Journal of the American Medical Informatics Association, 29(7), 2022, 1208–1216
https://doi.org/10.1093/jamia/ocac040
Advance Access Publication Date: 25 March 2022
Research and Applications

Research and Applications

CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records

Sicheng Zhou [1], Nan Wang [2], Liwei Wang [3], Hongfang Liu [3], and Rui Zhang [1,4]

[1]Institute for Health Informatics, University of Minnesota, Minneapolis, Minnesota, USA, [2]School of Statistics, University of Minnesota, Minneapolis, Minnesota, USA, [3]Department of AI and Informatics Research, Mayo Clinic, Rochester, Minnesota, USA, and [4]Department of Pharmaceutical Care & Health Systems, University of Minnesota, Minneapolis, Minnesota, USA

Corresponding Author: Rui Zhang, PhD, 8-100 Phillips-Wangensteen Building, 516 Delaware Street SE, Minneapolis, MN 55455, USA; zhan1386@umn.edu

Received 15 December 2021; Revised 6 March 2022; Editorial Decision 7 March 2022; Accepted 9 March 2022

ABSTRACT

Objective: Accurate extraction of breast cancer patients' phenotypes is important for clinical decision support and clinical research. This study developed and evaluated cancer domain pretrained CancerBERT models for extracting breast cancer phenotypes from clinical texts. We also investigated the effect of customized cancer-related vocabulary on the performance of CancerBERT models.

Materials and Methods: A cancer-related corpus of breast cancer patients was extracted from the electronic health records of a local hospital. We annotated named entities in 200 pathology reports and 50 clinical notes for 8 cancer phenotypes for fine-tuning and evaluation. We kept pretraining the BlueBERT model on the cancer corpus with expanded vocabularies (using both term frequency-based and manually reviewed methods) to obtain CancerBERT models. The CancerBERT models were evaluated and compared with other baseline models on the cancer phenotype extraction task.

Results: All CancerBERT models outperformed all other models on the cancer phenotyping NER task. Both CancerBERT models with customized vocabularies outperformed the CancerBERT with the original BERT vocabulary. The CancerBERT model with manually reviewed customized vocabulary achieved the best performance with macro *F*1 scores equal to 0.876 (95% CI, 0.873–0.879) and 0.904 (95% CI, 0.902–0.906) for exact match and lenient match, respectively.

Conclusions: The CancerBERT models were developed to extract the cancer phenotypes in clinical notes and pathology reports. The results validated that using customized vocabulary may further improve the performances of domain specific BERT models in clinical NLP tasks. The CancerBERT models developed in the study would further help clinical decision support.

Key words: natural language processing, CancerBERT, cancer phenotyping, electronic health record, name entity recognition

INTRODUCTION

Breast cancer is one of the most prevalent and lethal cancers for women in the United States. It is estimated that there will be about 250 000 patients diagnosed with breast cancer each year and around 40 000 deaths due to breast cancer.[1] The development of precision medicine has contributed new approaches to the better diagnosis, prognosis, and treatments of breast cancers, with the ultimate goal of selecting optimal treatments for individual patients.[2–4] A repre-

sentative example is the targeted therapy for breast cancer, which uses different medications to treat patients with different hormone receptors status, such as human epidermal growth factor receptor 2 (HER2) and estrogen receptor (ER). The application of precision medicine and its related translational research need the support of large amounts of cancer-specific patient clinical information. The widely adopted electronic health record systems (EHRs) are fundamental sources to provide longitudinal and multiperspective patient clinical data, which includes patients' demographics, lab results, disease progress, treatments, and outcomes. Some of these data are stored in the codified (structured) part of EHRs; however, a large amount of information is distributed in the narrative text data, such as the clinical notes and lab reports.[5] How to effectively extract the target information from the narrative text part of the EHRs data remains an important research topic.[6,7] Previous studies (detailed in "Background" section) focused mainly on the rule-based and conventional machine learning methods. The state-of-the-art language models such as bidirectional encoder representations from transformers (BERT)[8] have shown significant improvement in many NLP tasks; however, there is no cancer domain-specific BERT model for downstream clinical NLP tasks, such as cancer phenotype extraction. In addition, there is no investigation of out-of-vocabulary (OOV) issue for BERT-based model in cancer domain.

Objective

To address these gaps in current status of the cancer phenotyping extraction, our contributions in this study include:

1. We developed and evaluated cancer domain-specific BERT models (CancerBERT) that are able to extract comprehensive collections of breast cancer-related phenotypes (ie, *Hormone receptor type, Hormone receptor status, Tumor size, Tumor site, Cancer grade, Histological type, Tumor laterality,* and *Cancer stage*) from both clinical notes and pathology reports in EHRs. Our CancerBERT models significantly outperformed other existing BERT-based models (eg, BlueBERT, BioBERT, and CharBERT) on our name entity recognition (NER) task to extract target cancer phenotypes for breast cancer patients.
2. We also evaluated different methods to address the OOV issue of original BERT models. Specifically, we used 2 approaches (domain knowledge-based and statistics-based) to generate and add additional cancer-specific words that were missing in the original BERT vocabulary. We found that additional cancer-specific words can further improve the performance of CancerBERT model on the NER task.

Background

Previous works have developed different approaches to extract information from narrative data in EHRs. Manual chart review is a feasible approach to extract the phenotypes from the clinical texts; however, it is time-consuming and not cost-effective.[9,10] Researchers have developed approaches based on natural language processing (NLP) to finish the task automatically. Before the deep learning era, established studies mainly focused on rule-based, traditional machine learning-based methods, depending on the characteristic of the data and specific tasks. Nguyen et al developed a rule-based pipeline in 2015 to extract cancer-related phenotypes, including histological type, cancer grade, primary site, and laterality, from the textual contents of the pathology reports in EHRs.[11] The *F*1 scores for different variables range from 0.61 to 0.93, and a message producer/consumer module was integrated into the pipeline to enable the real-time processing of the reports. Yala et al[12] developed a machine learning algorithm to classify phenotypes of breast cancers. N-grams were used as features and boosting algorithm was applied to do the classification of phenotype status. The performances were robust, with *F*1 scores ranging from 0.57 to 1 for different categories. However, the pipeline was designed for judging the status of phenotypes, mainly binary classification; no detail information in the text can be captured. The DeepPhe software was developed in 2017 to extract cancer phenotypes from clinical records.[13] It could extract a wide range of breast cancer phenotypes from the EHRs through different approaches, such as, rules, domain knowledge bases, and machine learning methods. The interannotator agreement of the DeepPhe range from 0.2 to 0.96.[13] Qiu et al[14] developed a convolutional neural network (CNN) model to extract the cancer primary site from pathology reports, and the CNN model outperformed the traditional frequency vector space approach with a micro-*F* score of 0.722. A coarse-to-fine multi-task CNN model was further proposed to extract the cancer primary site, laterality, and grade from the pathology reports at the same time; this model obtained an $F-1$ score of 0.775 for extracting cancer primary site.[15]

These studies focused mainly on the rule-based and conventional machine learning methods. The latest BERT-based models have been developed in recent years and show great advantages in NLP tasks compared to the traditional feature-based machine learning approaches.[8] For BERT-based models, previous works have shown that using biomedical domain-specific text as training data can obtain better performance compared to models trained on general-domain language for tasks related to the biomedical domain.[16–19] In clinical domain, studies have explored using the advanced BERT-based models to solve clinical information extraction tasks.[16,17,20] These studies are mainly focusing on testing the BERT model on clinical benchmark datasets, such as Informatics for Integrating Biology and the Bedside, SemEval, and MedSTS. Currently, only few studies have applied the advanced deep learning models include BERT to extract the cancer phenotypes for cancer patients. For example, BERT models were applied to extract the clinical information for breast cancer patients from Chinese clinical texts and achieved *F*1 scores of 0.786 to 1 for different clinical concepts.[21] It is known that Chinese and English are different, and there are no existing studies that have explored the BERT-based models to extract cancer phenotypes from clinical notes and pathology reports in English. Furthermore, most BERT-based models deal with the OOV issue by tokenizing an unknown word into multiple subwords that exist in the vocabulary. In this case, the subword representations may not capture the semantics of the whole word.[22] Several studies explored to improve the OOV issue by either using character-level word embedding[22,23] or building a brand-new domain specific vocabulary to best match the training corpus.[24] These studies obtained promising results in benchmark tasks, but they need to train the models from scratch, which need a huge training corpus and computing sources.

METHODS

Data collection and annotation

The data used in this study were obtained from the EHRs of the University of Minnesota (UMN) Clinical Data Repository. The EHRs of UMN contain the health records of 21 291 breast cancer patients from year 2001 to 2018. We obtained the data with the approval of UMN Institutional Review Board under #1210M22601.

To develop the phenotype extraction algorithms, the reference standards (ie, an annotated corpus) needed to be obtained through chart review. To obtain the standard annotations, an annotation guideline was first developed through iterative discussions. The UMN team reviewed some of the pathology reports and clinical notes to collect the different descriptions of the phenotypes in the EHRs and formed the annotation guideline. We randomly sampled 200 pathology reports and 50 clinical notes of breast cancer patients that contain 9685 sentences; 221 356 tokens were manually annotated by 2 annotators (graduate students with clinical or pharmacy background). The target entities were annotated in entity level. Cohen's kappa scores were calculated to ensure the consistency between the annotators. INCEpTION was used as the annotation tool.[25] Figure 1 shows several examples of the annotation.

We focused on 8 breast cancer phenotypes that describe the characteristics of breast cancer, including the *Hormone receptor type, Hormone receptor status, Tumor size, Tumor site, Cancer grade, Histological type, Tumor laterality,* and *Cancer stage*. Targeted breast cancer phenotypes, their potential values, and the according examples in clinical text are shown in Table 1.

In total, 200 pathology reports and 50 clinical notes were annotated by 2 annotators, the Cohen kappa score for annotations was calculated to be 0.91. The annotation statistics are shown in Table 2.

Our model: CancerBERT

Pretraining on cancer domain-specific corpus

In this study, we trained cancer domain-specific BERT models (CancerBERT) that are expected to better capture the semantics in cancer-specific clinical notes and pathology reports, thus improving the performance of the task for extracting breast cancer-related phenotypes. The CancerBERT models training process is illustrated in Figure 2.

The BERT-origin model was trained using Wikipedia and book corpus[15] using the original BERT vocabulary, which was also generated from Wikipedia and book corpus. The BlueBERT was further pretrained on PubMed and MIMIC III data based on the BERT-origin model using the same vocabulary.[17] We kept pretraining the CancerBERT models based on the BlueBERT model. For the pretraining corpus, we extracted the 4 543 184 clinical notes and 1 278 805 pathology reports (about 1 billion tokens) for 21 291 breast cancer patients from UMN EHRs. The corpus was changed to lower case, no other preprocessing was needed. Hereinafter, we called the CancerBERT trained with original BERT vocabulary as $\text{CancerBERT}_{\text{OrigVoc}}$ to differentiate other CancerBERT variants using customized vocabulary described below.

Constructing cancer domain-specific vocabulary for improving $\text{CancerBERT}_{\text{OrigVoc}}$ models

As outlined above, the vocabulary of $\text{CancerBERT}_{\text{OrigVoc}}$ is identical to the BERT-origin model[15]; thus, many special words and abbreviations in the clinical narratives cannot be covered. The OOV issue may influence the performance of the language model. In the BERT-based models, the WordPiece tokenizer[26] was applied to deal with the OOV issue. It tokenizes an unknown word into multiple subwords that exist in the vocabulary. For instance, the word "HER2", a breast cancer-related cell receptor gene, is not in the original BERT vocabulary. It will be tokenized into "HER" and "2" by the WordPiece tokenizer; and the model will then use the average (or the first part) of their word embeddings to represent "HER2". However, the word embeddings of "HER" and "2" cannot correctly represent the semantics of the term "HER2". Thus, it may be helpful to add the unknown word "HER2" into the original vocabulary to train its own word embeddings. Thus, we further explored different approaches to generate and incorporate cancer-specific vocabulary in an attempt to further improve the performance of the $\text{CancerBERT}_{\text{OrigVoc}}$ model.

In this study, we explored 2 ways to generate additional cancer-related vocabularies to include in the original BERT vocabulary.

Method 1—Domain knowledge-based:

1. SpaCy[27] tokenizer was used to tokenize the breast cancer training corpus from EHRs to produce a new list of words.
2. All unique words in the new word list that did not appear in the original vocabulary were identified.

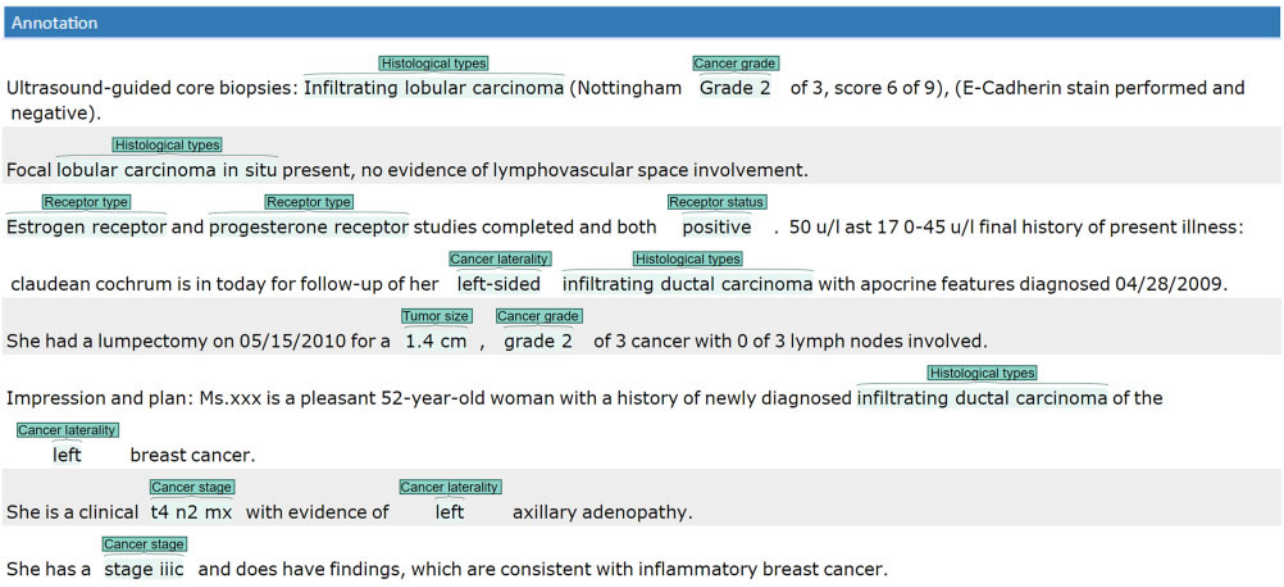

Figure 1. Examples of annotation in INCEpION.

Table 1. Breast cancer phenotypes, their potential values, and examples in clinical texts

Phenotypes	Values	Examples of descriptions in clinical text
Hormone receptor type	Positive, negative	HER2 gene was amplified; estrogen receptor: positive (95%, strong staining); tumor is PR negative (0% staining)
Tumor size	Numeric values describe volumes	Tumor size: 1.0×0.5×0.7 cm
Tumor site	Description of positions	Tumor is at 12 o'clock position and 2 cm from the nipple.
Cancer grade	Numerical values: (1–3)	Histologic grade: 1 of 3; Sample shows Nottingham grade 2 lesions
Histological type	Ductal carcinoma in situ (DCIS); lobular carcinoma in situ (LCIS), etc.	Histologic type of invasive carcinoma: ductal carcinoma in situ
Tumor laterality	Right, left	Specimen laterality: right breast; Laterality: left tumor
Cancer stage	TNM staging: TX, Tis, T1-4; NX, N0, N1-3; M0, M1	Pathologic stage is pT4 NX MX

HER2: human epidermal growth factor receptor 2.

Table 2 . Annotation statistics

		Total number	Total unique entities
Annotated statistics	Documents	200	NA
	Total sentences	9685	NA
	Total tokens	221 356	NA
Name entity statistics	Hormone receptor type	1673	29
	Hormone receptor status	436	14
	Tumor size	540	305
	Tumor site	329	173
	Cancer grade	271	15
	Tumor laterality	1192	4
	Cancer stage	173	38
	Histological type	1070	95

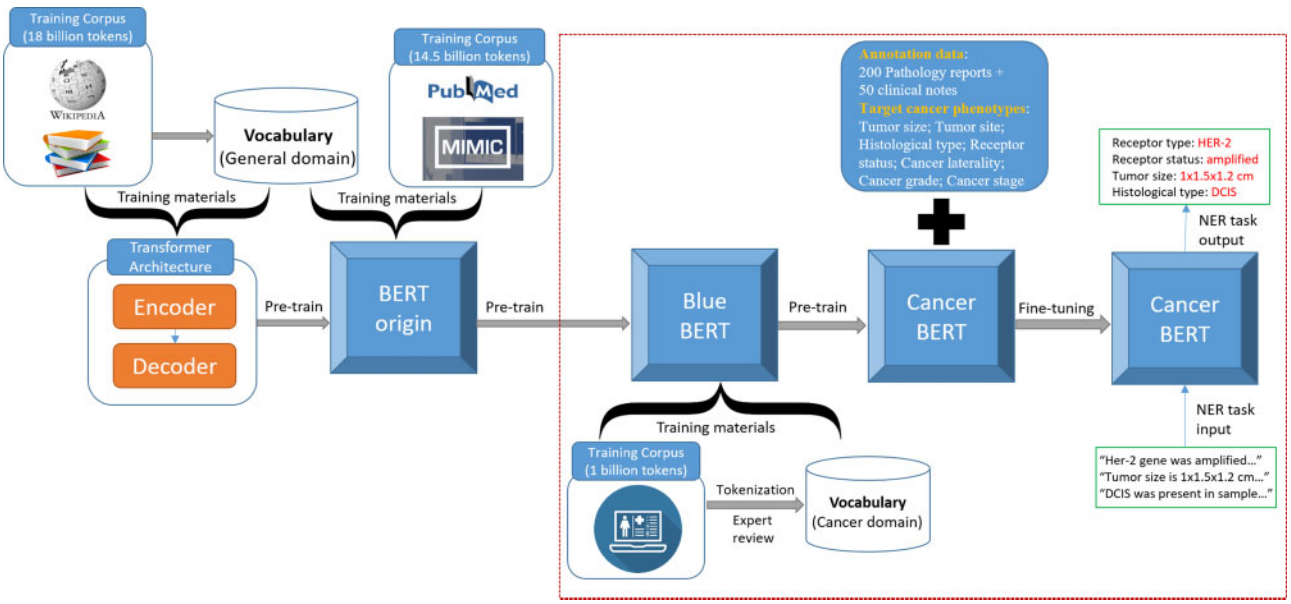

Figure 2. The training process of CancerBERT models. The CancerBERT models were pretrained based on the BlueBERT model. The process in the red box was implemented in this study. BERT: bidirectional encoder representations from transformers.

3. A researcher with clinical background reviewed the newly identified words and selected 397 cancer-related words for vocabulary expansion.

Method 2—Frequency-based:

1. SpaCy[27] tokenizer was used to tokenize the breast cancer training corpus from EHRs to produce a new list of words.
2. All unique words in the new word list that did not appear in the original vocabulary were identified.
3. The 997 most frequent words in the new word list were selected for vocabulary expansion (the original BERT vocabulary permits a maximum of 997 new words).

To compare the effect of using the cancer domain-customized vocabulary on the CancerBERT performance, we pretrained 3 CancerBERT models with different sets of vocabulary (described below) based on BlueBERT model:

1. $\text{CancerBERT}_{\text{OrigVoc}}$: used the original BERT vocabulary.
2. $\text{CancerBERT}_{\text{CustVoc\_397}}$: used the original BERT vocabulary + 397 cancer-related words based on domain knowledge.
3. $\text{CancerBERT}_{\text{CustVoc\_997}}$: used the original BERT vocabulary + 997 cancer-related words based on the term frequency.

Fine-tuning CancerBERT models along with other BERT-based models

BERT-based models could be further fine-tuned on annotation data to solve specific downstream tasks. The extraction of breast cancer-related phenotypes from texts can be framed as an NER task. The NER is one of the most important tasks in information extraction of text data. It classifies every token in the text into predefined entity classes.[28] In this study, we have 8 types of name entities: *Hormone receptor type, Hormone receptor status, Tumor size, Tumor site, Cancer grade, Histological type, Tumor laterality*, and *Cancer stage*. The annotated training set was used for the BERT-based model fine-tuning. For major hyperparameters, the max sequence length was set to 128, the training batch size was set to 32, and training epoch was set to 10. The hyperparameters were chosen based on the memory and computing power of our GPU resources. We fine-tuned our $\text{CancerBERT}_{\text{OrigVoc}}$, $\text{CancerBERT}_{\text{CustVoc\_397}}$, and $\text{CancerBERT}_{\text{CustVoc\_997}}$ models, along with the original BERT-large,[15] BioBERT,[16] BlueBERT,[17] CharBERT,[22] and character-BERT[23] models on the NER task. All the BERT-based models in this study are uncased. The original BERT-large model was pretrained on Wikipedia and BookCorpus.[15] The BioBERT model was pretrained on PubMed abstract and PMC full articles that contain about 18 billion words.[16] BlueBERT model was pretrained on PubMed abstract and MIMIC-III that contain about 4.5 billion words.[17] The CharBERT was pretrained on Wikipedia corpus that contains 2.5 billion words.[22] Character-BERT was pretrained on Wikipedia corpus then further pretrained on MIMIC-III clinical notes and PMC OA biomedical paper abstracts.[23]

Evaluation

We evaluated the 3 CancerBERT models along with other models on the NER task for cancer phenotyping extraction. We applied name entity level evaluation for the NER task. Twenty of the annotated data were used as a test set. The *F*1 score was used as an evaluation metric. For overall performance, the microaverage *F*1 (calculating precision and recall by counting the sums of the true positives, false negatives, false positives for all classes, and then calculating *F*1 score) and macroaverage *F*1 (arithmetic mean of all per-class *F*1 scores) were used. Following the n2c2 evaluation metrics,[29] we evaluated in both exact match and lenient match ways for name entities. For exact match, the entity boundary of predicted entity and gold standard should be same, while for lenient match, the entity boundary of predicted entity and gold standard can be overlapped.

We developed BiLSTM-CRF models as the baseline models for the NER task comparison. The input features for the BiLSTM-CRF models are the pretrained word embeddings. We compared the 4 different word embeddings (ie, Word2Vec model pretrained on Google News,[30] Word2Vec model pretrained on our breast cancer corpus, Global Vectors for Word Representation (GloVe) model pretrained on Wikipedia,[31] and GloVe model pretrained our breast cancer corpus) as features and finally chose the GloVe trained on Wikipedia corpus[31] as the baseline for further comparison since it obtained the best performance among the 4 word embeddings for cancer phenotype extraction task.

RESULTS

Performance comparison of CancerBERT models with other BERT models on the cancer phenotyping NER task

The evaluation results for CancerBERT and other BERT-based models pretrained in the general biomedical and clinical corpora are shown in Table 3. The strict and lenient match *F*1 scores are shown in the table (lenient match *F*1 in parenthesis). The scores were averaged scores based on 10 runs, numbers in bold indicate the highest score and asterisk indicates the number is statistically higher than other methods (CI: 0.95). All 3 CancerBERT models outperformed the baseline models and other BERT models. The $\text{CancerBERT}_{\text{CustVoc\_397}}$ model obtained the best performance on 4 of 8 entities and obtained the best overall macro *F*1 scores (0.876 for strict match and 0.904 for lenient match) and micro *F*1 scores (0.909 for strict match and 0.933 for lenient match). Overall, the performance of $\text{CancerBERT}_{\text{CustVoc\_397}}$ model significantly surpassed other models; the $\text{CancerBERT}_{\text{CustVoc\_997}}$ model also obtained good performance.

We explored the total number of unique annotated tokens for each name entity category and how many of them could be identified in the original BERT vocabulary, customized BERT vocabulary based on frequency (for $\text{CancerBERT}_{\text{CustVoc\_997}}$ model), and customized BERT vocabulary based on domain knowledge (for $\text{CancerBERT}_{\text{CustVoc\_397}}$ model). The results are shown in Table 4.

We extracted the word embeddings of all unique tokens for each name entity category from the $\text{CancerBERT}_{\text{CustVoc\_997}}$ and $\text{CancerBERT}_{\text{CustVoc\_397}}$ models and visualized the word embeddings for these tokens in a 2-dimensional plot using *t*-distributed stochastic neighbor embedding (t-SNE).[32] Examples of token clusters are shown in Figure 3. Figure 3a and b shows the clusters of *Cancer stage* (eg, ptis, n2a, pn1a, and t1c), *Hormone receptor status* (eg, er-positive, er-negative, and receptor-positive), and *Tumor laterality* (eg, left-sided, right-sided, and b-left) obtained from $\text{CancerBERT}_{\text{CustVoc\_397}}$ model. Figure 3c and d shows the clusters of *Hormone receptor type* (eg, estrogen, progesterone, and her2), *Hormone receptor status* (eg, equivocal, amplified, and nonamplified), and *Histological type* (eg, dcis, lcis, lobular, and adenocarcinoma) obtained from $\text{CancerBERT}_{\text{CustVoc\_997}}$ model. Some words in the

Table 3. BERT fine-tuning NER entity level evaluation by exact match *F*1 score (lenient match *F*1 scores are shown in parenthesis)

Entity type	BiLSTM-CRF	BERT-large origin	BlueBERT (PubMed + MIMIC III)	BioBERT (PubMed)	CharBERT (Wiki)	Character-BERT (Medical)	$\text{CancerBERT}_{\text{Orig-Voc}}$ (EHRs corpus)	$\text{CancerBERT}_{\text{Cust-Voc\_997}}$ (EHRs corpus)	$\text{CancerBERT}_{\text{Cust-Voc\_397}}$ (EHRs corpus)
Hormone receptor type	0.953 (0.957)	0.976 (0.985)	0.979 (0.984)	0.982 (0.987)	0.982 (**0.988**)	0.972 (0.983)	**0.984** (**0.988**)	0.979 (0.985)	0.982 (0.985)
Hormone receptor status	0.856 (0.856)	0.846 (0.846)	0.885 (0.885)	0.859 (0.859)	0.878 (0.878)	0.851 (0.851)	**0.901*** (**0.901***)	0.887 (0.887)	0.891 (0.891)
Tumor size	0.664 (0.709)	0.663 (0.767)	0.781 (0.819)	**0.785** (0.821)	0.727 (0.797)	0.674 (0.684)	0.765 (0.813)	0.784 (0.824)	0.781 (**0.827***)
Tumor site	0.562 (0.771)	0.696 (0.769)	0.711 (0797)	0.749 (0.799)	**0.761*** (0.806)	0.688 (0.762)	0.733 (0.792)	0.715 (0.787)	0.727 (**0.824***)
Cancer grade	0.910 (0.910)	0.857 (0.857)	0.891 (0.891)	0.886 (0.886)	0.856 (0.856)	0.833 (0.833)	0.891 (0.891)	0.898 (0.898)	**0.915*** (**0.915***)
Tumor laterality	0.935 (0.935)	0.926 (0.926)	0.931 (0.931)	0.943 (0.943)	0.948 (0.948)	0.934 (0.934)	0.939 (0.939)	0.947 (0.947)	**0.953*** (**0.953***)
Cancer stage	0.908 (0.908)	0.804 (0.804)	0.870 (0.870)	0.869 (0.869)	**0.909** (**0.909**)	0.907 (0.907)	0.870 (0.870)	0.885 (0.885)	0.898 (0.898)
Histological type	**0.885*** (0.938)	0.823 (0.918)	0.843 (0.922)	0.855 (0.934)	0.850 (0.927)	0.861 (**0.943***)	0.849 (0.922)	0.862 (0.937)	0.862 (0.938)
Macro average	0.834 (0.873)	0.824 (0.859)	0.862 (0.887)	0.868 (0.889)	0.864 (0.888)	0.840 (0.862)	0.867 (0.889)	0.871 (0.896)	**0.876*** (**0.904***)
Micro average	0.876 (0.905)	0.873 (0.907)	0.898 (0.921)	0.904 (0.926)	0.899 (0.923)	0.883 (0.906)	0.903 (0.925)	0.906 (0.930)	**0.909*** (**0.933***)

Note. The scores were averaged scores based on 10 runs. The numbers in bold indicate the highest score.

BERT: bidirectional encoder representations from transformers; BiLSTM: bidirectional long short-term memory; EHRs: electronic health record systems; NER: name entity extraction.

*Indicates statistically higher than other methods (CI: 0.95).

Table 4. Coverage of unique annotated tokens for different BERT vocabularies stated as token count (percentage of total number of unique annotated tokens)

	Total number of unique annotated tokens	Exist in original BERT vocabulary	Exist in customized BERT vocabulary based on frequency	Exist in customized BERT vocabulary based on domain knowledge
Hormone receptor type	33	14 (42.4%)	22 (66.7%)	26 (78.8%)
Hormone receptor status	11	4 (36.4%)	6 (54.5%)	8 (72.7%)
Tumor size	160	62 (38.7%)	62 (38.7%)	62 (38.7%)
Tumor site	146	88 (60.3%)	95 (65.1%)	95 (65.1%)
Cancer grade	20	15 (75.0%)	15 (75.0%)	18 (90.0%)
Tumor laterality	10	4 (40.0%)	6 (60.0%)	8 (80.0%)
Cancer stage	58	12 (20.7%)	18 (31.0%)	52 (89.7%)
Histological type	72	28 (38.9%)	53 (73.6%)	58 (80.6%)
Total	426	178 (41.8%)	227 (53.3%)	274 (64.3%)

BERT: bidirectional encoder representations from transformers.

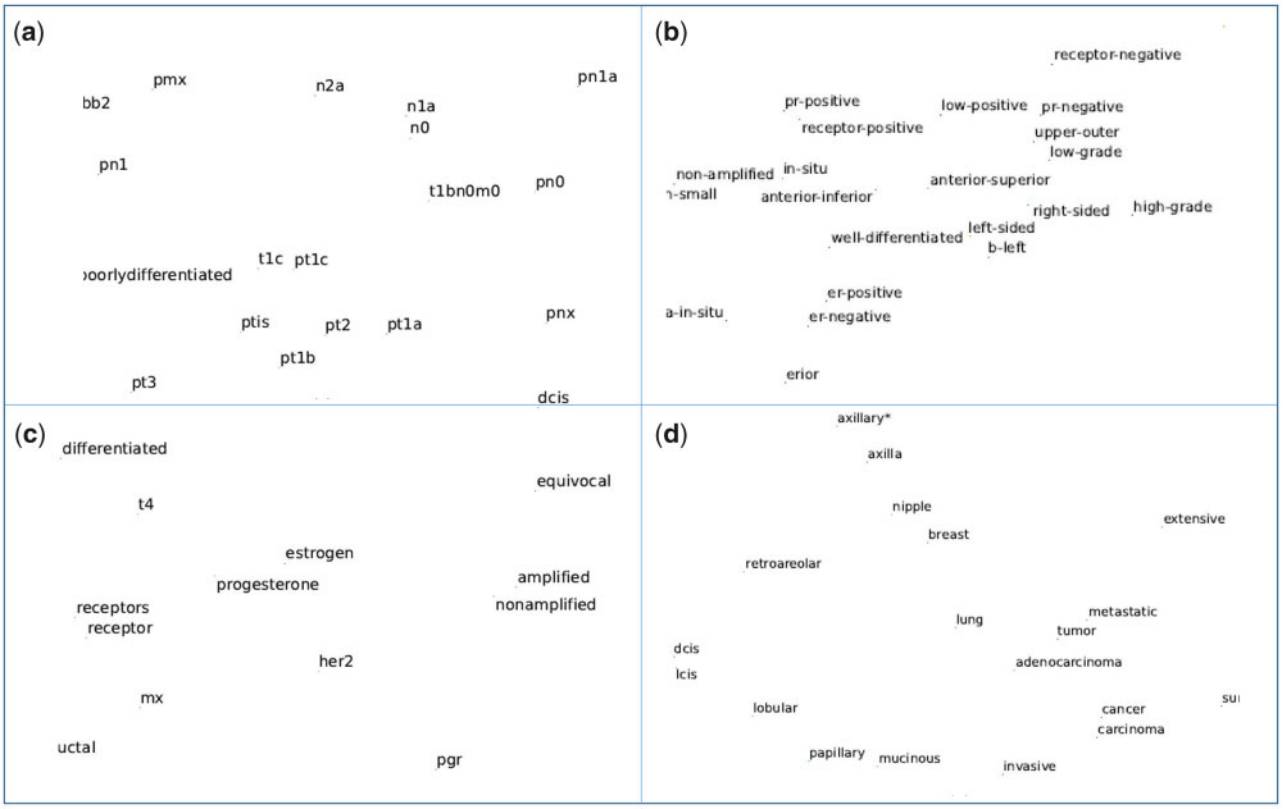

Figure 3. Examples of token clusters in the visualization of word embeddings obtained from $\text{CancerBERT}_{\text{CustVoc\_397}}$ (a and b) and $\text{CancerBERT}_{\text{CustVoc\_997}}$ (c and d) models using t-SNE. BERT: bidirectional encoder representations from transformers; t-SNE: *t*-distributed stochastic neighbor embedding.

plots are the newly added words that were not in the original BERT vocabulary, for example, "ptis", "n2a", and "pn1a" in *Cancer stage;* "er-positive", "er-negative", and "receptor-positive" in *Hormone receptor status*; "left-sided" and "right-sided" in *Tumor laterality*; "estrogen", "progesterone", and "her2" in *Hormone receptor type*, and "dcis" and "lcis" in *Histological type*. The visualization of the entire 426 unique tokens is provided in the Supplementary file.

DISCUSSION

Unstructured EHRs data contain valuable information of patients that can be used for clinical decision support, translational research. In our breast cancer patient corpus, each patient has about 60 pathology reports and over 200 clinical notes. The density of the targeted information is relatively low. As shown in Table 2, all name entities have been annotated for more than 100 cases. The *Hormone receptor type*, *Tumor laterality*, and *Histological type* are the most frequent entities with more than 1000 cases. The language usage for some entities is uniform, only several unique entities exist. For example, for the 1241 *Tumor laterality* cases, most of them are either "left" or "right". It is relatively easy for the NER models to identify those uniform entities. Some entities are various in the clinical texts. For instance, the *Tumor site* and *Tumor size* have 173 and 305 unique expressions, respectively, in the anno-

tated data, which make them relatively difficult to extract for all models.

For the NER task, the BiLSTM + CRF model outperformed BERT-large origin model for several phenotypes. All other BERT models trained on clinical domain corpus significantly outperformed the BERT-large origin model and baseline BiLSTM-CRF models. In this study, all 3 CancerBERT models outperformed the baseline model and other state-of-the-art BERT models. It is within expectation that pretraining BERT models on domain-specific corpus could improve the performance of downstream tasks. In addition, we found that adding the cancer domain specific words to the dictionary of the CancerBERT model can further improve the performance. In this study, we applied 2 methods to add cancer domain-specific words to expand the original BERT vocabulary, adding words based on domain knowledge or frequency. Both methods improve the performance for the cancer phenotype extraction task. Two character-based BERT models (character-BERT and CharBERT)[22,23] were also evaluated and compared with our models for the NER task. The character-based BERT models do not tokenize each word into subwords; instead, they use additional layers (eg, CNN, gated recurrent unit) to represent each word using character-level embeddings to avoid the OOV issue. Though the character-based BERT models may improve the robustness (eg, better handle misspelling issues) compared to the word-based BERT models, they are relatively slower to pretrain and need more computing sources. Our $\text{CancerBERT}_{\text{CustVoc\_397}}$ and $\text{CancerBERT}_{\text{CustVoc\_997}}$ models both significantly outperformed the character-based BERT models for the cancer phenotype extraction task, which indicate the advantage of using domain specific vocabulary for specific downstream tasks compared to the character-based BERT models. Table 4 indicates that both methods to build customized vocabularies could better cover the unique tokens for all name entity categories (except *Tumor size*) compared to the original BERT vocabulary. The domain knowledge-based vocabulary expansion approach covers more tokens compared to the frequency-based expansion approach. Figure 3 visualizes partial word embeddings of the annotated tokens obtained from the $\text{CancerBERT}_{\text{CustVoc\_397}}$ and $\text{CancerBERT}_{\text{CustVoc\_997}}$ models and clear clusters of different cancer name entity categories could be identified from the figure. It indicates that the pretrained CancerBERT models with customized vocabulary could capture the semantics of different name entities.

We also analyzed the prediction errors produced by our $\text{CancerBERT}_{\text{CustVoc\_397}}$ model and found that there are mainly 3 error types. The first is boundary mismatch, which usually happens when the name entity contains multiple tokens, eg, *Tumor size* and *Tumor site*. For example, in a sentence "The tumor measures 2 cm in length and 1 cm in width", the whole *Tumor size* entity is "2 cm in length and 1 cm in width", but our model only captured "2 cm in length". The second type of error is missing (false negative). For example, "mx" should be predicted as *Cancer stage*, but it was predicted as label "O". And the third error type is mixing up the entities. For example, the number "3" could refer to both *Cancer stage* and *Cancer grade*, sometimes the model could not differentiate them.

This study has certain limitations. We trained the CancerBERT models with customized vocabulary using the breast cancer patient narrative corpus extracted from the EHRs. The corpus contains 5.8 million documents (1 billion tokens); however, it was extracted from a single hospital (UMN), the corpus may not be comprehensive enough to reflect all characteristics of clinical narratives. Another limitation is that all the models were only evaluated on our NER task. In the future, we will further improve our model by integrating corpus from other healthcare institutions and evaluate its generalizability. We will also try different methods (eg, using MedSpaCy[33] for tokenization) to generate new words. We plan to annotate more data to evaluate our models on other downstream clinical NLP tasks, such as relation extraction and text classification.

CONCLUSIONS

In this study, a CancerBERT model and its 2 variations with cancer domain vocabulary were developed to extract the 8 breast cancer-related phenotypes from clinical notes and pathology reports in the UMN EHRs. They all outperformed all other existing models; the best model had average macro *F*1 scores of 0.876 for exact match and 0.904 for the lenient match. We also validated that customized vocabulary may further improve the performance of domain specific BERT models in clinical NLP tasks.

FUNDING

This work is partially supported by the National Center for Complementary and Integrative Health (NCCIH) under grant number R01AT009457 (to RZ); and the University of Minnesota Clinical and Translational Science Institute (CTSI), supported by the National Center for Advancing Translational Sciences under grant number UL1TR002494. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

AUTHOR CONTRIBUTIONS

SZ conducted the main experiments. NW, SZ, and LW participated in the data collection and annotation. HL and RZ guided the study design, data collection, and analysis. All authors participated in the writing of the manuscript and critical revisions of the manuscript for important intellectual content.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

The authors thank Dr Rubina Rizvi for guiding the annotation process, and thank Ms Lindsay B. Nichols for proofreading the draft.

COFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The data underlying this article cannot be shared publicly due to the privacy of patient health information. The pretrained CancerBERT models will be available at https://github.com/zhang-informatics/CancerBERT after the UMN approval.

REFERENCES

1. DeSantis C, Ma J, Goding Sauer A, Newman L, Jemal A. Breast cancer statistics, 2017, racial disparity in mortality by state. *CA Cancer J Clin* 2017; 67 (6): 439–48.
2. Arnedos M, Vicier C, Loi S, *et al*. Precision medicine for metastatic breast cancer – limitations and solutions. *Nat Rev Clin Oncol* 2015; 12 (12): 693–704.
3. Carels N, Spinassé L, Tilli T, Tuszynski JA. Toward precision medicine of breast cancer. *Theor Biol Med Model* 2016; 13 (1): 7–46.
4. Bettaieb A, Paul C, Plenchette S, Shan J, Chouchane L, Ghiringhelli F. Precision medicine in breast cancer: reality or utopia? *J Transl Med* 2017; 15 (1): 1–3.
5. Carroll R, Thompson W, Eyler A, *et al*. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J Am Med Inform Assoc* 2012; 19 (e1): e162–9.
6. Breitenstein M, Liu H, Maxwell K, Pathak J, Zhang R. Electronic health record phenotypes for precision medicine: perspectives and caveats from treatment of breast cancer at a single institution. *Clin Transl Sci* 2018; 11 (1): 85–92.
7. Zhou S. Extracting phenotypes of cancer patients from electronic health records. In: 2019 IEEE international conference on healthcare informatics (ICHI); June 10, 2019; IEEE, Xi'an China; pp. 1–2.
8. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
9. Wei Q, Ji Z, Li Z, *et al*. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *J Am Med Inform Assoc* 2020; 27 (1): 13–21.
10. Wang L, Luo L, Wang Y, Wampfler J, Yang P, Liu H. Natural language processing for populating lung cancer clinical research data. *BMC Med Inform Decis Mak* 2019; 19 (5): 239.
11. Nguyen A, Moore J, O'Dwyer J, Philpot S. Assessing the utility of automatic cancer registry notifications data extraction from free-text pathology reports. *AMIA Annu Symp Proc* 2015; 2015: 953.
12. Yala A, Barzilay R, Salama L, *et al*. Using machine learning to parse breast pathology reports. *Breast Cancer Res Treat* 2017; 161 (2): 203–11.
13. Savova GK, Tseytlin E, Finan S, *et al*. DeepPhe: a natural language processing system for extracting cancer phenotypes from clinical records. *Cancer Res* 2017; 77 (21): e115–8.
14. Qiu J, Yoon H, Fearn P, Tourassi G. Deep learning for automated extraction of primary sites from cancer pathology reports. *IEEE J Biomed Health Inform* 2018; 22 (1): 244–51.
15. Alawad M, Yoon H, Tourassi G. Coarse-to-fine multi-task training of convolutional neural networks for automated information extraction from cancer pathology reports. *IEEE EMBS Int Conf Biomed Health Inform BHI* 2018: 218–21.
16. 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.
17. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*, 2019.
18. Gu Y, Tinn R, Cheng H, *et al*. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*, July 31, 2020.
19. Du J, Xiang Y, Sankaranarayanapillai M, *et al*. Extracting postmarketing adverse events from safety reports in the vaccine adverse event reporting system (VAERS) using deep learning. *J Am Med Inform Assoc* 2021; 28 (7): 1393–400.
20. Fan Y, Zhou S, Li Y, Zhang R. Deep learning approaches for extracting adverse events and indications of dietary supplements from clinical text. *J Am Med Inform Assoc* 2021; 28 (3): 569–77.
21. Zhang X, Zhang Y, Zhang Q, *et al*. Extracting comprehensive clinical information for breast cancer using deep learning methods. *Int J Med Inform* 2019; 132: 103985.
22. Ma W, Cui Y, Si C, Liu T, Wang S, Hu G. CharBERT: character-aware pre-trained language model. *arXiv preprint arXiv:2011.01513*, November 3, 2020.
23. Boukkouri H, Ferret O, Lavergne T, Noji H, Zweigenbaum P, Tsujii J. CharacterBERT: reconciling ELMo and BERT for word-level open-vocabulary representations from characters. *arXiv preprint arXiv:2010.10392*, October 20, 2020.
24. Beltagy I, Lo K, Cohan A. Scibert: a pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, March 26, 2019.
25. Klie J, Bugert M, Boullosa B, Eckart de Castilho R, Gurevych I. The inception platform: machine-assisted and knowledge-oriented interactive annotation. In: proceedings of the 27th international conference on computational linguistics: system demonstrations; August 20, 2018; Santa Fe, New Mexico; pp. 5–9.
26. Wu Y, Schuster M, Chen Z, *et al*. Google's neural machine translation system: bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.
27. Honnibal M, Montani I, Van Landeghem S, Boyd A. spaCy: industrial-strength natural language processing in python. Zenodo, 2020.
28. Ritter A, Clark S, Etzioni O. Named entity recognition in tweets: an experimental study. In: proceedings of the 2011 conference on empirical methods in natural language processing; July 27, 2011; Edinburgh, UK; pp. 1524–34.
29. Yang X, Bian J, Hogan W, Wu Y. Clinical concept extraction using transformers. *J Am Med Inform Assoc* 2020; 27 (12): 1935–42.
30. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems*. Stroudsburg, PA: Association for Computational Linguistics; 2013: 3111–9.
31. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); October 25, 2014; Doha, Qatar; pp. 1532–43.
32. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008; 9 (11): 2579–605.
33. Eyre H, Chapman AB, Peterson KS, *et al*. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. *arXiv preprint arXiv:2106.07799*, June 14, 2021.