveals symptoms of impending COVID-19 diagnosis. Elife 2020 Jul 07;9:1-12 [FREE Full text] [doi: 10.7554/eLife.58227] [Medline: 32633720]
11. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc 2011 Oct;18(5):540-543 [FREE Full text] [doi: 10.1136/amiajnl-2011-000465] [Medline: 21846785]
12. Friedman C, Rindflesch TC, Corn M. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. J Biomed Inform 2013 Oct;46(5):765-773 [FREE Full text] [doi: 10.1016/j.jbi.2013.06.004] [Medline: 23810857]
13. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, et al. Using clinical Natural Language Processing for health outcomes research: overview and actionable suggestions for future advances. J Biomed Inform 2018 Dec;88:11-19 [FREE Full text] [doi: 10.1016/j.jbi.2018.10.005] [Medline: 30368002]
14. Patel R, Tanwani S. Application of machine learning techniques in clinical information extraction. In: Smart Techniques for a Smarter Planet. Cham: Springer; 2019:145-165.
15. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform 2020 Mar 31;8(3):e17984 [FREE Full text] [doi: 10.2196/17984] [Medline: 32229465]

16. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc 2018 Oct 01;25(10):1419-1428 [FREE Full text] [doi: 10.1093/jamia/ocy068] [Medline: 29893864]
17. Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. ACL AFNLP 2009. [doi: 10.3115/1690219.1690287]
18. Naseem U, Khushi M, Khan SK, Shaukat K, Moni MA. A comparative analysis of active learning for biomedical text mining. Appl Syst Innov 2021 Mar 15;4(1):23. [doi: 10.3390/asi4010023]
19. Kholghi M, Sitbon L, Zuccon G, Nguyen A. Active learning: a step towards automating medical concept extraction. J Am Med Inform Assoc 2016 Mar;23(2):289-296 [FREE Full text] [doi: 10.1093/jamia/ocv069] [Medline: 26253132]
20. Chen Y, Lasko TA, Mei Q, Denny JC, Xu H. A study of active learning methods for named entity recognition in clinical text. J Biomed Inform 2015 Dec;58:11-18 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.010] [Medline: 26385377]
21. Ratner A, De SC, Wu S, Selsam D, Ré C. Data programming: creating large training sets, quickly. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016 Presented at: 30th International Conference on Neural Information Processing Systems; December 5 - 10, 2016; Barcelona Spain p. 3574-3582 URL: https://dl.acm.org/doi/10.5555/3157382.3157497
22. Banerjee I, Li K, Seneviratne M, Ferrari M, Seto T, Brooks JD, et al. Weakly supervised natural language processing for assessing patient-centered outcome following prostate cancer treatment. JAMIA Open 2019 Apr;2(1):150-159 [FREE Full text] [doi: 10.1093/jamiaopen/ooy057] [Medline: 31032481]
23. Fries JA, Varma P, Chen VS, Xiao K, Tejeda H, Saha P, et al. Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. Nat Commun 2019 Jul 15;10(1):3111 [FREE Full text] [doi: 10.1038/s41467-019-11012-3] [Medline: 31308376]
24. Dunnmon JA, Ratner AJ, Saab K, Khandwala N, Markert M, Sagreiya H, et al. Cross-modal data programming enables rapid medical machine learning. Patterns (N Y) 2020 May 08;1(2):100019 [FREE Full text] [doi: 10.1016/j.patter.2020.100019] [Medline: 32776018]
25. Fries J, Wu S, Ratner A, Ré C. SwellShark: a generative model for biomedical named entity recognition without labeled data. arXiv 2017 [FREE Full text]
26. Lygrisse KA, Roof MA, Keitel LN, Callaghan JJ, Schwarzkopf R, Bedard NA. The inaccuracy of ICD-10 coding in revision total hip arthroplasty and its implication on revision data. J Arthroplasty 2020 Oct;35(10):2960-2965. [doi: 10.1016/j.arth.2020.05.013] [Medline: 32507451]
27. Logan R, Davey P, De Souza N, Baird D, Guthrie B, Bell S. Assessing the accuracy of ICD-10 coding for measuring rates of and mortality from acute kidney injury and the impact of electronic alerts: an observational cohort study. Clin Kidney J 2020 Dec;13(6):1083-1090 [FREE Full text] [doi: 10.1093/ckj/sfz117] [Medline: 33391753]
28. McIsaac DI, Hamilton GM, Abdulla K, Lavallée LT, Moloo H, Pysyk C, et al. Validation of new ICD-10-based patient safety indicators for identification of in-hospital complications in surgical patients: a study of diagnostic accuracy. BMJ Qual Saf 2020 Mar;29(3):209-216. [doi: 10.1136/bmjqs-2018-008852] [Medline: 31439760]
29. Samannodi M, Hansen M, Hasbun R. Lack of accuracy of the international classification of disease, ninth (ICD-9) codes in identifying patients with encephalitis. J Neurol 2019 Apr;266(4):1034-1035. [doi: 10.1007/s00415-019-09229-9] [Medline: 30729315]
30. Horsky J, Drucker E, Ramelson H. Accuracy and completeness of clinical coding using ICD-10 for ambulatory visits. AMIA Annu Symp Proc 2017;2017:912-920 [FREE Full text] [Medline: 29854158]
31. Weng W, Wagholikar KB, McCray AT, Szolovits P, Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Med Inform Decis Mak 2017 Dec 01;17(1):155 [FREE Full text] [doi: 10.1186/s12911-017-0556-8] [Medline: 29191207]
32. Xu K, Lam M, Pang J, Gao X, Band C, Mathur P, et al. Multimodal machine learning for automated ICD coding. arXiv 2018 [FREE Full text]
33. Haendel MA, Chute CG, Robinson PN. Classification, ontology, and precision medicine. N Engl J Med 2018 Dec 11;379(15):1452-1462. [doi: 10.1056/NEJMra1615014] [Medline: 30304648]
34. Stanfill MH, Williams M, Fenton SH, Jenders RA, Hersh WR. A systematic literature review of automated clinical coding and classification systems. J Am Med Inform Assoc 2010;17(6):646-651 [FREE Full text] [doi: 10.1136/jamia.2009.001024] [Medline: 20962126]
35. Baumel T, Nassour-Kassis J, Cohen R, Elhadad M, Elhadad N. Multi-label classification of patient notes a case study on ICD code assignment. arXiv 2017 [FREE Full text]
36. Shi H, Xie P, Hu Z, Zhang M, Xing E. Towards automated ICD coding using deep learning. arXiv 2017 [FREE Full text]
37. Huang J, Osorio C, Sy LW. An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. Comput Methods Programs Biomed 2019 Aug;177:141-153. [doi: 10.1016/j.cmpb.2019.05.024] [Medline: 31319942]
38. Campbell S, Giadresco K. Computer-assisted clinical coding: a narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals. Health Inf Manag 2020 Jan;49(1):5-18. [doi: 10.1177/1833358319851305] [Medline: 31159578]

39. Goldstein I, Arzrumtsyan A, Uzuner O. Three approaches to automatic assignment of ICD-9-CM codes to radiology reports. AMIA Annu Symp Proc 2007 Oct 11:279-283 [FREE Full text] [Medline: 18693842]
40. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform 2018 Dec;22(5):1589-1604. [doi: 10.1109/JBHI.2017.2767063] [Medline: 29989977]
41. Xie P, Xing E. A neural architecture for automated ICD coding. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018 Presented at: 56th Annual Meeting of the Association for Computational Linguistics; July, 2018; Melbourne, Australia p. 1066-1076. [doi: 10.18653/v1/p18-1098]
42. Ratner A, Bach S, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: rapid training data creation with weak supervision. VLDB J 2020;29(2):709-730. [doi: 10.1007/s00778-019-00552-1] [Medline: 32214778]
43. Honnibal M, Montani I, Van LS, Boyd A. Industrial-strength Natural Language Processing in Python. spaCy. 2020. URL: https://spacy.io/ [accessed 2022-02-03]
44. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst 2013;3111:3119. [doi: 10.18653/v1/d16-1146]
45. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci Data 2019 May 10;6(1):52 [FREE Full text] [doi: 10.1038/s41597-019-0055-0] [Medline: 31076572]
46. Le Q, Mikolov T. Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning. 2014 Presented at: 31st International Conference on Machine Learning; June 21–26, 2014; Beijing, China p. 1188-1196 URL: https://proceedings.mlr.press/v32/le14.html
47. Tamang S. CLEVER base terrminology. GitHub. URL: https://github.com/stamang/CLEVER [accessed 2022-02-03]
48. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019 Presented at: 18th BioNLP Workshop and Shared Task; August, 2019; Florence, Italy. [doi: 10.18653/v1/w19-5034]
49. Wei C, Peng Y, Leaman R, Davis A, Mattingly C, Li J, et al. Overview of the BioCreative V Chemical Disease Relation (CDR) Task. In: Proceedings of the Fifth BioCreative Challenge Evaluation Workshop. 2015 Presented at: Fifth BioCreative Challenge Evaluation Workshop; 2015; Spain p. 154-166 URL: https://biocreative.bioinformatics.udel.edu/media/store/files/2015/BC5CDR_overview.final.pdf
50. Breiman L. Random forests. Statistics Department, University of California, Berkeley, CA. 2001. URL: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf [accessed 2022-02-10]
51. Hastie T, Friedman J, Tibshirani R. The elements of statistical learning: mining, inference, and prediction. In: Springer Series in Statistics. New York: Springer; 2001.
52. Steinkamp JM, Bala W, Sharma A, Kantrowitz JJ. Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes. J Biomed Inform 2020 Feb;102:103354 [FREE Full text] [doi: 10.1016/j.jbi.2019.103354] [Medline: 31838210]
53. Uzuner O, Solti I, Xia F, Cadag E. Community annotation experiment for ground truth generation for the i2b2 medication challenge. J Am Med Inform Assoc 2010;17(5):519-523 [FREE Full text] [doi: 10.1136/jamia.2010.004200] [Medline: 20819855]
54. Moons E, Khanna A, Akkasi A, Moens M. A comparison of deep learning methods for ICD coding of clinical records. Appl Sci 2020 Jul 30;10(15):5262. [doi: 10.3390/app10155262]
55. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
56. Miranda-Excalada A, Gonzalez-Agirre A, Armengol-Estapé J, Krallinger M. Overview of automatic clinical coding: annotations, guidelines, and solutions for non-English clinical cases at CodiEsp track of eHealth CLEF 2020. CLEF (Working Notes) 2020. 2020. URL: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=1UFCgX0AAAAJ&citation_for_view=1UFCgX0AAAAJ:wbdj-CoPYUoC [accessed 2022-02-03]
57. Alsentzer E, Murphy J, Boag W, Weng W, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019 Presented at: 2nd Clinical Natural Language Processing Workshop; June, 2019; Minneapolis, Minnesota, USA p. 72-78. [doi: 10.18653/v1/w19-1909]
58. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
59. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019 Presented at: 18th BioNLP Workshop and Shared Task; August, 2019; Florence, Italy p. 58-65. [doi: 10.18653/v1/w19-5006]
60. Gao S, Alawad M, Young MT, Gounley J, Schaefferkoetter N, Yoon H, et al. Limitations of transformers on clinical text classification. IEEE J Biomed Health Inform 2021 Sep;25(9):3596-3607. [doi: 10.1109/jbhi.2021.3062322]

Abbreviations

AUROC: area under the receiver operating characteristic curve
BERT: bidirectional encoder representations from transformers
BOW: bag-of-words
CBOW: continuous bag-of-words
EHR: electronic health record
ICD-10: International Classification of Disease–10th revision
i2b2: Informatics for Integrating Biology and the Bedside
MIMIC: Medical Information Mart for Intensive Care
NLP: natural language processing
PV-DBOW: paragraph vector–distributed bag-of-words
SHC: Stanford Health Care
TF-IDF: term frequency–inverse document frequency

Edited by J Hefner; submitted 13.08.21; peer-reviewed by J Coquet, H Park, X Dong; comments to author 03.09.21; revised version received 12.11.21; accepted 16.12.21; published 14.03.22

Please cite as:
Humbert-Droz M, Mukherjee P, Gevaert O
Strategies to Address the Lack of Labeled Data for Supervised Machine Learning Training With Electronic Health Records: Case Study for the Extraction of Symptoms From Clinical Notes
JMIR Med Inform 2022;10(3):e32903
URL: https://medinform.jmir.org/2022/3/e32903
doi: 10.2196/32903
PMID: