CReSE: Enhancing Clinical Trial Design via Contrastive Learning and Rephrasing-based and Clinical Relevance-preserving Sentence Embedding

Anonymous ACL submission

Abstract

Eligibility criteria (EC) refer to a set of conditions an individual must meet to participate in a clinical trial, defining the study population and minimizing potential risks to patients. Previous research in clinical trial design has been primarily focused on searching for similar trials and generating EC within manual instructions, employing similarity-based performance metrics, which may not fully reflect human judgment. In this study, we propose a novel task of recommending EC based on clinical trial information, including trial titles, and introduce an automatic evaluation framework to assess the clinical validity of the EC recommendation model. Our new approach, known as **CReSE** (**C**ontrastive learning and **Re**phrasing-based and **C**linical **Re**levance-preserving **S**entence **E**mbedding), represents EC through contrastive learning and rephrasing via large language models (LLMs). The CReSE model outperforms existing language models pre-trained on the biomedical domain in EC clustering. Additionally, we have curated a benchmark dataset comprising 3.2M high-quality EC-title pairs extracted from 270K clinical trials available on ClinicalTrials.gov. The EC recommendation models achieve commendable performance metrics, with 49.0% precision@1 and 44.2% MAP@5 on our evaluation framework. We expect that our evaluation framework built on the CReSE model will contribute significantly to the development and assessment of the EC recommendation models in terms of clinical validity.

1 Introduction

Eligibility criteria (EC) consist of statements that outline the characteristics participants must possess to be included in a randomized controlled trial (RCT) (FDA, 2020). EC are typically divided into inclusion and exclusion criteria, covering diverse clinical factors such as age, sex, medical history, disease severity, previous treatments, and other physiologic parameters (Duggal et al., 2021). They

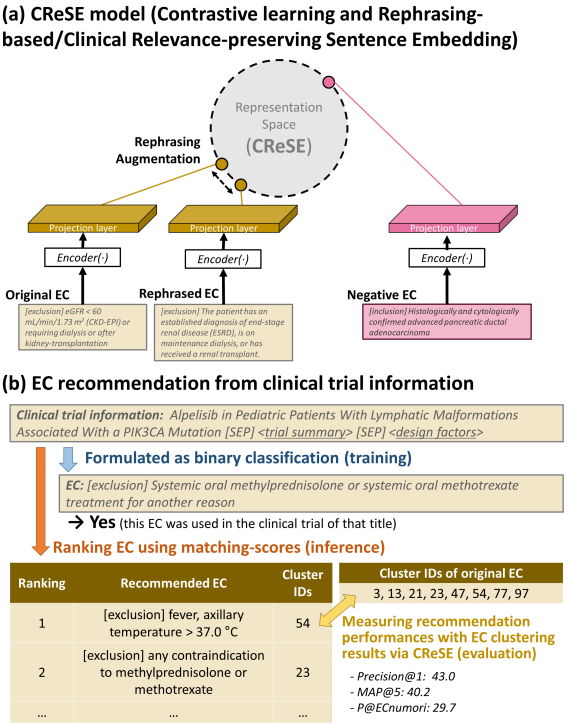

Figure 1: Study overview. a) We develop the CReSE model using contrastive learning and text rephrasing via LLMs to obtain a sentence embedding that preserves clinical relevance between EC. b) We introduce a task of recommending EC from clinical trial information, including trial titles, and provide an automatic evaluation framework to assess the clinical validity of the EC recommendation model using the CReSE model.

are a key design factor of RCTs, along with randomization and blinding, which contribute to the production of causal evidence between intervention and outcome (Akobeng, 2005; Listl et al., 2016). Moreover, EC are an important component of the enrichment strategy and minimize potential risk to study participants (Kim et al., 2017; FDA, 2023).

However, there are concerns that EC are overly restrictive (Breithaupt-Groegler et al., 2017; Osarogiagbon et al., 2021). While restrictive EC ensure homogeneity in the study population (Kim et al.,

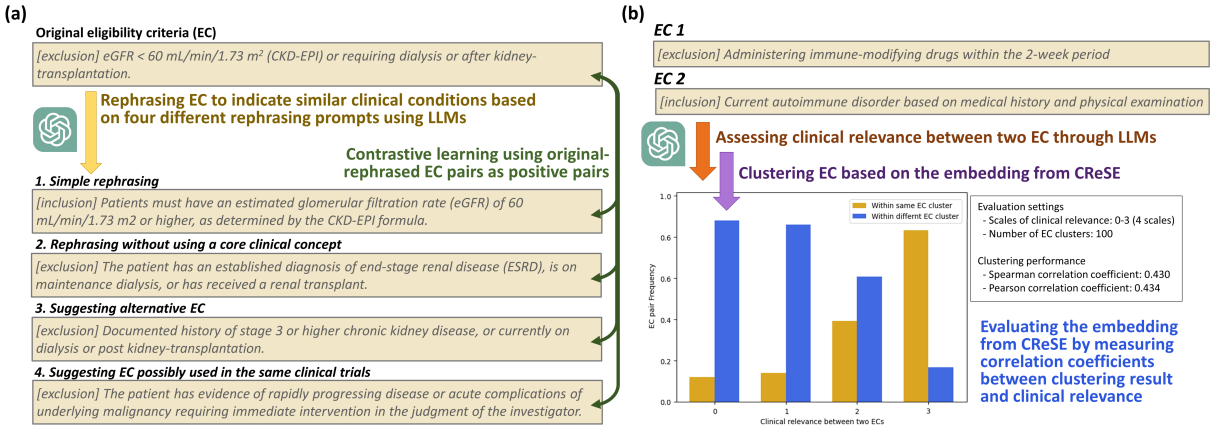

Figure 2: Overview of the development and evaluation of the CReSE model. a) Original EC and their rephrased counterparts generated from four different rephrasing prompts are used as positive pairs in contrastive learning. b) Correlation coefficients between clustering results and clinical relevance assessed by LLMs are employed as the clustering performance measures.

2021), they may also limit the generalizability of clinical findings and impede the translation of research results into clinical practice. Furthermore, the EC used by previous RCTs are often employed as templates for new trials without appropriate modifications (FDA, 2020). This practice can perpetuate issues such as the under-representation of specific patient subgroups (e.g., children, the elderly, and individuals with infections like HIV infection) (Humphreys et al., 2007; Uldrick et al., 2017).

To overcome these problems, previous studies attempted to automate EC generation or search for similar trials to aid in clinical trial design (Wang et al., 2023b,a; Wang and Sun, 2022). However, these studies relied on similarity-based performance metrics, which do not account for human judgment and clinical semantic similarity (Gehrmann et al., 2023; Moramarco et al., 2022). Furthermore, certain EC, such as age requirements, are widely employed across studies and are less specific to the purposes and designs of clinical trials (Jin et al., 2017; Magnuson et al., 2021). The presence of these common EC may have led to an overestimation of the model's performance.

In response, this study aims to recommend EC from clinical trial information, such as titles and summaries, to meet the needs of drug development and clinical evidence generation (Figure 1b). In addition, we propose an automatic evaluation framework to assess the clinical validity of EC recommendation models. To accomplish this, we develop sentence embedding, called **CReSE** (**C**ontrastive learning and **Re**phrasing-based and **C**linical **Re**levance-preserving **S**entence **E**mbedding) (Figures 1a and 2). Lastly, we investigate the characteristics that EC recommendation models should possess to be useful in clinical trial design for drug development, as discerned through human evaluation.

To the best of our knowledge, this study is the first attempt to formulate the EC recommendation task. Additionally, in this study, we explored the diverse utility of LLMs in handling biomedical texts in a clinically plausible manner, including rephrasing EC to develop sentence embedding, assessing clinical relevance for model evaluation, and streamlining the EC recommendation model into an end-to-end recommendation system.

The main contributions of this paper are as follows:[1]

- We propose a task of recommending EC from clinical trial information without any manual instruction.
- We develop a sentence embedding preserving clinical relevance between EC, called CReSE, through contrastive learning and text rephrasing via LLMs.
- Based on the CReSE model, we develop an automatic evaluation framework assessing the clinical validity of the EC recommendation models.

[1]All data and code used in this study are available at https://anonymous.4open.science/r/clinical_trial_eligibility_criteria_recommendation-4B86.

2 Related Works

Natural language processing research on EC has taken two main paths. The first approach focuses on converting free-text EC into structured criteria or queries using information extraction or context-free grammars (Weng et al., 2011; Kang et al., 2017; Yuan et al., 2019). These studies, known as 'patient-trial matching', ultimately aim to estimate the number of patients who match a proposed trial design based on in-hospital electronic medical records (EMRs) before patient enrollment (Zhang et al., 2020). However, a challenge in this approach is the lack of consensus on a universal query grammar for EC (Tu et al., 2009; Boland et al., 2012; Hao et al., 2016).

The second research stream involves studies that generate EC with manual instruction or search for similar trials to aid in clinical trial design (Zhang et al., 2020; Wang and Sun, 2022; Wang et al., 2023b,a; Jin et al., 2023). The AutoTrial study, for instance, proposed a hybrid approach that combines discrete and neural prompting in generating EC (Wang et al., 2023b). Furthermore, the PyTrial study aimed to create a unified Python package that incorporates diverse AI algorithms for tasks related to clinical trials (Wang et al., 2023a). However, to this date, no study has endeavored to recommend EC exclusively from clinical trial information without manual instructions. Moreover, previous studies have relied on traditional summarization metrics, such as BLEU or ROUGE, and EC parsers in evaluating their models (FAIR, 2022). However, these metrics are still insufficient for measuring clinical semantic similarity between EC, and clinical trial parsers have limited performances on complex EC (Gehrmann et al., 2023; Moramarco et al., 2022).

3 Method

3.1 Common EC classification

In clinical trials, certain EC, such as "age over 18" or "Patients must provide written, informed consent before any study procedures" are widely used in clinical trials, irrespective of the trial's objectives or designs. (Duggal et al., 2021). We refer to these commonly used EC as 'common EC.' Throughout this study, we exclude common EC to prevent potential overestimation of the EC recommendation model's performance and to enhance the heterogeneity of the EC dataset for contrastive learning (Appendix B.1, D.1, E.1).

3.2 The CReSE model

3.2.1 Prompts for rephrasing EC

We employed contrastive learning and rephrasing via LLMs as text augmentation to develop the CReSE model. To capture the diverse clinical relevance within EC, we devised four different types of rephrasing prompts (Figure 2a), each serving a specific purpose:

- **Simple rephrasing** This prompt involves a direct rewording of the input EC. Its purpose is to account for differences in EC description across clinical trials, even when conveying the same content.
- **Rephrasing without core clinical concepts** With this prompt, we aimed to integrate the meaning and context of clinical concepts frequently used in EC into the CReSE model.
- **Suggesting alternative EC** This prompt explores clinical relevance based on the epidemiological co-occurrence among different patient conditions.
- **Suggesting EC possibly used in the same clinical trial** This prompt aids in generating EC variations that might be used within the same clinical trial.

We utilized the ChatGPT model, specifically gpt-3.5-turbo, for EC rephrasing. We obtained a total of 50K original-rephrased EC pairs, which were used as positive pairs during contrastive learning (Appendix A.1).

3.2.2 Contrastive learning

The CReSE model consists of a text encoder and a projection layer. We utilized the embedding of the [CLS] token, which was obtained after passing through both the text encoder and the projection layer, as the EC embedding. The training process of the text encoder was initialized from pre-trained checkpoints of BioLinkBERT (Yasunaga et al., 2022), which exhibited superior performance in classifying common EC among diverse language models (LMs) used in fine-tuning (Appendix D.1).

The CReSE model was trained by maximizing the cosine similarity between embeddings of N positive pairs and minimizing the cosine similarity of $N^2 - N$ negative pairs within a batch of N EC pairs. This training methodology follows the approach used in the CLIP study (Radford et al.,

2021). The symmetric cross-entropy loss was used during this training process. Given the notable diversity in the original EC dataset, already achieved through the exclusion of common EC, we chose not to introduce additional techniques for sampling negative pairs.

3.3 EC Recommendation Model

We formulated the EC recommendation task as a binary classification, where a pair of individual EC and free-text clinical trial information served as input. The objective is to predict whether a given EC was used in a clinical trial with a specific title and trial information. The positive EC-title pairs consisted of 1.6M non-common EC selected from ClinicalTrials.gov.

The negative EC-title pairs were basically generated by random sampling of EC and trial titles. However, to ensure the quality of negative EC-title pairs, a random EC-title pair was included only if the sampled EC did not correspond to any EC used in the specified trial in terms of EC cluster. Here, EC clustering was conducted using EC embeddings derived from the CReSE model, described in Section 4.2.

Moreover, because relying solely on the title might not provide sufficient information to predict whether an EC was used in a clinical trial, we explored four different types of clinical trial information as input: 1) title only, 2) title + summary, 3) title + key design factors, and 4) title + summary + key design factors (Appendix C.2).

4 Experiments

4.1 Dataset

In this study, we collected trial information of 445K clinical trials registered on ClinicalTrials.gov from March 2002 to May 2023. From this initial dataset, we selected trials that satisfied several conditions (Appendix B.3) to ensure the quality of reported clinical trial information, resulting in a subset of 270K trials and 3M EC (Table 1). We set the positive-negative sample ratio to 1:1, so the total number of EC-title pairs used in training is 3.2M.

4.2 EC clustering

For EC clustering, we randomly selected a subset of 0.1M EC from the training dataset. To address randomness in the EC selection, we carried out each experiment 20 times using different seed numbers. The results were summarized using the median and the 95% confidence interval of clustering performances. Additionally, due to the significance of the cluster number on performance metrics, we evaluated EC clustering across different numbers of EC clusters (100, 200, and 300).

4.2.1 TF-IDF

To provide a simple baseline, we employed the TF-IDF (Term Frequency-Inverse Document Frequency) approach along with K-means clustering. Stopwords frequently used in EC were excluded before clustering.

4.2.2 Clustering using EC embeddings

For obtaining EC embeddings, we applied mean pooling to the token embeddings of each individual EC. Subsequently, we performed K-means clustering using cosine similarity as the distance measure between EC embeddings. We compared the CReSE model against several LMs pre-trained on the biomedical domain: BioLinkBERT (Yasunaga et al., 2022), BioGPT (Luo et al., 2022), TrialBERT (Wang and Sun, 2022), and BioSimCSE (Kanakarajan et al., 2022).

4.2.3 BERTopic

To further explore the potential of using text embeddings for clustering, we adopted the BERTopic model, specifically designed for topic clustering based on transformer-based sentence embeddings (Grootendorst, 2022). In the default configuration of BERTopic, text embeddings generated by sentence-transformer (Reimers and Gurevych, 2019) undergo dimensional reduction with UMAP (McInnes et al., 2018) and are subsequently clustered using HDBSCAN (McInnes et al., 2017).

4.3 Evaluation Strategy

4.3.1 CReSE

To assess EC embeddings from the CReSE model, we measured the correlation coefficients between the clinical relevance scores of EC pairs and whether they were assigned to the same EC cluster (Figure 2b). We utilized two correlation measures, Spearman's and Pearsons's, with a preference for Spearman's ranking correlation as the primary performance metric. To obtain the clinical relevance scores, we utilized ChatGPT, specifically gpt-3.5-turbo. In our prompt, we instructed ChatGPT to evaluate the clinical relevance scores for a given EC pair based on a 4-point scale ranging from 0 to 3 (Appendix A.2).

	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

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

4.3.2 EC recommendation model

We evaluated the EC recommendation model in two ways. Firstly, we assessed its performance as a binary classifier, using metrics like accuracy, precision, recall, and F1-score. This evaluation aimed to determine the model's ability to predict whether a given EC was used in a clinical trial of a given title.

Secondly, we evaluated the model's recommendation performance based on the EC clustering results. Here, the objective was to determine how accurately the models suggest the most relevant EC cluster from clinical trial information. We reported precision@1, MAP@5 (mean average precision at top 5), and precision@ECnumori as performance measures. ECnumori denotes the number of EC originally used in clinical trials. By definition, precision@ECnumori is equivalent to recall@ECnumori. In evaluating EC recommendation performances, the true labels are the identifiers of EC clusters that correspond to EC actually used in clinical trials.

4.3.3 Human evaluation

We conducted a human evaluation to assess the feasibility of the current EC recommendation model in providing a complete EC set to aid in clinical trial design. Two experienced senior physicians working in a pharmaceutical company, with extensive knowledge in clinical trial design and execution, participated in the assessment. The evaluation encompassed four categories: 1) Protecting patient safety, 2) Clearly defining the study population, 3) Avoiding overly restrictive, 4) Clinically valid and realistic (Appendix E.2). For comparison, we prepared two types of complete EC sets for given trial titles: 1) the original EC set used in clinical trials and 2) the EC set recommended by our model.

Since our EC recommendation model primarily focuses on non-common EC and ranks candidate EC based on given trial information, there was a limitation in using it to create a complete EC set. To address this issue, we engaged in prompt engineering to propose a complete EC set that would complement the non-common EC recommended by our model (Appendix A.3). The evaluation covered 20 clinical trials uploaded on ClinicalTrials.gov after September 2021, which was the knowledge-cutoff date of ChatGPT.

4.4 Results

4.4.1 CReSE

Regardless of the clustering method or the number of EC clusters, the CReSE model consistently exhibited superior performance in EC clustering performance compared to other LMs pre-trained in the biomedical domain (Table 2 and Appendix D.3). Moreover, within the BIOSSES dataset, the CReSE model demonstrated the second-highest semantic similarity performance, ranking just below BioSimCSE (Table 3).

In the ablation study, we observed the CReSE model was generally improved when using a more diverse range of rephrasing prompts for the same size of the training dataset (Figure 3). Meanwhile, it was noted that the performance of the CReSE model decreased when using all four rephrasing prompts as the dataset size increased beyond 20K while using three prompts yielded better results than using all four prompts for a dataset size of 40K. In addition, an inverse correlation between validation loss in contrastive learning and clustering performance was observed, although it is not distinctly evident (Appendix D.2).

These findings imply that while rephrasing through LLMs does indeed function as an effective text augmentation method in contrastive learning, aimed at incorporating medical knowledge from LLMs into embedding systems, there remains a need to discover the optimal composition of the

dataset containing the original-rephrased text pairs. Furthermore, it is clear that there is a difference between the objectives of contrastive learning, where a model predicts whether an EC pair is generated through rephrasing or not, and the assessment of clinical relevance between an EC pair. Thus, when employing rephrasing-via-LLMs as a text augmentation technique, the design of diverse rephrasing prompts becomes crucial.

Clustering methods	Spearman
TF-IDF	27.7 [23.4, 31.4]
Only embeddings	
Base-BERT	25.3 [21.4, 29.4]
BioLinkBERT	29.9 [24.9, 34.3]
TrialBERT	29.0 [24.7, 33.0]
BioSimCSE	34.7 [31.1, 38.1]
BioGPT	32.3 [28.8, 33.9]
CReSE (ours)	**43.0 [40.3, 45.3]**
BERTopic	
Package default	40.9 [35.6, 46.2]
BioLinkBERT	36.2 [29.7, 42.5]
TrialBERT	37.6 [33.4, 44.7]
BioSimCSE	40.8 [35.5, 43.3]
BioGPT	32.2 [25.4, 37.8]
CReSE (ours)	**44.9 [40.9, 48.4]**

Table 2: Comparison of the CReSE model and other biomedical language models in EC clustering.

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

4.4.2 EC recommendation model

In binary classification, we achieved an accuracy of 81.6% and an F1-score of 82.0% when using only titles as input (Table 4). Moreover, providing additional trial information to trial titles resulted in a significant improvement, pushing the accuracy and F1-score to over 92%.

When evaluating recommendation performances using our evaluation framework, we achieved precision@1, MAP@5, and precision@ECnumori of

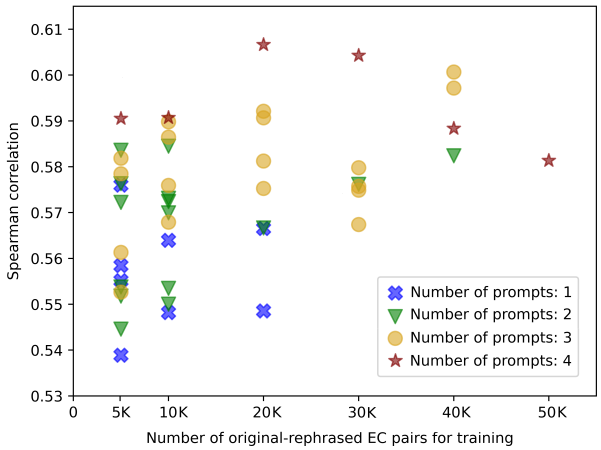

Figure 3: Clustering performance of the CReSE model by the number of rephrasing prompts used to generate a dataset of original-rephrased EC pairs and the size of the dataset

49.0%, 44.2%, and 31.5%, respectively (Table 4). These scores outperformed random recommendations by a substantial margin.

Moreover, when comparing the EC recommendation performance across time periods, we observed that the recommendation model exhibited better results for more recent clinical trials (Table 5). Furthermore, the model performances varied significantly depending on the therapeutic area of trials. These variations are not attributed to the number or distribution of EC within each category, because the performance of random recommendation showed no significant difference within categories. Instead, we attribute these differences to the fact that recent trials provide more specific titles and summaries for guessing EC used in the trials, while EC might be used in a more predictable manner in certain therapeutic areas.

4.4.3 Human evaluation

In the three remaining categories, except the one related to overly restrictive, the EC set proposed by our model demonstrated inadequacy when compared to the original EC set (p-value < 0.05, Figure 4). To be specific, the EC set recommended by our model performed poorly in properly protecting patient safety and building a clinically valid EC set, with statistically significant differences of 0.638 and 0.675, respectively. (Appendix D.4)

Furthermore, through consulting with the evaluators, we identified several features that can enhance the practicability of EC recommendation models for clinical trial design in the context of drug development. These proposed features are outlined as

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
random recommendation	NA	NA	NA	NA	11.3 [6.0, 19.0]	11.5 [8.3, 15.0]	11.6 [10.1, 13.6]

Table 4: Performances of the EC recommendation models using different input types on binary classification and EC recommendation. The evaluation metrics for EC recommendation were P@1 (precision at 1), MAP@5 (mean average precision at 5), and P@ECnumori (precision at the number of original EC in trials). We present the median and 95% confidence interval of performances achieved by randomly recommending EC, which helps gauge the task's difficulty.

follows:

- Incorporating the drug's mode of action (MoA) and findings from pre-clinical trials into the recommendation model becomes essential to assist in facilitating clinical trial design for drug development.
- Recognizing the sensitivity of the clinical trial design to regulatory shifts, it would be advantageous for the EC recommendation model to integrate regulatory guidance as one of its inputs.
- Developing a model to propose a suitable standard-of-care (SoC) treatment as a comparator along with suggesting the relevant supporting documents would carry significant value.

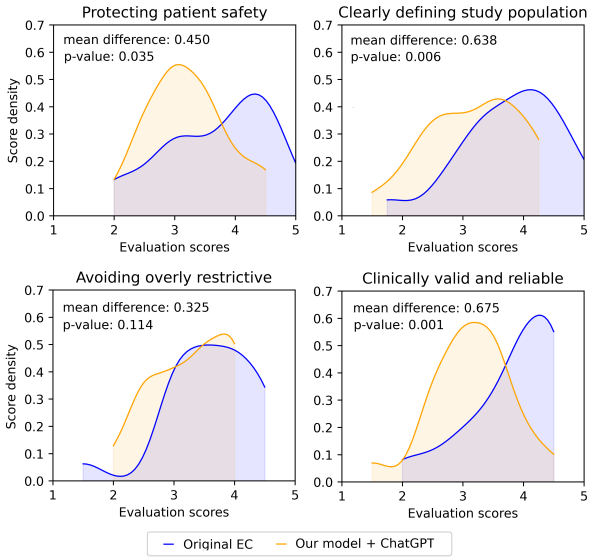

Figure 4: Distribution of human evaluation scores for original EC and EC recommended by our model with ChatGPT in four evaluation categories

5 Limitations

Despite these achievements, we want to underscore several considerations for evaluating the EC recommendation models and applying the automatic evaluation framework in a more clinically valid manner.

First of all, since the evaluation framework heavily relies on EC clustering results, researchers must be aware of the conditions under which clustering was executed. Our evaluation framework is based on all EC used in clinical trials, irrespective of the trial's therapeutic area. Thus, for example, exclusion criteria about cancer diagnosis before trial participation were mainly grouped into the same cluster. However, if you plan to employ EC recommendation in designing an oncology trial for an anticancer drug, a more finely-grained clustering result in terms of previous cancer diagnosis might be necessary. In such cases, it would be more fitting to develop EC recommendation and evaluation framework exclusively based on EC used in oncology clinical trials.

Secondly, as the EC recommendation functions as a 'recommendation' model, the quality of candidate EC for model inference holds substantial sway over the practical usefulness of the recommendation models. Once again, improving the quality of candidate EC necessitates domain expertise in a specific therapeutic area. Further, given that EC defining the intervention and study population exhibit greater diversity than those used to protect

	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)

Table 5: Performances of the EC recommendation model using title, summary, and design factors as input according to time periods and therapeutic areas of clinical trials. The numbers in parentheses represent the performances when EC topics were randomly recommended.

patient safety, it might be more effective for the EC generation model, rather than the recommendation model, to obtain these defining EC. In such scenarios, the EC recommendation model could serve to filter the generated EC in terms of clinical relevance. In this context, we believe that the gap between our model and the original EC in human evaluation could be bridged by designing a streamlined pipeline from the recommended EC to a complete EC set.

6 Conclusion

In this study, we introduce the task of recommending EC from clinical trial information and develop the CReSE model, designed to preserve clinical relevance between EC, by employing contrastive learning and using rephrasing via LLMs as text augmentation. We also demonstrate the importance of varied rephrasing prompts for developing the CReSE model through the ablation study. Additionally, we establish the automatic evaluation framework which assesses the clinical validity of the EC recommendation model based on the CReSE model.

In addition, we define common EC and exclude them from the dataset to prevent an overestimation of the EC recommendation model's performances and to align the EC recommendation task in accordance with actual needs in trial design. Furthermore, due to inconsistent quality in EC reporting on ClinicalTrials.gov, despite its extensive database, we employ the EC clustering outcomes from the CReSE model to enhance the quality of negative EC-title pairs. Through this refinement, we achieve a high-performance EC recommendation model with precision@1 of 48.0% and MAP@5 of 42.7%, without requiring specialized architecture modeling.

While the primary motivation of this study is to provide an appropriate EC template from limited trial information such as trial titles, we also envision the EC recommendation model as a clinical inference tool for exploring new therapeutic strategies and safety concerns by recommending EC. Although this work does not conclusively determine the potential of LMs as clinical inference tools, we expect that our automatic evaluation framework based on the CReSE model could enhance the development and evaluation of EC recommendation models in terms of clinical validity.

References

Al K Akobeng. 2005. Understanding randomised controlled trials. *Archives of disease in childhood*, 90(8):840–844.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-

Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Mary Regina Boland, Samson W Tu, Simona Carini, Ida Sim, and Chunhua Weng. 2012. Elixr-time: a temporal knowledge representation for clinical research eligibility criteria. *AMIA summits on translational science proceedings*, 2012:71.

Kerstin Breithaupt-Groegler, Christoph Coch, Martin Coenen, Frank Donath, Katharina Erb-Zohar, Klaus Francke, Karin Goehler, Mario Iovino, Klaus Peter Kammerer, Gerd Mikus, et al. 2017. Who is a 'healthy subject'?—consensus results on pivotal eligibility criteria for clinical trials. *European journal of clinical pharmacology*, 73:409–416.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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*.

Mili Duggal, Leonard Sacks, and Kaveeta P Vasisht. 2021. Eligibility criteria and clinical trials: An fda perspective. *Contemporary Clinical Trials*, 109:106515.

FAIR. 2022. Library for converting clinical trial eligibility criteria to a machine-readable format.

FDA. 2020. Enhancing the diversity of clinical trial populations — eligibility criteria, enrollment practices, and trial designs: Guidance for industry. Clinical/Medical.

FDA. 2023. Criteria for IRB approval of research.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Tianyong Hao, Hongfang Liu, and Chunhua Weng. 2016. Valx: a system for extracting and structuring numeric lab test comparison statements from text. *Methods of information in medicine*, 55(03):266–275.

Keith Humphreys, Kenneth R Weingardt, and Alex HS Harris. 2007. Influence of subject eligibility criteria on compliance with national institutes of health guidelines for inclusion of women, minorities, and children in treatment research. *Alcoholism: Clinical and Experimental Research*, 31(6):988–995.

Qiao Jin, Zifeng Wang, Charalampos S Floudas, Jimeng Sun, and Zhiyong Lu. 2023. Matching patients to clinical trials with large language models. *arXiv preprint arXiv:2307.15051*.

Susan Jin, Richard Pazdur, and Rajeshwari Sridhara. 2017. Re-evaluating eligibility criteria for oncology clinical trials: analysis of investigational new drug applications in 2015. *Journal of clinical oncology*, 35(33):3745.

Kamal raj Kanakarajan, Bhuvana Kundumani, Abhijith Abraham, and Malaikannan Sankarasubbu. 2022. BioSimCSE: BioMedical sentence embeddings using contrastive learning. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 81–86, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tian Kang, Shaodian Zhang, Youlan Tang, Gregory W Hruby, Alexander Rusanov, Noémie Elhadad, and Chunhua Weng. 2017. Eliie: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association*, 24(6):1062–1071.

Edward S Kim, Suanna S Bruinooge, Samantha Roberts, Gwynn Ison, Nancy U Lin, Lia Gore, Thomas S Uldrick, Stuart M Lichtman, Nancy Roach, Julia A Beaver, et al. 2017. Broadening eligibility criteria to make clinical trials more representative: American society of clinical oncology and friends of cancer research joint research statement. *Journal of Clinical Oncology*, 35(33):3737.

Edward S Kim, Thomas S Uldrick, Caroline Schenkel, Suanna S Bruinooge, R Donald Harvey, Allison Magnuson, Alexander Spira, James L Wade, Mark D Stewart, Diana Merino Vega, et al. 2021. Continuing to broaden eligibility criteria to make clinical trials more representative and inclusive: Asco–friends of cancer research joint research statement. *Clinical Cancer Research*, 27(9):2394–2399.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Stefan Listl, Hendrik Jürges, and Richard G Watt. 2016. Causal inference from observational data. *Community dentistry and oral epidemiology*, 44(5):409–415.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Allison Magnuson, Suanna S Bruinooge, Harpreet Singh, Keith D Wilner, Shadia Jalal, Stuart M Lichtman, Paul G Kluetz, Gary H Lyman, Heidi D Klepin, Mark E Fleury, et al. 2021. Modernizing clinical trial eligibility criteria: recommendations of the asco-friends of cancer research performance status work group. *Clinical Cancer Research*, 27(9):2424–2429.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. *arXiv preprint arXiv:2204.00447*.

Raymond U Osarogiagbon, Diana Merino Vega, Lola Fashoyin-Aje, Suparna Wedam, Gwynn Ison, Sol Atienza, Peter De Porre, Tithi Biswas, Jamie N Holloway, David S Hong, et al. 2021. Modernizing clinical trial eligibility criteria: recommendations of the asco–friends of cancer research prior therapies work group. *Clinical Cancer Research*, 27(9):2408–2415.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Samson Tu, Mor Peleg, Simona Carini, Daniel Rubin, and Ida Sim. 2009. Ergo: A templatebased expression language for encoding eligibility criteria. Technical report, Technical report.

Thomas S Uldrick, Gwynn Ison, Michelle A Rudek, Ariela Noy, Karl Schwartz, Suanna Bruinooge, Caroline Schenkel, Barry Miller, Kieron Dunleavy, Judy Wang, et al. 2017. Modernizing clinical trial eligibility criteria: recommendations of the american society of clinical oncology–friends of cancer research hiv working group. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 35(33):3774.

Zifeng Wang and Jimeng Sun. 2022. Trial2vec: Zero-shot clinical trial document similarity search using self-supervision. *arXiv preprint arXiv:2206.14719*.

Zifeng Wang, Brandon Theodorou, Tianfan Fu, Cao Xiao, and Jimeng Sun. 2023a. Pytrial: A comprehensive platform for artificial intelligence for drug development. *arXiv preprint arXiv:2306.04018*.

Zifeng Wang, Cao Xiao, and Jimeng Sun. 2023b. Autotrial: Prompting language models for clinical trial design. *arXiv preprint arXiv:2305.11366*.

Chunhua Weng, Xiaoying Wu, Zhihui Luo, Mary Regina Boland, Dimitri Theodoratos, and Stephen B Johnson. 2011. Elixr: an approach to eligibility criteria extraction and representation. *Journal of the American Medical Informatics Association*, 18(Supplement_1):i116–i124.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*.

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

Xingyao Zhang, Cao Xiao, Lucas M Glass, and Jimeng Sun. 2020. Deepenroll: patient-trial matching with deep embedding and entailment prediction. In *Proceedings of the web conference 2020*, pages 1029–1037.

A Prompts

In our study, we utilized large language models (LLMs) to handle biomedical free texts in a manner that aligns with clinical validity. Specifically, we rephrased the original eligibility criteria (EC) used in clinical trials using LLMs to develop the CReSE model. Additionally, we assessed the clinical relevance between pairs of EC and streamlined the EC recommendation model through LLMs, transforming it into the end-to-end recommendation system. This section provides an overview of all the prompts that were utilized in our study.

A.1 Prompts for rephrasing

We developed four different rephrasing prompts in a 2-shot manner for ChatGPT. The aim was to generate an original-rephrased EC dataset for training the CReSE model (Table 6).

Common introduction for rephrasing prompts
You are a world-renowned clinical specialist with expertise in clinical trial design and implementation. **{*Prompt-specific instructions*}** The proposed new EC must start with either “[Inclusion]" or “[Exclusion]." Here’s an example: **{*Examples*}** Original EC: **{*EC*}** Rephrased EC:
Simple rephrasing
Prompt-specific instructions: Please suggest different eligibility criteria (EC) that can identify patients who clinically resemble those already screened using a given EC. **{*Examples*}**: Original EC: “[Exclusion] previous bariatric or gastric surgery" Rephrased EC: “[Inclusion] Eligible patients must have a body mass index (BMI) of 30 or higher." Explanation: A new eligibility criteria for patients with a BMI of 30 or higher has been proposed as an alternative to the original exclusion criteria for bariatric or gastric surgery. This new criterion can help identify patients who are at risk of obesity-related health issues and may benefit from interventions aimed at reducing their BMI. Original EC: ’[Exclusion] gastrointestinal disorders affecting absorption’ Rephrased EC: “[Inclusion] Eligible patients must not be taking medications that interfere with gastrointestinal absorption." Explanation: A new eligibility criterion has been proposed to replace the old exclusion criterion of gastrointestinal disorders affecting absorption. This new criterion helps to identify patients without significant gastrointestinal problems that could affect the investigational product’s absorption.

Table 6: Prompts for rephrasing EC

Rephrasing without using a core clinical concept
Prompt-specific instructions: Please rephrased an eligibility criteria (EC) without using any core clinical concept words from the original EC. **{*Examples*}**: Original EC: “[Inclusion] International Prostate Symptom Score (IPSS) $<$ 7" Rephrased EC: “[Inclusion] Participants who report mild or no symptoms related to urination, as assessed by a standardized questionnaire." Explanation: The rephrased EC avoids using the specific term "international Prostate Symptom Score (IPSS)" and instead describes the symptoms that would be used to assess the severity of the participant’s urinary issues. Original EC: “[Exclusion] primary uveal or mucosal melanoma" Rephrased EC: “[Exclusion] Individuals with a history of melanoma in areas other than the skin." Explanation: The rephrased EC avoids using the specific clinical terms "uveal" and "mucosal" melanoma and instead describes the location of the melanoma that would make a participant ineligible for the trial.
Suggesting alternative EC
Prompt-specific instructions: Please suggest alternative eligibility criteria (EC) that can serve as substitutes for a given EC when there is not enough patient data to determine whether the current EC is met or not. **{*Examples*}**: Original EC: “[Inclusion] hbA1c 7.0% - 10.0%" Aim of original EC: To determine if the patient has diabetes Alternative EC: “[Inclusion] Documented history of type 2 diabetes in the past year." Original EC: “[Inclusion] platelet count $>=$ 100,000" Aim of original EC: To ensure the patient has a sufficient platelet count for safe treatment Alternative EC: “[Inclusion] No history of thrombocytopenia or related conditions in the past year."
Suggesting EC possibly used in the same clinical trial
Prompt-specific instructions: Please suggest an alternative eligibility criteria (EC) that can be utilized in the same clinical trial where a previous EC has already been employed. **{*Examples*}**: Original EC: “[Exclusion] cardiac ventricular arrhythmias requiring anti-arrhythmic therapy" Clinical Trial: "A Phase III Randomized Controlled Trial Evaluating the Efficacy and Safety of Carvedilol in Patients with Chronic Heart Failure" Suggested EC possibly from the same clinical trial: “[Exclusion] The patient has a history of sustained ventricular tachycardia or ventricular fibrillation, or is at high risk of these conditions as determined by the investigator." Original EC: “[Exclusion] history of major organ transplant" Clinical Trial: "Phase II Study Investigating the Safety and Efficacy of Pembrolizumab in Patients with Advanced Melanoma" Suggested EC possibly from the same clinical trial: “[Exclusion] The patient is currently on or requires systemic immunosuppressive therapy within two weeks prior to the first dose of study drug."

Table 6: (continued) Prompts for rephrasing EC

A.2 Prompt for assessing clinical relevance

A prompt was created to evaluate the clinical relevance between two EC using ChatGPT (Table 2). In this prompt, LLMs will be asked to assess the clinical relevance on a scale from 0 to 3, using four different levels.

Prompt for assessing clinical relevance between an EC pair
As an expert in clinical trial design and execution, please evaluate the clinical relevance of the following two eligibility criteria on a 4-point scale. Use the following guidelines to rate their relevance: Clinical relevance 3: The two eligibility criteria are essentially identical clinically. For example: EC1: “[exclusion] serum albumin is 2.4 g/dL or less" EC2: “[inclusion] serum albumin is 2.4 g/dL or more" Clinical relevance 2: The two eligibility criteria have strong relevance due to factors such as disease progression, or epidemiology. For example: Clinical relevance 1: The two eligibility criteria are not directly related, but still have some relevance due to factors such as general treatment plan, disease progression, or epidemiology. For example: EC1: “[inclusion] no concurrent major surgery" EC2: “[inclusion] histologically confirmed transitional cell carcinoma (TCC) of the urothelium" Clinical relevance 0: The eligibility criteria are irrelevant from a clinical perspective. For example: EC1: “[exclusion] history of a severe allergic reaction with generalized urticaria, angioedema, or anaphylaxis in the 2 years prior to enrollment" EC2: “[inclusion] male condoms with spermicide" Here are more examples: EC1: “[exclusion] Administration of long-acting immune-modifying drugs at any time during the study period" EC2: “[inclusion] Current autoimmune disorder (based on medical history and physical examination), for which the participant has received immune-modifying therapy within 6 months, before study vaccination" Clinical relevance: 1 EC1: “[exclusion] Antibiotic exposure within the past 4 weeks of helicobacter pylori diagnosis" EC2: “[exclusion] Prior helicobacter pylori treatment failure" Clinical relevance: 2 EC1: “[inclusion] 1 focal lesions on MRI (magnetic resonance imaging) studies; Each focal lesion must be 5 mm or more in size" EC2: “[exclusion] kellgren and Lawrence grade $\geq$ 3” Clinical relevance: 3 EC1: **{*EC1*}** EC2: **{*EC2*}** Clinical relevance:

Table 7: Prompt for assessing clinical relevance between given two EC

A.3 Prompts for recommending a complete EC set from the clinical trial title

To provide a baseline system for comparison, we devise a prompt for GPT-4 that request to recommend a complete EC set from the clinical trial titles (Table 3). However, since the EC recommendation model we developed was designed to handle only non-common EC, an additional system to generate a complete EC set from the clinical trial title when using our EC recommendation model was required. To solve this challenge, we integrated ChatGPT into our approach, creating an end-to-end recommendation system, starting from the clinical trial title and effectively suggesting the full set of EC.

B Detailed methodology

B.1 Development of common EC classifier

We employed the BertForSequenceClassification model from Huggingface as the classification model for common EC. In the biomedical domain, we utilized several pre-trained language models (LMs), namely BioClinicalBERT (Alsentzer et al., 2019), BioBERT (Lee et al., 2020), and BioLinkBERT (Yasunaga et al., 2022). Additionally, we adopted BaseBERT (Devlin et al., 2018), ELECTRA (Clark et al., 2020), and XLM-RoBERTa (Conneau et al., 2019) as baseline model for fine-tuning.

B.2 Original-rephrased EC pairs dataset

After performing the rephrasing, we notice that the two rephrasing prompts, one suggesting alternative EC and one suggesting EC possibly used in the same clinical trial, have a more varied rephrasing pattern than the former two prompts, one about simple rephrasing and one without using a core clinical concept (Table 1). In order to efficiently utilize the ChatGPT API, we rephrased 20K EC using the first two prompts and 5K EC using the second two prompts, thus obtaining a total of 50K original rephrased EC pairs for training the CReSE model. This difference in the total number of rephrased ECs resulted in an imbalance in the composition of training data for the ablation study (Table 4).

B.3 Selection of clinical trials and evaluation datsets

In this study, we selected trials that satisfied the following five conditions from 445K clinical trials registered on ClinicalTrials.gov from March 2002 to May 2023: 1) the date of information upload was reported, 2) a brief summary and official title were provided, 3) the trials were classified as 'interventional' (excluding observational trials), 4) at least two EC were reported, and 5) the intervention investigated in the trial was categorized as 'Drug' or 'Biological' (excluding 'Device' and 'Behavior' interventions). Additionally, for EC, we excluded studies where an individual EC was either too short (less than 3 characters) or too long (more than 353 characters).

To ensure a fair comparison with top performing LLMs including ChatGPT and GPT-4, the test dataset consisted of each 5K trials uploaded before and after September 2019, the knowledge cut-off date for ChatGPT and GPT-4. Therefore, the test dataset contains more recent trials than the training dataset, which is why we believe the test dataset has an overall higher number of ECs and longer EC lengths than the training dataset (Table 1 in the main manuscript).

In addition, we categorized clinical trials into three periods to explore the recommendation performance by the time periods of clinical trials: 1) May 2002 to December 2009, 2) January 2010 to the outbreak of COVID-19 (March 11th, 2020, the declaration of COVID-19 outbreak as a pandemic by WHO), and 3) COVID-19 outbreak to May 2023. Furthermore, recognizing that the EC recommendation performance might vary due to EC compositions and the number of EC used in clinical trials, we also reported the performance measures when EC clusters were randomly recommended. In all the evaluation settings and categories of clinical trials (Tables 4 and 5 in the main manuscript), we randomly sampled 100 clinical trials for each category and used them as the evaluation dataset.

While evaluating the CReSE model, we constructed the evaluation EC pairs datasets by randomly sampling 200, 300, 300, and 200 EC pairs for clinical relevance scores 0, 1, 2, and 3, respectively, to ensure a balanced distribution of clinical relevance scores.

Prompt for generating a complete EC set from the clinical trial title and recommend EC by our recommendation model (ChatGPT)

As an acclaimed specialist in clinical trial design and execution, your task involves drafting an exhaustive list of participant selection guidelines for a specific clinical trial. The details about the trial including its title, summary, and suggested eligibility criteria will be given by the user. Your task is to expand these criteria with a more comprehensive set. When crafting the eligibility criteria, ensure to consider potential risk factors, such as contraindications and possible interactions between the drug and the intervention. Clearly and professionally outline the intervention (as well as any control group treatment) and patient conditions. It's also crucial to confirm that the patient is in a mental and physical state where they can give informed consent. The selection criteria should not unduly narrow the prospective participant pool without medically valid reasoning, such as unjustified exclusion of HIV or HCV patients. Also, verify the patient's clinical and social circumstances to accurately assess the outcome during the follow-up period of the trial, like the presence of a measurable lesion or proximity to the trial location. For inclusion parameters, phrase them as '[inclusion] To be eligible, the patient must ...', and for exclusion parameters, use '[exclusion] To be eligible, the patient must not ...'. An example is provided below for better understanding. Note: the final selection criteria should be provided without duplicating the clinical trial's information or its summary. Also, the explanation for the final selection criteria set should not be included.

Clinical trial title: A Phase 1, Open-Label, Multicenter Study of KYV-101, an Autologous Fully-Human Anti-CD19 Chimeric Antigen Receptor T-Cell (CD19 CAR T) Therapy, in Subjects With Refractory Lupus Nephritis
Recommended eligibility criteria:
[Inclusion] have a confirmed diagnosis of systemic lupus erythematosus (SLE) according to the American College of Rheumatology (ACR) or Systemic Lupus International Collaborating Clinics (SLICC) criteria.
[Inclusion] lupus nephritis, defined by proteinuria, microscopic hematuria, and varying degrees of renal insufficiency.
[Inclusion] To be eligible, the patient must have SLE and lupus nephritis that is refractory to standard therapy, including at least one cycle of an immunosuppressant regimen (e.g., cyclophosphamide or mycophenolate mofetil) and corticosteroids.
[Exclusion] not have received prior treatment with any gene therapy medicinal product or any CAR T-cell therapy.
Complete the eligibility criteria set including the recommended eligibility criteria:
[Inclusion] To be eligible, the patient must have a confirmed diagnosis of systemic lupus erythematosus (SLE) according to the American College of Rheumatology (ACR) or Systemic Lupus International Collaborating Clinics (SLICC) criteria.
[Inclusion] To be eligible, the patient must have lupus nephritis, defined by proteinuria, microscopic hematuria, and varying degrees of renal insufficiency.
[Exclusion] To be eligible, the patient must not have a concurrent malignancy or history of malignancy within the last 5 years, to prevent potential complications or interactions with the trial treatment.
[Exclusion] To be eligible, the patient must not be pregnant or breastfeeding, due to potential unknown effects of the therapy on a developing fetus or infant.
[Exclusion] To be eligible, the patient must not have a history of severe allergic reactions to any of the components of the KYV-101 formulation.
[Exclusion] To be eligible, the patient must not have had a live vaccine administered within 30 days of initiating study treatment.

Clinical trial title: ***{Clinical_trial_title}***
Recommended eligibility criteria: ***{Recommended_EC***
Complete the eligibility criteria set including the recommended eligibility criteria:

Table 8: Prompts for generating a complete EC set from the clinical trial title and the recommended non-common EC

C Details on model development

In this section, we provide a comprehensive description of the training conditions for the common EC classifier, the CReSE model, and the EC recommendation model developed as part of this study. All experiments, except for the largest training of the EC recommendation model, were carried out using an RTX 4080 with 16GB of VRAM. For training the EC recommendation model with the entire training dataset, we employed 16 V100 GPUs in parallel.

The maximum token length was restricted to 256, and we ensured reproducibility by fixing all random seeds to 42. During hyper-parameter tuning, we experimented with learning rates of 5e-5, 2e-5, and 5e-6, and batch sizes of 32 and 64. We employed the AdamW optimizer and linear warmup scheduler with an epsilon value of 1e-8 for updating model parameters. The total number of training epochs was set to 25.

C.1 Development of the CReSE model

In the CReSE model training, we employed BioLinkBERT as the baseline model, which demonstrated superior performance in classifying common EC across various pre-trained LMs. This decision aimed to save time and computation resources. For hyper-parameter tuning, we conducted experiments with the different projection dimensions (256, 512, and 768), batch sizes (16 and 32), learning rates for the text encoder (5e-6 and 1e-6) and for the projection layer (5e-4, 1e-5, 5e-6, and 1e-6). The dropout probability of the projection layer was consistently set to 0.1.

During hyper-parameter tuning, we utilized the entire original-rephrased EC dataset comprising 50K examples with the four rephrasing prompts. The model underwent a total of 3 training epochs. We employed the AdamW optimizer with a weight decay of 1e-4 and implemented a ReduceLROnPlateau scheduler with patience of 1 and a reduction factor of 0.8. The CReSE model is trained for 10 epochs

For the ablation study, which aimed to investigate the CReSE model's performance variation concerning changes in the composition and size of the training dataset, we kept the hyper-parameters fixed. Specifically, we used a projection dimension of 256, a batch size of 32, and learning rates of 1e-5 and 5e-4 for the text encoder and projection layer, respectively.

C.2 Development of the EC recommendation model

In the EC recommendation model, the input text was constructed by combining EC and clinical trial information with the [SEP] token. Among the four types of clinical trial information available for input, we utilized the 'official title' from ClinicalTrials.gov as the title and the 'brief summary' as the summary. The key design factors, written in the free text but in a semi-structured form, encompassed important trial design elements, including the investigated condition, investigational drug or treatment, study phase, number of enrolled patients, and primary outcome measures. When multiple types of trial information were employed as input, each piece of information was concatenated with the [SEP] token.

During the development of the CReSE model, we adopted BioLinkBERT as the baseline LM for the EC recommendation model. For fine-tuning, we added a linear-ReLU stack of two layers with dimensions 768*2 $\times$ 512 with a drop-out of 0.1 as the classification layer above the text encoder. Throughout both the main model training and ablation studies, we maintained fixed hyper-parameters values such as a learning rate of 256, a hidden layer dimension of 512, and a dropout probability of 0.1 for the classification layer. Additionally, we applied gradient clipping with a maximum norm of 1.0 during model training. In the main training setting, we set the threshold for the minimum number of EC occurrences in the clinical trials to generate negative EC-title pairs as 8. Moreover, the maximum token length was set to 512 during the main training, while it was set to 256 in the ablation studies to accommodate computation resource limitations. In addition, we increased the batch size to 128, effectively reducing training times. This adjustment resulted in each model training involving 3 epochs taking approximately 3 hours to complete, utilizing 16 V100 GPUs in parallel.

D Supplementary results

D.1 Performances of common EC classifiers

After fine-tuning several types of LMs to develop a common EC classifier, we achieved an accuracy of up to 97.99% and an F1-score of 97.78% when using BioLinkBERT (Table 9). In order to minimize the overall computational demands in this study, we used the BioLinkBERT checkpoint in all subsequent experiments as the initial parameter settings of text encoders.

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

D.2 Correlation between validation loss for contrastive learning and EC clustering performances

An ablation study trained the CReSE model on training datasets with different configurations and found an inverse relationship between validation loss in contrastive learning and final EC clustering performance (Figure 5). This result suggests that utilizing LLMs for rephrasing indeed serves as an effective method for text augmentation in the context of contrastive learning to integrate medical knowledge from LLMs into embedding systems. However, it's important to note that there is a need to identify the optimal composition of the dataset containing the original-rephrased text pairs. Furthermore, a distinction becomes apparent between the goals of contrastive learning, where the model determines whether an EC pair was made by rephrasing or not, and the evaluation of clinical relevance between EC pairs. Therefore, when employing rephrasing via LLMs as a text augmentation method, the design of diverse rephrasing prompts becomes crucial.

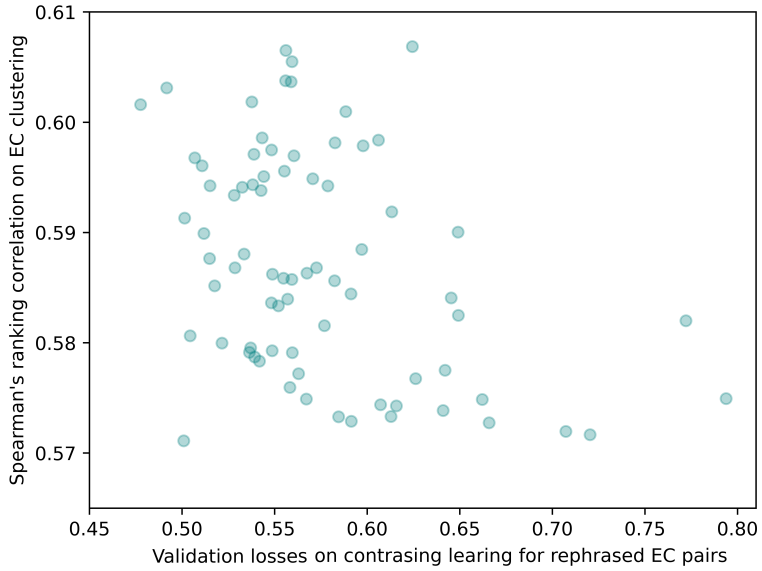

Figure 5: Scatter plot of validation losses and EC clustering performances of the CReSE model trained on diverse compositions of training datasets in the ablation study

D.3 Performances of the CReSE model

Regardless of the clustering method or the number of EC clusters, the CReSE model consistently exhibited superior performance in EC clustering performance compared to other LMs pre-trained in the biomedical domain (Table 10). Furthermore, when training the EC recommendation model with increasing dataset size, the binary classification performance continues to increase up to 1M positive EC-title pairs, while the recommendation performance stops increasing after 0.2M (Figure 6).

Clustering methods	**Spearman**				**Pearson**			
	50	100	200	300	50	100	200	300
TF-IDF	25.0 [16.4, 28.6]	27.0 [23.8, 30.3]	27.7 [23.4, 31.4]	26.6 [21.7, 31.3]	25.1 [16.5, 28.8]	27.1 [24.0, 30.5]	28.2 [23.6, 31.5]	26.7 [21.6, 31.7]
Only embedding								
Base-BERT	25.3 [21.0, 27.7]	25.3 [21.4, 29.4]	26.7 [21.9, 31.9]	26.9 [20.6, 29.8]	24.5 [20.2, 27.5]	25.5 [21.9, 29.9]	26.9 [20.6, 29.8]	27.0 [20.8, 30.5]
BioLinkBERT	27.4 [23.4, 32.3]	29.9 [24.9, 34.3]	28.3 [25.2, 33.4]	26.9 [21.4, 31.6]	27.1 [23.4, 32.2]	30.0 [25.3, 34.8]	28.6 [25.4, 33.7]	27.3 [21.4, 31.6]
TrialBERT	27.6 [23.1, 32.4]	29.0 [24.7, 33.0]	28.4 [24.0, 31.2]	28.0 [21.4, 35.6]	27.4 [22.8, 32.2]	29.2 [24.9, 33.3]	28.7 [24.4, 31.3]	28.4 [21.4, 35.6]
BioSimCSE	31.5 [29.0, 36.4]	34.7 [31.1, 38.1]	34.2 [28.4, 39.7]	30.4 [25.7, 36.3]	31.2 [28.5, 35.7]	35.0 [31.6, 38.1]	34.4 [28.6, 39.9]	30.7 [25.5, 36.5]
BioGPT	28.7 [24.4, 34.6]	32.3 [28.8, 33.9]	28.4 [23.3, 32.1]	27.8 [24.3, 34.5]	28.8 [24.5, 34.5]	32.0 [28.8, 33.9]	28.5 [23.3, 32.5]	29.0 [24.0, 34.9]
CREEP (ours)	**43.6 [41.8, 46.2]**	43.0 [40.3, 45.3]	42.4 [37.3, 45.1]	39.0 [35.2, 43.4]	**43.7 [42.2, 46.4]**	43.4 [40.7, 45.5]	42.8 [37.8, 45.9]	39.7 [35.9, 43.3]
BERTopic								
Package default	36.8 [30.4, 43.8]	40.9 [35.6, 46.2]	44.0 [40.5, 49.0]	43.7 [40.0, 47.6]	36.9 [30.7, 43.5]	40.8 [40.9, 46.2]	44.5 [40.9, 49.7]	44.2 [40.5, 48.4]
BioLinkBERT	32.5 [26.3, 35.9]	36.2 [29.7, 42.5]	37.6 [34.1, 42.0]	37.2 [33.4, 42.3]	32.5 [26.0, 36.3]	36.4 [29.9, 42.6]	37.8 [34.4, 42.4]	37.7 [34.0, 42.3]
TrialBERT	31.5 [25.3, 37.4]	37.6 [33.4, 44.7]	40.6 [38.3, 44.1]	40.2 [37.2, 44.3]	31.9 [25.6, 37.7]	38.2 [34.1, 45.1]	41.2 [39.2, 44.9]	41.1 [38.1, 45.2]
BioSimCSE	27.6 [15.8, 34.7]	40.8 [35.5, 43.3]	40.6 [37.9, 43.8]	41.2 [38.0, 44.1]	27.6 [16.0, 34.6]	40.6 [35.1, 43.4]	40.9 [37.9, 43.6]	41.4 [38.0, 44.4]
BioGPT	21.9 [14.2, 28.9]	32.2 [25.4, 37.8]	37.7 [33.8, 42.9]	39.9 [35.8, 42.5]	22.2 [14.5, 29.1]	32.3 [25.5, 38.0]	38.0 [34.3, 42.9]	39.9 [36.1, 42.7]
CREEP (ours)	42.1 [37.9, 47.0]	**44.9 [40.9, 48.4]**	**45.0 [41.7, 46.9]**	**45.7 [43.4, 47.5]**	42.0 [38.3, 46.7]	**45.3 [41.1, 48.5]**	**45.5 [42.2, 47.0]**	**46.4 [44.0, 48.0]**

Table 10: Comparison of the CReSE model and other biomedical LMs on EC clustering

D.4 Human Evaluation Results

Within the three remaining categories, excluding the one pertaining to overly restrictive recommendations, our model's proposed EC set exhibited insufficiency in comparison to the original EC set (p-value $<$ 0.05, Table 11). To elaborate, the EC set suggested by our model displayed suboptimal performance in effectively ensuring patient safety and constructing a clinically valid EC set. These differences were statistically significant, measuring 0.638 and 0.675, respectively.

	Original EC	**Our model + ChatGPT**	**Mean difference**	**P-value**
Overall	3.7±0.8	3.2±0.7	0.522	0.010
Protecting patient safety	3.7±0.9	3.2±0.7	0.450	0.035
Defining the study population	3.8±0.8	3.2±0.8	0.638	0.006
Avoiding overly restrictive	3.6±0.7	3.3±0.6	0.325	0.114
Clinically valid and realistic	3.8±0.7	3.2±0.7	0.675	0.001

Table 11: Human evaluation results on four evaluation categories

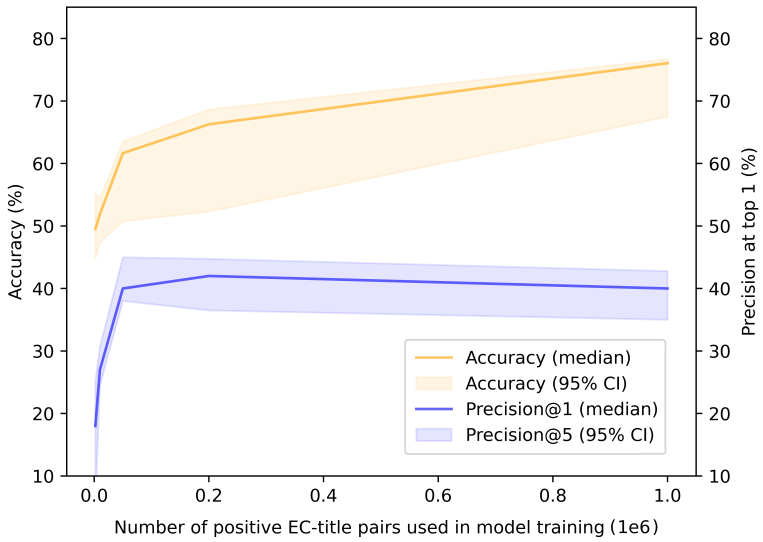

Figure 6: Performances of the EC recommendation models by the size of the training dataset containing EC-title pairs

E Guideline documents

E.1 Annotation guideline for classifying common EC

This document serves as an annotation guideline for classifying 'common EC' from the entire set of EC. Common EC are defined as EC that have been commonly accepted over time or used as templates across trials, often excluding certain populations from participation without strong clinical or scientific justification (e.g., older adults, those at the extremes of the weight range, those with malignancies or certain infections such as HIV, and children) (FDA, 2020). Additionally, common EC include poorly defined criteria in clinical trials, regardless of the clinical characteristics of investigational drugs and patient conditions. The annotation guideline elaborates on the different types of common EC and provides relevant examples.

1. Common EC universally used in clinical trials
We refer to EC universally used in clinical trials regardless of their purpose and design factors as 'common EC' and developed the classifier for common EC. Here are the detailed types of common EC and their definitions and examples (Table 12).
2. EC used to ensure the smooth conduct of the clinical trial
Some common EC were used in clinical trials to ensure the smooth operation of the process, such as assessing the trial location's accessibility and the communication abilities of enrolled patients (Table 13).

E.2 Human evaluation guideline for assessing the appropriateness of EC sets

This document aims to evaluate the appropriateness of the eligibility criteria for the given information of clinical trials. The purpose of this evaluation is to assess the extent to which the eligibility criteria adequately address the following points (Table 14): **1) Protecting patient safety, 2) Clearly defining the study population (and study intervention), 3) Avoiding overly restrictive, and 4) Clinically valid and realistic**. Evaluators rated questions from each category on a scale of 1 to 5. By conducting this evaluation, we aim to ensure that the eligibility criteria meet the highest standards of quality and align with the needs of clinical trials. Below is a detailed guideline for each evaluation category and question.

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"*
Unapproved Drug/Herbal Supplement	Taking unapproved drugs or herbal supplementary before the trial. *Ex) "[Exclusion] use of herbal supplements within 7 days or 5 half-lives (whichever is longer) before the first dose of study intervention"*
Hepatic and Renal Function	Excluding patients with reduced hepatic or renal function without adequate clinical and scientific justification - Includes defining hepatic or renal impairment based on a normal range of laboratory values (e.g., AST, ALT, bilirubin, creatinine clearance) *Ex) "[Inclusion] there was no previous severe renal dysfunction", "[Exclusion] if a liver lesion is the site of injection: All AST, ALT and bilirubin greater than 2.5 ULN"* *Hormonal and hematological test values such as TSH, PTT, INR, and ANC, as well as cardio tests like QT interval and ECG, are not considered as common EC.
Reduce Patient Risk	Used to reduce the patient risk, but without a clear and appropriate clinical justification: HIV, hepatitis, tuberculosis infection, prior organ transplant, any major infection, any immunodeficiency (not heart disease), active autoimmune disease, no previous malignancy, etc. *Exclusion based on previous surgery is considered as non-common EC *Ex) "[Exclusion] any known immunosuppressive condition or immune deficiency disease (including human immunodeficiency virus [HIV] infection), or ongoing receipt of any immunosuppressive therapy", " [Exclusion] subject positive for hepatitis B virus (HBV) surface antigen, hepatitis B virus core antibody with a negative hepatitis B surface antibody or with detectable serum hepatitis B DNA"*

Table 12: Types of common EC and their definitions and examples

Common EC Type	Definitions and Examples
Life expectancy or performance status	Life expectancy or performance status for checking the general health of a patient. *Ex) "[Inclusion] life expectancy >= 12 weeks as judged by the Investigator", " [Inclusion] Eastern Cooperative Oncology Group (ECOG) performance status of 0 to 1 at trial entry"*
Contraindication	Contraindication, allergy or hypersensitivity to investigational drug, or previous exposure to investigational drug. *Ex) "[Exclusion] known allergies, hypersensitivity, or intolerance to monoclonal antibodies or hyaluronidase", "[Exclusion] use any investigational drug within 28 days before the start of trial treatment"*
Drug Interaction	Intake of drugs that possibly interact with investigational drugs. *Ex) "[Inclusion] maintained on modern therapeutic regimen utilizing non-CYP interacting agents (e.g. excluding ritonavir)"*
Conflict of Interest	If there is a conflict of interest through family... *Ex) "[Exclusion] family member or household contact who was an employee of the research center or otherwise involved with the conduct of the study"*
Mental Illnesses/Informed Consent Form	Broad range of mental illnesses which may harm the ability to make an informed consent or understand a study purpose and protocol by the patient self. *Ex) "[Exclusion] mental conditions rendering a subject unable to understand the nature, scope, and possible consequences of the study"*
Prior use of (other) investigational drug	If a patient has received any other investigational drug before randomization.. *"Ex) [Exclusion] prior treatment with 89Strontium or 153Samarium containing compounds (e.g. Metastron®, Quadramet®)", " [Exclusion] prior thiopurine therapy"* *Prior use of clinically substitutable drugs with the investigational drug is not considered as common-EC. *Ex) "Exclusion: Received previous therapy with capecitabine, neratinib, lapatinib, or any other HER2-directed tyrosine kinase inhibitor."*
Patient adequate to measure outcome	measurable disease (mainly in oncology trial), Refrain from blood donation, have some contra-indication for measurement. *Ex) " [Inclusion] patients must have evaluable disease, either with informative tumor markers or with the measurable disease on imaging, by RECIST (Response Evaluation Criteria in Solid Tumors) criteria (Appendix II)", " [Exclusion] agreement to refrain from blood donation during the course of the study"*

Table 12: (continued) Types of common EC and their definitions and examples

Common EC Type	Description and Examples
Area of Residence	To ensure that participants reside in a particular geographical location that allows them easy access to the study site for regular investigations, measurements, or follow-up visits *Ex) "[Inclusion] patients followed in the Rheumatology Department at the hospital of St Etienne"*
Limit Language	Limit speaking language to control for language barriers in the study. *Ex) " [Exclusion] speaks a language other than English"*
Limit Patient Ethnicity	include or exclude specific ethnic groups. *Ex) "[Exclusion] Limited to individuals of Asian ethnicity"*
Informed consent	Informed consent and agree to comply with the protocol: to ensure that potential participants fully understand the study's purpose, procedures, risks, and benefits before they decide to participate. *Ex) "[Inclusion] study subjects must obtain informed consent to this study and voluntarily sign a written informed consent before screening for enrollment."*
Past or Duplicated Participation	Do not enroll in other studies or previous participation in the same study: to maintain the integrity of the study and avoid potential confounding effects, researchers may exclude individuals who are already participating in other clinical trials or have previously taken part in the same study. *Ex) "[Exclusion] participation in other clinical trials (pharmaceutical trials)"*
Commitment of Participant	Confirmation of the patient's ongoing and good faith participation in the study: to ensure that participants are committed to actively participating in the study and completing all study requirements. *Ex) "[Inclusion] be willing and able to follow study instructions and likely to complete all study requirements"*

Table 13: Types of common EC used to ensure the smooth conduct of the clinical trials and their definitions and examples

Category	Question	Descriptions/Examples
Protecting patient safety	**[1]** Do eligibility criteria adequately exclude contraindications of the interventions/drugs being used and minimize potential harm to subjects during the course of the trial? (No 1 - 2 - 3 - 4 - 5 Yes)	This question is to review whether the criteria adequately account for potential risks, contraindications, and precautions that may affect patient safety. *Ex) Exclusion criteria: History of cancer and/or any known primary immunodeficiency disorder (e.g., HIV)*
Defining the study population	**[2-1]** Are the eligibility criteria clearly defining the study population being tested as appropriate to evaluate the given research hypothesis? (No 1 - 2 - 3 - 4 - 5 Yes)	This question is to assess whether the eligibility criteria align with the specific objectives of the study, ensuring that only suitable patients are included, and the study outcomes can be effectively evaluated. *Ex) Trial title: A Randomised, Double-blind, Placebo-controlled, Phase 3 Trial to Evaluate the Efficacy and Safety of Tralokinumab Monotherapy in Subjects With Moderate to Severe Atopic Dermatitis Who Are Candidates for Systemic Therapy* *Inclusion Criteria: Diagnosis of AD as defined by the Hanifin and Rajka (1980) criteria for AD, Diagnosis of AD for 1 year, AD involvement of 10 of body surface area at screening and baseline (visit 3), An EASI score of 12 at screening and 16 at baseline*
Defining study intervention	**[2-2]** Are the eligibility criteria clearly define the intervention? (No 1 - 2 - 3 - 4 - 5 Yes)	This question is to assess whether the eligibility criteria for the intervention are explicitly stated and well-defined. *Ex) Trial title: A Randomised, Double-blind, Placebo-controlled, Phase 3 Trial to Evaluate the Efficacy and Safety of Tralokinumab Monotherapy in Subjects With Moderate to Severe Atopic Dermatitis Who Are Candidates for Systemic Therapy* *Inclusion criteria: Subjects with documented systemic treatment for AD in the past year are also considered as inadequate responders to topical treatments and are potentially eligible for treatment with tralokinumab after appropriate washout.*

Table 14: Evaluation category for assessing the appropriateness of EC sets

Category	Question	Descriptions/Examples
Avioding overly restrictive	**[3]** Are eligibility criteria based on appropriate clinical evidence and do not unduly limit the study population? (No 1 - 2 - 3 - 4 - 5 Yes)	This question is to evaluate whether the eligibility criteria ensure the patient population is diverse and accurately reflects the target population for the study. *Ex) ECs that limit the study population* *Inclusion criteria: Participants between the ages of 25 and 30.* *Exclusion criteria: Participants with any other chronic condition*
Clinically valid and realistic	**[4]** Are the eligibility criteria consistent with current medical knowledge and clinical guidelines (standards of care)? (No 1 - 2 - 3 - 4 - 5 Yes)	This question is to evaluate the accuracy, reliability, and consistency of the eligibility criteria against established medical knowledge and accepted clinical guidelines. *Ex) Trial title: A Phase 3, Multi-Center, Open-Label Study to Assess the Diagnostic Performance and Clinical Impact of 18F-DCFPyL PET/CT Imaging Results in Men With Suspected Recurrence of Prostate Cancer* *Suspected recurrence of prostate cancer based on rising PSA after definitive therapy on the basis of: - Post-radical prostatectomy: Detectable or rising PSA that is 0.2 ng/mL with a confirmatory PSA 0.2 ng/mL (American Urological Association)*

Table 14: (continued) Evaluation category for assessing the appropriateness of EC sets