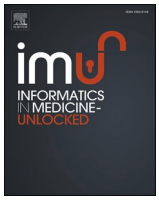
Informatics in Medicine Unlocked 36 (2023) 101139

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu



Comparison of BERT implementations for natural language processing of narrative medical documents

Alexander Turchin [a,b,*], Stanislav Masharsky [c], Marinka Zitnik [b]

[a] Brigham and Women's Hospital, Boston, MA, USA
[b] Harvard Medical School, Boston, MA, USA
[c] First Line Software, Cambridge, MA, USA

ARTICLE INFO

Keywords:
Natural language processing
Neural networks
Representation learning
Embeddings
Medical named entity recognition
BERT

ABSTRACT

Background and objectives: Bidirectional Encoder Representations from Transformers (BERT) word embedding models have been successfully used for many natural language processing (NLP) tasks, including medical named entity recognition. However, there are many more linguistically complicated concepts in healthcare documentation, often reflecting medical decision-making processes or complex patient characteristics, where performance of transformer-based models has not been as well investigated. Furthermore, the dataset on which a BERT model has been pre-trained could affect performance.

Methods: We compared accuracy of identification of three linguistically complex medical concepts – a) discussion of bariatric surgery between patients and their healthcare providers; b) non-acceptance of statin treatment recommendation by patients; and c) tobacco use status documentation – by three BERT implementations: regular BERT; BioBERT and ClinicalBERT. For each of the three NLP tasks, all three BERT implementations were trained on a manually annotated training dataset of outpatient provider notes and then evaluated on a held-out manually annotated test dataset. All datasets were obtained from the electronic health record system of Mass General Brigham. Filtering by keywords was used to improve class balance by undersampling the null class.

Results: Prevalence of study labels (concepts) ranged from 1.3% to 11.8% and was similar between training and held-out validation datasets within each task-model combination. Over 80% of NLP tasks achieved recall and 75% of tasks achieved precision between 0.4 and 0.9. Among different study evaluation categories, F1 score ranged from 0.0 to 0.860. Macro-averaged F1 score ranged from 0.466 to 0.854.

Overall, ClinicalBERT achieved best performance (by F1-macro score) in the Bariatric Surgery task, BioBERT in the Tobacco Use task and regular BERT in the Statin Non-Acceptance task. The mean macro-F1 score across all task-model pairs was 0.761 for ClinicalBERT, 0.735 for BioBERT and 0.699 for regular BERT.

Conclusions: BERT implementations trained on documents from biomedical domain – both BioBERT and ClinicalBERT – achieve superior NLP performance for identifying a range of complex medical concepts compared to regular BERT. Neither of the two biomedical BERT implementations we tested attained clearly greater accuracy than the other.

1. Introduction

Electronic health records are now near universal in the U.S [1–3]. and are a rich data source for research, quality improvement and population management [4–6]. Amount of data available in electronic health records has been growing exponentially [7] and therefore efficient and effective computational analytical methods are needed to fully realize its potential benefits.

An important component of electronic health records are narrative documents [8–10]. These contain a large amount of data not found in structured database tables: nuanced assessments of the patient's condition, reasoning behind choice of treatment, documentation of patient-provider discussions, etc. Natural language processing (NLP) has been effectively employed to study narrative EHR data [11–13]. NLP has been particularly successful in the area of named entity recognition, enabling identification of diagnoses, medications and other concepts that are described by a single word or several closely spaced words (e.g. *pneumonia* or *myocardial infarction*). On the other hand, linguistic

* Corresponding author. Division of Endocrinology, Brigham & Women's Hospital, 221 Longwood Avenue, Boston, MA, 02115, USA.
E-mail address: aturchin@bwh.harvard.edu (A. Turchin).

https://doi.org/10.1016/j.imu.2022.101139
Received 16 August 2022; Received in revised form 25 November 2022; Accepted 29 November 2022
Available online 30 November 2022

constructs comprised of terms spaced apart from each other in the sentence or even in different sentences, can be more challenging [14]. Complex concepts of this nature therefore represent a critical next frontier in analysis of EHR data [15,16].

One of the recent advances in NLP has been development of transfer learning and contextual word embedding models, such as Bidirectional Encoder Representations from Transformers (BERT) [17]. While the original (base) BERT was developed for analysis of general domain text, subsequently specialized versions have been proposed for analysis of narrative biomedical text, including BioBERT [18] and ClinicalBERT [19]. Performance of different versions of BERT models has been compared on several NLP tasks, including named entity recognition (NER) [20], biomedical entity normalization [21] and next sentence prediction [19]. However, less is known about these models' relative performance in identification of more complicated linguistic structures, including the ones that consist of components that may be spaced apart from each other in text.

At the same time, these complex linguistic constructs can carry important information about patient care. For example, they may describe discussion of treatment options between the healthcare provider and the patient or the patient's opinion about the treatment proposed by their clinician. Recently studied examples include documentation of discussion of bariatric (weight loss) surgery, shown to be associated with subsequent receipt of the surgery and weight loss [22]; and non-acceptance of medications recommended to the patient, often followed by poor disease control [23,24]. We therefore conducted this study to compare performance of regular BERT and its biomedical versions on a range of complex linguistic concepts in narrative electronic health record data.

2. Materials and methods

2.1. Objective

To compare accuracy of identification of a range of biomedical concepts in narrative medical documents using general and biomedical implementations of the BERT text representation model.

2.2. Study settings

The study was conducted at Mass General Brigham (formerly known as Partners Healthcare) – an integrated healthcare delivery system in eastern Massachusetts that includes several tertiary care, specialty and community hospitals and multiple affiliated outpatient practices. Mass General Brigham has been using an electronic health record since 2000 (most recently Epic). All study analyses were conducted using outpatient provider notes extracted from Mass General Brigham electronic health records. The study was reviewed by Mass General Brigham institutional review board and the requirement for informed consent was waived.

2.3. NLP tasks

The study involved three groups of NLP tasks: a) discussion of bariatric (weight loss) surgery between providers and patients; b) non-acceptance of HMG-CoA reductase inhibitors (statins) medications by patients and c) tobacco use status. The ultimate goal of task (a) was to develop an NLP tool that could be used to study how commonly clinicians discuss bariatric surgery with eligible patients and clinical outcomes of patients who do vs. do not have these discussions with their clinicians. This task included identification of documentation of two concepts in provider notes: i) whether the patient had a previous bariatric surgery (e.g. *she had a gastric bypass 10 years ago*) and ii) whether the clinician discussed the possibility of bariatric surgery with the patient (e.g. *he has class 2 obesity and I recommended that he consider surgery*). The ultimate goal of task (b) was to develop an NLP tool that could be used to study how commonly patients do not accept statins (a class of cholesterol-lowering medications) recommended to them by their clinicians. This task involved identification of a single concept in provider notes: documentation of non-acceptance of statin recommendations by patients (e.g. *she adamantly is not interested in any statins as she is already taking too many medications*). Task (c) involved identification of documentation of three concepts in provider notes: i) the patient never used tobacco products (e.g. *Smoking: never*); ii) the patient previously, but not currently, used tobacco products (e.g. *Mr. Johnson quit smoking after his MI five years ago*) and iii) the patient is currently using tobacco products (e.g. *she smokes 1 pack per day*).

Each NLP task utilized its own set of provider notes, selected at random from among patients for whom documentation of the task concept could be expected. Specifically, the bariatric surgery discussion NLP task utilized a training set of 2623 notes and a held-out validation set of 1498 notes of patients with body mass index (BMI) $\geq$ 35 kg/m^2 (and thus potentially eligible for bariatric surgery). The statin non-acceptance NLP task utilized a training set of 20,974 notes and a held-out validation set of 1996 notes of patients at high cardiovascular risk (e.g. with coronary artery disease or diabetes mellitus). The tobacco use task utilized a training set of 2910 notes and a held-out validation set of 972 notes of patients at high cardiovascular risk (for whom it was therefore clinically important to document tobacco use status).

2.4. BERT implementations

BERT is a contextualized text representation model based on a masked language model and is pre-trained using bidirectional transformer encoder architecture [25]. Three BERT implementations were used in the study. The first one was the original (base) BERT [17] pre-trained on BookCorpus [26] and English Wikipedia. The second was BioBERT, which is initialized with weights from the original BERT, and then pre-trained on PubMed abstracts and PubMed Central full-text articles [18]. The third model was ClinicalBERT [19], which is pre-trained on Medical Information Mart for Intensive Care III (MIMIC-III) [27]. All BERT models used a SoftMax layer but no other classifiers.

2.5. Training and evaluation of NLP tools

Documents in the training and held-out validation datasets for all three tasks were manually annotated by trained clinicians for the labels (concepts) being sought. Annotators were trained by a senior practicing clinician (AT) who developed the definition (gold standard) of each study concept. Annotated documents were cleaned using a custom-designed automated process that created uniform representations for common abbreviations (e.g. changing *Ph.D.* to *PhD*), removed non-ASCII characters, inserted missing spaces between words (e.g. changing *ended. New* to *ended. New*), replaced repeating characters with a single character for both non-whitespace (e.g. changing ******* to *) and white-space (e.g. replacing five tabs in a row with a single tab) characters. Cleaning was carried out in part to reduce the number of unique tokens BERT would have to consider that did not represent distinct semantic values. Removing of numbers and stop words (e.g. "a" or "the") as well as stemming were not carried out because tokenization by BERT is a more effective approach to generation of contextually representative word embeddings.

For each task, two document selection streams were utilized. In the unselected stream, all documents in the dataset were used for training of the BERT models. In the filtered stream, in order to achieve closer balance of document classes (with vs. without labels), documents in the training set were further processed (using regular expressions) to identify expressions potentially indicative of the concept of interest; only the documents where these tokens were found were used for training of the BERT models. Regular expressions used in each of the tasks are listed in Table 1 (Python implementation of regular expressions was used). No enrichment was carried out on the held-out validation dataset.

To fine-tune the models, we first conducted 200 cycles of a random

Table 1
Regular expressions used to enrich training documents.

Task	Regular Expressions
Bariatric Surgery	bariatric
	obesity.*surg
	gastric
	obesity treatment
	RYGB
	SMP
	duodenal shunt
	gastrectomy
	weight loss.*surg
	surg.*weight loss
	reduc.*surgery
	surgery.*reduc
	Lap.{,10}Band
	LAGB
	WLS
Statin non-acceptance	Statin
	decline|refuse|avoid|cholesterol|lipid.*therapy
	cholesterol|lipid.*wish|untreated
	chol.{,10} meds
	history of MI.*compliant.*medication
	severe CAD
Tobacco use	smok
	chantix
	vareniclin
	nicotin
	tob
	bacco
	cig

search hyperparameter optimization using 3-fold cross-validation. Ten top-performing hyperparameter combinations for each task and model were selected and then 10-fold cross-validation was used to identify the best performing hyperparameter combination to be used in the model. Macro-averaged F1 score was used as the optimization target. Hyperparameters that were optimized for each model are listed in Table 2.

To train the BERT models, we initialized the encoder's weight parameters from their respective pre-trained weight parameters; a randomly initialized linear layer head was used for the classification task. Separate learning rates were used for main BERT and for the classifier layer. Gradient clipping was used to reduce exploding gradients [28] and gradient accumulation to enable splitting sample batches into smaller mini-batches (8 sequences of 512 tokens each) to optimize utilization of GPU memory [29]. Adam algorithm with decoupled weight decay modification [30] and Cosine Annealing learning rate scheduler [31] were used to train the models. During training, over-represented classes (e.g. empty class) were downsampled and under-represented classes were oversampled to achieve balance of all classes under analysis. Text was processed by BERT in chunks of 500 tokens, with an overlap of 100 tokens between chunks.

To evaluate the models' performance, we calculated recall (sensitivity) and precision (positive predictive value) [32] for each label class. We used these to calculate F1 score for each model, and macro- and micro-averaged F1 scores [33] for each task.

Table 2
Hyperparameters optimized in BERT models.

Hyperparameter	Lower Limit	Upper Limit
Number of epochs	2	8
Main BERT learning rate	2×10^{-5}	10^{-4}
Classifier learning rate	10^{-5}	3×10^{-4}
Gradient accumulation steps	1, 2 or 4	
Balance	True or False	

3. Results

3.1. Study data

The number of documents in the training datasets ranged from 2623 to 20,974, and the mean number of words per document from 346.5 to 389.5 (Table 3). Prevalence of study labels (concepts) ranged from 1.3% to 11.8% and was similar between training and held-out validation datasets within each task-model combination (Table 4). The filtered processing stream, as expected, selected a smaller subset of documents for training with a higher prevalence of documents with positive labels (Table 4 and Table 5).

3.2. Performance of BERT models

In evaluation against the held-out validation dataset (Table 7, Table 9 and Table 11), among different study evaluation categories, over 80% of categories achieved recall and 75% of categories achieved precision between 0.4 and 0.9. F1 score ranged from 0.0 to 0.860. Macro-averaged F1 score ranged from 0.466 to 0.854. Three out of five numerically highest metrics were in the Tobacco Use task, and three out five numerically lowest metrics were in the Statin Non-Acceptance task. Both the highest (0.854) and the lowest (0.466) mean macro-averaged F1 score were in the Statin Non-Acceptance task. Performance of BERT NLP tools was consistent between cross-validation (Table 6, 8 and 10) and the held-out validation dataset.

For most tasks and models, in 27 out of 36 evaluated categories, recall was lower than precision. In some cases, such as the regular BERT model of current tobacco use and ClinicalBERT model of statin non-acceptance utilizing filtered training dataset, it was more than two-fold lower. On the other hand, recall was higher than precision for all three BERT models of identification of both a) no previous history of tobacco use and b) past history of tobacco use that utilized the unselected training dataset.

Enriching the training dataset to rebalance positive and negative labels resulted in a higher F1 score in 9 of the 18 evaluated categories. In four of the nine categories where enrichment led to a higher F1 score, it resulted in a higher precision only; and in five categories in both higher recall and precision. There were no categories where enrichment led to a higher F1 score solely through a higher recall.

Overall, ClinicalBERT achieved best performance (by F1-macro score) in the Bariatric Surgery task, BioBERT in the Tobacco Use task and regular BERT in the Statin Non-Acceptance task (Table 7, Table 9, Table 11 and Fig. 1). The mean macro-F1 score across all tasks/models was 0.761 for ClinicalBERT, 0.735 for BioBERT and 0.699 for regular BERT.

4. Discussion

In this study of natural language processing of a broad range of linguistically complex medical concepts we found that BERT implementations focused on biomedical terminology performed better than general BERT. On the other hand, differences between two biomedical BERT implementations – one focused on scientific literature (BioBERT)

Table 3
Characteristics of training and validation datasets.

Task	Training Dataset		Validation Dataset	
	Number of documents	Mean words per document	Number of documents	Mean words per document
Bariatric surgery	2623	377.1	1498	322.7
Statin non-acceptance	20,974	346.5	1996	376.5
Tobacco use	2910	389.5	972	386.1

Table 4
Prevalence of positive labels in unfiltered study datasets.

Task	Model	Positive Labels	
		Training Dataset	Validation Dataset
Bariatric surgery	Past surgery	63 (2.4%)	30 (2.0%)
	Discussion	132 (5.0%)	49 (3.3%)
Statin non-acceptance	Non-acceptance	263 (1.3%)	24 (1.2%)
Tobacco use	Never	344 (11.8%)	107 (11.0%)
	Past	158 (5.4%)	60 (6.2%)
	Current	99 (3.4%)	37 (3.8%)

Table 5
Prevalence of positive labels in filtered study datasets.

Task	Model	Positive Labels	
		Training Dataset [1]	Validation Dataset
Bariatric surgery	Past surgery	63 (15.9%)	27 (16.9%)
	Discussion	132 (33.3%)	48 (30.0%)
Statin non-acceptance	Non-acceptance	263 (4.7%)	24 (4.4%)
Tobacco use	Never	344 (47.6%)	107 (43.5%)
	Past	158 (21.9%)	60 (24.4%)
	Current	99 (13.7%)	37 (15.0%)

and the other on medical documentation (ClinicalBERT) – were smaller and inconsistent.

Superior performance in recognition of medical concepts of BERT implementations that were trained on biomedical texts is intuitive and consistent with previously published investigations that found that domain-specific BERT implementations outperformed general BERT in their respective areas of focus [34–36]. Based on similar assumptions, it might have been expected that ClinicalBERT would achieve higher accuracy than BioBERT in analysis of medical documentation. However, ClinicalBERT's superiority was not uniform. Several factors could have accounted for this finding. On the one hand, ClinicalBERT was trained on the MIMIC-III dataset that includes inpatient documentation on patients hospitalized in intensive care units, whereas all documents analyzed in this study were authored in ambulatory settings. It is therefore possible that the dataset used to train ClinicalBERT was missing some of the relevant medical terminology (e.g. neither bariatric surgery nor treatment of high cholesterol are commonly discussed in intensive care units).

On the other hand, it is worth noting that tobacco use was the domain where BioBERT performance significantly exceeded that of ClinicalBERT. This domain is less clinically focused than the other two (bariatric surgery and statin therapy) used in the study. Terminology from this domain may be more commonly found in scientific literature on which BioBERT was trained. Overall, our finding of small differences in performance between BERT implementations trained on different

Table 6
BERT model cross-validation performance: Bariatric surgery.

Model	Measure	BERT		BioBERT		ClinicalBERT	
		Unselected	Filtered	Unselected	Filtered	Unselected	Filtered
Past Surgery	Precision	0.650	0.378	0.724	0.878	0.787	0.704
	Recall	0.255	0.262	0.571	0.502	0.510	0.647
	F1	0.330	0.299	0.620	0.611	0.605	0.656
Surgery Discussion	Precision	0.698	0.649	0.617	0.860	0.785	0.782
	Recall	0.645	0.621	0.779	0.610	0.583	0.788
	F1	0.665	0.625	0.685	0.700	0.665	0.782
	F1-micro	0.951	0.949	0.951	0.965	0.957	0.970
	F1-macro	0.657	0.634	0.760	0.765	0.749	**0.808**

F1-micro and F1-macro scores include model performance on the "empty" class with no labels detected.

Table 7
BERT model held-out test performance: Bariatric surgery.

Model	Measure	BERT		BioBERT		ClinicalBERT	
		Unselected	Filtered	Unselected	Filtered	Unselected	Filtered
Past Surgery	Precision	0.700	0.733	0.731	0.944	0.483	0.850
	Recall	0.467	0.367	0.633	0.567	0.467	0.567
	F1	0.560	0.489	0.678	**0.708**	0.475	0.680
Surgery Discussion	Precision	0.651	0.821	0.667	0.935	0.587	0.793
	Recall	0.571	0.653	0.612	0.592	0.551	0.939
	F1	0.609	0.727	0.638	0.725	0.568	**0.860**
	F1-micro	0.967	0.975	0.969	0.977	0.957	0.983
	F1-macro	0.718	0.735	0.767	0.807	0.674	**0.844**

F1-micro and F1-macro scores include model performance on the "empty" class with no labels detected.

Table 8
BERT model cross-validation performance: Statin non-acceptance.

Measure	BERT		BioBERT		ClinicalBERT	
	Unselected	Filtered	Unselected	Filtered	Unselected	Filtered
Precision	0.673	0.669	0.655	0.655	0.812	0.840
Recall	0.433	0.285	0.498	0.498	0.437	0.396
F1	0.524	0.355	0.536	0.536	0.566	0.531
F1-micro	0.990	0.986	0.988	0.988	0.992	0.991
F1-macro	0.759	0.674	0.765	0.765	**0.780**	0.763

F1-micro and F1-macro scores include model performance on the "empty" class with no labels detected.

Table 9
BERT model held-out test performance: Statin non-acceptance.

Measure	BERT		BioBERT		ClinicalBERT	
	Unselected	Filtered	Unselected	Filtered	Unselected	Filtered
Precision	0.762	0.044	0.800	0.0	0.867	1.0
Recall	0.667	1.0	0.500	0.0	0.542	0.375
F1	**0.711**	0.084	0.615	0.0	0.667	0.545
F1-micro	0.993	0.738	0.992	0.988	0.993	0.992
F1-macro	0.854	0.466	0.806	0.497	**0.832**	0.770

F1-micro and F1-macro scores include model performance on the "empty" class with no labels detected.

Table 10
BERT cross-validation model performance: Tobacco use.

Model	Measure	BERT		BioBERT		ClinicalBERT	
		Unselected	Filtered	Unselected	Filtered	Unselected	Filtered
Never	Precision	0.585	0.947	0.597	0.963	0.623	0.947
	Recall	0.790	0.521	0.724	0.691	0.753	0.502
	F1	0.670	0.660	0.653	0.785	0.679	0.628
Past	Precision	0.534	0.871	0.556	0.876	0.513	0.858
	Recall	0.652	0.492	0.654	0.640	0.721	0.551
	F1	0.585	0.611	0.590	0.728	0.594	0.645
Current	Precision	0.667	0.838	0.699	0.827	0.733	0.865
	Recall	0.604	0.495	0.583	0.617	0.514	0.577
	F1	0.625	0.613	0.617	0.702	0.588	0.688
	F1-micro	0.867	0.895	0.872	0.928	0.874	0.900
	F1-macro	0.704	0.707	0.700	**0.794**	0.701	0.727

F1-micro and F1-macro scores include model performance on the "empty" class with no labels detected.

Table 11
BERT model held-out test performance: Tobacco use.

Model	Measure	BERT		BioBERT		ClinicalBERT	
		Unselected	Filtered	Unselected	Filtered	Unselected	Filtered
Never	Precision	0.585	0.947	0.597	0.963	0.623	0.947
	Recall	0.790	0.521	0.724	0.691	0.753	0.502
	F1	0.670	0.660	0.653	**0.785**	0.679	0.628
Past	Precision	0.534	0.871	0.556	0.876	0.513	0.858
	Recall	0.652	0.492	0.654	0.640	0.721	0.551
	F1	0.585	0.611	0.590	**0.728**	0.594	0.645
Current	Precision	0.667	0.838	0.699	0.827	0.733	0.865
	Recall	0.604	0.495	0.583	0.617	0.514	0.577
	F1	0.625	0.613	0.617	**0.702**	0.588	0.688
	F1-micro	0.894	0.882	0.895	0.927	0.888	0.899
	F1-macro	0.756	0.667	0.738	**0.792**	0.712	0.736

F1-micro and F1-macro scores include model performance on the "empty" class with no labels detected.

biomedical datasets without clear superiority of any particular implementation are consistent with the previously published results [20].

Our findings of superiority of biomedical BERT implementations over general BERT in analyses of medical texts, but lack of clear advantage of any particular biomedical implementation over another, are similar to the previously reported results. The study by Alsentzer et al. that compared several versions of clinically trained BERT (all different from the one tested in our investigation) to BioBERT and general BERT across several NLP tasks also demonstrated consistently higher accuracy of biomedical versions of BERT over the general version but generally similar performance between all biomedical versions [20]. Similarly, the original publication that described the particular implementation of ClinicalBERT that was tested in our study, also found consistent superiority of ClinicalBERT over general BERT (BioBERT was not evaluated in that study) [19].

No single task stood out for uniformly higher or uniformly lower performance across all BERT implementations and mean macro-averaged F1 scores across all BERT implementations were similar between the three tasks. This indicates that the NLP tasks had comparable level of complexity and thus well-suited for the comparative evaluation study we conducted.

Across the NLP tasks and BERT implementations, recall tended to be lower than precision. One possible explanation is that the limited training samples did not include all possible vocabulary (e.g. abbreviations/acronyms, misspellings and indirect references) that could represent the study concepts. Another potential reason is that complex concepts, like the ones analyzed in this study, are often represented by components located some distance from each other (e.g. anaphora or cataphora). While BERTs' ability to analyze sequences up to 512 tokens should improve handling these situations, high degree of variability of text between the concept components likely continues to present a challenge.

BERT and other pre-trained transformer models have emerged as silver bullets for many NLP tasks [37]. While they are one of the most successful deep learning models for NLP, their core limitation is that they can only process sequences in continuous 512-token spans because of quadratic dependency (mainly in terms of computational and memory requirement) on the sequence length. This sequence length, which provides the context in an NLP task, sacrifices the possibility that very distant tokens "pay attention" to each other. An interesting future

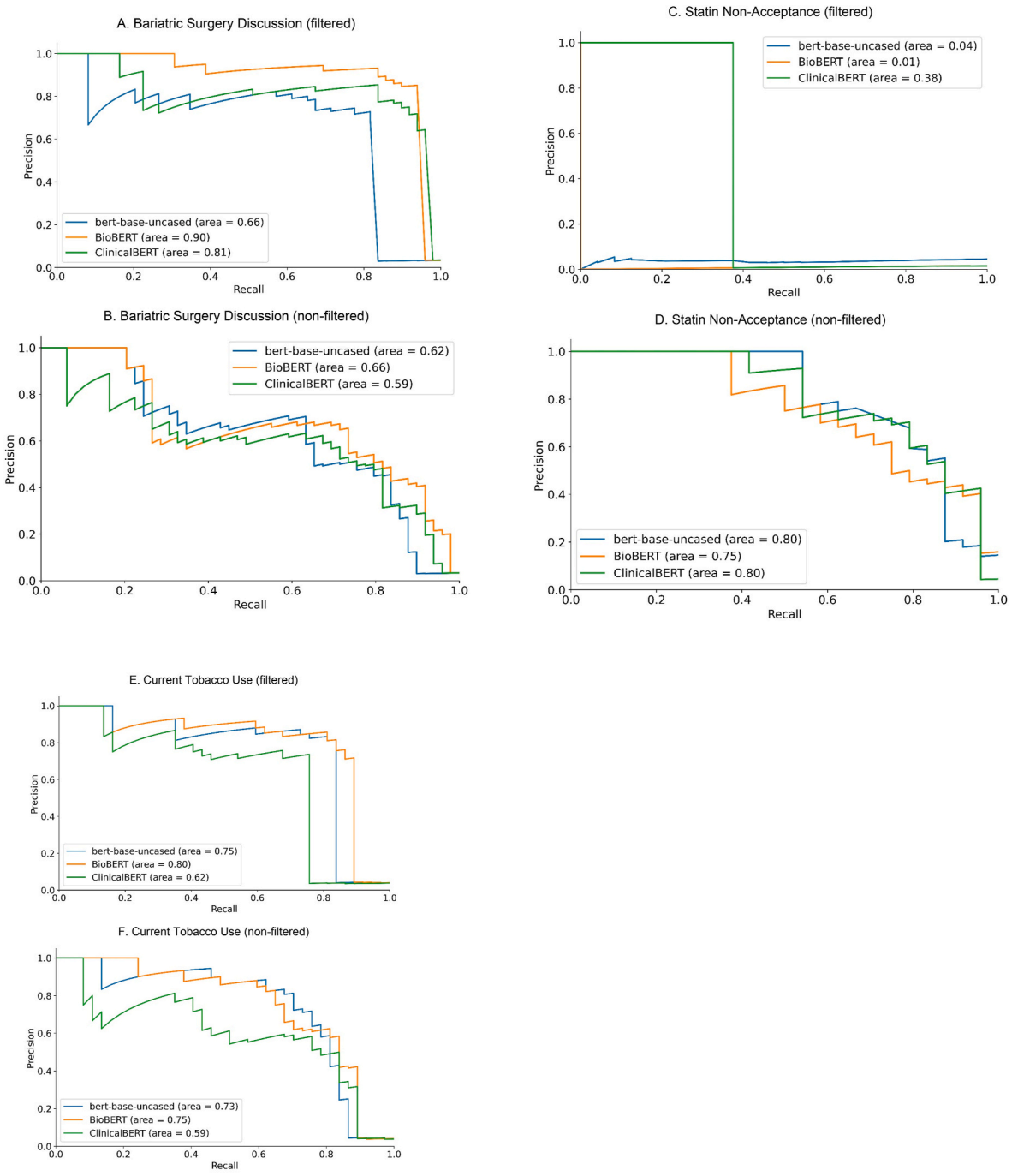


Fig. 1. Comparative performance of BERT models: Precision-recall Curves.

direction is to slice the text by a sliding window or first identify key sentences and then concatenate them for reasoning in a multi-step BERT analysis. Another worthy future direction would be to adapt general algorithms, such as CogLTX or BigBird, for medical named entity recognition [38,39].

Rebalancing of classes in the training dataset led to an increased accuracy of identification of non-null classes in half of the tasks. This finding is consistent with previously published studies and is a known weakness of many machine-learning NLP methodologies [40–42]. Keyword-based filtering to achieve undersampling of the most common (typically null) classes was effective in rebalancing the training datasets and improving performance. Increase in accuracy was more pronounced for NLP tasks where filtering led to greater prevalence of non-null classes (e.g. identification of discussion of bariatric surgery and documentation of past smoking). On the other hand, filtering resulted in no increase in accuracy for the task of identification of statin non-acceptance by patients where prevalence of the non-null class remained low (<5%) even after filtering. Biomedical implementations of BERT appear to have been able to take greater advantage of the class rebalancing achieved by filtering.

Findings of this study should be interpreted in the light of its limitations. We only tested one of several available ClinicalBERT implementations, and our findings may not generalize to the ones not included in this analysis. Furthermore, BERT implementations trained on different documents (e.g. with a greater focus on ambulatory/outpatient documentation) may have obtained different results. Our results may also not generalize to other biomedical concepts. Finally, all test data came from a single integrated healthcare delivery system in the

U.S. and therefore the findings may not apply to data from other settings.

5. Conclusions

In conclusion, we have found that both BERT implementations trained on documents from biomedical domain – both BioBERT and ClinicalBERT – achieve superior NLP performance in identifying a range of complex medical concepts compared to regular BERT trained only on Wikipedia. Neither of the two biomedical BERT implementations we tested attained clearly greater accuracy than the other.

Sources of funding

Patient-Centered Outcomes Research Institute (PCORI), Washington, DC.

Study sponsor played no role in collection, analysis or interpretation of the data; writing of the manuscript; and in the decision to submit the manuscript for publication.

Author contribution

Study design: Alexander Turchin, Stanislav Masharsky, Data collection: Alexander Turchin, Data analysis: Stanislav Masharsky, Manuscript drafting: Alexander Turchin, Critical review of the manuscript: Stanislav Masharsky, Marinka Zitnik, Funding: Alexander Turchin

Consent

Study was reviewed by the Mass General Brigham IRB and the requirement for patient consent was waived.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Henry J, Pylypchuk Y, Searcy T, Patel V. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008–2015. ONC data brief 2016;35(35):2008–15.
[2] Jamoom E, Yang N, Hing E. Adoption of certified electronic health record systems and electronic information sharing in physician offices: United States, 2013 and 2014. 2016.
[3] Hecht J. The future of electronic health records. Nature 2019;573(7775). S114-S114.
[4] Kruse CS, Stein A, Thomas H, Kaur H. The use of electronic health records to support population health: a systematic review of the literature. J Med Syst 2018;42(11):214.
[5] Rudrapatna VA, Butte AJ. Opportunities and challenges in using real-world data for health care. J Clin Investig 2020;130(2):565–74.
[6] Wikström K, Toivakka M, Rautiainen P, Tirkkonen H, Repo T, Laatikainen T. Electronic health records as valuable data sources in the health care quality improvement process. Health Serv Res Manag Epidemiol 2019;6:2333392819852879.
[7] Cyganek B, Graña M, Krawczyk B, et al. A survey of big data issues in electronic health record analysis. Appl Artif Intell 2016;30(6):497–520.
[8] Hicks J. The potential of claims data to support the measurement of health care quality. San Diego, CA: RAND; 2003 [PhD].
[9] Turchin A, Shubina M, Breydo E, Pendergrass ML, Einbinder JS. Comparison of information content of structured and narrative text data sources on the example of medication intensification. J Am Med Inf Assoc 2009;16(3):362–70.
[10] Skentzos S, Shubina M, Plutzky J, Turchin A. Structured vs. unstructured: factors affecting adverse drug reaction documentation in an EMR repository. AMIA Annu Symp Proc 2011:1270–9. 2011.
[11] Friedman C, Hripcsak G. Natural language processing and its future in medicine. Acad Med 1999;74(8):890–5.
[12] Ohno-Machado L. Realizing the full potential of electronic health records: the role of natural language processing. J Am Med Inf Assoc 2011;18(5):539.
[13] Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inf Assoc 2004;11(5):392–402.
[14] Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inf Assoc 2011;18(5):544–51.
[15] Groves P, Kayyali B, Knott D, Kuiken SV. The'big data'revolution in healthcare: accelerating value and innovation. 2016.
[16] Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inf Assoc 2012;20(1):117–21.
[17] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. 2018. *arXiv preprint arXiv:181004805*.
[18] Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40.
[19] Huang K, Altosaar J, Ranganath R. Clinicalbert: modeling clinical notes and predicting hospital readmission. 2019. *arXiv preprint arXiv:190405342*.
[20] Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. 2019. *arXiv preprint arXiv:190403323*.
[21] Ji Z, Wei Q, Xu H. Bert-based ranking for biomedical entity normalization. AMIA Summits Transl Sci Proc 2020:269. 2020.
[22] Chang LS, Malmasi S, Hosomura N, et al. Patient-provider discussions of bariatric surgery and subsequent weight changes and receipt of bariatric surgery. Obesity 2021;29(8):1338–46.
[23] Hosomura N, Malmasi S, Timerman D, et al. Decline of insulin therapy and delays in insulin initiation in people with uncontrolled diabetes mellitus. Diabet Med 2017;34(11):1599–602.
[24] Turchin A, Hosomura N, Zhang H, Malmasi S, Shubina M. Predictors and consequences of declining insulin therapy by individuals with type 2 diabetes. Diabet Med 2020;37(5):814–21.
[25] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Paper presented at: advances in neural information processing systems. 2017.
[26] Zhu Y, Kiros R, Zemel R, et al. Aligning books and movies: towards story-like visual explanations by watching movies and reading books. Proceedings of the IEEE international conference on computer vision; 2015. Paper presented at:.
[27] Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3(1):1–9.
[28] Zhang J, He T, Sra S, Jadbabaie A. Why gradient clipping accelerates training: a theoretical justification for adaptivity. 2019. *arXiv preprint arXiv:190511881*.
[29] Andersson A, Koriakina N, Sladoje N, Lindblad J. End-to-end multiple instance learning with gradient accumulation. 2022. *arXiv preprint arXiv:220303981*.
[30] Loshchilov I, Hutter F. Decoupled weight decay regularization. 2017. *arXiv preprint arXiv:171105101*.
[31] Loshchilov I, Sgdr Hutter F. Stochastic gradient descent with warm restarts. 2016. *arXiv preprint arXiv:160803983*.
[32] Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. J Am Med Inf Assoc 1994;1(2):142–60.
[33] Takahashi K, Yamamoto K, Kuchiba A, Koyama T. Confidence interval for micro-averaged F1 and macro-averaged F1 scores. Appl Intell 2022;52(5):4961–72.
[34] Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. 2019. *arXiv preprint arXiv:190310676*.
[35] Lee J-S, Hsiang J. Patentbert: patent classification with fine-tuning a pre-trained bert model. 2019. *arXiv preprint arXiv:190602124*.
[36] Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: the muppets straight out of law school. 2020. *arXiv preprint arXiv:201002559*.
[37] McDermott M, Yap B, Szolovits P, Zitnik M. Structure inducing pre-training. *arXiv e-prints*. 2021. arXiv: 2103.10334.
[38] Ding M, Zhou C, Yang H, Tang J. Cogltx: applying bert to long texts. Adv Neural Inf Process Syst 2020;33:12792–804.
[39] Zaheer M, Guruganesh G, Dubey KA, et al. Big bird: Transformers for longer sequences. Adv Neural Inf Process Syst 2020;33:17283–97.
[40] Casula C, Tonelli S. Hate speech detection with machine-translated data: the role of annotation scheme, class imbalance and undersampling. Paper presented at: Seventh Italian Conference on Computational Linguistics; 2020. CLiC-it 2020.
[41] Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. Neural Network 2018;106:249–59.
[42] Subramanian S, Rahimi A, Baldwin T, Cohn T, Frermann L. Fairness-aware class imbalanced learning. 2021. *arXiv preprint arXiv:210910444*.