CReSE: Benchmark Data and Automatic Evaluation Framework for Recommending Eligibility Criteria from Clinical Trial Information

EACL 2024

Siun Kim[1], Jung-Hyun Won[2], David Seung U Lee[2],
Renqian Luo[3], Lijun Wu[3], Tao Qin[3], Howard Lee[1,2]

[1] Seoul National University Hospital, [2] Seoul National University, [3] Microsoft Research

Motivation

Restrictive eligibility criteria (EC) in clinical trials could limit diverse participation, affecting the generalizability of trial findings and health equity.

There is a lack of an evaluation framework to assess the performance of EC recommendation and generation models from a clinical perspective.

What did we do?

- Developing benchmark data for EC recommendation task
- Suggesting an automatic evaluation framework for evaluating EC recommendation model from a clinical perspective

Clinical trial information: *Alpelisib in Pediatric Patients With Lymphatic Malformations Associated With a PIK3CA Mutation [SEP] <trial summary> [SEP] <design factors>*

→ **Formulated as binary classification (training)**

EC: *[exclusion] Systemic oral methylprednisolone or systemic oral methotrexate treatment for another reason*

→ **Yes** (this EC was used in the clinical trial of that title)

→ **Ranking EC using matching-scores (inference)**

Ranking	Recommended EC	Cluster IDs
1	[exclusion] fever, axillary temperature > 37.0 °C	54
2	[exclusion] any contraindication to methylprednisolone or methotrexate	23
...	...	...

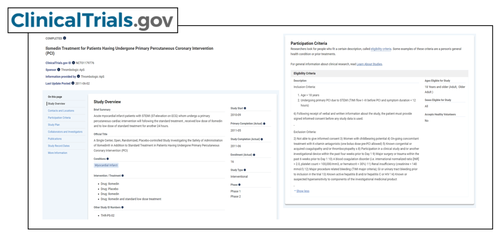
Cluster IDs of original EC
3, 13, 21, 23, 47, 54, 77, 97

→ **Measuring recommendation performances with EC clustering results via CReSE (evaluation)**

- *Precision@1: 43.0*
- *MAP@5: 40.2*
- *P@ECnumori: 29.7*

Introduce a task of recommending EC from clinical trial information, including trial titles, and provide an automatic evaluation framework to assess the clinical validity of the recommendation model

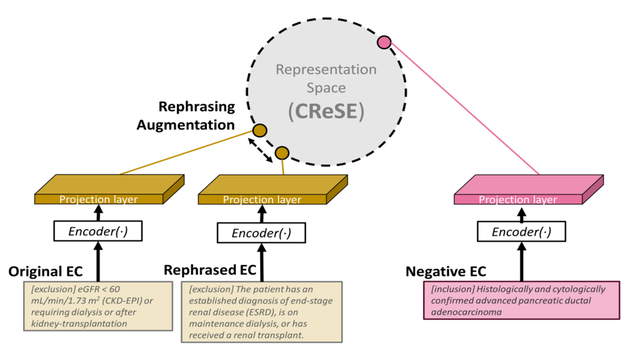
Benchmark data

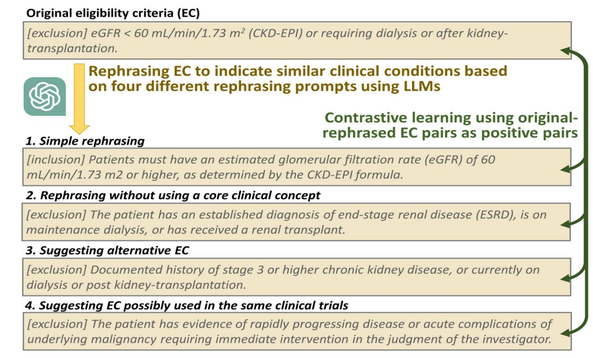


	Train-Valid	Test
Number of clinical trials	260K	10K
Number of EC (%)		
Total	2.8M (100.0)	176K (100.0)
Common	1.2M (44.4)	78K (44.3)
Non-common	1.6M (55.6)	98K (55.7)
Average number of EC per clinical trial	10.7	17.6
Length of EC in characters (mean ± SD)	117.8 ± 70.7	123.7 ± 73.0

Table 1: Statistics of clinical trials and eligibility criteria (EC) used in this study

CReSE: Contrastive learning and Rephrasing-based and Clinical Relevance-preserving Sentence Embedding





- ***Develop sentence embedding (CReSE) that preserve clinical relevance through contrastive learning***
- ***Use 4 different rephrasing prompts to obtain diverse original-rephrased EC pairs***

Key insights

EC clustering performance of CReSE model

Clustering methods	Spearman
TF-IDF	32.8 [26.8, 37.9]
Only embeddings	
BioLinkBERT	40.7 [37.5, 46.0]
TrialBERT	39.8 [34.6, 43.2]
BioSimCSE	46.2 [41.0, 50.4]
BioGPT	44.0 [40.6, 48.3]
CReSE (ours)	**59.9 [56.3, 63.3]**
BERTopic	
BioLinkBERT	46.1 [40.3, 51.4]
TrialBERT	47.4 [43.4, 50.1]
BioSimCSE	45.5 [39.6, 54.9]
BioGPT	37.7 [32.5, 46.1]
CReSE (ours)	**60.4 [53.0, 64.7]**

Model	Spearman	Pearson
BioSimCSE	86.7	86.7
CReSE (ours)	**84.7**	**80.7**
BioSentVec	78.0	81.7
BioGPT	72.1	70.2
BioBART	69.5	67.7
BioClinicalBERT	65.2	65.2
BioBERT	63.8	66.2

Table 3: Results on BIOSSES

- ***Our model outperformed other biomedical LMs in EC clustering and well represented semantics in the biomedical domain (BIOSSES).***

Ablation study results on CReSE model

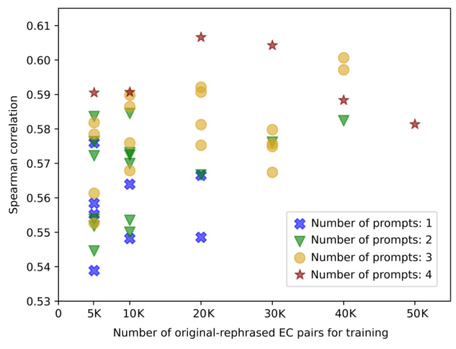
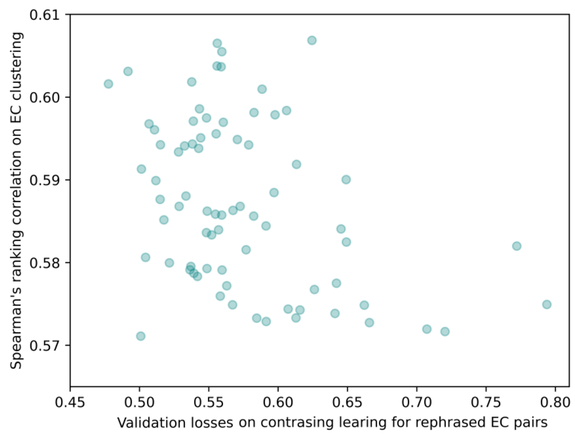


Figure 3: Clustering performance of the CReSE model by the number of rephrasing prompts used to generate a dataset of original-rephrased EC pairs and the size of the dataset



- ***Utilizing multiple rephrasing prompts is important for training CReSE model, rather than having a larger number of original-rephrase EC data.***

High-quality benchmark dataset for EC recommendation

- ***Based on expertise in clinical trials, we processed clinical trials and ECs intended for use in benchmark data to ensure that the EC recommendation task is defined within a consistent context, leading to the provision of a valid clinical advice.***

- **Data preprocessing**
  - **Clinical trial selection**
    1) trials where trial information was uploaded between March 2002 and May 2023
    2) trials categorized as interventional
    3) trials where 'brief summary' information and 'official title' are available
    4) trials with at least two eligibility criteria reported
    5) trials where the intervention is classified as 'Drug' or 'Biological' (excluding trials on medical devices and behavioral therapies)
  - **EC preprocessing**
    1) Exclude too short or too long EC from training data (3 or more characters, 353 or less)
    2) Define common ECs and develop a classification model to exclude them from training data
    3) Try to improve the quality of negative EC-title pairs

- **Efforts to improve the quality of negative EC-title pairs in benchmark data for EC recommendations**
  - generated negative EC-title pairs basically by random sampling of EC and trial titles
  - Since an identical or similar EC are used in different clinical trials, simply applying random sampling to obtain a negative sample cause a quality issue
  - **two steps to obtain a quality negative sample:**
    1) selected trials where the number of ECs exceeds a predefined threshold (i.e., 8) to ensure the quality of EC reporting
    2) created an EC-title negative sample by randomly sampling EC whose clusters do not overlap with EC used in a selected trial (clustering was based on the CReSE model)

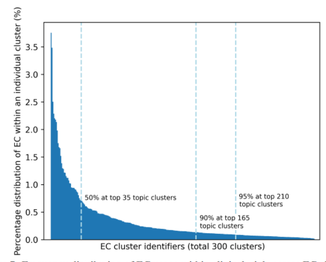


Figure 7: Frequency distribution of EC usage within clinical trials across EC clusters.

Common EC Type	Definitions and Examples
Used as a template over time	All age restrictions, about patient sex, weight, or BMI range restriction without clinical justification. *Ex) "[Inclusion] age 18 years", "[Inclusion] males and females", "[Inclusion] Body Mass Index (BMI) 18.5 kg/m2 and 28 kg/m2"*
Infant/Child Protection	To protect infant and child from the investigational drug (mostly exclusion criteria): pregnancy, breast-feeding, willing to take contraceptives. *Ex) "[Exclusion] pregnancy or breastfeeding", "[Inclusion] males and females of childbearing potential must agree to utilize highly effective contraception methods from screening"*
Drug addiction and alcoholism	To exclude patients with a current or past history of drug addiction. *Ex) "[Exclusion] excessive alcohol, opiate, or barbiturate use; history of drug abuse or dependence"*

Table 15: (continued) Types of common EC and their definitions and examples

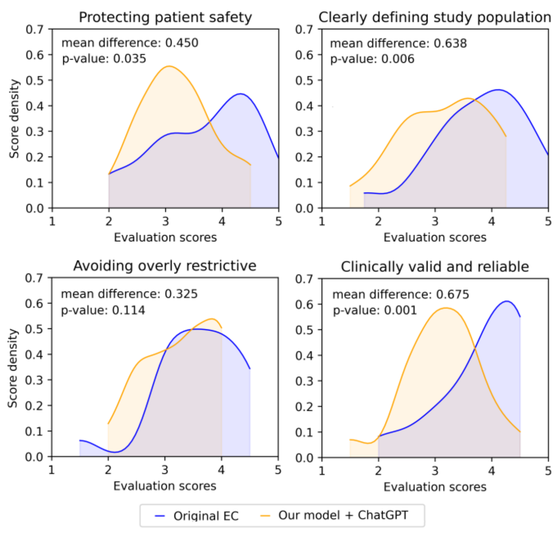
Model name	Binary classification performances (%)			
	Accuracy	Precision	Recall	F1
BERT-base	89.30	83.56	93.85	88.41
BioClinicalBERT	95.99	98.36	92.31	95.24
BioBERT	97.32	95.41	95.38	96.88
BioLinkBERT	97.99	98.51	97.06	97.78
ELECTRA	82.61	86.26	76.88	81.29
XLM-RoBERTa	85.28	79.49	82.30	80.87

Table 9: Performances of common eligibility criteria classifiers

Baseline EC recommendation performance

Input type	Binary classification				EC recommendation		
	Accuracy	Precision	Recall	F1	P@1	MAP@5	P@ECnumori
title only	81.6	80.3	83.8	82.0	37.0	29.5	23.7
title + summary	93.1	92.6	93.7	93.1	47.0	41.2	30.0
title + design factors	92.2	91.8	92.7	92.2	46.0	40.4	31.5
title + summary + design factors	93.1	92.6	93.7	93.1	49.0	44.2	29.6
ChatGPT	42.3	78.6	13.9	23.7	NA	NA	NA
GPT-4	75.6	92.9	31.0	46.4	NA	NA	NA
random recommendation	NA	NA	NA	NA	11.3 [6.0, 19.0]	11.5 [8.3, 15.0]	11.6 [10.1, 13.6]

	P@1	MAP@5	P@ECnumori
Posted date			
May 2002 - Dec 2009	25.0 (8.6)	20.8 (8.9)	18.2 (9.0)
Jan 2010 - COVID	31.0 (10.0)	25.4 (9.9)	19.0 (9.7)
COVID - May 2023	59.0 (8.9)	48.6 (9.3)	33.4 (9.3)
Therapeutic area			
Oncology	56.0 (9.9)	42.1 (10.2)	28.7 (10.5)
Neurology	52.0 (9.0)	38.6 (8.9)	29.0 (9.0)
Metabolic disease	49.0 (9.1)	44.8 (9.0)	33.1 (8.8)
Cardiology	47.0 (8.1)	37.5 (8.2)	27.7 (8.1)
Rheumatology	46.0 (8.5)	30.9 (8.6)	20.6 (8.5)
Infectious disease	45.0 (8.1)	38.3 (8.2)	25.8 (8.3)
Hematology	40.0 (9.2)	32.6 (9.1)	23.1 (9.0)
Immunology	34.0 (9.2)	29.2 (9.6)	22.9 (9.6)
Dermatology	33.0 (7.4)	26.5 (7.7)	23.6 (8.0)
Nephrology	32.0 (8.6)	31.2 (8.6)	24.7 (8.7)
Pulmonology	28.0 (8.5)	26.6 (9.7)	29.5 (8.8)
Gastroenterology	21.0 (8.9)	23.2 (9.0)	20.6 (9.1)



- ***The EC recommendation model demonstrated an accuracy of up to 90% and a P@1 close to 50% in binary classification and recommendation settings, respectively, with 100 EC clusters, greatly surpassing the performance of ChatGPT and GPT-4.***
- ***However, in physician evaluation, the full EC sets recommended by our model were not clinically valid and did not ensure patient safety, unlike those used in real clinical trials.***