Journal of Big Data

RESEARCH Open Access

Adapting transformer-based language models for heart disease detection and risk factors extraction

Essam H. Houssein[1*], Rehab E. Mohamed[1], Gang Hu[2] and Abdelmgeid A. Ali[1]

*Correspondence:
essam.halim@mu.edu.eg

[1] Faculty of Computers and Information, Minia University, Minia, Egypt
[2] Department of Applied Mathematics, Xi'an University of Technology, Xi'an 710054, China

Abstract

Efficiently treating cardiac patients before the onset of a heart attack relies on the precise prediction of heart disease. Identifying and detecting the risk factors for heart disease such as diabetes mellitus, Coronary Artery Disease (CAD), hyperlipidemia, hypertension, smoking, familial CAD history, obesity, and medications is critical for developing effective preventative and management measures. Although Electronic Health Records (EHRs) have emerged as valuable resources for identifying these risk factors, their unstructured format poses challenges for cardiologists in retrieving relevant information. This research proposed employing transfer learning techniques to automatically extract heart disease risk factors from EHRs. Leveraging transfer learning, a deep learning technique has demonstrated a significant performance in various clinical natural language processing (NLP) applications, particularly in heart disease risk prediction. This study explored the application of transformer-based language models, specifically utilizing pre-trained architectures like BERT (Bidirectional Encoder Representations from Transformers), RoBERTa, BioClinicalBERT, XLNet, and BioBERT for heart disease detection and extraction of related risk factors from clinical notes, using the i2b2 dataset. These transformer models are pre-trained on an extensive corpus of medical literature and clinical records to gain a deep understanding of contextualized language representations. Adapted models are then fine-tuned using annotated datasets specific to heart disease, such as the i2b2 dataset, enabling them to learn patterns and relationships within the domain. These models have demonstrated superior performance in extracting semantic information from EHRs, automating high-performance heart disease risk factor identification, and performing downstream NLP tasks within the clinical domain. This study proposed fine-tuned five widely used transformer-based models, namely BERT, RoBERTa, BioClinicalBERT, XLNet, and BioBERT, using the 2014 i2b2 clinical NLP challenge dataset. The fine-tuned models surpass conventional approaches in predicting the presence of heart disease risk factors with impressive accuracy. The RoBERTa model has achieved the highest performance, with micro F1-scores of 94.27%, while the BERT, BioClinicalBERT, XLNet, and BioBERT models have provided competitive performances with micro F1-scores of 93.73%, 94.03%, 93.97%, and 93.99%, respectively. Finally, a simple ensemble of the five transformer-based models has been proposed, which outperformed the most existing methods in heart disease risk fan, achieving a micro F1-Score of 94.26%. This study demonstrated the efficacy of transfer

learning using transformer-based models in enhancing risk prediction and facilitating early intervention for heart disease prevention.

Keywords: Coronary artery disease, Electronic health records, Natural language processing, Bidirectional encoder representations from transformers, Heart disease, Transformer-based models

Introduction

Heart disease, chronic respiratory disease, and diabetes are among the many non-communicable diseases associated with the modern lifestyle. One of the highest death rates is caused by heart disease [1]. Heart disease is a term used to describe abnormalities of the heart. It is regarded as one of the world's most powerful killers, surpassing Alzheimer's and cancer in power. The prevention of heart disease has become a serious issue in today's world that needs to be addressed. It is estimated that one American dies of heart disease every 30 s [2]. Each year, 647,000 Americans suffer from heart disease [3]. Approximately 17.8 million deaths were caused by heart disease worldwide in 2017, an increase of 21.2% compared to 2007 [4]. In addition, heart disease can increase the need for hospital treatment by acting as a risk factor for other diseases. For example, they have been associated with a poor prognosis in the setting of COVID-19, threatening the ability of healthcare systems around the world [5]. Half of those who have a heart attack are not 'at risk'. These concerns require automatic prediction of heart disease and earlier identification, which is a critical issue. It is essential to prevent this life-threatening disease before it leads to millions of deaths. It is important to identify various risk factors to diagnose and prevent this disease earlier, such as Coronary Artery Disease (CAD), Diabetes, Hypertension, Hyperlipidemia, Smoking, Medications, Family history of CAD, and Obesity [6–9].

All other heart risk factors must be identified with indicators and temporal features, except CAD in the family and smoking status. Each characteristic of the indicator indicates the clinical significance of the risk factor. A significant difficulty in the field of heart disease detection and prevention is the identification of risk factors reported in clinical notes.

That means it is a difficult problem in clinical data analysis to create a fully automated method to predict heart disease from EHR [10, 11]. Natural language used in clinical narratives stored in EHRs is sometimes described as idiosyncratic, with considerable variability in format and quality [12]. Structured data are commonly created for administrative purposes only in electronic health records, so the data are biased toward diagnoses and procedures that are relevant to billing purposes. Unstructured clinical notes are the most in-depth source of data, but semantic labeling is not common because it requires advanced planning and analysis [13]. Although unstructured data has numerous uses, there is a growing need to unlock them for primary and secondary purposes [14]. Secondary use of such data can include supporting observational studies, such as cohorts, cross-sections, and case–control research [15]. By developing systems for analyzing narrative clinical notes to register patients according to selection criteria, sampling bias could be reduced [16]. Using NLP techniques, we can convert the meaning of human language into machine-readable representations that can be used for secondary purposes. NLP models from the general domain cannot be easily applied to clinical text

due to significant linguistic differences since it is likely to be simple terms, often referred to as the telegraphic style. Developing these systems for the clinical field is challenging because there are few publicly accessible annotated clinical narrative datasets. During the big data revolution, neural networks (NNs) were trained to model a variety of human languages with high accuracy as a result of the availability of large amounts of data in the general domain, but there has not been the same in small data scenarios, where models are often trained from scratch. Consequently, transfer learning methods have become increasingly popular, allowing previously trained models to be applied in new contexts with minimal annotation and labeling [17].

Transfer learning is a deep learning technique that refers to the process of adapting a model originally pre-trained for a specific task and is used as a basis for training a model to perform a different task using a new dataset [18, 19]. Although transfer learning has received much research in the field of medical image analysis, its application to text-clinical data is still lacking. Therefore, this scoping study aimed to investigate the feasibility of applying transfer learning to non-image data in the clinical text. Many of the most recent advances in generalizable and adaptable techniques are based on transfer learning. When data is scarce, knowledge of fields, tasks, or languages with large data is applied [20]. Several clinical studies highlighted the potential of transfer learning to reuse models in a wide range of prediction tasks, data types, and even species. Transfer learning was apparent to be particularly effective when applied to smaller datasets, rather than when machine learning algorithms were trained from scratch in terms of prediction [18].

Different methods can be used to transfer knowledge from a large dataset depending on the availability of the data source, task labels, and reused data [21]. Feature representation transfer is one of the most common methods in which an input representation strategy that is trained unsupervised on a large dataset is transferred to a smaller annotated sample [21]. However, Goodfellow et al. [22] suggest that the application of this strategy has decreased since deep learning provides human intervention with large labeled datasets, while Bayesian methods perform better when small data are available. Mikolov et al. [23] promoted feature representation transfer in the NLP area, by releasing word2vec embeddings trained on approximately 100 billion words extracted from a Google News corpus. However, this model has a low coverage rate for clinical text due to uncommon words and misspellings, prompting the search for other input representation strategies [24]. Bojanowski et al. The [25] suggested including sub-word information in word vectors to accommodate morphology. Although deep learning has become common for text classification, Joulin et al. [26] have developed fastText, a quick and accurate application of multinomial logistic regression that makes text classification on a large scale possible. More recently, the National Center for Biotechnology Information developed BioWordVec, which was trained using fastText on more than 30 million documents from the MIMIC-III (Medical Information Mart for Intensive Care) [27] and the PubMed clinical dataset. When combined, these tools can facilitate the transfer of learning from the large data domain to the clinical field by addressing the unique challenges posed by clinical settings [28, 29].

Motivation There is promise in the detection of heart disease risk factors using transformer-based models based on transfer learning approaches to learn bidirectional

relationships in EHR. We proposed a heart disease risk factor identification model by comparing five transformer-based models using EHR data. We modeled the task as NER task according to [30, 31]. The study uses several statistical criteria and evaluation measurements to support these findings. The evaluation included measures of precision, recall, and F1 at the micro-level when comparing the results of the fine-tuned transformer-based models to the document-level gold standard. The primary contributions of this paper can be summarized as follows:

1. Developing a model that identifies heart disease risk factors in EHRs using transfer learning models.
2. In this study, we explored transfer learning using openly available biomedical contextual embeddings.
3. Implement a transfer learning technique that could effectively use these embeddings to identify risk factors for heart disease.
4. The fine-tuned transformer-based models outperformed the 2014 i2b2/UTHealth shared task systems and models.
5. In this study, we applied five transformer-based models, which are contextual embeddings of BERT [32], BioBERT [33], BioClinicalBERT [34], RoBERTa [35], and XLNet [36] contextual embeddings.
6. Ensembling strategies help improve the performance of all eight risk factors extraction challenging.

The remaining sections of the paper are structured as follows, Section "Related work", provides a literature review of several recent related works on the 2014 i2b2UTHealth shared task track 2 and adaptation of transfer learning in clinical EHR. Section "Materials and methods", demonstrates the objectives of the proposed task, the description of the dataset, the description of the research problem, the pre-processing steps, the transfer learning models, and the transformer-based models. Pre-training and fine-tuning process. Section "Experimental results and simulations", shows the evaluation and results of the proposed study. Finally, the conclusion and future works are discussed in Section "Conclusion and future work".

Related work

The proposed study is motivated by the challenges of the 2014 i2b2/UTHealth heart disease risk factor detection task, as well as some previous Information Extraction (IE) research in the clinical domain with the adaptation of transfer learning techniques.

Track 2 of the 2014 i2b2/UTHealth shared task

The National Center (https://www.i2b2.org/) for Biomedical Computing has organized the Informatics for Integrating Biology and Bedside (i2b2) (https://www.i2b2.org/) Challenges since 2006 to encourage NLP study in the health domain. Track 2 of the 2014 i2b2/UTHealth shared task proposed the challenge of text classification in the clinical domain with limited data and requested the participating teams to categorize patients based on eight risk factors for heart disease: CAD, diabetes, hypertension, hyperlipidemia, obesity, smoking, medications, and family history.

The teams investigated a wide range of approaches, from rule-based to hybrid, using a wide variety of feature combinations and machine learning methods [30]. Participants could not clearly agree on the optimal approach to the challenging task because so many different hybrid systems were proposed. Most teams had discovered a challenging issue with the encoded pseudo-tables and heart disease risk indicators in clinical notes, which led to low F1 scores. Using SVM models based on custom-built lexica, the best team participating in 2014 achieved an F1-score of 0.9276 after reannotating a large portion of the training corpus [37].

A preprocessing step was performed to extract headings from sections, negation markers, modalities, and other output using the ConText tool [38], but no other syntactic or semantic signals were used. They demonstrated that other automated systems can be improved with fine-grained annotations.

Kotfila and Uzuner [39] investigated the effectiveness of SVM classifiers trained on the shared dataset by comparing the size of training data, features, weighting schemes, and kernels.

The authors indicated that limited feature spaces with lowercase alphabetic tokens were equivalent to combinations of lexically normalized tokens and extracted semantic concepts using MetaMap [40], and linear kernels were not significantly less effective than radial kernels.

Furthermore, they demonstrated that the use of SVM models may not require large corpora to achieve high efficiency.

Chen et al. [41] developed a hybrid pipeline system with three modules for tag extraction to extract tags based on phrases, logic, and discourse, as well as a module for identifying time attributes with temporal indicators using SVM.

The system achieved significant efficiency among Information Extraction (IE) systems that do not require more annotations by treating phrase-based tagging as a Name Entity Recognition (NER) task and identifying time attributes as a temporal relationship extraction task.

In addition, Urbain [42] has used various techniques such as conditional random fields (CRFs) to identify risk factors, regular expressions to identify time attributes, and a semantic distribution model to classify specific risk factors. Torii et al. [43] have developed three classifiers for various identifications: a general classifier, a smoking status classifier, and a sequence labeling-based classifier, using hot-spot features (phrases annotated as risk factor evidence) in conjunction with several open machine learning tools such as MedEx [44], Weka [45], LibSVM [46] and Stanford NER [47].

Related work in transfer learning and domain adaptation for NLP of EHRs

Several studies have proposed applications that applied text-based transfer learning. These applications have proposed the prediction of morbidity, mortality, and adverse events from oncological radiation [48–50], and the assessment of the risk of psychological stressors, diseases, and drug abuse [51–54]. Transfer learning methods have been applied to the clinical domain by sequentially training several tasks. The researchers pre-trained a convolutional neural network (CNN) in PubMed-indexed biomedical articles to identify medical subject headings and then transferred this model to predict International Classification of Diseases (ICD) codes in EHRs [55].

A similar approach uses unlabeled data from three institutions and applies self-training and transfer learning to classify radiological reports using a small labeled data set [56]. Pre-trained word embeddings are commonly transferred to downstream tasks, such as applying medical embeddings to NER in the clinical domain [57]. Embeddings have been trained in both the general and clinical domains, and as a result, many methods have been developed to adapt embeddings, such as concatenation and fine-tuning [58]. Another proposed method pre-trained embeddings on the relation extraction task of the 2009 i2b2 challenge [59] and then transferred them to NNs for the extraction of clinical terms in the shared dataset [60].

Pre-trained transformers methods Transfer learning with transformer-based models has become a standard approach in NLP due to its effectiveness and efficiency in leveraging pre-existing knowledge for various downstream tasks. Recently, transformer architectures [33] (e.g., BERT) applying self-attention mechanisms [61] have achieved the best results on many NLP tasks [62, 63]. A transformer-based NLP model has achieved significant performance in several areas, such as NER [64, 65], relation extraction [66, 67], sentence similarity [68, 69], natural language inference [69, 70], and question answering [69, 71–73]. Transformer training involves two phases: (1) pretraining, where the language model is learned based on self-supervised training on a large unlabeled dataset; and (2) fine-tuning when the pre-trained model is applied to labeled training data to address specific tasks. Fine-tuning is the process of applying a pre-trained language model to address several NLP tasks, which is known as transfer learning. Transfer learning is a technique for transferring knowledge from one task to another [74]. The sample space for human language is enormous; there are an infinite number of possible permutations of tokens, sentences, and their grammar and meaning. According to recent studies, the emergence and homogenization of large transformer models trained on large text data have been significantly superior to previous NLP models [74].

Biomedical models based on BERT include BioBERT [33], BlueBERT [75], and ClinicalBERT [34]. These models use a continuous pretraining method, initializing the model weights using weights from BERT pre-trained on Book Corpus and Wikipedia while using the same vocabulary. Pre-training from scratch using domain-specific corpora and vocabulary improves the performance of models SciBERT [76], PubMedBERT [77], and Biolm [78].

The BERT model has been applied to the scientific, clinical, and biomedical domains. BERT is pre-trained in PubMed and PubMed Central articles in BioBERT [33]. In BlueBERT [75], BERT is pre-trained on PubMed, PMC, and MIMIC III data [27]. ClinicalBERT [34] is pre-trained in MIMIC III data using BioBERT weights, while SciBERT [76], PubMedBERT [77] and Bio-lm [78] train BERT with domain-specific data. SciBERT pre-trained on Semantic Scholar data. PubMed and PMC data are used to pretrain PubMedBERT. PubMed, PMC, and MIMIC III are used to pretrain Bio-lm data [78]. BlueBERT and PubMedBERT have launched benchmarks for biomedical NLP-BLUE (Biomedical Language Understanding Evaluation) and BLURB (Biomedical Language Understanding & Reasoning Benchmark). Table 1 summarizes the state-of-the-art transformer-based models with their pre-trained dataset and training weights.

Table 1 The recent pre-trained transformers models

Model	Pre-trained dataset	The training weights	References
BERT model	Book Corpus Wikipedia	–	[33]
BioBERT model	PubMed PubMed Central articles	BERT weights	[33]
BlueBERT model	PubMed PMC MIMIC III data	BERT weights	[75]
ClinicalBERT model	MIMIC III data	BioBERT weights	[34]
SciBERT model	Semantic Scholar data	BERT weights	[76]
PubMedBERT model	PubMed PMC data	BERT weights	[77]
Bio-lm model	PMC PubMed MIMIC III	BERT weights	[78]

Materials and methods

Objective task

We proposed a high-performance heart disease risk factor identification model, so we turned to open source NLP frameworks as part of our work to bring together clinical decision support functionalities and note taking interfaces. We modeled the task as a NER task according to [30, 31] and explored transfer learning using openly available biomedical contextual embeddings. Our main objective was to get a transfer learning process working with these embeddings. The context in which this is performed is as follows:

1. In this study, we examined transfer learning models that employed BERT [32], BioBERT[33], BioClinicalBERT [34], RoBERTa [35], and XLNet [36] contextual embeddings pre-trained on PubMed abstracts [79] which are the best deep language models that based on encoder-decoder transformer architectures [61] and have been pre-trained on massive unstructured text datasets.
2. Our research examines embedding-specific methods in order to improve performance, including language-model finetuning, scalar mix, and aggregation of subword tokens.
3. We develop a model for the identification of heart disease risk factors based on the performance of transfer learning models. Risk factors can be extracted more effectively with sentence enhancement at prediction time. Furthermore, it allows for a better understanding of the behavior of the embeddings. Ensembling strategies helps improve the performance of all eight risk factors that make extraction challenging.

Hypothesis

We proposed that transfer learning methods are deep learning methods using a pre-training/fine-tuning learning architecture. Transfer learning methods have superior performance for heart disease risk factors prediction, and pre-trained embeddings can enhance classification efficiency in the clinical domain.

We proposed systematically investigating five widely used transformer-based models, including BERT, BioBERT, BioClinicalBERT, RoBERTa, and XLNet, to develop a model for the detection of heart disease risk factors that can identify diseases, risk factors, medications, and the time of occurrence.

Dataset

The proposed model uses a data set provided by Partners HealthCare [http://www.partners.orghttps://www.i2b2.org/NLP/HeartDisease/], which includes clinical notes and discharge summaries. The shared task dataset includes 1304 patient records that identify 296 diabetics with heart disease risk factors and temporal attributes based on DCT. According to the challenge provider, the dataset is divided into a training set with 60% of the records (790 records) and a test set with 40% of the records (514 records). The organizers of the i2b2 NLP shared task provided two annotated datasets, namely SET1 and SET2, for development and training purposes. SET1 contained 521 de-identified clinical notes, while SET2 consisted of 269 de-identified notes. Therefore, a combined total of 790 documents were accessible for training. The test set consisted of 514 de-identified clinical notes. The document annotation guidelines for annotating data can be used to identify the presence of diseases (including CAD, diabetes, and heart disease), eight relevant evidence risk factors (including hyperlipidemia, hypertension, obesity, smoking status, and family history), and associated medications. There are a number of indicators to determine whether there is a disease or risk factor in the patient at the time of the DCT (before, during, or after). Table 2 provides a summary of the tag types and their corresponding attribute values provided in the challenge data. There were two versions of the data released by the challenge organizers: complete and gold. Each clinical record in the Gold version is presented in XML format, and XML tags are used to annotate target concepts that are mentioned anywhere in the record (such as <*DIABETES time=”before DCT” indicator=”mention’* >) if they are present. Additionally, the complete version includes evidence annotations made by three clinicians in the text segments. Therefore, in this data set, each concept annotated at the document level is linked to the relevant text segment at the record level to provide evidence of heart disease (e.g., < *DIABETES start = ”4401’ end = ’4422’ text = ”HbA1c 03/05/2074 6.6’ time = ’during DCT” indicator = ’A1C’* >). Figure 1 shows an example of the tag extracted from training data (220–05.xml) in both versions, together with its evidence in the text field.

Research problem description

The research problem identified each type of tag as follows: First, identify the available evidence by its type and indicator. Then identify the time attribute (if it exists).

Risk factor tags can be categorized into three groups by analyzing the evidence of the tag based on the terminology used by Chen et al. [30]:

1. Phrase-based risk factors are identified by detecting relevant phrases in the clinical note, such as ’diabetes’ or the name of a specific drug.

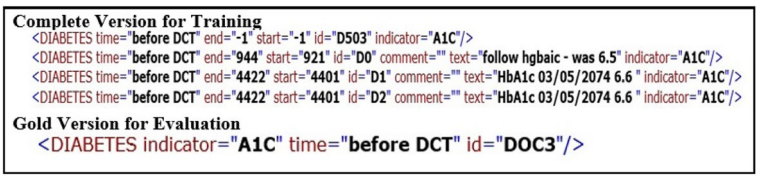


Fig. 1 Example of a sample of the Heart Disease Risk Factor Tags included in the complete and gold versions

Table 2 A summary of the shared task dataset's risk factor tags

Risk factor tags	Indicator	Time		
		Before DCT	During DCT	After DCT
(a) Tag: CAD Indicator	Mention	260	261	259
	Event	224	20	2
	Symptom	54	24	3
(b) Tag: Diabetes indicator	Mention	518	524	518
	Glucose	16	9	0
	A1C	89	21	0
(c) Tag: Hyperlipdemia indicator	High LDL	23	10	0
	High chol.	5	1	0
	Mention	340	340	340
(d) Tag: Hypertension indicator	High bp	41	322	0
	Mention	523	521	519
(e) Tag: Obese indicator	DMI	3	15	2
	Mention	133	147	133
(f) Tag: Medication type (type1)	Thienopyridine	97	98	97
	Statin	436	427	438
	Thiazolidinedione	43	41	40
	Aspirin	424	6	424
	Metformin	187	176	181
	Insulin	204	218	212
	Fibrate	22	20	22
	Ezetimibe	12	12	12
	Diuretic	113	99	106
	Anti diabetes	1	1	1
	ARB	98	93	97
	Sulfonylureas	159	155	157
	DPP4 inhibitors	1	0	0
	ACE inhibitor	326	318	323
(g) Tag: Family_history indicator	Not present	NA	NA	768
	Present			22
(g) Tag: Smoker status	Current	NA	NA	58
	Ever			9
	Never			184
	Past			149
	Unknown			371

Annotation-level training and testing set sizes, as well as indicators for each heart risk factor

2. Logic-based risk factors are based on the analysis of the detected relevant phrase; for example, determining whether or not high blood pressure is a risk factor requires locating a blood pressure measurement and comparing the numbers.
3. Discourse-based risk factors that require sentence parsing because they are embedded in clinical text fragments, such as the identification of family history or smoking status.

After classifying all tags into the three groups shown in Table 3, we presented a standard organizing principle for each category. The following Fig. 2 illustrates the proposed model modules, which include pre-processing, tag extraction, identifying time attributes, and post-processing. Initially, the preprocessing module detected sentence boundaries and tokenized the clinical notes in the raw data file. Then, the tag extraction module identified the type and indicator of the tag in each of the three categories in Table 3. Next, the module for identifying time attributes determined whether or not there was evidence for the time attribute. Fine-tuning of the

Table 3 Examples of evidence types for risk factors

Evidence category	Risk factor indicator	Example
Phrase-based indicators	- CAD: mention - Medication	- 'Coronary arteriosclerosis', 'CAD', '3-vessel coronary artery disease' - 'Insulin 70/30 HUMAN 70-30', 'lisinopril', 'Zestril (LISINOPRIL)'
Logic-based indicators	- Diabetes: alc -Hyperlimidemia: high LDL	- 'hgba1c 7.3%', 'last 1/2137' hgba1c 7.3% - 'Her last LDL was over 100', 'Cholesterol-LDL 12/15/2105 110
Discourse-based indicators	- CAD: event - Smoker: current	- 's/p ant SEMI + stent LAD', 's/p ant SEMI + stent LAD 2/67', 'MI in 2092' - 'Has smoked 1/2ppd for 35 years, 'Tobacco abuse'

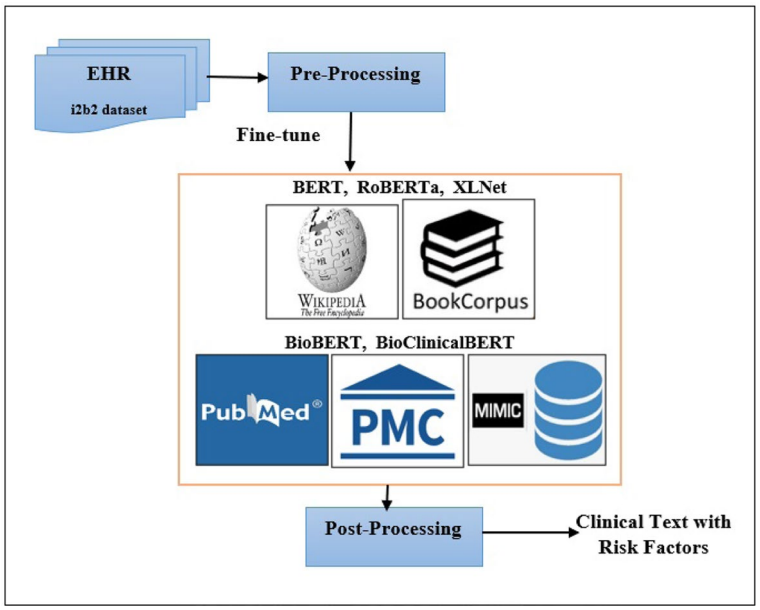


Fig. 2 The proposed model for heart disease risk factor identification by fine-tuning transformer-based models

proposed transformer-based models. For evaluation, the post-processing module transformed the tags from the complete version to gold version tags.

Preprocessing

Preprocessing involves first splitting full-text clinical records into separate sentences. Using Metamap [40], tokens and sentences from clinical notes were assigned to concepts. The next step is to tokenize the sentence and add context-sensitive features to each token and occurrence of the token. Meanwhile, we applied Splitta [80], an open-source machine learning model, to split sentences. As soon as a token or sentence is mapped to one of the targeted concepts (such as disease or syndrome, family group, smoke, etc.), its sentence is identified as one of the candidates for further processing. The annotation set is processed using Metamap to identify the target concepts.

The proposed model for identifying heart disease risk factors based on transfer learning models

This study aimed to systematically investigate five widely used transformer-based models, including BERT, BioClinicalBERT, RoBERTa, BioBERT, and XLNet to detect risk factors for heart disease by identifying diseases, risk factors, medications, and the time of occurrence.

1. **BERT model** was pre-trained by masked language modeling, and then the next sentence prediction was used to optimize it. The number of transformer layers is 12, with a hidden layer of size 768, and the number of attention heads is 8 in the base model architecture using 110 million parameters.
2. **RoBERTa model** (A Robustly Optimized BERT Pretraining Approach) is a transformer-based model based on the BERT's architecture but is pre-trained using a dynamic masked language modeling approach and optimized by removing the next sentence prediction.
3. **BioClinicalBERT model** is based on the BERT architecture but is pre-trained on a large data set of biomedical and clinical texts, including EHR and biomedical texts. This allows the model to capture the specific language and terminology used in these domains. BioClinicalBERT also incorporates additional inputs, such as segment labels to indicate the source of each token (e.g. medication, diagnosis, test result), and position labels to indicate the position of each token within a segment.
4. **BioBERT model** was pre-trained on a large corpus of biomedical text, including biomedical research articles, clinical notes, and EHRs. This pre-training process allows the model to learn the patterns and structures of biomedical language, which can then be fine-tuned for specific downstream tasks, such as NER, text classification, and question answering. One of the main advantages of BioBERT is its ability to handle the complex and domain-specific language used in the biomedical field, which can often be challenging for traditional NLP models. Additionally, BioBERT has achieved state-of-the-art performance on several biomedical language processing benchmarks, demonstrating its effectiveness in various tasks.
5. **XLNet model** is based on the transformer architecture, which has been used in other successful language models such as BERT. However, XLNet introduces a new training objective called permuted language modeling, which differs from the masked language modeling used in BERT to predict tokens in a randomly permuted sequence of the input.

Fine-tuning the previous models, they can then be trained on the pre-processed data to detect the risk factor for heart disease as shown in Fig. 2. This involves fine-tuning the model's weights to optimize its accuracy. Feature extraction: Once the model is trained, it can be used to extract meaningful features from new EHRs. By analyzing the text of the records and extracting meaningful features using fine-tuned transformer-based models, the model can predict the risk factor for heart disease with a high degree of accuracy. Validation and Testing: To ensure that the model is accurate and generalizable, it is important to validate and test it on diverse datasets. This involves evaluating the model's performance on a validation dataset and testing

it on a separate test dataset to assess its accuracy and generalizability. Deployment: Once the model is validated and tested, it can be deployed to identify heart disease risk factors from EHRs in real-world settings. This may involve integrating the model into EHR systems or developing a standalone application that can analyze EHR data and provide risk assessments to healthcare providers.

Transformer-based models incorporate two processes, the pre-training process and the fine-tuning process. The Pretraining Process: Previous research [33] demonstrated that pretraining on a clinical dataset improved the clinical concept extraction efficiency; therefore, we investigated both general models pre-trained with a general English data set and clinical models pre-trained with a clinical data set for each of the 5 transformer-based models. We employed state-of-the-art transformer-based models pre-trained on general English domain datasets, such as bert-base-uncased (BERT-general), RoBERTa-base (RoBERTa-general), and XLNet. We used clinical transformer-based models that were pre-trained using the MIMIC-III [27] dataset, which contains clinical notes, such as BioBERT and BioClinicalBERT.

The Fine-tuning Process: To predict heart disease risk factors from clinical concepts annotated in the training dataset, we built upon transformer-based models pre-trained on the MIMIC data by adding a linear classification layer. It will be necessary to optimize the classification layer parameters and the transformer-based models to achieve significant results from the extraction of clinical concepts.

Experimental results and simulations

In this section, we describe in detail the weighted-averaged proposed model results that are achieved by the fine-tuned transformer-based models compared to the most recent systems and models from the 2014 i2b2 shared task, as shown in Table 4.

Table 4 illustrated a comparison between the fine-tuned transformer-based models' results and the top-ranked systems [37, 41, 81] which use a hybrid of knowledge-and

Table 4 The weighted-averaged evaluation results of fine-tuned transformer-based models and the most recent models and systems from 2014 i2b2 shared task

Model	Precision	Recall	F1-score	Micro F1-score (Accuracy
BERT	0.9251	0.9373	0.9284	0.9373
RoBERTa	0.9390	0.9427	**0.9394**	**0.9427**
BioBERT	0.9337	0.9399	0.9357	0.9399
BioClinicalBERT	0.9338	0.9403	0.9357	0.9403
XLNet	0.9361	0.9397	0.9371	0.9397
Roberts et al. [37]	**0.9625**	0.8951	**0.9276**	0.9276
Chen et al. [41]	0.9436	**0.9106**	0.9268	0.9268
Cormack et al. [82]	0.9375	0.8975	0.9171	0.9171
Yang and Garibaldi [81]	0.9488	0.8847	0.9156	0.9156
Khalifa and Meystre [83]	0.8951	0.8552	0.8747	0.8747
Chokkwijitkul et al. [10]	0.9180	0.8983	0.9081	0.9081

Bold indicates the best value

data-driven techniques, and systems [10, 82, 83] that only use knowledge-driven techniques, such as lexicon and rule-based classifiers.

Evaluation metrics

The performance of the proposed model using transformer-based models was evaluated based on the evaluation script provided by the shared task organizers. We used recall, precision, and the F1-measure as primary measurements. Both macro and micro-averages are included in the overall averages. Micro-averages are provided for each class indicator pair [The official evaluation script: https://github.com/kotfic/i2b2_evaluation_scripts].

Results and discussion

We applied transfer learning to develop a model that detects risk factors for heart disease from clinical texts over time using the 2014 i2b2 clinical NLP challenge dataset. The most recent models chosen for fine-tuning in the classification task included five transformer-based models: BERT, BioBERT, RoBERTa, BioClinicalBERT, and XLNet. Our objective was to identify diseases, risk factors, medications, and time factors based on DCT. First, the proposed transformer-based models retrieved these risk indicators and then determined their temporal attributes.

Data augmentation is applied to the i2b2 dataset for heart disease risk factor detection from EHR by generating variations of the existing data to increase the size of the training set. Data augmentation can help prevent overfitting and improve the generalization of the pre-trained models. The augmented data is validated to ensure semantic integrity. Then the original dataset is integrated with the augmented data to generate a new training set. The annotation process for risk factors is applied to the new training set. The training process is performed on the new augmented dataset, then the fine-tuned models are validated using a validation set and finally tested using a test dataset to assess their performance.

The data augmentation is performed to address the issue of under-representation of a specific class and to ensure adequate representation of the minority class, in this case, the 'glucose' class in the Diabetes-Indicator classification. When a particular class is under-represented in a dataset, it can lead to imbalanced training data, potentially causing the model to struggle in accurately learning and predicting the minority class.

To address this challenge, data augmentation techniques are employed. In this context, data augmentation involves duplicating instances belonging to the minority class within the training set. By creating additional copies of the minority class samples, the augmented dataset now contains more instances representing the under-represented class.

After performing data augmentation and adding duplicated instances, the next step described is shuffling the entire training set. Shuffling the data ensures that the duplicated examples of the minority class are evenly distributed throughout the training data. This process helps to promote a more balanced representation by preventing the model from encountering batches or patterns that are biased towards the majority class.

By employing data augmentation and shuffling, the training data is modified to have a more balanced representation of the minority class, such as 'glucose' in this case. This

approach aims to improve the model's ability to learn and generalize well for all classes, including the under-represented class.

After developing and fine-tuning the NLP techniques and transformer-based models using the training corpus (SET1 and SET2), they were applied to the testing corpus and the results were compared to the shared task organizers for analysis. The outputs of the transformer-based models were compared with the primary metric of the i2b2 challenge evaluation script provided by the shared task organizers, and all extracted tags are classified as true positive (i.e., the result matches the primary metric), false positive (i.e., the result does not match the primary measure), or false negative. The tables below show the results for each class of risk factors and their indicators.

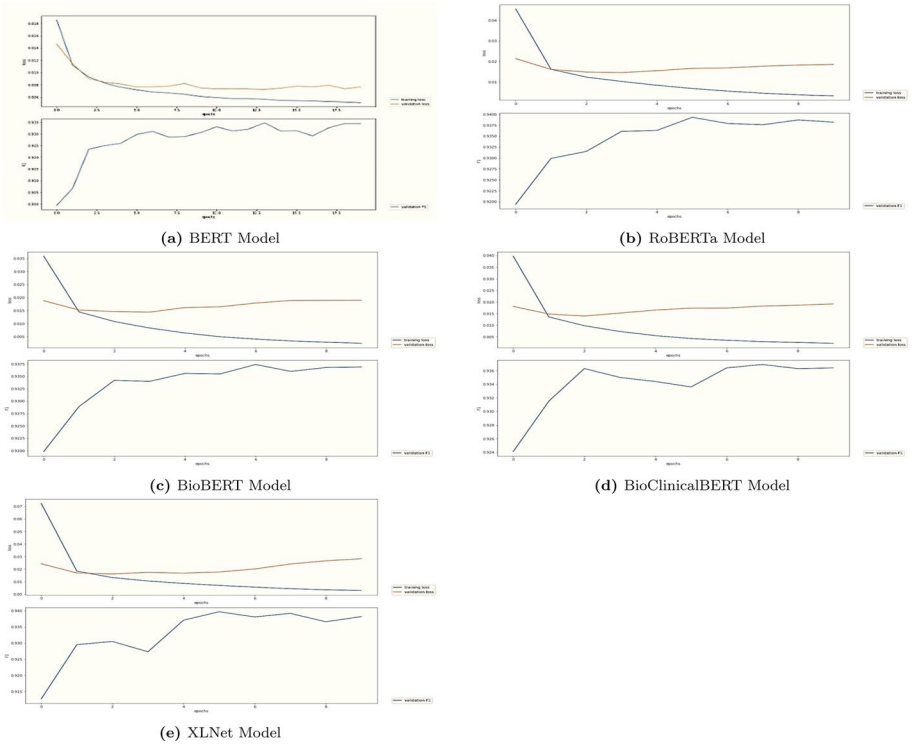
We used the filter option provided by the evaluation script to determine the results for each class of risk factors. Using the option Conjunctive, it is also possible to determine specific risk factors and their attribute value indicators, such as the tag CAD and the attribute value of indicator = 'mention'.

According to the annotation standard, we give the results for each disease category separately, for general mention and disease-specific indicators. Results for the SMOKING category are reported as status only, while results for the MEDICATION categories are combined and are accurately identified on the EHRs. For each heart disease risk factor class, the results in the tables below were generated for all temporal information tags and an attempt was made to categorize the (before, during, and after) DCT results.

The best-performing model in most cases for risk factor identification was RoBERTa with F-measure of 93.94%, a precision of 93.90%, and a recall of 94.27% at the weighted-averaged level. According to the BERT prevailed in cases with higher numbers of categories with micro -precision, -recall, and -F1-scores of 92.51%, 93.73%, and 92.84%, respectively. BioBERT obtained -precision, -recall, and -F1-scores of 93.37%, 93.99%, and 93.57%, respectively in identifying risk factors. BioClinicalBERT attained a precision of 93.38%, recall of 94.03%, and F1-measure of 93.57% at the weighted-averaged level. XLNet achieved an F-measure of 93.71%, a recall of 93.97%, and a precision of 93.61% at the weighted level.

Furthermore, the best results achieved at the risk indicator level by applying the BioBERT model to identify hypertension, diabetes, and SMOKER were 0.91, 0.83, and 0.90, respectively, using micro-averaged F1-measures. The BERT model performs best on Hypertension (0.90), Smoker (0.88), and FamilyHist (0.88). BioClinicalBERT performs best on Hypertension (0.92), Smoker (0.94), and FamilyHist (0.88). RoBERTa performs best on Hypertension (0.92), Smoker (0.90), and FamilyHist (0.88). XLNET performs best on Hypertension (0.91), Smoker (0.90), and FamilyHist (0.88). The FAMILY_HIST is a simple task because there are few records containing evidence of family members who have been diagnosed with CAD.

It is shown in Tables 5, 6, 7, and 8 how BERT, RoBERTa, BioClinicalBERT, and XLNet performed on the i2b2 test data with respect to F1-measure, Recall, and Precision at the risk indicator level. The overall performance of the transformer-based models at the level of time attributes associated with the presentation of the risk indicator based on DCT is shown in Tables 9, 10, 11 and 12. Table 13 shows the overall performance of the eight risk factor categories based on the i2b2 test data. Figure 3 shows the F1-Plot curve of five transformer-based models using the final dataset after the augmentation process.



(a) BERT Model (b) RoBERTa Model

(c) BioBERT Model (d) BioClinicalBERT Model

(e) XLNet Model

Fig. 3 The F1-plot of the train- and validation-learning curves of five transformer-based models using the final dataset

In this study, we evaluated model-ensembling techniques for improving the proposed risk factor detection model's performance for heart disease. The provided predictions are based on ensembling fine-tuned transformer-based models, which achieved F1-scores of 94.26%. The overall F1-scores for five different ensemble models are shown in Table 14. The ensemble model provides the best performance in all risk factor detections, demonstrating the efficiency of the technique. We proposed to ensemble the five transformer-based models to get the benefits of each model in the word embedding technique. The reason to use an ensemble of five transformer-based models is to apply the performance of different models. Every model within the ensemble can be particularly good at handling particular kinds of datasets or capturing various linguistic patterns based on a pre-training dataset. We might be able to improve overall performance and get more reliable results by integrating their outputs. Furthermore, ensembles can provide a type of model averaging that decreases overfitting and improves generalization. Therefore, the ensemble approach has a powerful improvement in many NLP tasks [84].

The optimized hyperparameters of the most recent models chosen for fine-tuning are presented in Table 15.

Error analysis

We classified the risk factors into one of three groups, as mentioned in Table 3, determining the type of evidence for each risk factor. Evaluation of the test dataset provided by the shared task organizers showed that the proposed model based on transformer-based models performed effectively, with the best micro F1-score being

Table 5 BERT evaluation metrics for each risk factor indicator

Risk Factor	Indicator	Precision	Recall	F1-score	Support
CAD	Mention	0.84	0.93	0.88	260
	Event	0.71	0.76	0.74	250
	Test	0.89	0.12	0.21	68
	Symptom	0.89	0.80	0.84	94
Diabetes	Mention	0.98	1.00	0.99	693
	A1c	0.67	0.92	0.77	65
	Glucose	1.00	0.07	0.13	42
Hypertension	Mention	1.00	0.99	0.99	548
	High bp	0.98	0.99	0.98	212
Obese	Mention	0.93	1.00	0.97	127
	Obese_BMI	0.00	0.00	0.00	9
Hyperlipidemia	Mention	0.98	1.00	0.99	240
	High chol.	0.00	0.00	0.00	8
	High LDL	0.85	0.74	0.79	31
Smoker	Smoker_never	0.93	0.96	0.94	115
	Smoker_ever	0.00	0.00	0.00	3
	Smoker_current	0.00	0.00	0.00	39
	Smoker_past	0.78	0.85	0.81	123
	Smoker_unknown	0.99	0.97	0.98	203
Medication		0.89	0.68	0.77	7225
Family history	NA	0.0000	0.0000	0.0000	13
Weighted average		0.9251	0.9373	0.9284	42946
Macro average		0.3144	0.2846	0.2736	42946
Micro average				0.9373	42946
Accuracy				**0.9373**	42946

Bold indicates the best value

94.27% using RoBERTa model. Despite using fewer annotations, the fine-tuned transformer-based models performed better than the highest performing system participating in the shared task. Although the efficiency of the fine-tuned transformer-based models was exceptional, they did not achieve significant results with several types of tags, such as obesity, CAD, and smoking status.

Discourse-based indicators evidence had an unusually large number of negative samples (not an indicator) for the two types of tags that contain the most discourse-based indicators (CAD and smoking status). An example would be the number of negative samples of CAD: events. The results may have been poor due to the large number of negative samples. There are only 2.4% of obesity status tags in the test data set. Due to the class imbalance issue associated with transfer learning techniques, their low frequency makes them challenging to recognize. The summary report of the i2b2 2014 challenge reveals that other participating systems also had similar results with these three types of tags.

In terms of disease indicators, hyperlipidemia had the lowest recall, and obesity had the lowest precision. Due to inaccurate chunking, some clinical notes containing hyperlipidemia indicators appearing as 'high cholesterol', 'elevated lipids', and 'elevated serum cholesterol' failed to be recognized by our proposed model. Furthermore, our

Table 6 RoBERTa Evaluation Metrics for each risk factor indicator

Risk factor	Indicator	Precision	Recall	F1-score	Support
CAD	Mention	0.82	0.79	0.81	228
	Event	0.72	0.70	0.71	278
	Test	0.00	0.00	0.00	70
	Symptom	0.90	0.39	0.54	95
Diabetes	Mention	0.92	0.99	0.95	549
	A1c	0.89	0.79	0.84	71
	Glucose	0.00	0.00	0.00	40
Hypertension	Mention	0.99	0.97	0.98	525
	High bp	0.93	0.99	0.96	210
OBESE	Mention	0.95	1.00	0.97	115
	Obese_BMI	0.00	0.00	0.00	6
Hyperlipidemia	Mention	0.87	0.99	0.93	232
	High chol.	0.00	0.00	0.00	7
	High LDL	0.50	0.10	0.17	30
Smoker	Smoker_never	0.80	0.72	0.76	114
	Smoker_ever	0.00	0.00	0.00	3
	Smoker_current	0.00	0.00	0.00	37
	Smoker_unknown	0.93	0.90	0.91	202
	Smoker_past	0.85	0.59	0.69	121
Medication		0.87	0.74	0.80	7329
Family history		0.2000	0.0769	0.1111	13
Weighted average		0.9390	0.9427	0.9394	42946
Macro average		0.4195	0.4037	0.3915	42946
Micro average				0.9427	42946
Accuracy				**0.9427**	42946

Bold indicates the best value

dictionary lookup module did not contain the associated ICD codes for some of these items.

In the testing corpus, there were at least two instances in which hyperlipidemia was mentioned directly following a word without a space, such as 'hemodialysis Hyperlipidemia', which our proposed model failed to recognize. As a result of including the concept of 'overweight' in the Unified Medical Language System (UMLS) in the list of ICD codes for obesity, we experienced low precision. Although the obesity indicator 'overweight' appeared in one record of the training dataset, this generated a large number of false positives. Additionally, our proposed model generated false positives when the 'obese' indicator was used as a mention instead of the 'obesity' indicator (e.g., 'abdomen is moderately obese' and 'abdomen is slightly obese'). Regular expressions at the lexical level were not always effective in addressing other indicators of disease and risk factors.

The following issues are associated with heart disease indicators:

– Many different lexical forms and acronyms are used to refer to the same set of laboratory indicators for heart disease. In the case of diabetes and hypertension, regular expressions are applied to determine blood glucose and pressure levels. Blood pressure can be stated using BP and b/p, and glucose levels can be described using BG, BS, FS and FG. This is an example of some of the shortcomings of our proposed

Table 7 BioClinicalBERT Evaluation Metrics for each risk factor indicator

Risk Factor	Indicator	Precision	Recall	F1-score	Support
CAD	Mention	0.86	0.92	0.88	259
	Event	0.80	0.81	0.80	258
	Test	0.93	0.42	0.57	65
	Symptom	0.85	0.80	0.82	99
Diabetes	Mention	0.99	0.99	0.99	601
	A1c	0.75	0.92	0.83	66
	Glucose	1.00	0.38	0.55	40
Hypertension	Mention	1.00	0.99	1.00	602
	High bp	0.99	1.00	0.99	206
OBESE	Mention	0.96	1.00	0.98	117
	Obese_BMI	0.00	0.00	0.00	5
Hyperlipidemia	Mention	0.97	0.99	0.98	286
	High chol.	0.00	0.00	0.00	10
	High LDL	0.79	0.79	0.79	29
Smoker	Smoker_never	0.94	0.95	0.94	114
	Smoker_ever	0.00	0.00	0.00	3
	Smoker_current	0.00	0.00	0.00	40
	Smoker_past	0.88	0.77	0.82	122
	Smoker_unknown	0.99	0.99	0.99	202
Medication		0.88	0.74	0.80	7253
Family history		0.0000	0.0000	0.0000	13
Weighted average		0.9338	0.9403	0.9357	42946
Macro average		0.3715	0.3614	0.3505	42946
Micro average				0.9403	42946
Accuracy				**0.9403**	**42946**

Bold indicates the best value

model, and it would be necessary to develop an integrated approach to address this problem to achieve improved accuracy.

- The numerical results of a laboratory must be extracted accurately. After the proposed model has found the matching terms for laboratory or test indicators, the model must extract the numerical values associated with those terms and compare them with threshold levels that indicate an abnormality. When the numbers follow the term and are expressed as a single unit, like in 'GLU 230(H)', it may be easy to extract them. There are, however, some phrases that may be more difficult to detect, such as 'FS in the AM ~ 90–180'; now 80–175, mostly 90–180'. This example is a case where '_' is used to indicate value ranges, and several units can be denoted by using time and frequency terms such as 'now' and 'mostly'.
- Training data sparsity: Sometimes, there were not enough training examples available to allow an adequate generalization of the proposed model. In the case of the cholesterol indicator for hyperlipidemia, there were only nine annotations available in the entire set of 790 training sets. For the LDL indicator, there were approximately 33 annotations.
- Analysis of complex time attributes: Another issue with our proposed model is that indicators for laboratory tests require additional analysis for temporal information.

Table 8 XLNET Evaluation Metrics for each risk factor indicator

Risk Factor	Indicator	Precision	Recall	F1-score	Support
CAD	Mention	0.93	0.92	0.92	301
	Event	0.69	0.78	0.73	241
	Test	0.00	0.00	0.00	63
	Symptom	0.94	0.49	0.64	94
Diabetes	Mention	0.93	0.99	0.96	530
	A1c	0.91	0.81	0.86	75
	Glucose	0.00	0.00	0.00	40
Obese	Mention	0.94	1.00	0.97	114
	Obese_BMI	0.00	0.00	0.00	7
Hypertension	Mention	0.99	0.99	0.99	515
	High bp	0.97	0.98	0.97	205
Hyperlipidemia	Mention	0.86	1.00	0.93	221
	High chol.	0.00	0.00	0.00	7
	High LDL	0.00	0.00	0.00	28
Smoker	Smoker_never	0.80	0.72	0.76	114
	Smoker_ever	0.00	0.00	0.00	3
	Smoker_current	0.00	0.00	0.00	37
	Smoker_past	0.85	0.59	0.69	121
	Smoker_unknown	0.93	0.90	0.91	202
Medication		0.87	0.76	0.81	7352
Family history		0.2000	0.0769	0.1111	13
Weighted average		0.9361	0.9397	0.9371	42946
Macro average		0.3988	0.3848	0.3772	42946
Micro average				0.9397	42946
Accuracy				**0.9397**	42946

Bold indicates the best value

There are various time-attribute values used in the annotations of most laboratory tests and vital signs; in contrast, the time-attribute value 'continuing' is primarily used in the annotation of chronic disease mention (i.e. during, before, and after DCT). As an example, glucose and A1c tests were typically performed on a previous visit and are therefore labelled 'during DCT'. BP is often taken at the time of the patient's visit and labeled 'during DCT'.

Conclusion and future work

In conclusion, we developed a model to identify heart disease risk factors and demonstrated that transfer learning can be effectively applied to detect heart disease risk factors and the time they are presented in EHRs. Our research highlighted that the application of transfer learning has increased dramatically in recent years. Several studies have identified and demonstrated the significant role of transfer learning based on transformers in the extraction of clinical concepts from clinical notes and other clinical NLP tasks through fine-tuning. Using the shared dataset of heart disease risk factors i2b2, transformer-based models outperformed conventional models in terms of precision in predicting the presence of risk factors. Furthermore, it identified novel risk factors that were not captured by traditional models. Our experiments

Table 9 BERT Evaluation Metrics for each risk factor indicator based on time attribute Identification

Risk factor	Time attribute	Precision	Recall	F-Score	Support
CAD	Before_DCT	0.78	0.93	0.85	462
	During_DCT	0.00	0.00	0.00	94
	After_DCT	0.67	0.63	0.65	116
Diabetes	During_DCT	0.49	0.33	0.39	134
	Before_DCT	0.78	0.85	0.81	557
	After_DCT	0.00	0.00	0.00	109
Hypertension	Before_DCT	0.00	0.00	0.00	33
	During_DCT	0.79	0.87	0.83	363
	After_DCT	0.89	0.79	0.84	364
Obese	Before_DCT	0.00	0.00	0.00	35
	During_DCT	0.89	0.75	0.82	65
	After_DCT	0.73	0.67	0.70	36
Medication	Before_DCT	0.62	0.42	0.50	910
	During_DCT	0.67	0.34	0.45	785
	After_DCT	0.61	0.26	0.36	715
Hyperlipidemia	Before_DCT	0.00	0.00	0.00	85
	During_DCT	0.66	0.95	0.78	160
	After_DCT	0.00	0.00	0.00	34
Weighted average		0.9251	0.9373	0.9284	42946
Macro average		0.3144	0.2846	0.2736	42946
Micro average				0.9373	42946
Accuracy				**0.9373**	42946

Bold indicates the best value

investigate the effectiveness of the five models (BERT, RoBERTa, BioBERT, BioClinicalBERT, XlNet, and BioBERT) in terms of the extraction of risk factors for heart disease. The RoBERTa model achieved state-of-the-art performance with micro F1-scores of 94.27%, while the BERT, BioClinicalBERT, XlNet, and BioBERT models have provided significant performance with micro F1-scores of 93.73%, 94.03%, 93.97%, and 93.99%, respectively. The results showed that a simple ensemble of the five transformer-based models is an effective strategy that significantly improved the performance of the proposed heart disease risk factor identification model with a micro F1-score of 94.26%. Using transformer-based models, our study demonstrated the effectiveness of transfer learning to improve the prediction of heart disease risk.

As part of our future work, we will focus on analyzing embedding-specific issues such as misclassification as well as the incorporation of fine-tuning processes into other clinical NLP tasks.

Table 10 RoBERTa Evaluation Metrics for each risk factor indicator based on time attribute Identification

Risk factor	Time attribute	Precision	Recall	F-Score	Support
CAD	Before_DCT	0.78	0.89	0.83	436
	During_DCT	0.67	0.44	0.53	201
	After_DCT	0.00	0.00	0.00	34
Diabetes	During_DCT	0.00	0.00	0.00	107
	Before_DCT	0.67	0.58	0.62	288
	After_DCT	0.65	0.64	0.65	265
Hypertension	Before_DCT	0.76	0.82	0.79	159
	During_DCT	0.91	0.95	0.93	521
	After_DCT	0.00	0.00	0.00	55
Obese	Before_DCT	0.77	0.41	0.54	41
	During_DCT	0.80	0.11	0.20	36
	After_DCT	0.00	0.00	0.00	44
Medication	Before_DCT	0.68	0.51	0.58	995
	During_DCT	0.58	0.43	0.50	722
	After_DCT	0.62	0.45	0.52	727
Hyperlipidemia	Before_DCT	1.00	0.11	0.20	88
	During_DCT	0.81	0.57	0.67	77
	After_DCT	0.82	0.73	0.77	104
Weighted average		0.9390	0.9427	0.9394	42946
Macro average		0.4195	0.4037	0.3915	42946
Micro average				0.9427	42946
Accuracy				**0.9427**	42946

Bold indicates the best value

Table 11 BioClinicalBERT Evaluation Metrics for each risk factor indicator based on time attribute Identification

Risk factor	Time attribute	Precision	Recall	F-Score	Support
CAD	Before_DCT	0.83	0.81	0.82	422
	During_DCT	0.00	0.00	0.00	82
	After_DCT	0.64	0.92	0.76	177
Diabetes	During_DCT	0.61	0.66	0.63	267
	Before_DCT	0.72	0.75	0.73	329
	After_DCT	0.50	0.02	0.03	111
Hypertension	Before_DCT	0.76	0.79	0.77	186
	During_DCT	0.89	0.95	0.92	452
	After_DCT	0.80	0.59	0.68	170
Obese	Before_DCT	0.00	0.00	0.00	13
	During_DCT	0.00	0.00	0.00	47
	After_DCT	0.53	0.63	0.57	62
Medication	Before_DCT	0.58	0.57	0.58	1002
	During_DCT	0.39	0.08	0.13	609
	After_DCT	0.67	0.42	0.52	808
Hyperlipidemia	Before_DCT	0.81	0.75	0.78	151
	During_DCT	0.71	0.71	0.71	141
	After_DCT	0.00	0.00	0.00	33
Weighted average		0.9338	0.9403	0.9357	42946
Macro average		0.3715	0.3614	0.3505	42946
Micro average				0.9403	42946
Accuracy				**0.9403**	42946

Bold indicates the best value

Table 12 XLNET Evaluation Metrics for each risk factor indicator based on time attribute Identification

Risk factor	Time attribute	Precision	Recall	F-Score	Support
CAD	Before_DCT	0.80	0.83	0.81	406
	During_DCT	0.78	0.61	0.69	222
	After_DCT	0.66	0.63	0.65	71
Diabetes	During_DCT	0.67	0.49	0.57	272
	Before_DCT	0.72	0.57	0.64	209
	After_DCT	0.49	0.39	0.44	164
Hypertension	Before_DCT	0.70	0.64	0.67	116
	During_DCT	0.89	0.95	0.92	553
	After_DCT	0.10	0.08	0.09	51
Obese	Before_DCT	0.46	0.32	0.38	41
	During_DCT	0.77	0.52	0.62	69
	After_DCT	0.00	0.00	0.00	11
Medication	Before_DCT	0.67	0.51	0.58	863
	During_DCT	0.54	0.36	0.43	695
	After_DCT	0.67	0.61	0.64	894
Hyperlipidemia	Before_DCT	1.00	0.02	0.05	82
	During_DCT	0.75	0.72	0.73	65
	After_DCT	0.88	0.81	0.84	109
Weighted average		0.9361	0.9397	0.9371	42946
Macro average		0.3988	0.3848	0.3772	42946
Micro average				0.9397	42946
Accuracy				**0.9397**	42946

Bold indicates the best value

Table 13 Transformer Models

Model/risk factor	BERT	BioBERT	BioClinical Bert	RoBERTa	XLNET
Diabetes	0.82	0.83	0.80	0.75	0.75
Hypertension	0.90	0.91	0.92	0.92	0.91
CAD	0.77	0.78	0.79	0.71	0.77
Medication	0.77	0.80	0.80	0.80	0.81
Smoker	0.88	0.90	0.94	0.90	0.90
Obese	0.81	0.81	0.71	0.70	0.74
Hyperlipidemia	0.78	0.80	0.83	0.75	0.77
FamilyHist	0.88	0.88	0.88	0.88	0.88
F-score (micro)	0.9373	0.9399	0.9403	0.9427	0.9397

Table 14 Ensembles

All Ensembles (BERT+BioBERT+BioClinicalBERT+RoBERTa+XLNet)	Precision	Recall	F1-score	Support
Macro average	0.3218	0.3330	0.3129	42946
Weighted average	0.9337	0.9426	0.9366	42946
Accuracy			**0.9426**	42946

Bold indicates the best value

Table 15 Hyperparameters optimized via training

Hyperparameter	BERT	RoBERTa	BioBERT	BioClinicalBERT	XLNet
Hidden size	768	768	768	768	144
Number of layers	12	12	12	13	6
Number of attention heads	12	12	12	12	6
Feed-forward layer hidden size	128	128	128	128	128
Learning rate	1×10^{-6}	5×10^{-7}	5×10^{-5}	5×10^{-6}	5×10^{-6}
Batch size	16	16	16	16	16
Dropout	0.5	0.1	0.1	0.4	0.4

Acknowledgements
The authors would like to thank the Science, Technology & Innovation Funding Authority (STDF) in cooperation with the Egyptian Knowledge Bank (EKB).

Author contributions
E.H.H., participated in the supervision, sorting of the experiments and analyzed the results, E.H.H., R.E.M., and G.H., performed the experiments, visualization, formal analysis, discussed/analyzed the results and wrote the article. A.A.A., participated in the supervision. All authors approved the work in this article.

Funding
Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Availability of data and materials
The data sets provided during the current study are available: http://www.partners.org and https://www.i2b2.org/NLP/HeartDisease/.

Declarations

Ethics approval and consent to participate
This article does not contain any studies with human participants or animals carried out by any of the authors.

Consent for publication
Not applicable.

Competing interests
The authors declare that there is no Competing interests.

Received: 9 July 2023 Accepted: 14 March 2024
Published online: 04 April 2024

References

1. World Health Organization et al. Global status report on noncommunicable diseases 2014. Number WHO/NMH/NVI/15.1. World Health Organization. 2014.
2. Herron MP. Cdc national vital statistics reports. Deaths: Leading Causes for 2017. Statistics. 2017;66:5.
3. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Das SR, et al. Heart disease and stroke statistics-2019 update: a report from the American heart association. Circulation. 2019;139(10):e56–528.
4. Roth GA, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, Abbastabar H, Abd-Allah F, Abdela J, Abdelalim A, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017. Lancet. 2018;392(10159):1736–88.
5. Zhao M, Wang M, Zhang J, Ye J, Yao X, Wang Z, Ye D, Liu J, Wan J. Advances in the relationship between coronavirus infection and cardiovascular diseases. Biomed Pharmacother. 2020;127: 110230.
6. Hajar R. Risk factors for coronary artery disease: historical perspectives. Heart Views. 2017;18(3):109.
7. U.S. Department of Health and Human Services. National institute of diabetes and digestive and kidney diseases. 2021. https://www.niddk.nih.gov/health-information/diabetes. Accessed 27 Nov 2021.
8. National Heart Lung and Blood Institute. Coronary heart disease | nhlbi, nih. 2016. https://www.nhlbi.nih.gov/health-topics/coronary-heart-disease. Accessed 27 Nov 2021.
9. Dokken BB. The pathophysiology of cardiovascular disease and diabetes: beyond blood pressure and lipids. Diabetes Spectr. 2008;21(3):160–5.

10. Chokwijitkul T, Nguyen A, Hassanzadeh H, Perez S. Proceedings of the identifying risk factors for heart disease in electronic medical records: a deep learning approach. In: Demner-Fushman D, Cohen KB, Ananiadou S, Tsujii J, editors. BioNLP 2018 workshop. Melbourne: Association for Computational Linguistics; 2018. p. 18–27.
11. Zhan X, Humbert-Droz M, Mukherjee P, Gevaert O. Structuring clinical text with AI: old versus new natural language processing techniques evaluated on eight common cardiovascular diseases. Patterns. 2021;2(7): 100289.
12. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008;17(01):128–44.
13. Hebal F, Nanney E, Christine Stake ML, Miller GL, Barsness KA. Automated data extraction: merging clinical care with real-time cohort-specific research and quality improvement data. J Pediatr Surg. 2017;52(1):149–52.
14. Safran C, Meryl Bloomrosen W, Hammond E, Labkoff S, Markel-Fox S, Tang PC, Detmer DE. Toward a national framework for the secondary use of health data: an American medical informatics association white paper. J Am Med Inform Assoc. 2007;14(1):1–9.
15. Mann CJ. Observational research methods. Research design ii: cohort, cross sectional, and case-control studies. Emerg Med J. 2003;20(1):54–60.
16. Geneletti S, Richardson S, Best N. Adjusting for selection bias in retrospective, case-control studies. Biostatistics. 2009;10(1):17–31.
17. Oleynik M, Kugic A, Kasáč Z, Kreuzthaler M. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. J Am Med Inform Assoc. 2019;26(11):1247–54.
18. Ebbehoj A, Thunbo MØ, Andersen OE, Glindtvad MV, Hulman A. Transfer learning for non-image data in clinical research: a scoping review. PLOS Digit Health. 2022;1(2): e0000014.
19. Alyafeai Z, AlShaibani MS, Ahmad I. A survey on transfer learning in natural language processing. 2020. arXiv preprint arXiv:2007.04239.
20. Laparra E, Mascio A, Velupillai S, Miller T. A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. Yearb Med Inform. 2021;30(01):239–44.
21. Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2010;22(10):1345–59.
22. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT press; 2016.
23. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. 2013. arXiv preprint arXiv:1301.3781.
24. Arnold S, Gers FA, Kilias T, Löser A. Robust named entity recognition in idiosyncratic domains. 2016. arXiv preprint arXiv:1608.06757.
25. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist. 2017;5:135–46.
26. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. arXiv preprint. 2016. arXiv:1607.01759.
27. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. Mimic-iii, a freely accessible critical care database. Sci Data. 2016;3(1):1–9.
28. Zhang Y, Chen Q, Yang Z, Lin H, Zhiyong L. Biowordvec, improving biomedical word embeddings with subword information and mesh. Sci Data. 2019;6(1):52.
29. Chen Q, Peng Y, Lu Z. Biosentvec: creating sentence embeddings for biomedical texts. In: Chen Q, editor. 2019 IEEE International Conference on Healthcare Informatics (ICHI). Xi'an: IEEE; 2019. p. 1–5.
30. Stubbs A, Kotfila C, Hua X, Uzuner Ö. Identifying risk factors for heart disease over time: overview of 2014 i2b2/uthealth shared task track 2. J Biomed Inform. 2015;58:S67–77.
31. Stubbs A, Uzuner Ö. Annotating risk factors for heart disease in clinical narratives for diabetic patients. J Biomed Inform. 2015;58:S78–91.
32. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint. 2018. arXiv:1810.04805.
33. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40.
34. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, McDermott M. Publicly available clinical bert embeddings. 2019. arXiv preprint arXiv:1904.03323.
35. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. Roberta: A robustly optimized bert pretraining approach. 2019. arXiv preprint arXiv:1907.11692.
36. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: generalized autoregressive pretraining for language understanding. Adv Neural Inform Process Syst. 2019:32.
37. Roberts K, Shooshan SE, Rodriguez L, Abhyankar S, Kilicoglu H, Demner-Fushman D. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. J Biomed Inform. 2015;58:S111–9.
38. Harkema H, Dowling JN, Thornblade T, Chapman WW. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform. 2009;42(5):839–51.
39. Kotfila C, Uzuner Ö. A systematic comparison of feature space effects on disease classifier performance for phenotype identification of five diseases. J Biomed Inform. 2015;58:S92–102.
40. Aronson AR. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In: Aronson AR, editor. Proceedings of the AMIA Symposium. Bethesda: American Medical Informatics Association; 2001. p. 17.
41. Chen Q, Li H, Tang B, Wang X, Liu X, Liu Z, Liu S, Wang W, Deng Q, Zhu S, et al. An automatic system to identify heart disease risk factors in clinical texts over time. J Biomed Inform. 2015;58:S158–63.
42. Urbain J. Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models. J Biomed Inform. 2015;58:S143–9.
43. Torii M, Fan J, Yang W, Lee T, Wiley MT, Zisook DS, Huang Y. Risk factor detection for heart disease by applying text analytics in electronic medical records. J Biomed Inform. 2015;58:S164–70.

44. Hua X, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. Medex: a medication information extraction system for clinical narratives. J Am Med Inform Assoc. 2010;17(1):19–24.
45. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The weka data mining software: an update. ACM SIGKDD Explorat Newsl. 2009;11(1):10–8.
46. Chang C-C, Lin C-J. Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol (TIST). 2011;2(3):1–27.
47. Manning CD, Surdeanu M, Bauer J, Finkel JR, Bethard S, McClosky D. The stanford corenlp natural language processing toolkit. In: Manning CD, editor. Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. Association for Computational Linguistics: Baltimore; 2014. p. 55–60.
48. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, Zhu Y, Rahimi K, Salimi-Khorshidi G. Behrt: transformer for electronic health records. Sci Rep. 2020;10(1):1–12.
49. Si Y, Roberts K. Patient representation transfer learning from clinical notes based on hierarchical attention network. AMIA Summits Transl Sci Proc. 2020;2020:597.
50. Syed K, William Sleeman IV, Hagan M, Palta J, Kapoor R, Ghosh P. Automatic incident triage in radiation oncology incident learning system. Healthcare. 2020;8:272.
51. Dai H-J, Chu-Hsien S, Lee Y-Q, Zhang Y-C, Wang C-K, Kuo C-J, Chi-Shin W. Deep learning-based natural language processing for screening psychiatric patients. Front Psychiatry. 2021;11: 533949.
52. Al-Garadi MA, Yang Y-C, Cai H, Ruan Y, O'Connor K, Graciela G-H, Perrone J, Sarker A. Text classification models for the automatic detection of nonmedical prescription medication use from social media. BMC Med Inform Decis Mak. 2021;21(1):1–13.
53. Jingcheng D, Zhang Y, Luo J, Jia Y, Wei Q, Tao C, Hua X. Extracting psychiatric stressors for suicide from social media using deep learning. BMC Med Inform Decis Mak. 2018;18:77–87.
54. Howard D, Maslej MM, Lee J, Ritchie J, Woollard G, French L. Transfer learning for risk classification of social media posts: model evaluation study. J Med Internet Res. 2020;22(5): e15371.
55. Rios A, Kavuluru R. Neural transfer learning for assigning diagnosis codes to EMRs. Artif Intell Med. 2019;96:116–22.
56. Hassanzadeh H, Kholghi M, Nguyen A, Chu K. Clinical document classification using labeled and unlabeled data across hospitals. In: Hassanzadeh H, editor. AMIA annual symposium proceedings, vol. 2018. Bethesda: American Medical Informatics Association; 2018. p. 545.
57. Ji B, Li S, Jie Yu, Ma J, Tang J, Qingbo W, Tan Y, Liu H, Ji Y. Research on Chinese medical named entity recognition based on collaborative cooperation of multiple neural network models. J Biomed Inform. 2020;104: 103395.
58. Newman-Griffis D, Zirikly A. Embedding transfer for low-resource medical named entity recognition: a case study on patient mobility. 2018. arXiv preprint arXiv:1806.02814.
59. Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc. 2010;17(5):514–8.
60. Gligic L, Kormilitzin A, Goldberg P, Nevado-Holgado A. Named entity recognition in electronic health records using transfer learning bootstrapped neural networks. Neural Netw. 2020;121:132–9.
61. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Adv Neural Inform Process Syst. 2017:30.
62. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR. Glue: a multi-task benchmark and analysis platform for natural language understanding. 2018. arXiv preprint arXiv:1804.07461.
63. Tay Y, Dehghani M, Bahri D, Metzler D. Efficient transformers: a survey. 2020. arxiv. arXiv preprint arXiv:2009.06732.
64. Yu J, Bohnet B, Poesio M. Named entity recognition as dependency parsing. 2020. arXiv preprint arXiv:2005.07150.
65. Li X, Sun X, Meng Y, Liang J, Wu F, Li J. Dice loss for data-imbalanced NLP tasks. 2019. arXiv preprint arXiv:1911.02855.
66. Benfeng X, Wang Q, Lyu Y, Zhu Y, Mao Z. Entity structure within and throughout: modeling mention dependencies for document-level relation extraction. Proc AAAI conf Artif Intelli. 2021;35:14149–57.
67. Wang J, Lu W. Two are better than one: joint entity and relation extraction with table-sequence encoders. 2020. arXiv preprint arXiv:2010.03851.
68. Jiang H, He P, Chen W, Liu X, Gao J, Zhao T. Smart: robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. 2019. arXiv preprint arXiv:1911.03437.
69. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(1):5485–551.
70. Zhang Z, Yuwei W, Zhao H, Li Z, Zhang S, Zhou X, Zhou X. Semantics-aware bert for language understanding. Proc AAAI Conf Artif Intell. 2020;34:9628–35.
71. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. Albert: A lite bert for self-supervised learning of language representations. 2019. arXiv preprint arXiv:1909.11942.
72. Zhang Z, Yang J, Zhao H. Retrospective reader for machine reading comprehension. Proc AAAI Conf Artif Intell. 2021;35:14506–14.
73. Garg S, Thuy V, Moschitti A. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. Proc AAAI Conf Artif Intell. 2020;34:7780–8.
74. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E et al. On the opportunities and risks of foundation models. 2021. arXiv preprint arXiv:2108.07258.
75. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. 2019. arXiv preprint arXiv:1906.05474.
76. Beltagy I, Lo K, Cohan A. Scibert: a pretrained language model for scientific text. arXiv preprint. 2019. arXiv:1903.10676.
77. Yu G, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comp Healthc (HEALTH). 2021;3(1):1–23.
78. Lewis P, Ott M, Du J, Stoyanov V. Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art. In: Rumshisky A, Roberts K, Bethard S, Naumann T, editors. Proceedings of the 3rd Clinical Natural Language Processing Workshop. Stroudsburg: Association for Computational Linguistics; 2020. p. 146–57.

79. Fiorini N, Leaman R, Lipman DJ, Zhiyong L. How user intelligence is improving pubmed. Nat biotechnol. 2018;36(10):937–45.
80. Gillick D. Sentence boundary detection and the problem with the us. In: Gillick D, editor. Proceedings of human language technologies: the 2009 annual conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. Boulder: Association for Computational Linguistics; 2009. p. 241–4.
81. Yang H, Garibaldi JM. A hybrid model for automatic identification of risk factors for heart disease. J Biomed Inform. 2015;58:S171–82.
82. Cormack J, Nath C, Milward D, Raja K, Jonnalagadda SR. Agile text mining for the 2014 i2b2/uthealth cardiac risk factors challenge. J Biomed Inform. 2015;58:S120–7.
83. Khalifa A, Meystre S. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. J Biomed Inform. 2015;58:S128–32.
84. Kumar Vivek, Recupero Diego Reforgiato, Riboni Daniele, Helaoui Rim. Ensembling classical machine learning and deep learning approaches for morbidity identification from clinical notes. IEEE Access. 2020;9:7107–26.

Publisher's Note