Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220403

Pretrained Neural Networks Accurately Identify Cancer Recurrence in Medical Record

Hussam Kaka [a], George Michalopoulos [a], Sujan Subendran [a], Kathleen Decker [b], Pascal Lambert [b], Marshall Pitz [b], Harminder Singh [b], Helen Chen [a,1]

[a] *University of Waterloo*
[b] *CancerCare Manitoba*

Abstract. Cancer recurrence is the diagnosis of a second clinical episode of cancer after the first was considered cured. Identifying patients who had experienced cancer recurrence is an important task as it can be used to compare treatment effectiveness, measure recurrence-free survival, and plan and prioritize cancer control resources. We developed BERT-based natural language processing (NLP) contextual models for identifying cancer recurrence incidence and the recurrence time based on the records in progress notes. Using two datasets containing breast and colorectal cancer patients, we demonstrated the advantage of the contextual models over the traditional NLP models by overcoming the laborious and often unscalable tasks of composing keywords in a specific disease domain.

Keywords. Natural language processing, Real-world data, Breast cancer, Colorectal cancer, Cancer recurrence, BERT, ClinicalBioBert architecture

1. Introduction

Breast and colorectal cancer are two of the most common cancer types globally. They are the second and third leading causes of cancer death in North America [4]. Both diseases carry a risk of recurrence after treatment, and recurrent cases need to be flagged expediently for further treatment to increase the chances of survival. Cancer patients tend to have many interactions with clinicians throughout their treatment, resulting in many unstructured medical records that are difficult to search. Consequently, quality control interventions such as identifying patients with recurrence who are not yet on a treatment plan require manual chart abstraction, which is time-consuming and sometimes inaccurate [6]. Moreover, retrieval of information on recurrence becomes difficult during clinical encounters, potentially affecting the patient experience.

Previous approaches for cancer recurrence detection include using treatment codes, such as those for chemotherapy or radiation, detecting breast and lung cancer recurrence [8]. Still, treatment codes can vary by jurisdiction and over time. Other approaches in-

[1]Corresponding Author, Helen Chen, University of Waterloo, School of Public Health Sciences; E-mail: helen.chen@uwaterloo.ca

clude generating a vocabulary of terms specific to the cancer of interest by specialized oncologists, which can then be used for parsing the notes and recurrence detection [5].

In the past decade, the field of natural language processing, which studies information extraction from text, has undergone significant advances, culminating in the development of robust transformer-based models [12]. These models consist of large neural networks that are "pre-trained" on large datasets and then transferred to the task at hand. Bidirectional Encoder Representations from Transformers (BERT)[7] is a transformer-based model which has shown excellent results in many language tasks, including classification and named entity recognition. It is pre-trained on general-purpose English text, and the resulting pre-trained model, BERT-base, can be subsequently used to detect cancer recurrence. A variation of this, ClinicalBioBERT, was trained on medical data to improve its performance on medical tasks [1]. Other published models include BioBERT [10], which is trained on PubMed abstracts and full texts, and UmlsBERT [11], which incorporates information from the Unified Medical Language System (UMLS) [3].

In this study, two models named BERT-base and ClinicalBioBERT, are used to detect cancer recurrence on unstructured medical notes. They are trained and evaluated for the detection of breast cancer recurrence and colorectal cancer recurrence without the need for expert input in the form of specialized vocabulary or decision rules.

2. Material

In this study, we used two cancer datasets (breast and colorectal) which were curated from electronic medical record notes at Cancer Care Manitoba.

Each of these datasets was split to an internal and an external set. The internal dataset was used for hyperparameter tuning and model training, and it consisted of all notes dated between 2004 and 2007. The external dataset was used for final model evaluation, which consisted of notes generated between 2008 and 2012.

Both breast and colorectal cancer datasets demonstrated a skew towards negative instances. Of the 112,285 notes in the breast cancer internal dataset, only 5,082 (4.3%) were positive for cancer recurrence. Of the 116,146 notes in the colorectal cancer internal dataset, only 4,207 (3.6%) were positive. We observed similar trends for the external breast and colorectal cancer datasets. Full dataset characteristics are shown in Table 1.

	Breast Cancer		Colorectal Cancer	
	Internal	External	Internal	External
N (notes)	112,285	78,469	116,146	122,262
Recurrent Cancer	5,082	2,985	4,207	4,245
# patients	897	615	536	589

Table 1. statistics of the datasets

3. Method

We used two pre-training models: BERT-base and ClinicalBioBert. They are identical in network architecture but differ in their pre-training datasets. BERT-base is pre-trained on

general English texts (Books Corpus and English Wikipedia), while ClinicalBioBERT is further trained using biomedical texts (PubMed Abstracts and PubMed Central Papers) and clinical texts (the MIMC-III dataset [9]). Therefore, ClinicalBioBert, with medical context, serves as the target model and BERT-base, with general English pre-training, serves as the control.

BERT models convert input text into tokens before processing. The tokenization process converts input words into a numeric representation based on a provided dictionary. Since the dictionaries are limited in size (approximately 30,000 words in BERT-base), the tokenizer will inevitably encounter dictionary words that are split into subcomponents. Thus, a single word can be represented by one or more tokens. The vocabulary used in this study consisted of 28,996 words and was originally published by Devlin et al. [7] and subsequently used by ClinicalBioBERT. The attention mechanism used by BERT exhibits quadratic runtime complexity with respect to the input length, necessitating a limit to input length. Pre-trained BERT-base and ClinicalBioBERT models used in this study set the input length limit to 512 tokens. Since medical notes can easily exceed this length, a mechanism to shorten notes is required for these models to be usable. We hypothesized that the most important information in medical notes would be included at the beginning or the end of a note. Thus the input document of the model consists of the first 256 and last 256 tokens of each note.

Each breast cancer and the colorectal cancer datasets were divided 90%-10% for training and validation, respectively. A grid search was performed on the learning rate and batch size with 5-fold cross-validation to obtain the optimal hyperparameters for the model. Binary cross-entropy with class weights corresponding to dataset prevalence was used as the loss function. The best-performing model hyperparameters were then selected and used to train a final model for each dataset using 100% of the data. The final models were then evaluated on the external datasets. The selection of the optimal hyperparameters was based on the area under the receiver operating characteristic curve (ROC-AUC).

It should be noted that independent dataset metrics were used for model evaluation. These metrics were: (i) ROC-AUC (ii) sensitivity (iii) specificity (iv) the modified Brier score. Dataset-dependent metrics were used for the assessment of model suitability for clinical use. In this setting, the unbalanced accuracy, positive predictive value (PPV) and the negative predictive value (NPV) were used.

4. Results-Discussion and Conclusion

The results of the 5-fold cross-validation are shown in Table 2. As seen by the ROC-AUC results, ClinicalBioBERT and BERT-base showed very close results, within one standard deviation of each other.

	Breast Cancer Dataset	Colorectal Cancer Datase
ClinicalBioBERT	0.9955 ± 0.0006	0.9921 ± 0.0074
BERT-base	0.9948 ± 0.0014	0.9912 ± 0.0091

Table 2. ROC-AUC values after 5-fold cross validation on the training dataset. Mean and standard deviation of ROC-AUC values are reported

Finally, both models were trained on the entire internal dataset and evaluated on the external dataset. It can be observed in Table 3 that for the breast cancer dataset, the results on the external dataset closely matched those in the internal dataset, with only a minor reduction in the ROC-AUC to 0.9892 for ClinicalBioBERT and 0.9883 for BERT-base. Using a threshold cutoff of 0.01 and BERT-base, we can estimate that for a randomly selected sample of 1,000 notes, only 54 would be flagged as positive and would require manual review. This is a 94.6% reduction in the volume of notes requiring manual review, at the expense of only 3 missed recurrences. Furthermore, on the colorectal cancer dataset, it can be observed that ClinicalBioBERT had a small edge as it achieved a ROC-AUC of 0.9810. Using a threshold cutoff of 0.01 and ClinicalBioBERT, we estimate that for a randomly selected sample of 1,000 notes, only 52 would be flagged as positive, requiring manual review, which is a 94.8% reduction. This would be at the expense of only 7 missed recurrences.

			Cut-off: 0.01		Cut-off: 0.5	
	ROC-AUC	Scl Br	Sn (PPV)	Sp (NPV)	Sn (PPV)	Sp (NPV)
Breast Cancer						
ClinicalBioBERT	.9892	.419	.926 (.592)	.981 (.998)	.863 (.659)	.987 (.996)
BERT-base	.9883	.377	.929 (.629)	.985(.996)	.863 (.639)	.986 (.996)
Colorectal Cancer						
ClinicalBioBERT	.9810	.251	.806 (.544)	.976 (.991)	.619 (.616)	.986 (.986)
BERT-base	.9694	.219	.751 (.531)	.976 (.991)	.577 (.612)	.987 (.985)

Table 3. Result values at multiple cutoffs on the breast cancer and colorectal cancer datasets. Scl Br: Scaled Brier Score. Sn: sensitivity. Sp: specificity. PPV: positive predictive value. NPV: negative predictive value

These results demonstrate that pre-trained transformer models can perform exceptionally well on detecting cancer recurrence from electronic medical record notes. These findings exceed previously reported results on cancer recurrence detection using classical machine learning methods [5], and earlier neural network architectures [2]. Results between BERT-base and ClinicalBioBERT showed only minor differences. In the internal datasets where cross-validation could be performed, the results were only a standard deviation away from each other. This suggests that the differences may not be statistically significant. The significance of this may lie in the fact that the dataset is of sufficient size for BERT to learn vocabulary associations with cancer recurrence and thus may take advantage of medical text pretraining less important.

A major advantage of the BERT approach is the model's ability to extract useful information without the need for expert knowledge. Earlier approaches using more traditional machine learning techniques required in-domain vocabulary compiled by oncologists [5], which is a time-consuming process. Moreover, since vocabulary and abbreviations may be local and institutional, BERT's approach eliminates the need for each institution to create its own vocabulary. These results pave the way for the use of this software in clinical work. Depending on clinical needs, a threshold cutoff can be selected to match the requirements. For example, to create automated tracking of the incidence of cancer recurrence, a lower threshold cutoff, such as 0.01, could be selected to achieve high sensitivity and specificity. Alternatively, to create a screening tool for patients with recurrence, generating curated lists that a human expert can then review, a higher threshold value closer to 0.5 would be selected to maximize specificity and minimize the likelihood of false negatives.

However, the main limitations of this study are that all data are obtained from the same institution, which may affect the generalizability of the trained model at other institutions. Moreover, it is possible that some cancer patients had notes spanning the periods of both the training and validation datasets, thereby causing similarities in some of their notes. Finally, the datasets were skewed towards negative occurrences, as is common in many medical datasets, and this resulted in a model with excellent specificity but relatively lower sensitivity. In conclusion, transformer-based models, such as BERT and its variants, can achieve excellent sensitivity and specificity on the task of detecting cancer recurrence from medical record notes. These models do not require rules or specialized knowledge from domain experts, yet nevertheless outperform earlier machine learning methods.

References

[1] Alsentzer E, Murphy J, Boag W, Weng W-H, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop [Internet]. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019. p. 72–8. Available from: https://www.aclweb.org/anthology/W19-1909

[2] Banerjee I, Bozkurt S, Caswell-Jin JL, Kurian AW, Rubin DL. Natural Language Processing Approaches to Detect the Timeline of Metastatic Recurrence of Breast Cancer. JCO Clin Cancer Inform. 2019 Dec;(3):1–12.

[3] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(90001):267D – 270.

[4] Brenner DR, Weir HK, Demers AA, Ellison LF, Louzado C, Shaw A, et al. Projected estimates of cancer in Canada in 2020. Can Med Assoc J. 2020 Mar 2;192(9):E199–205.

[5] Carrell DS, Halgrim S, Tran D-T, Buist DSM, Chubak J, Chapman WW, et al. Using Natural Language Processing to Improve Efficiency of Manual Chart Abstraction in Research: The Case of Breast Cancer Recurrence. Am J Epidemiol. 2014 Mar 15;179(6):749–58.

[6] Denny JC, Peterson JF, Choma NN, Xu H, Miller RA, Bastarache L, et al. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. J Am Med Inform Assoc. 2010 Jul;17(4):383–8.

[7] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv181004805 Cs [Internet]. 2019 May 24 [cited 2020 Apr 16]; Available from: http://arxiv.org/abs/1810.04805

[8] Hassett MJ, Uno H, Cronin AM, Carroll NM, Hornbrook MC, Ritzwoller D. Detecting Lung and Colorectal Cancer Recurrence Using Structured Clinical/Administrative Data to Enable Outcomes Research and Population Health Management. Med Care. 2017 Dec;55(12):e88–98.

[9] Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3(1):160035.

[10] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019 Sep 10;btz682.

[11] Michalopoulos G, Wang Y, Kaka H, Chen H, Wong A. UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus. NAACL 2021.

[12] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. ArXiv170603762 Cs [Internet]. 2017 Dec 5 [cited 2021 May 6]; Available from: http://arxiv.org/abs/1706.03762