MALNIS & EMA3 @ TREC 2023 Clinical Trials Track

Mozhgan Saeidi[†][*] **Aman Jaiswal**[†][*] **Abhishek Dhankar**[§][*] **Alan Katz**[§] **Evangelos Milios**[†]

[§] Manitoba Centre for Health Policy, University of Manitoba

[†] Department of Computer Science, Dalhousie University

{aman.jaiswal, mozhgan.saeidi, eem}@dal.ca

{abhishek.dhankar, alan.katz}@umanitoba.ca

Abstract

This paper describes the submissions of the EMA3[1] team from the MALNIS[2] lab to the TREC 2023 Clinical Trials Track. In our approach to the TREC clinical trial matching problem, we use a two-stage process for effectively ranking and re-ranking clinical trials pertaining to a specific disorder. First, we identify candidate trials by matching normalized medical terms and non-negated inclusion/exclusion criteria to the disorder. Then, we rank the candidates using weighted relevance scores based on cosine similarity between contextual embeddings of the disorder and trial criteria. We use three different weighting schemes to compute a matching score. The unique aspect of our approach lies in the innovative use of these criteria to filter clinical trials and in the weighted relevance scoring, which reflects the varying importance of inclusion and exclusion criteria. Once we have computed the weighted relevance score for each candidate clinical trial, we rank the clinical trials by their score. Our submission performs better in terms of Precision@10 and NDCG-cut-10 than the median scores of the TREC 2023 Clinical trials track.

1 Introduction

Clinical trials are essential for developing new medical treatments, but they are often delayed or even canceled due to difficulty recruiting enough patients. This is because traditional recruitment methods, such as direct contact with clinical specialists or searching the electronic health record, can be inefficient and time-consuming.

Recently, patients have become more involved in the clinical trial process, and they are increasingly using online resources to search for and enroll in trials. The 2023 TREC Clinical Trials track simulates this scenario by providing participants with a simulated questionnaire that a patient or clinician would fill out to identify eligible trials. Participants are then challenged to retrieve relevant clinical trials from ClinicalTrials.gov, a registry of clinical trials in the United States. This task is difficult because clinical trial descriptions can be quite long and complex, and the most important information for determining eligibility is often buried in the inclusion/exclusion criteria.

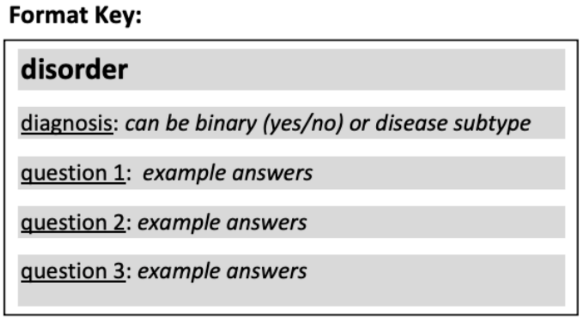
In this track, the evaluation will be broken down into three categories: eligible, excluded, and not relevant. This will allow participants to develop retrieval methods that can distinguish between patients who do not have enough information to qualify for a trial (not relevant) and those who are explicitly excluded (excludes).

The unique idea in our approach is to use the inclusion and exclusion criteria to filter the clinical trials and to weight the relevance score based on the importance of the inclusion and exclusion criteria. This allows us to rank the clinical trials more accurately and efficiently. Our proposed procedure has several benefits. First, it is comprehensive, as it ranks all clinical trials for a given disorder, regardless of their status (active, recruiting, completed, etc.). Second, it is accurate, as it uses transformer embeddings to compute the relevance score (Han et al., 2021). Transformer embeddings are a state-of-the-art word embedding technique that is known to produce accurate results. Third, it is efficient, as it uses a two-stage process to filter and rank the clinical trials. This allows us to reduce the number of clinical trials that need to be ranked, which makes the ranking process more efficient. Our proposed procedure can be used in a variety of applications. One application is in clinical trial matching: Our procedure can be used to match patients with suitable clinical trials. This can be helpful for patients who are looking for clinical trials to participate in, as well as for researchers who are recruiting patients for their clinical trials. The other application

[*] These authors contributed equally to this work.

[1] Evangelos (E), Mozhgan (M), Aman (A), Abhishek (A), and Alan (A).

[2] Machine Learning and Networked Information Spaces



Format Key:

disorder

diagnosis: *can be binary (yes/no) or disease subtype*

question 1: *example answers*

question 2: *example answers*

question 3: *example answers*

Figure 1: Questionnaire Template

is clinical trial prioritization: Our procedure can be used to prioritize clinical trials for funding or other resources. This can be helpful for funding agencies and other organizations that need to decide which clinical trials to support. The third application is in clinical trial landscape analysis: Our procedure can be used to analyze the clinical trial landscape for a given disorder. This can be helpful for researchers who are trying to identify gaps in the clinical trial landscape and to develop new clinical trials.

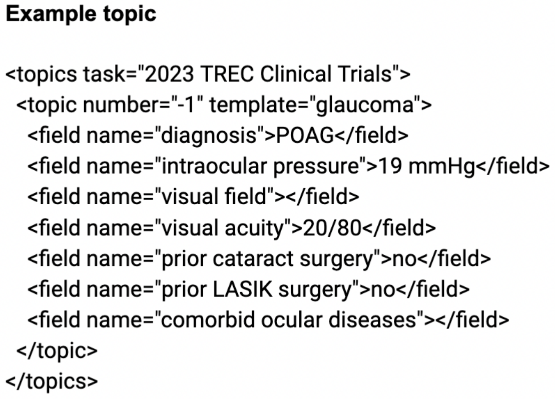
2 Problem Description and Dataset

The general problem in this study is information retrieval and ranking. Given 40 topics where each topic represents patient information, rank the provided clinical trials such that clinical trials that are suitable for the given patient (topic) are ranked higher. The only information available for a patient is the topic, which consists of the diagnosis, questions, and answers related to the diagnosis.

The topics for the track consist of synthetic patient descriptions based on questionnaire templates. They come in 8 formats covering 8 different disorders and are similar to the template shown in Fig 1. The answers to questions are optional. The questions/answers are specific to each disorder.

The topic covers the following disorders, a total of 40 topics, 5 each for every disorder: Glaucoma, Anxiety, Chronic obstructive pulmonary disease (COPD), Breast cancer, COVID-19, Rheumatoid arthritis, Sickle cell anemia, and Type 2 diabetes. Question templates for Glaucoma, COPD, COVID-19, Rheumatoid arthritis, Type 2 Diabetes, Anxiety, Breast cancer, and Sickle cell anemia are in the Appendix B.

The 40 topics are available in XML format, as Fig 2 demonstrates an example topic.



Example topic

```
<topics task="2023 TREC Clinical Trials">
  <topic number="-1" template="glaucoma">
    <field name="diagnosis">POAG</field>
    <field name="intraocular pressure">19 mmHg</field>
    <field name="visual field"></field>
    <field name="visual acuity">20/80</field>
    <field name="prior cataract surgery">no</field>
    <field name="prior LASIK surgery">no</field>
    <field name="comorbid ocular diseases"></field>
  </topic>
</topics>
```

Figure 2: An example of XML topic format

For the trials, the May 2023 snapshot of ClinicalTrials.gov is used as the corpus, which includes 52,130 clinical trials and is provided as 5 single zip files. The 40 topics need to be matched against these trials.

3 Background

Previous TREC tracks have focused on retrieving clinical trials, but the 2023 track is the most realistic and challenging to date. It provides participants with a simulated questionnaire that a patient or clinician would fill out and then challenges them to retrieve relevant clinical trials from ClinicalTrials.gov. This is a difficult task because clinical trial descriptions can be quite long and complex. The TREC Clinical Trials Track is not the first TREC track to focus on retrieving clinical trials. Previous iterations of TREC included the Clinical Decision Support (CDS) Tracks in 2014-2016 and the Precision Medicine Track in 2017-2020. These tracks focused on retrieving relevant abstracts of scientific publications from PubMed and evidence-based treatment literature and clinical trials, respectively (Simpson et al., 2014; Roberts et al., 2019). In the clinical trial task in 2016, they introduced a dataset with relevant judgments for topics from TREC CDS 2014 and a set of clinical trials from a snapshot of the ClinicalTrials.gov registry (Koopman and Zuccon, 2016). They also introduced the use of ad-hoc queries, which are constructed by asking domain experts to write down what they would normally use as queries when searching for potential trials that are suitable for a patient. Their empirical results showed that ad-hoc queries outperform full-text or summarized-text queries. In other words, the 2023 TREC Clinical Trials Track builds on previous TREC tracks by providing participants with a more realistic and challenging task:

retrieving relevant clinical trials from ClinicalTrials.gov based on a simulated questionnaire that a patient or clinician would fill out.

4 Our Model

Our proposed two-stage procedure ranks clinical trials for a given disorder based on their relevance to the disorder. We divide our approach into two stages.

4.1 First Stage

In the first stage, we identify all clinical trials that include the disorder in the conditions column. We then extract the negated & non-negated inclusion and exclusion criteria (using negspaCy (Pizarro, 2023)) from the eligibility criteria column and normalize them by converting them into UMLS Concept IDs (Bodenreider, 2004; Saeidi et al., 2023) by using ScispaCy (Neumann et al., 2019; Saeidi et al., 2022). Similarly, we converted the terms in the patient profiles into UMLS Concept IDs and searched for these IDs among those identified in the clinical trial description. The trials were ranked based on the number of matches of Concept IDs. The rank score was then calculated by finding the natural log of the index of each trial subtracted from the total number of documents. The output of this stage is a list of candidate clinical trials that are likely to be relevant to the disorder.

4.2 Second Stage

In the second stage, we compute a relevance score for each candidate clinical trial. The relevance score is based on the cosine similarity between the contextual embeddings of the topic sentences for the disorder and the inclusion and exclusion criteria. Figure 4 illustrates the two cosine similarity matrices we utilize to compute relevance scores. The complete pipeline for the second stage is illustrated and described in Figure 3. We use the contextual word embeddings from the Transformers model trained on a large corpus of text (Reimers and Gurevych, 2019; Devlin et al., 2018; Saeidi et al., 2021b). The relevance score weighs inclusion and exclusion criteria differently, with each strategy affecting the ranking of candidate trials. The relevance scoring functions are discussed in section 4.2.1. Once we have computed the weighted relevance score for each candidate clinical trial, we rank the clinical trials by their scores. The clinical trials with the highest scores are the most relevant to the disorder.

4.2.1 Relevance Scoring Functions

We consider three relevance scoring functions each of which operates on cosine similarity matrices. Let I be the inclusion matrix, where I_{ij} represents the similarity between extracted topic sentence i and inclusion criteria sentence j. Similarly, let E be the exclusion matrix, where E_{ij} represents the similarity between topic sentence i and exclusion criteria sentence j. Figure 4, illustrates an example of cosine similarity matrices.

Naive High Precision Score: It is denoted as S_{naive} and is based on setting a threshold for matching. For a given clinical trial is computed as follows:

1. For each inclusion criteria sentence j, we check if it is satisfied by one or more topic sentences. If at least one topic sentence has a similarity greater than or equal to a threshold (T) with the inclusion criteria sentence, it is considered satisfied.

2. Each satisfied inclusion criteria sentence is given a score of 1. This is represented as $I_{satisfied}$, where $I_{satisfied}$ is a binary matrix indicating which inclusion criteria sentences are satisfied.

3. If there are exclusion criteria sentences (i.e., E is no None and has at least one exclusion criteria sentence), we check if any topic sentence has a similarity greater than or equal to the threshold with any exclusion criteria sentence. If this condition is met, we set the exclusion score (E_{score}) to -1, indicating that the trial is excluded due to one or more exclusion criteria.

The final score (S_{naive}) for the clinical trial is computed as follows:

$$S_{naive} = \frac{1}{N} \sum_{i=1}^{N} I_{satisfied_i} + E_{score}$$

Where:

- N is the number of inclusion criteria sentences.
- $I_{satisfied_i}$ is 1 if the i-th inclusion criteria sentence is satisfied, 0 otherwise.
- E_{score} is -1 if any exclusion criteria are satisfied, 0 otherwise.

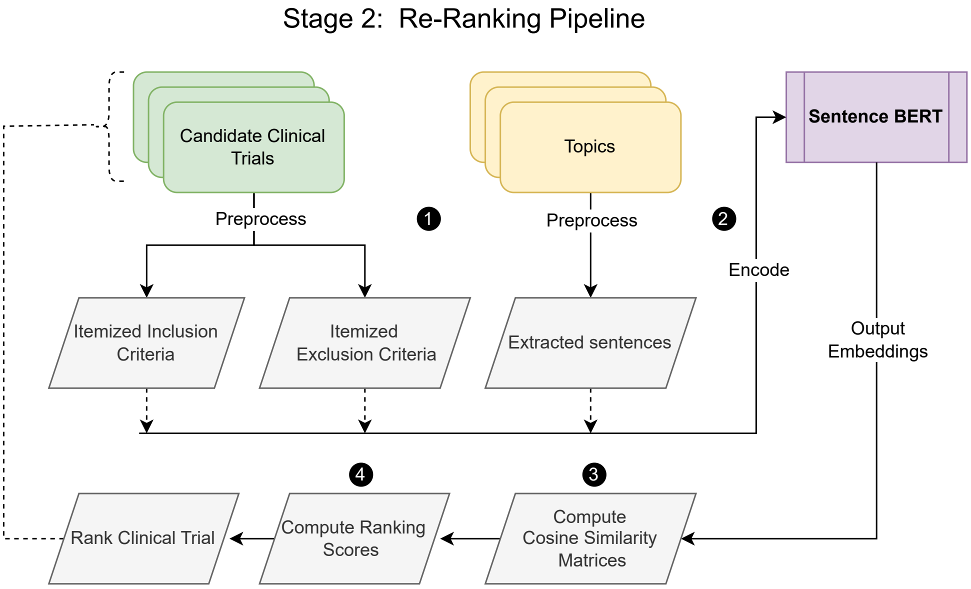

Stage 2: Re-Ranking Pipeline
Candidate Clinical Trials
Topics
Sentence BERT
Preprocess
Preprocess
1
2
Encode
Output Embeddings
Itemized Inclusion Criteria
Itemized Exclusion Criteria
Extracted sentences
4
3
Rank Clinical Trial
Compute Ranking Scores
Compute Cosine Similarity Matrices

Figure 3: *Stage 2 Re-ranking Pipeline* uses Cosine Similarity matrices for {Criteria, Topic Sentences}. Stage 2, re-ranking can be divided into 4 steps. The first step involves pre-processing candidate trials and disorder topics. The itemized inclusion and exclusion criteria are obtained using regex parsing and heuristics. The topics XML file is parsed, and the question responses are converted into sentences using the OpenAI (OpenAI, 2023) generative model. The prompt description and the outputs are described in appendix A. The second step is to convert the text from the 3 sources (inclusion, exclusion, and topic sentences) into embeddings using Sentence BERT(Reimers and Gurevych, 2019). Given the candidate trials, we compute two cosine similarity matrices I and E of dimensions $|T| \times |I|$ and $|T| \times |E|$, where $|T|$, $|I|$, and $|E|$ are the number of topic sentences, inclusion, and exclusion criteria, respectively. The relevance ranking score is a function of cosine similarity matrices I and E, which provide a score for each candidate trial based on the topic. The final step is to normalize the ranking scores and sort the clinical trials based on their relevance scores.

The score (S_{naive}) is a naive measure of the relevance of the clinical trial, where a negative score or zero score indicates exclusion due to the presence of exclusion criteria, and a positive score indicates inclusion based on the satisfaction of inclusion criteria. We use a threshold of 0.5 to compute this score.

Weighted Relevance Score: It is denoted as $S_{weighted}$ and does not require any threshold for matching. It assigns a higher weight to exclusion criteria by taking the maximum value in the exclusion matrix and subtracting it from the average cosine similarities in the inclusion matrix. A negative score indicates that the topic sentences are more closely aligned with the exclusion criteria than the inclusion criteria. The scoring process can be summarized as follows:

1. Compute the average cosine similarity in the inclusion matrix I, where I_{ij} represents the similarity between topic sentence i and inclusion criteria sentence j. The inclusion score is given by:

$$I_{Score} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} I_{ij}$$

   where N is the number of inclusion criteria sentences and T is the number of topic sentences.

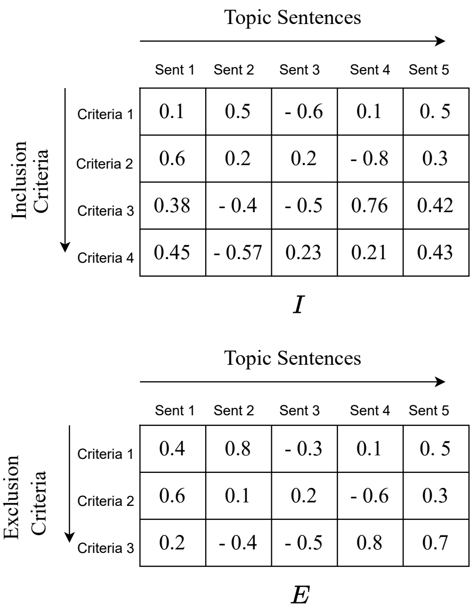
2. Compute the maximum cosine similarity in the exclusion matrix E, where E_{ij} represents the similarity between topic sentence i and exclusion criteria sentence j. The exclusion score is given by:

$$E_{Score} = \max_{i}(E_{ij})$$

3. Calculate the final score as the difference between the inclusion score and the exclusion score:

$$S_{weighted} = I_{Score} - E_{Score}$$

The $S_{weighted}$ score provides a measure of relevance for the clinical trial. A negative score indicates that the topic sentences are more aligned



	Sent 1	Sent 2	Sent 3	Sent 4	Sent 5
Criteria 1	0.1	0.5	- 0.6	0.1	0. 5
Criteria 2	0.6	0.2	0.2	- 0.8	0.3
Criteria 3	0.38	- 0.4	- 0.5	0.76	0.42
Criteria 4	0.45	- 0.57	0.23	0.21	0.43

	Sent 1	Sent 2	Sent 3	Sent 4	Sent 5
Criteria 1	0.4	0.8	- 0.3	0.1	0. 5
Criteria 2	0.6	0.1	0.2	- 0.6	0.3
Criteria 3	0.2	- 0.4	- 0.5	0.8	0.7

Figure 4: *Example Cosine Similarity matrices*, We utilize two cosine similarity matrices to compute ranking score, I and E for {inclusion criteria, topic sentences} similarity and {exclusion criteria, topic sentences} similarity respectively.

with the exclusion criteria than the inclusion criteria. This score gives more weight to the exclusion criteria and penalizes matching with the exclusion criteria.

Balanced Relevance Score: It is denoted as $S_{balanced}$ and does not require any threshold setting for matching. It calculates a balanced relevance score for a clinical trial by considering both inclusion and exclusion criteria equally. This scoring function aims to find a balance between the two criteria, with a negative score indicating that the topic sentences are more aligned with the exclusion criteria than the inclusion criteria.

The scoring process can be summarized as follows:

1. Compute the average cosine similarity in the inclusion matrix I, where I_{ij} represents the similarity between topic sentence i and inclusion criteria sentence j. The inclusion score is given by:

$$I_{Score} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} I_{ij}$$

   where N is the number of inclusion criteria sentences and T is the number of topic sentences.

Run	P-10	MAP	Reciprocal-rank
stage1ema	0.4595	0.0992	0.6296
nnrema3	0.4865	0.0949	0.6277
brsema3	**0.6270**	**0.1366**	**0.8242**
wrsema3	0.5676	0.1249	0.7803

Table 1: Mean Precision-10 (P-10), Mean Average Precision (MAP), and Reciprocal Rank for the submitted runs. 'stage1ema,' 'nrema3,' 'brsema3,' and 'wresema3' refer to stage1 (no re-ranking), naive relevance ranking, balanced relevance ranking, and weighted relevance ranking, respectively.

2. Compute the average cosine similarity in the exclusion matrix E, where E_{ij} represents the similarity between topic sentence i and exclusion criteria sentence j. The exclusion score is given by:

$$E_{Score} = \frac{1}{KT} \sum_{i=1}^{K} \sum_{j=1}^{T} E_{ij}$$

   where K is the number of exclusion criteria sentences and T is the number of topic sentences.

3. Calculate the combined score as the difference between the inclusion score and the exclusion score:

$$S_{balanced} = I_{Score} - E_{Score}$$

The $S_{balanced}$ provides a balanced assessment of the clinical trial's relevance, considering both inclusion and exclusion criteria. A negative score suggests that the trial may align more with the exclusion criteria, while a positive score indicates stronger alignment with the inclusion criteria. This scoring function is designed to offer a more balanced evaluation of clinical trial relevance, allowing for a nuanced assessment of trial suitability.

5 Evaluation and Results

In this work, we were allowed to submit up to 5 runs. For each topic, we had to provide a maximum of 1000 clinical trials for which the patient was most suitable. We submitted 4 runs, with three runs involving re-ranking the candidate trials from stage 1, while the first run represents the output from *stage 1*.

The results of our runs are summarized in Table 1, which reports TREC evaluation metrics: Precision at 10 (P-10), Mean Average Precision (MAP),

Runs	**NDCG-cut-10** **(min/median/max)**	**NDCG-cut-1000** **(min/median/max)**
Stage1	0.00 / 0.71 / 1	0.02 / 0.29 / 0.48
Naïve relevance rank	0.00 / 0.63 / 1	0.01 / 0.28 / 0.5
Balanced Relevance Rank	0.06 / 0.82 / 1	0.13 / 0.31 / 0.51
Weighted Relevance Rank	0.06 / 0.73 / 1	0.01 / 0.28 / 0.52
Across all Participants (Averaged over all topics)	**0.01 / 0.64 / 0.92**	**0.00 / 0.39 / 0.54**

Table 2: Minimum, median, and maximum of Mean Normalized Discounted Cumulative Gain (NDCG) scores for the submitted runs. 'stage1ema,' 'nrema3,' 'brsema3,' and 'wresema3' refer to stage1 (no re-ranking), naive relevance ranking, balanced relevance ranking, and weighted relevance ranking, respectively. We include the results across all participants.

Runs	**Precision@10** **(min/median/max)**	**Mean Average Precision** **(min/median/max)**	**Reciprocal rank** **(min/median/max)**
Stage1	0.00 / **0.5** / 1	0.00 / **0.07** / 0.26	0.00 / **1** / 1
Naïve relevance rank	0.00 / **0.5** / 1	0.00 / **0.08** / 0.31	0.00 / **0.5** / 1
Balanced Relevance Rank	0.00 / **0.70** / 1	0.00 / **0.14** / 0.36	0.00 / **1** / 1
Weighted Relevance Rank	0.00 / **0.60** / 1	0.00 / **0.12** / 0.36	0.00 / **1** / 1
Across all Participants (Averaged over all topics)	**0.00 / 0.39 / 0.88**	**0.00 / 0.09 / 0.25**	**0.01 / 0.53 / 1.00**

Table 3: Minimum, median, and maximum of Mean Precision-10 (P-10), Mean Average Precision (MAP), and Reciprocal Rank for the submitted runs. 'stage1ema,' 'nrema3,' 'brsema3,' and 'wresema3' refer to stage1 (no re-ranking), naive relevance ranking, balanced relevance ranking and weighted relevance ranking, respectively. We include the results across all participants.

Run	NDCG-C-10	NDCG-C-1000
stage1ema	0.6003	0.2978
nrema3	0.6780	0.3029
brsema3	0.6376	0.2991
wrsema3	**0.7246**	**0.3167**

Table 4: Mean Normalized Discounted Cumulative Gain (NDCG) scores for the submitted runs. 'stage1ema,' 'nrema3,' 'brsema3,' and 'wresema3' refer to stage1 (no re-ranking), naive relevance ranking, balanced relevance ranking, and weighted relevance ranking, respectively.

and Reciprocal Rank. The distribution of the P-10, MAP, and reciprocal rank is given in Table 3 and visualized in Figure 5. Three out of the four submissions outperform the median for all three evaluation metrics - P-10, Reciprocal Rank, and MAP. The balanced relevance ranking and weighted relevance ranking strategy run surpass the median trec evaluation metrics across all submissions. Table 4 presents the Normalized Discounted Cumulative Gain (NDCG) scores. The distribution of NDGC scores is given in Table 2 and visualized in Figure 6. The run with a balanced relevance ranking performs the best in terms of NDCG scores. Re-ranking using any of the proposed methods improves NDCG scores. In particular, re-ranking using naive relevance scoring and weighted relevance scoring outperforms the median scores of the TREC 23 Clinical Trial Track.

The distribution of TREC evaluation metrics and NDCG scores across topics is presented in Figure 5 and Figure 6.

6 Conclusion and Discussion

Our approach to the TREC clinical trial matching problem uses a two-step process to rank and rerank clinical trials for a particular disorder. First, we identify all clinical trials that include the target disorder in their conditions column. We then extract non-negated inclusion and exclusion criteria from the eligibility criteria field for each trial. To rank these trials against the given topics, we mapped medical terms in the eligibility criteria to standardized UMLS Concept IDs. We matched those to the UMLS Concept IDs found in the given topics, thus generating a list of candidate trials for the next step.

In the second step, we calculate a weighted relevance score for each candidate's clinical trial. This score is determined by measuring the cosine similarity between contextual embeddings of the topic sentences associated with the disorder and the cor-

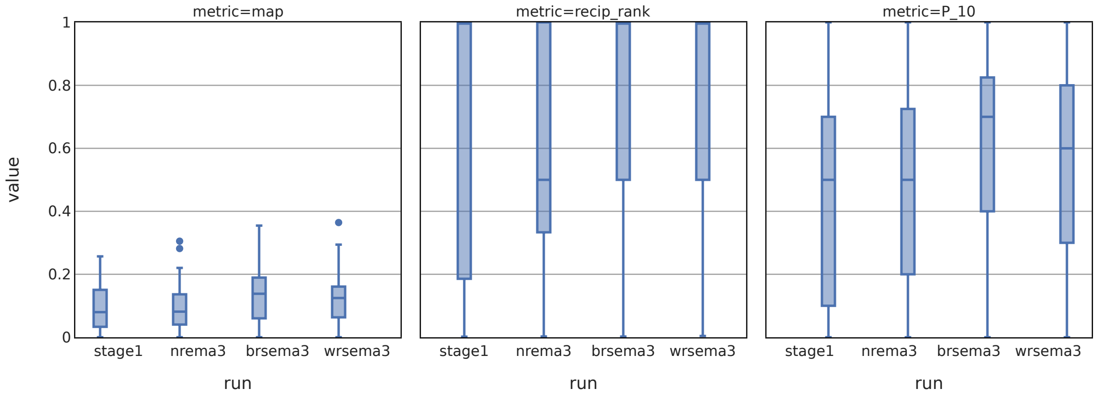


Figure 5: The distribution of TREC evaluation metrics, including Mean Precision-10 (P-10), Mean Average Precision (MAP), and Reciprocal Rank (Reciprocal Rank), is observed across various topics. 'stage1ema,' 'nrema3,' 'brsema3,' and 'wresema3' correspond to stage1 (no re-ranking), naive relevance ranking, balanced relevance ranking, and weighted relevance ranking, respectively.

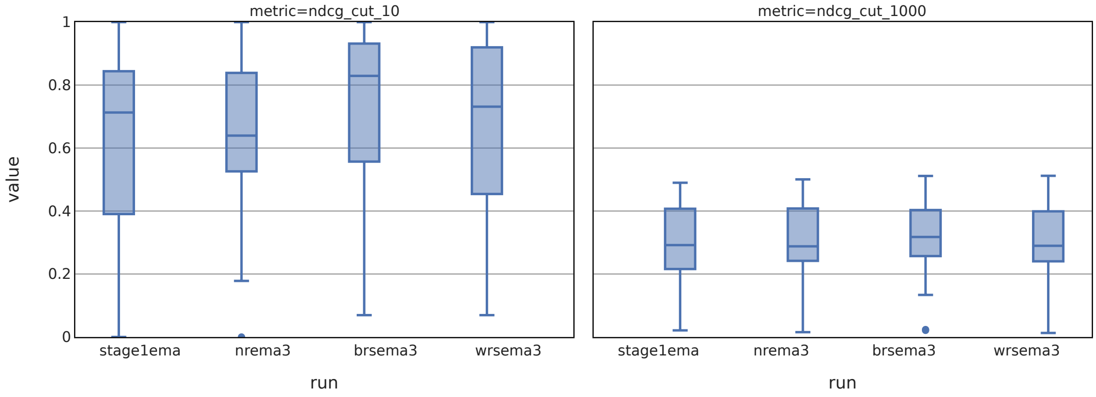


Figure 6: The distribution of Normalized Discounted Cumulative Gain (NDCG) scores for NDCG-10 and NDCG-1000 across various topics. 'stage1ema,' 'nrema3,' 'brsema3,' and 'wresema3' correspond to stage1 (no re-ranking), naive relevance ranking, balanced relevance ranking, and weighted relevance ranking, respectively

responding inclusion and exclusion criteria. We consider three strategies to weigh the inclusion and exclusion criteria, which reflect the varying importance of inclusion and exclusion criteria. Our submission performs better in terms of Precision@10 and NDCG-cut-10 than the median scores TREC 2023 Clinical trials track. *Balanced relevance ranking* outperforms the other approaches in TRECEVAL median metrics.

For future direction, we can continue solving this trial ranking problem using deep learning algorithms, such as GCN (Kipf and Welling, 2016; Saeidi et al., 2021a), while employing pretrained embeddings on biomedical text, such as BIOCBERT (Saeidi et al., 2022), and applying the overlapping windowing approach in trials understanding (Saeidi et al., 2023). The other interesting approach to follow is reinforcement learning which might improve the results.

References

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

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

Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. 2021. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Bevan Koopman and Guido Zuccon. 2016. A test collection for matching patients to clinical trials. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 669–672.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. *ArXiv*, abs/1902.07669.

OpenAI. 2023. GPT-3.5 by OpenAI. [Online; accessed 5-November-2023].

Jeno Pizarro. 2023. negspacy. `https://github.com/jenojp/negspacy`.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, Alexander J Lazar, Shubham Pant, and Funda Meric-Bernstam. 2019. Overview of the trec 2019 precision medicine track. In *The... text REtrieval conference: TREC. Text REtrieval Conference*, volume 1250. NIH Public Access.

Mozhgan Saeidi, Kaveh Mahdaviani, Evangelos Milios, and Norbert Zeh. 2023. Context-enhanced concept disambiguation in wikification. *Intelligent Systems with Applications*, page 200246.

Mozhgan Saeidi, Evangelos Milios, and Norbert Zeh. 2021a. Contextualized knowledge base sense embeddings in word sense disambiguation. In *Document Analysis and Recognition–ICDAR 2021 Workshops: Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 174–186. Springer.

Mozhgan Saeidi, Evangelos Milios, and Norbert Zeh. 2021b. Graph representation learning in document wikification. In *Document Analysis and Recognition–ICDAR 2021 Workshops: Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 509–524. Springer.

Mozhgan Saeidi, Evangelos Milios, and Norbert Zeh. 2022. Biomedical word sense disambiguation with contextualized representation learning. In *Companion Proceedings of the Web Conference 2022*, pages 843–848.

Matthew S Simpson, Ellen M Voorhees, and William R Hersh. 2014. Overview of the trec 2014 clinical decision support track. In *TREC*.

A Prompt Description

Task Description: You are an expert concept identifier and linker. You are especially well-versed in the clinical and medical domains. You will be given questionnaires filled by people with respect to some conditions they have, such as "glaucoma." In this questionnaire, the responses are either yes, no, or some numbers related to certain tests or diagnoses. Your job is to look at the responses and convert the structured responses into sentences, preserving the original meaning. The sentences should be useful for information retrieval using dense passage retrieval. Keep in mind that we need to convert the short responses into sentences to match with clinical trial summaries or inclusion and exclusion criteria of these trials extracted from clinicaltrials.gov. The general task we want to solve is clinical trial matching, so extrapolating information from the responses may be useful, but too much extrapolation can lead to false positives. Therefore, be careful when linking concepts and relevant terminology. //XML lines Here//

Figure 7: Prompt description used for converting the questionnaire responses to sentences.

Input XML

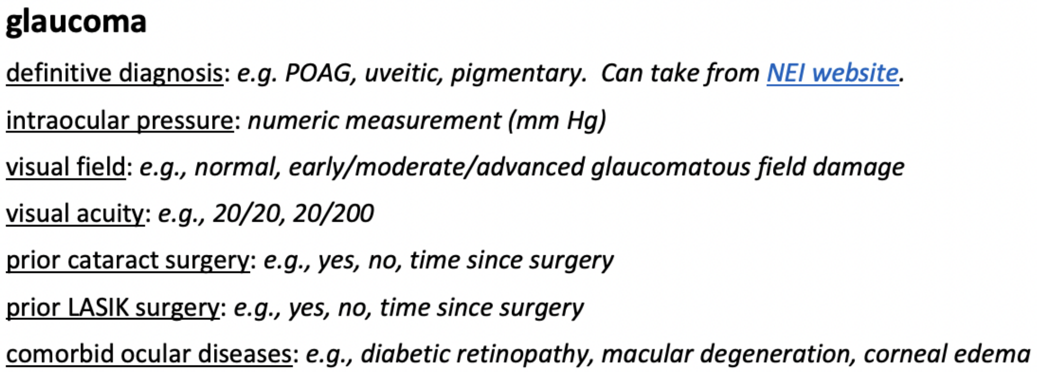
```
<topic number="2" template="glaucoma">
<field name="definitive diagnosis">pigmentary</field>
<field name="intraocular pressure">15 mmHg</field>
<field name="visual field">normal</field>
<field name="visual acuity">20/50</field>
<field name="prior cataract surgery">no</field>
<field name="prior LASIK surgery">yes</field>
<field name="comorbid ocular diseases">macular degeneration</field>
</topic>
```

Output Sentences

Definitive diagnosis: The patient has pigmentary glaucoma.
Intraocular pressure: The patient’s intraocular pressure is 15 mmHg.
Visual field: The patient’s visual field is normal.
Visual acuity: The patient’s visual acuity is 20/50.
Prior cataract surgery: The patient has not undergone prior cataract surgery.
Prior LASIK surgery: The patient has undergone prior LASIK surgery.
Comorbid ocular diseases: The patient has macular degeneration in addition to glaucoma.

Figure 8: The output sentences obtained using prompting.

B Questionnaire Templates



glaucoma

definitive diagnosis: *e.g. POAG, uveitic, pigmentary. Can take from NEI website.*

intraocular pressure: *numeric measurement (mm Hg)*

visual field: *e.g., normal, early/moderate/advanced glaucomatous field damage*

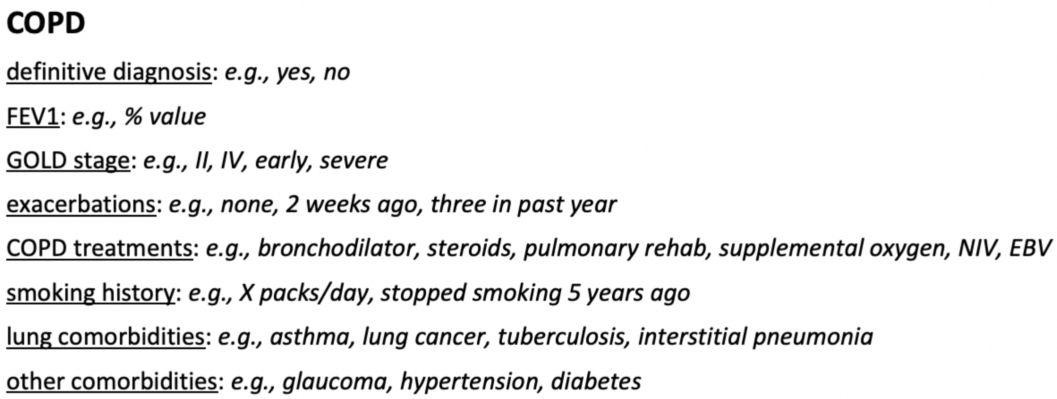
visual acuity: *e.g., 20/20, 20/200*

prior cataract surgery: *e.g., yes, no, time since surgery*

prior LASIK surgery: *e.g., yes, no, time since surgery*

comorbid ocular diseases: *e.g., diabetic retinopathy, macular degeneration, corneal edema*

Figure 9: Question Template for Glaucoma disorder



COPD

definitive diagnosis: *e.g., yes, no*

FEV1: *e.g., % value*

GOLD stage: *e.g., II, IV, early, severe*

exacerbations: *e.g., none, 2 weeks ago, three in past year*

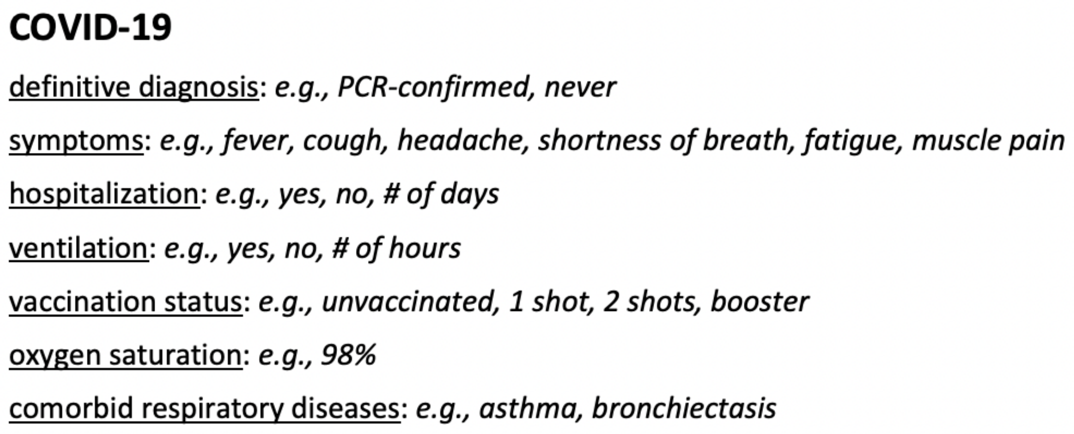
COPD treatments: *e.g., bronchodilator, steroids, pulmonary rehab, supplemental oxygen, NIV, EBV*

smoking history: *e.g., X packs/day, stopped smoking 5 years ago*

lung comorbidities: *e.g., asthma, lung cancer, tuberculosis, interstitial pneumonia*

other comorbidities: *e.g., glaucoma, hypertension, diabetes*

Figure 10: Question Template for Chronic obstructive pulmonary disease (COPD) disorder



COVID-19

definitive diagnosis: *e.g., PCR-confirmed, never*

symptoms: *e.g., fever, cough, headache, shortness of breath, fatigue, muscle pain*

hospitalization: *e.g., yes, no, # of days*

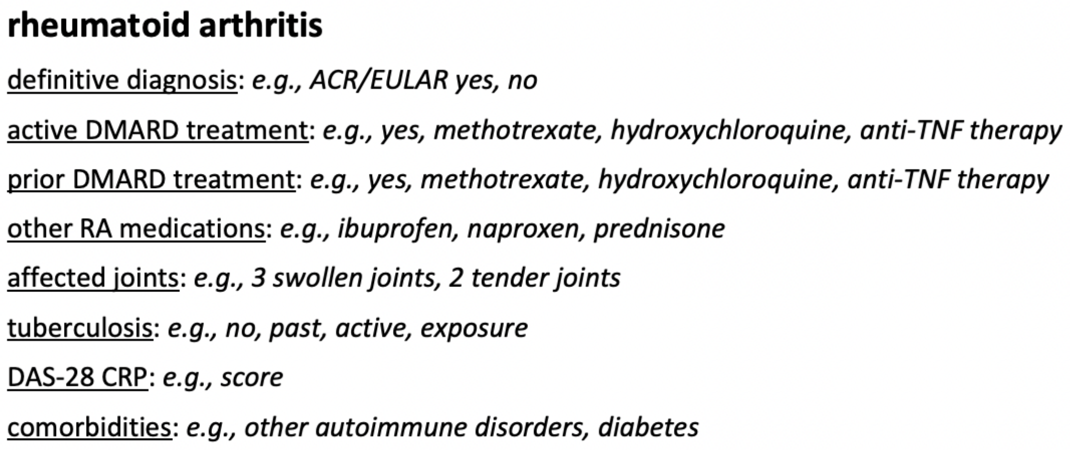
ventilation: *e.g., yes, no, # of hours*

vaccination status: *e.g., unvaccinated, 1 shot, 2 shots, booster*

oxygen saturation: *e.g., 98%*

comorbid respiratory diseases: *e.g., asthma, bronchiectasis*

Figure 11: Question Template for Covid19 disorder



rheumatoid arthritis

definitive diagnosis: *e.g., ACR/EULAR yes, no*

active DMARD treatment: *e.g., yes, methotrexate, hydroxychloroquine, anti-TNF therapy*

prior DMARD treatment: *e.g., yes, methotrexate, hydroxychloroquine, anti-TNF therapy*

other RA medications: *e.g., ibuprofen, naproxen, prednisone*

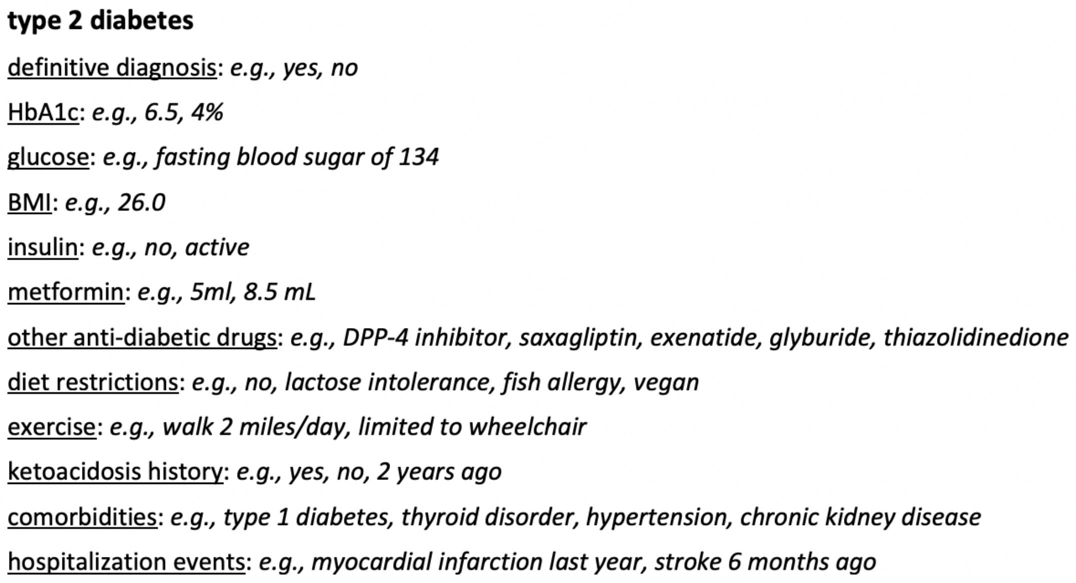
affected joints: *e.g., 3 swollen joints, 2 tender joints*

tuberculosis: *e.g., no, past, active, exposure*

DAS-28 CRP: *e.g., score*

comorbidities: *e.g., other autoimmune disorders, diabetes*

Figure 12: Question Template for Rheumatoid Arthritis disorder



type 2 diabetes

definitive diagnosis: *e.g., yes, no*

HbA1c: *e.g., 6.5, 4%*

glucose: *e.g., fasting blood sugar of 134*

BMI: *e.g., 26.0*

insulin: *e.g., no, active*

metformin: *e.g., 5ml, 8.5 mL*

other anti-diabetic drugs: *e.g., DPP-4 inhibitor, saxagliptin, exenatide, glyburide, thiazolidinedione*

diet restrictions: *e.g., no, lactose intolerance, fish allergy, vegan*

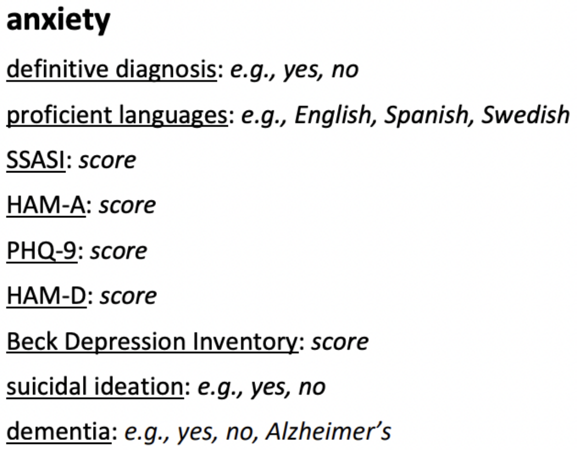
exercise: *e.g., walk 2 miles/day, limited to wheelchair*

ketoacidosis history: *e.g., yes, no, 2 years ago*

comorbidities: *e.g., type 1 diabetes, thyroid disorder, hypertension, chronic kidney disease*

hospitalization events: *e.g., myocardial infarction last year, stroke 6 months ago*

Figure 13: Question Template for Type 2 Diabetes disorder



anxiety

definitive diagnosis: *e.g., yes, no*

proficient languages: *e.g., English, Spanish, Swedish*

SSASI: *score*

HAM-A: *score*

PHQ-9: *score*

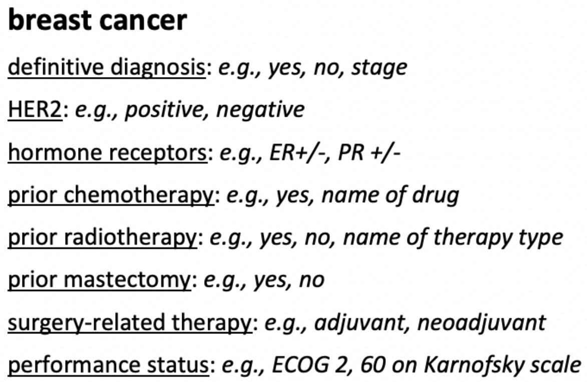
HAM-D: *score*

Beck Depression Inventory: *score*

suicidal ideation: *e.g., yes, no*

dementia: *e.g., yes, no, Alzheimer's*

Figure 14: Question Template for Anxiety disorder



breast cancer

definitive diagnosis: *e.g., yes, no, stage*

HER2: *e.g., positive, negative*

hormone receptors: *e.g., ER+/-, PR +/-*

prior chemotherapy: *e.g., yes, name of drug*

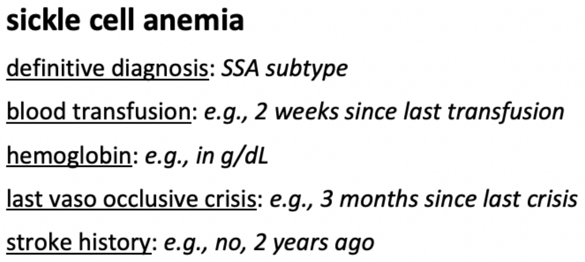
prior radiotherapy: *e.g., yes, no, name of therapy type*

prior mastectomy: *e.g., yes, no*

surgery-related therapy: *e.g., adjuvant, neoadjuvant*

performance status: *e.g., ECOG 2, 60 on Karnofsky scale*

Figure 15: Question Template for Breast Cancer disorder



sickle cell anemia

definitive diagnosis: *SSA subtype*

blood transfusion: *e.g., 2 weeks since last transfusion*

hemoglobin: *e.g., in g/dL*

last vaso occlusive crisis: *e.g., 3 months since last crisis*

stroke history: *e.g., no, 2 years ago*

Figure 16: Question Template for Sickle Cell Anemia disorder