Learning Representations of Eligibility Criteria in Clinical Trials Using Transformers

Stanford CS224N Custom Project

Kevin Wu
Department of Biomedical Data Science
Stanford University
kevinywu@stanford.edu

Josiah Aklilu
Department of Biomedical Data Science
Stanford University
josaklil@stanford.edu

Abstract

A clinical trial's eligibility criteria can have a significant impact on the successful completion of the study, as they determine essential factors such as the recruitment efficiency, patient withdrawal rates, and translational power. Most inclusion and exclusion criteria are written in free-text, which makes a systematic review and analysis of these criteria prohibitive on a large scale. In this paper, we address these issues by learning standardized representations of eligibility criteria using transformers. In particular, we pretrain a BERT model on a large unlabeled corpus of eligiblity criteria acquired from ClinicalTrials.gov. Using Named Entity Recognition (NER) as a proxy for the quality of our representations, we show that our pretrained model (ecBERT) outperforms other publicly available biomedical BERT models, suggesting the benefit domain-specific representations for eligiblity criteria.

1 Key Information to include

- Mentor: Akshay Smit
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

Clinical trials are essential for improving patient care and advancing scientific knowledge, as they are the primary mechanism through which drug and device companies can test the efficacy and safety of their treatments. Each clinical trial contains a set of eligibility criteria which outline rules for selecting clinical trial participants. These criteria are set by the study's investigators and subsequently implemented by coordinators during the recruitment phase. After the completion of a study, these criteria are used by physicians to screen for patients who may benefit from such treatment. Therefore, eligibility criteria are crucial for the effectiveness of clinical trials as well as downstream patient care.

In practice, eligibility criteria can contain a variety of shortcomings, such as being unnecessarily strict [1] or being unrepresentative of a real-world population [2]. When criteria do not adequately represent an intended patient population, the results of trials have limited generalizability to actual patient care. However, addressing this issue can be a time-consuming process, as eligibility criteria are written as free text, without strong standardization and uniformity [3]. Additionally, the process of comparing treatments between trials is similarly difficult given the lack of standard representation. Further work has shown that the choice of eligibility criteria can significantly slow down clinical trials, as ill-specified criteria require additional protocol amendments [4].

As such, it is crucial to develop strong standard representations of eligibility criteria. These representations can be used to compare the targeted patient populations behind clinical trials, as well as address inefficiencies due to ill-specified criteria.

Eligibility criteria are generally written in free-text format in bullet-point format, and are divided into **inclusion** criteria and **exclusion** criteria. Below is an selected. example of eligibility criteria used in a trial for a treatment for Stage II breast cancer:

- **Inclusion Criteria**:
- Women or men with stage II or stage III, early invasive breast cancer according to the UICC 8th edition for TNM classification
- Histologically confirmed Estrogen Receptor ER+ (at least 10 % of cells staining positive for ER)
- Age $\geq$ 70 years
- Eastern Cooperative Oncology Group (ECOG) Performance status 0-2
- Patient must have undergone breast +/- axillary surgery with curative intent for the current malignancy $\leq$ 8 weeks before randomization.
- **Exclusion Criteria**:
- Previous history of invasive breast cancer
- Systemic anticancer therapy prior to the breast cancer surgery
- Prior therapy with any Cyclin-Dependent Kinase (CDK) 4/6 inhibitor
- Concurrent investigational agent within 28 days of randomization

3 Related Work

Previous work has been done in applying NLP-based approaches to eligibility criteria, although they predate the recent breakthroughs in NLP [5], using non-transformer based models such as Conditional Random Fields (CRFs). While these models are powerful tools for Named Entity Recognition (NER), they do not leverage large amounts of unlabeled data, nor do they allow for fine-tuning on a variety of other related tasks.

Furthermore, previous work has been in done in organizing and annotating datasets of eligibility criteria [6] [7] [8]. Recently, [9] has released the largest annotated dataset to date, but work has yet to be done on how well an NLP model works on this data.

Our work aims to be the first to use a transformer-based approach to modeling representations of eligibility criteria in a way that would allow for other downstream tasks to be learned easily with limited annotated data.

4 Approach and Experiments

Our approach involves pretraining a BERT model on a large corpus of unlabeled eligibility criteria and fine-tuning this model on Named Entity Recognition (NER) Figure 1, which we will call ecBERT (eligibility criteria BERT).

Transformer-based model (BERT): One of the primary aims of our paper is in understanding how well transformers can be used for the task of eligibility criteria representations. Whereas previous approaches had used deep-learning based approaches to modeling named entities, they predate the use of transformers and large-scale pretraining [5].

- We start with a standard BERT pretrained model (bert-base-uncased), which we further pretrain on Masked Language Modeling (MLM).
- We fine-tune on the largest publicly available dataset of labeled eligibility criteria, Chia [9], which contains annotations for 1K studies on the task of Named Entity Recognition (NER).
- We compare the performance of our pretrained model to other popular pretrained models in the biomedical space, Clinical-bioBERT [10] and blueBERT [11] in addition to the base BERT model [12],

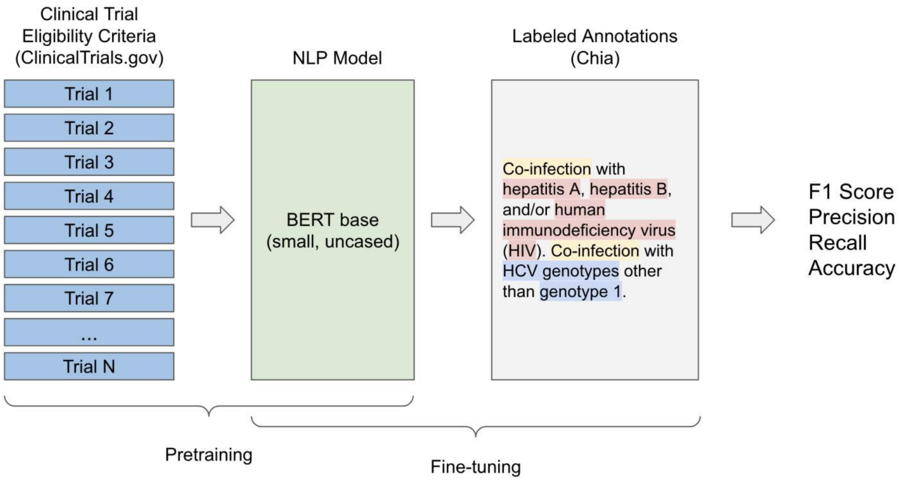

Figure 1: Process of training our model, from pretraining to finetuning.

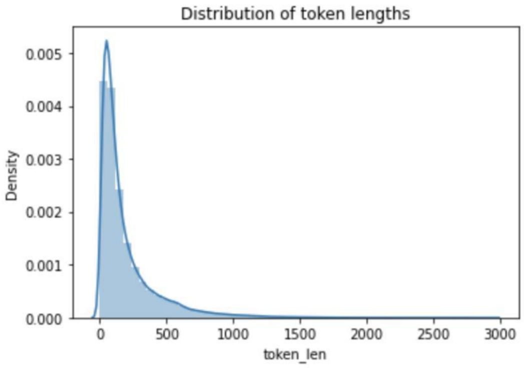

Figure 2: Distribution of token lengths in our pretraining dataset. A vast majority of texts are less than 512 in length.

Extracting and processing a large corpus of unlabeled data (AACT): Our paper also focuses on organizing and cleaning a corpus of over 4.6M eligibility criteria from over 360K clinical trials. This is a dataset of unprecedented size that was previously unused due to the lack of annotations.

4.1 Data

1. Chia[9]: The largest publicly-available dataset of annotated eligibility criteria to date. The data contain 1K interventional, Phase IV clinical trials from ClinicalTrials.gov, and includes 12,409 annotated individual eligibility criteria. In total, there are 30 entity types and 41,487 distinct entities. To preprocess the downloaded dataset, we wrote a script with the following procedure:

- Iterate through each annotation and extract annotation indices to produce a label vector that associates each word in the reference text to one of 30 possible categories.
- Convert the data to CoNLL format to be used in Huggingface's NER training script.
- We observed some ambiguous labels in the dataset, for which we developed rules to assign labels. For example, certain terms such as 'HIV+' are assigned labels as a whole, but also given sublabels according to 'HIV', and '+'. When dealing with such cases, we defaulted to labels given to the larger phrases in order to encourage our models to learn more general behavior.

2. Unannotated Eligibility Criteria:

- We downloaded Aggregate Content of ClinicalTrials.gov (AACT)[13]'s collection of clinical trials to get an up-to-date collection of clinical trials and an extensive feature list (phase, therapeutic area, study design, patient withdrawal rates, etc.).
- Convert AACT's stored datasets to Pandas using PostgreSQL.
- Clean and filter studies that either do not have eligibility criteria or do not specify which are inclusion or exclusion criteria.
- Parse and reformat each eligibility criteria into a standardized sentence format and attach a special token for inclusion and exclusion criteria ([INCL] and [EXCL]) at the beginning of each sentence.
- Notably, we had to decide whether to split documents longer than 512 tokens into separate documents. However, given the distribution of token lengths (shown above), we found that long documents were a tail end of the distribution, with less than 9% of documents were longer than 512 Figure 2. Therefore, we decided to leave truncation into the pretraining system.
- Random shuffle the order of each eligibility criteria so the model learns to model each criteria independent of order (and for truncation to not affect how the model trains).
- Split the dataset into training and evaluation sets with a 90/10% split.

4.2 Evaluation method

1. Masked Language Modeling (MLM)

For MLM, the loss function is the softmax function over possible word predictions. We monitor the loss over a held-out validation set.

2. Named Entity Recognition (NER)

Named Entity Recogntion is evaluated using the following metrics:

- Precision:
$$\textbf{Precision} = \frac{\textbf{True Positives}}{\textbf{True Positives} + \textbf{False Positives}}$$
- Recall:
$$\textbf{Recall} = \frac{\textbf{True Positives}}{\textbf{True Positives} + \textbf{False Negatives}}$$
- Accuracy:
$$\textbf{Accuracy} = \frac{\textbf{True Positives} + \textbf{True Negatives}}{\textbf{Total}}$$
- F1 score:
$$F1 = 2(\frac{\textbf{Precision} * \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}})$$

We report these metrics in aggregate across categories, as well as category-specific metrics.

4.3 Experimental details

1. Pretraining

We ran pretraining on top of a base BERT model over 12 epochs and on 2 Nvidia TITAN V GPUs. We use a batch size of 8 and a learning rate of 5e-5, with an Adam classifier with default beta parameters. We also use the default BERT tokenizer, and tokenize and collate the data for language modeling prior to training. In total, our initial run took approximately 5 days. We divide the data into training and test sets, using a 90/10% split.

As Figure 3 indicates, even after 5 days of training, pretraining loss was still dropping. Due to time constraints on our project, we were not able to let it run through to convergence. However, we can assume that given more time, our model's performance would increase over time.

2. Fine-tuning

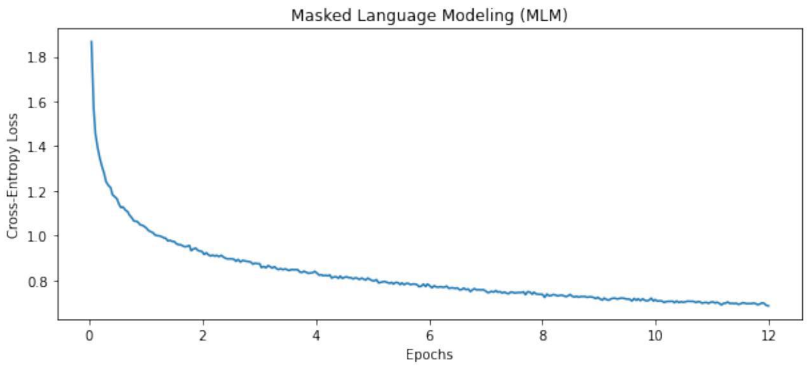

Figure 3: Pretraining loss over 12 epochs

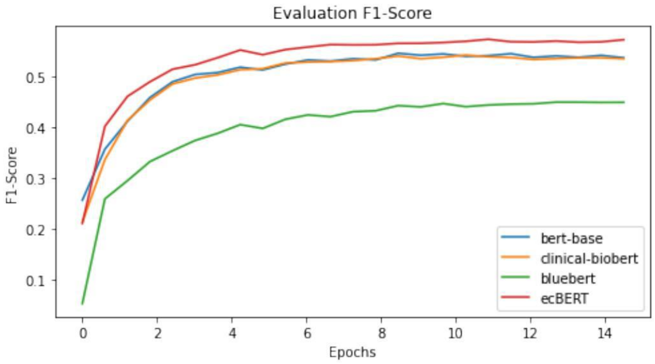

Figure 4: F1 score on the Chia validation set over 15 epochs, broken down by the choice of pretrained model.

We ran finetuning with a learning rate of 1e-05, using the Adam optimizer with default beta parameters. We trained a total of 750 steps using a batch size of 8. After training, we took the weights saved from the highest-performing point (by F1-score) to run on our test set. We divide our data into a training, validation, and test set, using a 70/10/20% split.

In total, fine-tuning each model took less than one hour on 2 Nvidia TITAN V GPUs.

4.4 Results

We report the results of fine-tuning each pretrained model on Chia's training set, shown in Figure 4. While the choice of pretrained model does not change final F1 score significantly, we do observe that our model, which was pretrained on eligibility criteria, performs better than other pretrained models. Additionally, we see that our model has better precision, recall, and accuracy as well, meaning the performance increase does not come at the cost of a certain metric.

Table 1: Performance on finetuning for Named Entity Recognition (NER) on Chia, with different pretrained models

Pretrained Model	F1 Score	Precision	Recall	Accuracy
BERT-base	0.517	0.480	0.560	0.632
blueBERT	0.431	0.367	0.521	0.589
Clinical-bioBERT	0.523	0.477	0.578	0.642
ecBERT (Ours)	**0.549**	**0.509**	**0.596**	**0.658**

Table 2: F1 score for NER with ecBERT, broken down by named entity.

Named Entity	**F1 Score**	**Count**
value	0.720	566
measurement	0.701	480
person	0.700	167
condition	0.687	1545
drug	0.640	455
no_label	0.577	4136
procedure	0.557	488
device	0.507	62
temporal	0.440	516
negation	0.440	100
informed_consent	0.367	48
qualifier	0.321	520
reference_point	0.313	112
multiplier	0.265	93
observation	0.206	184
visit	0.200	16
eligibility	0.178	80
competing_trial	0.174	21
pregnancy_considerations	0.149	47
mood	0.133	94
query-able	0.095	141
representable	0.048	54
subjective_judgement	0.000	23
undefined_semantics	0.000	36
line	0.000	2
intoxication_considerations	0.000	3
grammar_error	0.000	12
context_error	0.000	12
not_a_criteria	0.000	5

5 Analysis

Our results suggests that our model, ecBERT, produces a stronger representation than the BERT base model, Clinical-bioBERT, and blueBERT. Given that Clinical-bioBERT and blueBERT are pretrained on large corpuses of biomedical texts, this indicates that additional pretraining in the sub-domain of eligibility texts helps in downstream tasks such as NER. Further, we observe that our ecBERT model can still improve in performance given more training time. We expect that pretraining with more GPUs with longer alloted GPU allocation can lead to even higher performance increases.

5.1 Performance breakdown

We look at the F1 score for each named entity type to better understand the nature of the model. First, we observe that more data generally leads to higher predictions, where the lowest performing groups with F1 score of 0 are just poorly represented in the training dataset. Second, the model has issues with ambiguity. Whereas fields like value, measurement, condition, and drug are all entities that can be learned and memorized within the training set, other areas that are more ambiguously expressed such as mood, grammatical error, and context error are all much lower in performance. Finally, we see that the "no label" category has an F1 score of 0.577, indicating that there is still a relatively high false positive rate in our performance due to the density of the labels.

5.2 Analysis of the Chia dataset

As our report is the first time the Chia dataset has been publicly evaluated, we note the following findings:

- The Chia dataset is a relatively harder dataset, whereas reference performance for previous NER on eligibility criteria report F1 scores of around 0.7 [14].
- There is still a lot of ambiguity in the labeling of the dataset, as mentioned in the previous Data section of our report. This places a theoretical limit on the top performance our model can have on the dataset. Further work is required to clean this dataset.
- Overall, we observe that the Chia dataset is reasonably representative of the larger corpus of eligibility criteria, as our pretrained ecBERT performs favorably on it compared to other biomedical BERts.

6 Conclusion

We show that our pretrained model ecBERT outperforms other pretrained transformer models in named entity recognition for eligibility criteria. This illustrates the benefit of learning stronger representations of free-text eligibility criteria using transformers. Also, by pretraining on an unlabeled dataset of unprecedented size, we capitalize on the previously untapped wealth of unlabeled data. Our results indicate that pretraining with transformer based models on niche data helps models learn better representations of textual data. Our primary achievement with this project is therefore exemplifying the power of transformer based models for learning more comprehensive representations which can be used flexibly for downstream tasks.

6.1 Limitations

First, our corpus of eligibility criteria is taken from an unfiltered set of clinical trials from ClinicalTrials.gov. We believe that more careful selection of types of criteria (eg. only Phase III studies or only certain therapeutic areas) would perhaps learn a distribution that is more specific for our downstream task of NER.

Second, our parsing of the Chia dataset is imperfect, as some labeling is ambiguous. A more deliberate and methodological is needed to tap into the dataset's full list of intended labels.

Third, we are limited in our compute resources when pretraining ecBERT. A longer runtime for training can lead to a stronger representation, which in turn can improve downstream tasks.

6.2 Next Steps

Possible next steps include applying ecBERT to other downstream tasks such as relation extraction. We are also interested in how the embeddings extracted from our pretrained model can be used for self-supervised clustering of studies and measurement of trial similarities. This would allow for physicians and medical decision makers to evaluate clinical trials at scale, something that is currently time-prohibitive. Ultimately, we envision ecBERT being used for a variety of possible tasks related to the efficiency of clinical trials.

References

[1] David B Fogel. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. *Contemporary clinical trials communications*, 11:156–164, 2018.

[2] Juan M Banda, Martin Seneviratne, Tina Hernandez-Boussard, and Nigam H Shah. Advances in electronic phenotyping: from rule-based definitions to machine learning models. *Annual review of biomedical data science*, 1:53–68, 2018.

[3] Hongxing Luo, Yu Xu, Fengyang Yue, Cong Zhang, and Chang Chen. Quality of inclusion criteria in the registered clinical trials of heart failure with preserved ejection fraction: Is it time for a change? *International journal of cardiology*, 254:210–214, 2018.

[4] Kenneth A Getz, Rachael Zuckerman, Anne B Cropp, Anna L Hindle, Randy Krauss, and Kenneth I Kaitin. Measuring the incidence, causes, and repercussions of protocol amendments. *Drug information journal: DIJ/Drug Information Association*, 45(3):265–275, 2011.

[5] Chunhua Weng, Samson W Tu, Ida Sim, and Rachel Richesson. Formal representation of eligibility criteria: a literature review. *Journal of biomedical informatics*, 43(3):451–467, 2010.

[6] Pinal Patel, Disha Davey, Vishal Panchal, and Parth Pathak. Annotation of a large clinical entity corpus. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2033–2042, 2018.

[7] Sunil Mohan and Donghui Li. Medmentions: a large biomedical corpus annotated with umls concepts. *arXiv preprint arXiv:1902.09476*, 2019.

[8] Jake Luo, Min Wu, Deepika Gopukumar, and Yiqing Zhao. Big data application in biomedical research and health care: a literature review. *Biomedical informatics insights*, 8:BII–S31559, 2016.

[9] Fabrício Kury, Alex Butler, Chi Yuan, Li-heng Fu, Yingcheng Sun, Hao Liu, Ida Sim, Simona Carini, and Chunhua Weng. Chia, a large annotated corpus of clinical trial eligibility criteria. *Scientific data*, 7(1):1–11, 2020.

[10] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[11] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*, 2019.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Asba Tasneem, Laura Aberle, Hari Ananth, Swati Chakraborty, Karen Chiswell, Brian J McCourt, and Ricardo Pietrobon. The database for aggregate analysis of clinicaltrials. gov (aact) and subsequent regrouping by clinical specialty. *PloS one*, 7(3):e33677, 2012.

[14] Chi Yuan, Patrick B Ryan, Casey Ta, Yixuan Guo, Ziran Li, Jill Hardin, Rupa Makadia, Peng Jin, Ning Shang, Tian Kang, et al. Criteria2query: a natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics Association*, 26(4):294–305, 2019.