Text Classification of Cancer Clinical Trial Eligibility Criteria

Yumeng Yang, MS[1], Soumya Jayaraj, BAT[1], Ethan Ludmir, MD[2 3], Kirk Roberts, PhD[1]
[1]School of Biomedical Informatics
The University of Texas Health Science Center at Houston, Houston, TX, USA
[2]Department of Radiation Oncology
The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[3]Department of Biostatistics
The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Abstract

Automatic identification of clinical trials for which a patient is eligible is complicated by the fact that trial eligibility are stated in natural language. A potential solution to this problem is to employ text classification methods for common types of eligibility criteria. In this study, we focus on seven common exclusion criteria in cancer trials: prior malignancy, human immunodeficiency virus, hepatitis B, hepatitis C, psychiatric illness, drug/substance abuse, and autoimmune illness. Our dataset consists of 764 phase III cancer trials with these exclusions annotated at the trial level. We experiment with common transformer models as well as a new pre-trained clinical trial BERT model. Our results demonstrate the feasibility of automatically classifying common exclusion criteria. Additionally, we demonstrate the value of a pre-trained language model specifically for clinical trials, which yield the highest average performance across all criteria.

1 Introduction

Cancer has a high morbidity and mortality rate and threatens millions of people's lives. According to the American Cancer Society report, there will be a total of 1.9 million new cancer cases alone with more than 600,000 deaths in the US in 2022[1]. Clinical trials have always been recognized as significant for cancer treatment and anticancer drug development[2], but trial recruitment is still problematic, with incomplete accrual as the leading reason for non-informative trial closure/failure[2,3,4]. Barriers to trial accrual are multifactorial, including patient and provider concerns and biases, as well as availability of trials across different centers[5,6,7]. Critically, matching potentially eligible patients to relevant clinical trials is a key barrier to clinical trial accrual, as patients seeking trial options aim to identify candidate trials for which they are eligible. Eligibility criteria for clinical trials define which patients are eligible for a given study; however, trial eligibility criteria are often written in jargon and are often difficult for lay audiences to interpret, understand, and apply[8].

Eligibility criteria are a key component of any clinical trial or study protocol, as they define the requirements for participation, and indeed define the study population. Conventionally, eligibility criteria include both inclusion and exclusion criteria[9]. Eligibility criteria are generally listed in clinical trial registries as well; ClinicalTrials.gov is a public database that provides information on registered human trials. It is managed by the United States National Library of Medicine (NLM)[10], and federal mandates require that human trials in the US are registered with trial information in ClinicalTrials.gov[11]. As of February 2023, approximately half a million studies were registered in ClinicalTrials.gov, making it the largest available trials registry worldwide. Such a comprehensive database brings enormous potential for research, including identifies high-level trends in trial structure, evaluation of disparities[12,13], and developing tools to help facilitate trial recruitment.

Identifying eligible patients for these trials can be a time-consuming and challenging task, partly due to the non-standardized format of eligibility criteria. As a result, patients and clinicians may struggle to identify relevant trials, leading to potential delays in the accrual process. [14]. This underscores the urgent need for automated tools, such as text classification, to streamline and enhance the recruitment process. By developing an automatic classifier to identify key eligibility criteria from free-text records, we can potentially address the unmet needs of patients and clinicians by facilitating faster and more accurate identification of relevant clinical trials. Natural Language Processing (NLP) aims to enable machine understanding of human language. In the clinical field, NLP has a wide range of applications, including name entity recognition[15,16,17], text mining[18,19,20], and text classification[21,22,23]. NLP models can help extract and structure information from text data, raising from patients' clinical notes to trial eligibility criteria. Our project aims to develop classifiers that can automatically identify key exclusion criteria from eligibility descriptions listed on ClinicalTrials.gov. Such classification tools have the potential to streamline the accrual process for both clinicians and patients. For our study, we chose seven key eligibility criteria and used five state-of-the-art domain-specific language models trained on our data. Additionally, we pre-trained our own model based on ClinicalBERT by using approximately half a million eligibility criteria sections derived from ClinicalTrials.gov.

2 Related work

Numerous prior works have focused on text mining of eligibility criteria for clinical trials for the purpose of streamlining the recruitment process for both patients and clinicians. Criteria2Query is a NLP tool aiming to convert free-text eligibility criteria for a clinical trial to a structured query to aid in the identification of eligible patients from clinical data[24]. DQueST is a dynamic questionnaire to help find eligible clinical trials by asking trial criteria related questions[25]. There also exists many models to automatically structure eligibility criteria for clinical trials[26] and EHR data[27]. RuleEd is a web-based tool to revise and refine free text eligibility criteria[28]. EXTRACTS is a search tool that allows users to customize criteria and weight each criteria for potential trials[29].

Another aspect of emphasis is information extraction combined with machine learning techniques to match terms from eligibility criteria and patient records. Some tools used regular expression and machine learning models to identify certain criteria for cancer trials[30], and to satisfy specific department needs[31]. Another tool was developed to identify eligible trials using key terms and patterns matching[32]. Many other tools were developed based upon EHR data to further identify and match eligible patients from their medical records for specific diseases, such as cancer and Alzheimer's disease[33,34,35] with clinical trial criteria.

Along with texting mining and information extraction, some works aim to create a knowledge base of common eligibility criteria related annotated corpus and build knowledge base. EliIE[36] contains 230 trials from various phases with focus on Alzheimer's disease specifically, Chia[37] contains 1000 phase IV trials covering all diseases, and The Leaf Clinical Trials Corpus[38] contains 1006 trials cross all phases and diseases. A lexicon base for breast cancer clinical trial eligibility criteria was created to identify concepts related to eligibility[39]. This study shows that a specified lexicon can improve the accuracy of subjects in clinical trial eligibility criteria analysis. Another knowledge base used a hierarchical taxonomy to classify criteria into multiple categories, including disease, intervention, and condition.[40].

3 Method

Our dataset is based on PROTECTOR1, a manually-annotated database of 764 Phase III cancer trials collected from ClinicalTrials.gov covering the years 2000-2017. PROTECTOR1 has been used to analyze many aspects of cancer trials, including evaluate the relationship between sponsorship types with trial accural[41], measure impacts and trail-related factors for these exclude patients with brain metastases in cancer trials[42], and ascertain the transparency regarding cancer trials results reports[43]. The database was originally developed for manual analysis, which resulted in several important design decisions that will be described in detail later in Section 3.1 (notably, criteria were annotated at the trial level not the criterion level). Each trial was initially annotated by clinicians followed a two-person blinded annotation paradigm. In this study, we mainly focus on automatically classifying seven key exclusions for cancer trials, chosen based on their frequency of occurrence as well as clinical significance. The selected exclusions (with the abbreviations used throughout this paper) are:

- prior malignancy (Prior): exclude patients with a previous cancer history
- human immunodeficiency virus (HIV): exclude HIV/AIDS positive patients, including those with well-controlled HIV
- hepatitis B virus (HBV): exclude patients with a history of hepatitis B infection
- hepatitis C virus (HCV): exclude patients with a history of hepatitis C infection
- psychiatric illness (Psych): exclude patients with a history of a major psychiatric / mental health disorder
- substance abuse (Subst): exclude patients with a history of substance abuse, including drug and/or alcohol abuse

- autoimmune disease (Auto): exclude patients with a chronic autoimmune disease (e.g., lupus, rheumatoid arthritis, scleroderma)

These clinically-relevant criteria may facilitate selection of a more homogeneous study population for a given trial. However, some controversy surrounds application or misapplication of some of these criteria, which may inappropriately generate disparities by excluding specific populations of patients whose comorbid conditions may not be germaine to the intervention assessed in a given trial.

For the 764 trials in the dataset, eligibility criteria section was obtained from ClinicalTrials.gov for each trial and was further divided into individual criterion. This was done for two main reasons. First, most BERT-based models have a maximum token input limit of 512, and the original token length in our data ranged from 0 to 2355. Therefore, the eligibility criteria descriptions were split into individual criterion (which are almost always separated using numbered or unordered lists) to ensure that each criterion fits within this limit. Second, this approach helped to remove noisy data from the eligibility criteria descriptions using the keyword approach described below.

The eligibility section is often, but not always divided into inclusion criteria (a patient must meet all of these) and exclusion criteria (a patient must meet none of these). Sometimes these are not separated and a single nonspecific list of eligibility criteria are provided, but these are clear from reading the text. Importantly, an exclusion is not necessarily only stated in the exclusion criteria section. Here are some examples,

- NCT00095875: Inclusion: "No other malignancy within the past 5 years except adequately treated carcinoma in situ of the cervix, basal cell or squamous cell skin cancer, or other cancer curatively treated by surgery alone"
- NCT00057876: Exclusion: "Malignancy within the past 5 years except nonmelanoma skin cancer, carcinoma in situ of the cervix, or organ-confined prostate cancer (Gleason score no greater than 7)"
- NCT00048997: Eligibility: "No other malignancy within the past 3 years except nonmelanoma skin cancer"

All of these trials excluded patients with prior malignancy within a specific time frame, but using different terms under different subdivision. We thus use all criteria in the eligibility criteria section, but prepared each criterion with an indicator of the section if came from ("inclusion", "exclusion", "eligibility" for non-specific section) to provide context. Table 4 shows the sample text input for the classification model.

3.1 Keyword Filtering

Our dataset was annotated at the trial level, not the individual criterion level. However, it is the individual criterion that conveys the semantic constraint of the exclusion, so it would make the most sense to focus the classification at the individual criterion level. In order to accurately find the specific criterion that conveys the given exclusion condition, we created lists of keywords for all 7 exclusion types. This allows us to convert the task of binary classification at the trial level to binary classification at the criterion level by only classifying criterion that contain one of the selected keywords. Our goal in creating the keyword lists was recall: keywords alone are insufficient to classify a criterion according to each of the targeted exclusions, a downstream classifier (described later) will perform the binary classification. However, if a criterion that specifies a targeted exclusion were not to contain one of the specified keywords, then the downstream classifier takes a hit in terms of recall. Table 1 provides the list of keywords used for each exclusion.

While in theory it could be problematic to assume any criterion containing one of the above keywords is positive for the exclusion if the trial as a whole is positive for the exclusion, there is a more important limitation to consider. Another important feature of PROTECTOR1 is that the annotated exclusions were not based only on the clinical trial description. There were three primary sources the annotators consulted for whether a given exclusion applied to a clinical trial: (1) the description on ClinicalTrials.gov, (2) the original clinical trial protocol, (3) any publications associated with the trial. Ideally (and ethically), the eligibility criteria across all three of these would be consistent enough that any inclusion/exclusion stated in the protocol or trial would also be present on ClinicalTrials.gov. In practice, this is not the case, unfortunately, and PROTECTOR1 does not specify the source(s) of information for the

Table 1: Keywords for each criteria

Criteria	Keywords
Prior	prior malignancy, concurrent malignancy, prior invasive malignancy, other malignancy, known additional malignancy, squamous cell carcinoma, in-situ, cancer, 3 years, 5 years, five years
HIV	human immunodeficiency virus, acquired immunodeficiency syndrome, AIDS-defining malignancy, hiv, AIDS-related illness
HBV	hbv, hepatitis
HCV	hcv, hepatitis
Psych	psychosis, depression, psychiatric, psychological, psychologic, nervous, mental illness, mental disease
Subst	ethanol, abuse, alcohol, alcoholism, illicit substance, drug, drugs, medical marijuana, inadequate liver, illicit substance, addictive, substance misuse, cannabinoids, chronic alcoholism
Auto	uncontrolled systemic, autoimmune

Table 2: Criterion-level annotation summary

	Prior	HIV	HBV	HCV	Psych	Subst	Auto
Sample Size	529	200	130	282	281	523	54
Cohen's κ	0.95	0.74	0.16	0.89	0.93	0.98	0.22
Agreement Accuracy	0.99	0.96	0.85	0.95	0.97	0.99	0.89

exclusion annotation. Since our method is designed to work only on the description from ClinicalTrials.gov, manual annotation specific to this project became necessary, as described in the next sub-section. This did, however, at the same time solve the first problem and allow us to have a criterion-specific label for each of the exclusions.

3.2 Annotation

For each exclusion type, all criteria in the 764 trials that matched one of the associated keywords was annotated independently by two annotators using a double-blind paradigm (YY and SJ). The annotation rule for each criterion used the same standard as the original trial level annotation. All discrepancies were resolved through discussion and consensus, including the involvement of a subject-matter expert and curator of the PROTECTOR1 database (EL). A descriptive summary (sample size and annotation agreement) of the annotated dataset is shown in Table 2. Due to the small number of samples, the annotation proceeded in a single phase such that consistent disagreements (e.g., what level of granularity counts as an autoimmune disease) were not resolved until the end. This resulted in low agreement in HBV and Auto (the latter of which was also impacted by its low prevalence, and such imbalance skews the κ statistic). However, during reconciliation and consultation with the subject-matter expert, these disagreements were easily clarified, leading to a better gold standard than the agreement numbers suggest. The other exclusion types, meanwhile, had excellent levels of agreement. Table 3 shows some examples of annotated criterion for some criteria.

3.3 Machine Learning

We applied six BERT-based models on all exclusions. Five of the models are pre-existing BERT models, pre-trained on domain specific corpus:

1. **BioBERT**[44]: the original BERT model further pre-trained on PubMed abstracts and PMC full-text articles.
2. **ClinicalBERT**[45]: the BioBERT model further pre-trained on MIMIC-III[46] notes.
3. **BlueBERT**[47]: the original BERT model further pre-trained on PubMed abstracts and MIMIC-III clinical notes.
4. **PubMedBERT**[48]: a from-scratch BERT model pre-trained on PubMed abstracts (notably the from-scratch nature allowed for it to use a domain-specific word piece model).

Table 3: Examples of annotated criteria

ClinicalTrials.gov ID	Criterion Text	Classification
NCT00005047	eligibility: At least 5 years since other prior systemic chemotherapy	0: Prior not excluded
NCT00216060	exclusion: No prior history of malignancy in the past 5 years with the exception of basal cell and squamous cell carcinoma of the skin	1: Prior excluded
NCT00075803	exclusion: HIV positive	1: HIV excluded
NCT00114101	inclusion:Patients must not be human immunodeficiency virus (HIV), hepatitis B surface antigen (HBSag), or hepatitis (Hep) C positive	1: HBV / HCV /HIV excluded
NCT00262067	exclusion:Known brain or other central nervous system (CNS) metastases	0: Psych not excluded
NCT00022516	eligibility:No psychiatric or addictive disorders that would preclude study	1: Psych excluded

Table 4: Keyword performance metrics for all exclusion types

	Prior	HIV	HBV	HCV	Psych	Subst	Auto
Precision	0.87	0.90	0.98	0.96	0.68	0.27	0.62
Accuracy	0.82	0.88	0.74	0.95	0.67	0.27	0.57
Recall	0.98	0.97	0.98	1	0.99	1	0.89

5. **SciBERT**[49]: the original BERT model further pre-trained on 1.14M full-text papers from Semantic Scholar (which mainly focuses on computer scienice and biomedicine).

Additionally, since no pre-trained model specific to clinical trial descriptions exist, we further pre-trained the ClinicalBERT model using 442,370 eligibility criteria sections from ClinicalTrials.gov. We used a batch size of 64, a maximum sequence length of 512, and a learning rate of 2e-05. The model trained on all available text for 10,000 steps, and the masked language model probability = 0.15. We denote this model as **ClinicalTrialBERT**. We plan to release the ClinicalTrialBERT model on huggingface upon acceptance for others in the community to use.

3.4 Evaluation

Due to the small sample size (as shown in Table 2), we evaluate using 5-fold cross validation. To avoid data leakage, the fold splitting was performed at the trial level (as opposed to the criterion level) such that a single trial with multiple matching criteria does not end up in both the training and testing set for any iteration of cross validation.

We evaluate all classification models using precision, recall, and F1 metrics. Each metric was evaluated on both the criterion level (how well the model does at predicting each criterion) and at the trial level (how well the model does at predicting each trial, assuming that a single positive criterion means the trial is positive for that exclusion).

4 Results

4.1 Keyword Filtering

Table 4 shows the summary of performance metrics for all exclusion types. We conducted error analysis for psychiatric illness (Psych), substance abuse (subst), and autoimmune disease (Auto), as these have lower overall precision.

In Psych trials, we note the high frequency of keywords “psychiatric” (n = 124) and “nervous” (n = 96). The precision for “psychiatric” is 0.92, while the precision for “nervous” is 0.35. However, removing “nervous” would cause recall to drop from 0.99 to 0.90, so this keyword is left in for the downstream machine learning classifier to disambiguate. In Subst trials, the highest frequency keywords are “drug” (n = 269) and “drugs” (n = 143). These keywords, however,

Table 5: Evaluation Results of BERT-Based Models Across All Criteria

	Prior						HIV					
	Criterion Level			Trial Level			Criterion Level			Trial Level		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BioBERT	0.89	0.89	0.89	0.93	0.93	**0.93**	0.97	0.97	0.96	0.97	0.97	0.96
ClinicalBERT	0.90	0.90	0.90	0.89	0.89	0.89	0.97	0.97	0.96	0.97	0.97	0.96
PubMedBERT	0.91	0.91	**0.91**	0.92	0.91	0.91	0.97	0.97	0.96	0.97	0.97	0.96
BlueBERT	0.87	0.87	0.87	0.87	0.86	0.87	0.98	0.98	**0.98**	0.98	0.98	**0.98**
SciBERT	0.86	0.86	0.86	0.89	0.89	0.89	0.97	0.97	0.96	0.97	0.97	0.96
ClinicalTrialBERT	0.91	0.91	**0.91**	0.91	0.91	0.91	0.97	0.97	0.96	0.97	0.97	0.96

	Psych						HBV					
	Criterion Level			Trial Level			Criterion Level			Trial Level		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BioBERT	0.95	0.94	0.94	0.96	0.96	0.96	0.86	0.93	0.89	0.90	0.95	0.92
ClinicalBERT	0.97	0.96	**0.96**	0.97	0.97	**0.97**	0.86	0.93	0.89	0.90	0.95	0.92
PubMedBERT	0.97	0.96	0.96	0.97	0.97	0.97	0.86	0.93	0.89	0.90	0.95	0.92
BlueBERT	0.95	0.94	0.94	0.96	0.96	0.96	0.86	0.93	0.89	0.90	0.95	0.92
SciBERT	0.97	0.96	**0.96**	0.97	0.97	**0.97**	0.86	0.93	0.89	0.90	0.95	0.92
ClinialTrialBERT	0.96	0.95	0.95	0.96	0.96	0.96	0.86	0.93	0.89	0.90	0.95	0.92

	HCV						Auto					
	Criterion Level			Trial Level			Criterion Level			Trial Level		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BioBERT	0.89	0.89	**0.89**	0.90	0.90	0.89	1.00	1.00	1.00	1.00	1.00	1.00
ClinicalBERT	0.88	0.88	0.88	0.90	0.90	0.89	1.00	1.00	1.00	1.00	1.00	1.00
PubMedBERT	0.86	0.86	0.85	0.87	0.87	0.86	1.00	1.00	1.00	1.00	1.00	1.00
BlueBERT	0.84	0.84	0.84	0.86	0.85	0.85	1.00	1.00	1.00	1.00	1.00	1.00
SciBERT	0.84	0.83	0.83	0.85	0.84	0.83	1.00	1.00	1.00	1.00	1.00	1.00
ClinicalTrialBERT	0.89	0.89	**0.89**	0.92	0.91	**0.91**	1.00	1.00	1.00	1.00	1.00	1.00

	Subst					
	Criterion Level			Trial Level		
	P	R	F1	P	R	F1
BioBERT	0.98	0.98	0.98	0.98	0.98	0.98
ClinicalBERT	0.98	0.98	0.98	0.99	0.99	**0.99**
PubMedBERT	0.98	0.98	0.98	0.99	0.99	**0.99**
BlueBERT	0.97	0.97	0.97	0.98	0.98	0.98
SciBERT	0.92	0.92	0.92	0.95	1.00	0.98
ClinicalTrialBERT	0.99	0.99	**0.99**	0.99	0.99	**0.99**

are liable to be highly confused with the connotation of “drug” meaning treatment (since these are phase III clinical trials, they almost all are drug trials). As such the precision of “drug” is only 0.22 and the precision of “drugs” is only 0.23. Removing these would cause the overall recall will drop from 1 to 0.92. Finally, for Auto trials, we note there are only 36 trials in the dataset that have an autoimmune exclusion, and the keywords from Table 1 are only able to capture 32 of them. This results in the recall already being lower than 0.9, while the precision is around 0.6. So any further sacrifices for the sake of precision would result in an unacceptable loss of recall for this stage of the pipeline. In summary, we believe that our keyword lists are suitable for capturing the criteria with the considered exclusions, with some consideration for precision but the primary focus being on recall.

4.2 Exclusion Classification

Table 5 presents the results for all six BERT-based models evaluated on the seven exclusion types, including both criterion-level and trial-level assessment. Across all models, the overall F1 score for all criteria ranged from 0.83 to 1.0 for both evaluation levels. For the Prior exclusion, PubMedBERT and our own ClinicalTrialBERT achieved the highest F1 score of 0.91, while BioBERT outperformed other models with the highest F1 score of 0.93 for trial-level evaluation. BlueBERT achieved the highest F1 score of 0.98 on both evaluation levels for the HIV exclusion. ClinicalBERT and SciBERT both achieved the highest F1 score for the Psych exclusion, with 0.96 and 0.97 on criterion-level and trial-level evaluation, respectively. For the HBV exclusion, all models achieved 0.97 and 1.00 for criterion-level and trial-level evaluation, respectively, except for ClinicalBERT, which had a slightly lower score of 0.96. On the criterion-level, BioBERT and our ClinicalTrialBERT reached 0.89 F1 score for HCV, while it also outperformed other models with 0.91 F1 score on the trial-level. There was no difference across all models for the Auto exclusion. Lastly, on both evaluation levels, our pre-train model outperformed other models for the Subst exclusion, with ClinicalBERT and PubMedBERT also achieving an F1 score of 0.99 on the trial-level evaluation.

Our results indicate that, at the criterion-level evaluation, there is no differences among all models for HBV and Auto. Our pre-trained ClinicalTrialBERT model performed equally or better than other models for three out of five remaining exclusions (Prior, HCV, Subst), and performed comparably well against the best-performing model on the other two (0.96 vs. 0.98 on HIV, 0.95 vs. 0.96 on Psych). At the trial-level evaluation, our pre-trained ClinicalTrialBERT model performed equally or better than other models in two out of five criteria compared to other models (HCV and Subst), while also performing comparably well against the best-performing model on the other three (0.91 vs. 0.93 on Prior, 0.96 vs. 0.98 on HIV, 0.96 vs. 0.97 on Psych). Meanwhile, BlueBERT and SciBERT had several exclusions for which they performed particularly below the best-performing model. This is to be somewhat expected for SciBERT, which has additional non-biomedical training data.

The metrics for the Auto exclusion are all 1.0. The reasons behind such perfect performance is likely due to the nature of the tasks. The sample size for autoimmune is only 54, with only 7 negatives. Yet the keyword-only precision for Auto is just 0.62, so the model is learning something beyond the keywords. In examining the data, in many cases this exclusion is simply stated (e.g., "no autoimmune disease"). These is fairly simple criteria, then, compared with descriptions for many of the other exclusions.

Of all criteria, HCV yields the poorest performance from all models. We conducted an error analysis on SciBERT for HCV, since it performed the worst overall model performance. We found the model conflated HCV and HBV (hepatitis C and B viruses, respectively), even though they oftentimes show concurrently in the same sentence. However, some sentences only mentioned HBV without saying HCV, and the model wrongly recognized these as HCV.

5 Discussion

Our study demonstrates that using an automatic tool to classify key exclusions from clinical trial eligibility criteria description holds immense potential. The criterion-level evaluation provides insight into our model's overall performance, while the trial-level evaluation provides a more practical and informative end-user view as it gives a sense of how many trials will be missed or falsely recommended based on each exclusion. With this in mind, the best-performing criterion-level model ranged from 0.89 (Psych) to 1.0 (Auto) while the best trial-level model ranged from 0.91 (HCV) to 1.0 (Auto). These results are more than sufficient to enable the scaling up of the types of analyses performed on the 764 PROTECTOR1 trials[41,42,43] to much larger subsets of the 445,000 trials currently available from ClinicalTrials.gov.

Our results also suggests that such methods can be used as part of patient-trial matching methods since a large percentage of eligibility criteria are shared across many trials (especially within specific domains such as cancer). Many eligibility criteria are not applicable to this approach, instead requiring more fine-grained information extraction techniques such as those done by the Criteria2Query system[24]. Future work in the space of patient-trial matching then should focus on hybrid solutions: differentiating between criteria that require specific facts extracted (and automatically covered to structured queries) and those criteria that are are semantically common yet lexically diverse. The latter type may be better approached using text classification based on the large language models used here since modern language models are excellent at identifying paraphrase-like similarity between sentences that share few words in common. These types of criteria further do not generally have specific argument structures (e.g., substance abuse criteria do not specifically detail the extact type of substance, its regularity of use, or the exact length of use). Such criteria are, rather, loose descriptions of common patient features that will be easily recognized and differentiated by clinicians. Therefore for these kinds of criteria, the approach taken in this work is highly appropriate, so future work should concentrate on differentiating which criteria are of this type and, of those, which have the critical mass of frequency to approach with the types of methods studied here.

Prior to this study, to the best of our knowledge there had been no pre-trained large language model on clinical trial descriptions. By pre-training such a model on the eligibility criteria section (the main section targeted by NLP systems) from hundreds of thousands of trials, our to-be-released ClinicalTrialBERT model will be useful for clinical trial NLP tasks well beyond the current work. The specific results in this study indicate that ClinicalTrialBERT is a robust model for text classification for clinical trials. This is not a surprising conclusion, but demonstrating the efficacy of the model across seven tasks is empirically important as we plan to share this BERT model and use it for future clinical trial NLP tasks.

In our study, the original PROTECTOR1 trial level-annotation was conducted based from various sources, including the ClinicalTrials.gov, available study protocols, and publications. Another interesting finding is that many trials did not disclosure certain key criteria on ClinicalTrials.gov, but only mentioned key exclusion criteria in protocols or publications, which will be investigated in future studies. Further work includes extend the current text classification framework to include information from protocols or publications. This would also allow for identifying which trials publish inconsistent information in various sources, which will ultimately help to improve the quality of clinical trial information provided to the public on ClinicalTrials.gov.

Limitations Our study is limited by the size of its samples, the number of exclusions considered, and its focus on a specific subset of trials (that is, phase III cancer clinical trials). First, in terms of sample size, most of the exclusions had only a few hundred annotations. The high recall of the keywords ensured that the annotations were largely complete, but still such samples may not generalize well when one goes beyond the scope of trials in this work (see third limitation). It certainly could mean that the very high performance of our models (with the best-performing F1 ranging from 0.89 to 1.0) are likely overly positive. Second, practical considerations limited this work to just seven exclusions. PROTECTOR1 has many more exclusions (and inclusions) annotated, so we hope to overcome this limitation with follow-up studies that expand the considered criteria. Finally, this work was limited to phase III cancer trials (the scope of PROTECTOR1). Hardly any design considerations of the system are overly specific to either phase III trials or cancer trials, but this scope does limit our ability to generalize the results with high confidence. Notably, oncology trial specialists are familiar with how other oncology trials are described (and likewise less familiar with trials for other fields [such as cardiology]), so as a result the way patient characteristics are specified for cancer trials may be different than how the same basic information is specified for trials in other diseases or specialties. We do note, however, that cancer trials make up a substantial portion of all clinical trials, so there certainly is optimism that these methods will be applicable to non-oncology trials.

6 Conclusion

In conclusion, we have successfully trained automatic classifiers using domain-specific BERT-based models to identify seven different exclusion criteria in clinical trials. We conducted evaluations at both the trial and criterion levels to assess the performance of all models. These evaluations demonstrate the ability of BERT-based models in general, and our new pre-trained ClinicalTrialBERT model in particular, to identify these seven exclusions with high performance in precision, recall, and F1. Our immediate future plan is to create a more comprehensive and mature model that can identify a significant number desired criteria.

References

[1] Society AC. Cancer Facts & Figures 2022; 2022. Accessed: February 27, 2023. https://www.cancer.org/latest-news/facts-and-figures-2022.html.

[2] Cox K, McGarry J. Why patients don't take part in cancer clinical trials: an overview of the literature. European journal of cancer care. 2003;12(2):114-22.

[3] Kadam RA, Borde SU, Madas SA, Salvi SS, Limaye SS. Challenges in recruitment and retention of clinical trial subjects. Perspectives in clinical research. 2016;7(3):137.

[4] Ross S, Grant A, Counsell C, Gillespie W, Russell I, Prescott R. Barriers to participation in randomised controlled trials: a systematic review. Journal of clinical epidemiology. 1999;52(12):1143-56.

[5] Jones JM, Nyhof-Young J, Moric J, Friedman A, Wells W, Catton P. Identifying motivations and barriers to patient participation in clinical trials. Journal of Cancer Education. 2007;21(4):237-42.

[6] Jenkins V, Farewell D, Batt L, Maughan T, Branston L, Langridge C, et al. The attitudes of 1066 patients with cancer towards participation in randomised clinical trials. British journal of cancer. 2010;103(12):1801-7.

[7] Mills EJ, Seely D, Rachlis B, Griffith L, Wu P, Wilson K, et al. Barriers to participation in clinical trials of cancer: a meta-analysis and systematic review of patient-reported factors. The lancet oncology. 2006;7(2):141-8.

[8] Kang T, Elhadad N, Weng C. Initial readability assessment of clinical trial eligibility criteria. In: AMIA Annual Symposium Proceedings. vol. 2015. American Medical Informatics Association; 2015. p. 687.

[9] of Medicine USNL. ClinicalTrials.gov. Glossary of Common Site Terms; 2022. Retrieved from: `https://clinicaltrials.gov/ct2/about-studies/glossary`.

[10] of Medicine USNL. ClinicalTrials.gov: 157588; 2013. Available: `http://www.clinicaltrials.gov`.

[11] ClinicalTrials.gov: FDAAA 801 and the Final Rule;. Accessed on: February 27, 2023. `https://clinicaltrials.gov/ct2/manage-recs/fdaaa#WhoIsResponsibleForRegistering`.

[12] Grant SR, Lin TA, Miller AB, Mainwaring W, Espinoza AF, Jethanandani A, et al. Racial and ethnic disparities among participants in US-based phase 3 randomized cancer clinical trials. JNCI Cancer Spectrum. 2020;4(5):pkaa060.

[13] Corrigan KL, Kouzy R, Abi Jaoude J, Patel RR, Layman RM, Giordano SH, et al. Inclusion of premenopausal women in breast cancer clinical trials. The Breast. 2022;66:204-7.

[14] Bhattacharya S, Cantor MN. Analysis of eligibility criteria representation in industry-standard clinical trial protocols. Journal of Biomedical Informatics. 2013;46(5):805-13.

[15] Zhao S. Named entity recognition in biomedical texts using an HMM model. In: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP); 2004. p. 87-90.

[16] Ji B, Liu R, Li S, Yu J, Wu Q, Tan Y, et al. A hybrid approach for named entity recognition in Chinese electronic medical record. BMC medical informatics and decision making. 2019;19(2):149-58.

[17] Bhatia P, Celikkaya B, Khalilia M, Senthivel S. Comprehend medical: a named entity recognition and relationship extraction web service. In: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). IEEE; 2019. p. 1844-51.

[18] Eom JH, Zhang BT. PubMiner: machine learning-based text mining for biomedical information analysis. Genomics & Informatics. 2004;2(2):99-106.

[19] Huang ZX, Tian HY, Hu ZF, Zhou YB, Zhao J, Yao KT. GenCLiP: a software program for clustering gene lists by literature profiling and constructing gene co-occurrence networks related to custom keywords. BMC bioinformatics. 2008;9:1-9.

[20] Bucur A, Van Leeuwen J, Chen NZ, Claerhout B, De Schepper K, Perez-Rey D, et al. Supporting patient screening to identify suitable clinical trials. In: e-Health–For Continuity of Care. IOS Press; 2014. p. 823-7.

[21] Uzuner Ö, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. Journal of the American Medical Informatics Association. 2008;15(1):14-24.

[22] Nii M, Tsuchida Y, Kato Y, Uchinuno A, Sakashita R. Nursing-care text classification using word vector representation and convolutional neural networks. In: 2017 Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS). IEEE; 2017. p. 1-5.

[23] Yao L, Mao C, Luo Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC medical informatics and decision making. 2019;19(3):31-9.

[24] Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. Journal of the American Medical Informatics Association. 2019;26(4):294-305.

[25] Liu C, Yuan C, Butler AM, Carvajal RD, Li ZR, Ta CN, et al. DQueST: dynamic questionnaire for search of clinical trials. Journal of the American Medical Informatics Association. 2019;26(11):1333-43.

[26] Luo Z, Johnson SB, Lai AM, Weng C. Extracting temporal constraints from clinical research eligibility criteria using conditional random fields. In: AMIA annual symposium proceedings. vol. 2011. American Medical Informatics Association; 2011. p. 843.

[27] Jonnalagadda SR, Adupa AK, Garg RP, Corona-Cox J, Shah SJ. Text mining of the electronic health record: an information extraction approach for automated identification and subphenotyping of HFpEF patients for clinical trials. Journal of cardiovascular translational research. 2017;10(3):313-21.

[28] Olasov B, Sim I. RuleEd, a web-based semantic network interface for constructing and revising computable eligibility rules. In: AMIA Annual Symposium Proceedings. vol. 2006. American Medical Informatics Association; 2006. p. 1051.

[29] Miotto R, Jiang S, Weng C. eTACTS: a method for dynamically filtering clinical trial search results. Journal of biomedical informatics. 2013;46(6):1060-7.

[30] Zhang K, Demner-Fushman D. Automated classification of eligibility criteria in clinical trials to facilitate patient-trial matching for specific patient populations. Journal of the American Medical Informatics Association. 2017;24(4):781-7.

[31] Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T, et al. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. Journal of the American Medical Informatics Association. 2015;22(1):166-78.

[32] Petkov VI, Penberthy LT, Dahman BA, Poklepovic A, Gillam CW, McDermott JH. Automated determination of metastases in unstructured radiology reports for eligibility screening in oncology clinical trials. Experimental Biology and Medicine. 2013;238(12):1370-8.

[33] Kirshner J, Cohn K, Dunder S, Donahue K, Richey M, Larson P, et al. Automated Electronic Health Record–Based Tool for Identification of Patients With Metastatic Disease to Facilitate Clinical Trial Patient Ascertainment. JCO Clinical Cancer Informatics. 2021;5:719-27.

[34] Tissot HC, Shah AD, Brealey D, Harris S, Agbakoba R, Folarin A, et al. Natural language processing for mimicking clinical trial recruitment in critical care: a semi-automated simulation based on the LeoPARDS trial. IEEE Journal of Biomedical and Health Informatics. 2020;24(10):2950-9.

[35] Cai T, Cai F, Dahal KP, Cremone G, Lam E, Golnik C, et al. Improving the efficiency of clinical trial recruitment using an ensemble machine learning to assist with eligibility screening. ACR Open Rheumatology. 2021;3(9):593-600.

[36] Kang T, Zhang S, Tang Y, Hruby GW, Rusanov A, Elhadad N, et al. EliIE: An open-source information extraction system for clinical trial eligibility criteria. Journal of the American Medical Informatics Association. 2017;24(6):1062-71.

[37] Kury F, Butler A, Yuan C, Fu Lh, Sun Y, Liu H, et al. Chia, a large annotated corpus of clinical trial eligibility criteria. Scientific data. 2020;7(1):281.

[38] Dobbins NJ, Mullen T, Uzuner Ö, Yetisgen M. The Leaf Clinical Trials Corpus: a new resource for query generation from clinical trial eligibility criteria. Scientific Data. 2022;9(1):490.

[39] Jung E, Jain H, Sinha AP, Gaudioso C. Building a specialized lexicon for breast cancer clinical trial subject eligibility analysis. Health Informatics Journal. 2021;27(1):1460458221989392.

[40] Liu H, Chi Y, Butler A, Sun Y, Weng C. A knowledge base of clinical trial eligibility criteria. Journal of biomedical informatics. 2021;117:103771.

[41] Pasalic D, Tang C, Jagsi R, Fuller CD, Koong AC, Ludmir EB. Association of industry sponsorship with cancer clinical trial accrual. JAMA oncology. 2020;6(10):1625-7.

[42] Patel RR, Verma V, Miller AB, Lin TA, Jethanandani A, Espinoza AF, et al. Exclusion of patients with brain metastases from cancer clinical trials. Neuro-oncology. 2020;22(4):577-9.

[43] Patel RR, Verma V, Fuller CD, McCaw ZR, Ludmir EB. Transparency in reporting of phase 3 cancer clinical trial results. Acta Oncologica. 2021;60(2):191-4.

[44] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40.

[45] Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. arXiv preprint arXiv:190403323. 2019.

[46] Johnson AE, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific data. 2016;3(1):1-9.

[47] 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:190605474. 2019.

[48] Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH). 2021;3(1):1-23.

[49] Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:190310676. 2019.