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

EACL 2024

Siun Kim¹, Jung-Hyun Won², David Seung U Lee², Rengian Luo³, Lijun Wu³, Tao Qin³, Howard Lee^{1,2}



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





Motivation

K	
C	У

Microsoft[®]

Research

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

Benchmark data

TED 0				
edin Treatment for P)	Patients Having Undergone Primary Percutaneous Coronary Intervention		Participation Criteria Researchers look for people who fit a certain description, called <u>eligibility criteria</u> . Some examples of th health condition or prior treatments.	ese criteria are a person's general
calTrials.gov ID ① NCT01179	3776		For general information about clinical research, read Learn About Studies.	
• • Incompologic ApS	and		Eligibility Criteria	
Indate Posted @ 2011-06-02	2		Description	Ages Eligible for Study
2011-00-0			Inclusion Criteria:	18 Years and older (Adult, Older
his page	Study Commission		1. Age > 18 years 2. Undergoing primary PCI due to STEMI (TIMI flow I -III before PCI and symptom duration < 12	Adult) Sexes Eligible for Study
dy Overview	Study Overview	Stude Start	hours)	All
cts and Locations	Brief Summary	2010-09	6) Following receipt of verbal and written information about the study, the patient must provide	Accepts Healthy Volunteers
ticipation Criteria	percutaneous cardiac intervention will following the standard treatment, received low dose of llomedin		signed informed consent before any study data is used.	No
dy Plan	and to low dose of standard treatment for another 24 hours.	Primary Completion (Actual)		
aborators and Investigators	Official Title	2011-05	Exclusion Criteria:	
fications	A Single Center, Open, Randomized, Placebo-controlled Study Investigating the Safety of Administration	Study Completion (Actual)	 Not able to give informed consent 3) Women with childbearing potential 4) On-going concomitant treatment with Kwitemia anteoprists (one holes does one-PCI allowed) S) Kosura consectal or 	
dy Record Dates	of Ilomedin® in Addition to Standard Treatment in Patients Having Undergone Primary Percutaneous Coronary Intervention (PCI)	2011-06	acquired coagulopathy and/or thrombocytopathy s 8) Participation in a clinical study and/or another	
re Information	Conditions A	Enrollment (Actual) 🛛	investigational device within the past four weeks prior to Day 1 9) Major surgery or trauma within the	
	Myocardial Infarct	16	> 2.0, platelet count < 100,000/mm3, or hematocrit < 30%) 11) Renal insufficiency (creatinine > 140	
		Study Type 0	mmol/i) 12) Major procedure related bleeding (TIMI major criteria); GI or urinary tract bleeding prior	
	Intervention / Treatment 0	Interventional	to inclusion in the trial 13) Known active hepatitis B and/or hepatitis C or HIV 14) Known or suspected hypersensitivity to components of the investigational medicinal product.	
	Drug: Ilomedin	Phase 0		
	Drug: Placebo Drug: Ilemedia	Phase 1	- Show less	
	Drug: Ilomedin and standard low dose treatment	Phase 2		
	ound over a terminete •			

	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 \pm SD)	117.8 ± 70.7	123.7 ± 73.0

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] <<u>trial summary</u>> [SEP] <<u>design factors</u>>

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)



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

CReSE: Contrastive learning and <u>Rephrasing-based</u> and <u>Clinical Relevance-preserving Sentence Embedding</u>



- 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

Ablation study results on CReSE model

Clustering methods	Spearman			
TF-IDF	32.8 [26.8, 37.9]			
Only embeddings		Model	Spearman	Pearson
BioLinkBERT	40.7 [37.5, 46.0]	BioSimCSE	86.7	86.7
TrialBERT	39.8 [34.6, 43.2]	CReSE (ours)	84.7	80.7
BioSimCSE	46.2 [41.0, 50.4]	BioSentVec	78.0	81.7
BioGPT	44.0 [40.6, 48.3]	BioGPT	72.1	70.2
CReSE (ours)	59.9 [56.3, 63.3]	BioBART	69.5	67.7
BERTopic		BioClinicalBERT	65.2	65.2
BioLinkBERT	46.1 [40.3, 51.4]	BioBERT	63.8	66.2
TrialBERT	47.4 [43.4, 50.1]			
BioSimCSE	45.5 [39.6, 54.9]	Table 3: Res	sults on BIOSS	ES
BioGPT	37.7 [32.5, 46.1]			
CReSE (ours)	60.4 [53.0, 64.7]			

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

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.



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

Baseline EC recommendation performance

	Ri	nary classif	ication		E	[¬] recommen	dation		P@1	MAP@5	P@ECnumori	
Input type		nary classifi	ication					Posted date				
	Accuracy	Precision	Recall	F1	P@1	MAP@5	P@ECnumori	May 2002 - Dec 2009	25.0 (8.6)	20.8 (8.9)	18.2 (9.0)	
441	01 (00.2	02.0	00.0	27.0	20.5	22.7	Jan 2010 - COVID	31.0 (10.0)	25.4 (9.9)	19.0 (9.7)	
title only	81.0	80.3	83.8	82.0	37.0	29.5	23.7	COVID - May 2023	59.0 (8.9)	48.6 (9.3)	33.4 (9.3)	
title + summary	93.1	92.6	93.7	93.1	47.0	41.2	30.0	Therapeutic area				
the isometry	>5.1	2.0	20.7	<i>)).</i> 1	17.0	11.2	50.0	Oncology	56.0 (9.9)	42.1 (10.2)	28.7 (10.5)	
title + design factors	92.2	91.8	92.7	92.2	46.0	40.4	31.5	Neurology	52.0 (9.0)	38.6 (8.9)	29.0 (9.0)	
title + summary								Metabolic disease	49.0 (9.1)	44.8 (9.0)	33.1 (8.8)	
title + summary	93.1	92.6	93.7	93.1	49.0	44.2	29.6	Cardiology	47.0 (8.1)	37.5 (8.2)	27.7 (8.1)	
+ design factors	2011	12.0	20.1	<i>JJII</i>	12.0	11.2	29.0	Rheumatology	46.0 (8.5)	30.9 (8.6)	20.6 (8.5)	
	40.2	70 6	12.0	027	NIA	NTA	ΝΙΑ	Infectious disease	45.0 (8.1)	38.3 (8.2)	25.8 (8.3)	
ChatGPT	42.3	/8.0	13.9	23.7	INA	NA	NA	Hematology	40.0 (9.2)	32.6 (9.1)	23.1 (9.0)	
GPT-4	75.6	92.9	31.0	46.4	NA	NA	NA	Immunology	34.0 (9.2)	29.2 (9.6)	22.9 (9.6)	
	75.0	, 2.,	21.0	10.1	111	1.1.7		Dermatology	33.0 (7.4)	26.5 (7.7)	23.6 (8.0)	
random	NT A	NT A	NTA	NT A	11.3	11.5	11.6	Nephrology	32.0 (8.6)	31.2 (8.6)	24.7 (8.7)	
1	INA	INA	INA	INA	FC 0 10 01		F10 1 10 C1	Pulmonology	28.0 (8.5)	266(97)	29 5 (8 8)	

 $[6.0, 19.0] \quad [8.3, 15.0] \quad [10.1, 13.6]$

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

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
- 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 <u>randomly sampling EC whose clusters do not overlap</u> with EC used in a selected trial (clustering was based on the CReSE model)

- 5.0 distribution of EC	50% at top 35 topic clusters 90% at top 165 topic clusters 90% at top 165 topic clusters	
Per	EC cluster identifiers (total 300 clusters)	

[₩] 2.0 -

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 fe-
	males", "[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.
	<i>Ex) "[Exclusion] pregnancy or breastfeeding", "[Inclusion]</i>
	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 addic-
	tion.
	<i>Ex) "[Exclusion] excessive alcohol, opiate, or barbiturate use;</i>
	history of drug abuse or dependence"

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

Model nome	Binary classification performances (%)							
would name	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				



recommendation

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.

Pulmonology

Gastroenterology

28.0 (8.5)

26.6 (9.7)

21.0 (8.9) 23.2 (9.0)

29.5 (8.8)

20.6 (9.1)

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.

Table 9: Performances of common eligibility criteria classifiers