






Research and Applications

Analysis of eligibility criteria clusters based on large language models for clinical trial design

Alban Bornet , PhD^{1,*}, Philipp Khlebnikov , MSc², Florian Meer , MSc², Quentin Haas, PhD², Anthony Yazdani , MSc¹, Boya Zhang , MSc¹, Poorya Amini , PhD², Douglas Teodoro , PhD^{1,*}

¹Department of Radiology and Medical Informatics, University of Geneva, 1202 Geneva, Switzerland, ²Risklick AG, 3013 Bern, Switzerland

*Corresponding authors: Alban Bornet, PhD, Department of Radiology and Medical Informatics, University of Geneva, Campus Biotech, G6-N3, 9 Chemin des Mines, 1202 Geneva (CH), Switzerland (alban.bornet@unige.ch) and Douglas Teodoro, PhD, Department of Radiology and Medical Informatics, University of Geneva, Campus Biotech, G6-N3, 9 Chemin des Mines, 1202 Geneva (CH), Switzerland (douglas.teodoro@unige.ch)

Abstract

Objectives: Clinical trials (CTs) are essential for improving patient care by evaluating new treatments' safety and efficacy. A key component in CT protocols is the study population defined by the eligibility criteria. This study aims to evaluate the effectiveness of large language models (LLMs) in encoding eligibility criterion information to support CT-protocol design.

Materials and Methods: We extracted eligibility criterion sections, phases, conditions, and interventions from CT protocols available in the ClinicalTrials.gov registry. Eligibility sections were split into individual rules using a criterion tokenizer and embedded using LLMs. The obtained representations were clustered. The quality and relevance of the clusters for protocol design was evaluated through 3 experiments: intrinsic alignment with protocol information and human expert cluster coherence assessment, extrinsic evaluation through CT-level classification tasks, and eligibility section generation.

Results: Sentence embeddings fine-tuned using biomedical corpora produce clusters with the highest alignment to CT-level information. Human expert evaluation confirms that clusters are well structured and coherent. Despite the high information compression, clusters retain significant CT information, up to 97% of the classification performance obtained with raw embeddings. Finally, eligibility sections automatically generated using clusters achieve 95% of the ROUGE scores obtained with a generative LLM prompted with CT-protocol details, suggesting that clusters encapsulate information useful to CT-protocol design.

Discussion: Clusters derived from sentence-level LLM embeddings effectively summarize complex eligibility criterion data while retaining relevant CT-protocol details. Clustering-based approaches provide a scalable enhancement in CT design that balances information compression with accuracy.

Conclusions: Clustering eligibility criteria using LLM embeddings provides a practical and efficient method to summarize critical protocol information. We provide an interactive visualization of the pipeline [here](#).

Key words: clinical trials; eligibility criteria; natural language processing (NLP); LLMs; clustering; topic modeling.

Introduction

Clinical trials (CTs) are essential for advancing medical knowledge and improving patient care by systematically assessing the safety and efficacy of new treatments and interventions.^{1,2} A critical component of CTs is the protocol, which defines the eligibility criteria—specific characteristics that determine which participants are included or excluded from the study. Eligibility criteria balance the necessity to enroll enough participants with the need for a homogeneous study population and excluding individuals for whom the intervention could be unsafe.³

Developing the eligibility section of the study protocol is a critical and complex task in CT design. On the one hand, underrestrictive criteria may lead to heterogeneous study populations, which can compromise the validity and efficacy of the CT results^{4,5} and increase the risk of adverse events.⁶ On the other hand, overly restrictive criteria can significantly hinder participant recruitment.⁷ It was shown that 1 CT in 5 fails to recruit enough participants, and most of them

encounter delays because of the recruitment process.^{8,9} H analysis suggests that the number of participants is one of the main factors for clinical failure.^{10–12} Moreover, restrictive criteria have a negative impact on the generalizability of CT results, which limits their range of application in future patients^{13–15} and reduces their effectiveness in underrepresented populations.¹⁶ Hence, finding the optimal level of inclusivity and specificity in eligibility criteria is a major challenge in CT design.

Given this challenge, there is a growing research trend to develop data-driven solutions for optimizing and understanding CT data.¹⁷ Publicly accessible databases of clinical studies, such as ClinicalTrial.gov,¹⁸ can be exploited by machine learning (ML) algorithms to analyze and predict CT data. Machine learning research on eligibility criterion classification, recruitment prediction, patient-trial matching, or operational efficiency has met significant advancements in the last few decades (see references^{7,19} for recent reviews).

Background and significance

Classic ML techniques have been applied to exploit the diverse features found in CT protocols. One branch explored the representation of free-text eligibility criteria into structured data using information extraction ML methods.^{20–22} The goal is usually to help patient enrollment by automatically matching CTs based on patient medical records and estimate the number of eligible patients. Similarly, statistical modeling and various ML algorithms were used to predict recruitment rates and recruitment success in CTs.^{23–25} Moreover, gradient boosting algorithms were trained to predict more specific features of operational efficiency in CTs, such as screen failure ratio, dropout ratio, or preenrollment delay.²⁶ Despite these successes, classic ML methods often struggle with the vast and varied nature of CTs and eligibility criteria, which requires continuous algorithm adaptation to their format.

In response to these limitations, recent advancements in natural language processing greatly improved the vectorial representation of unstructured text data (ie, embeddings), such as the ones found in CT protocols, for classification and information extraction tasks. For example, pretrained word embeddings were used to learn and predict eligibility for cancer CTs,²⁷ and active learning was included in word embeddings to automatically classify eligibility criteria and enhance the efficiency of patient recruitment.²⁸

Attention-based models like BERT²⁹ and its large language model (LLM) variants better handle long-range dependencies and context-specific meanings, leading to improved performance with CT-protocol language. For example, BERT embeddings were used to match patients with appropriate CTs, and improved performance over classic language model baselines.³⁰ Large language models can be pretrained with large databases from the general domain and further pretrained with biomedical literature to generate embeddings that are more suited for clinical tasks, such as PubMedBERT³¹ or BioBERT.³² They can also be fine-tuned on more specific data domains or relevant tasks, which further improves the representation of CT data and patient records. For example, a study fine-tuned BioBERT on CT data and improved performance in a CT-retrieval task.³³ Large language models can also be flexibly integrated into complex architectures. For example, BERT was used to encode the content of leave nodes in a graph structure representing CT protocols for risk prediction and identified enrollment count as the main factor contributing to design-related risk.^{34,35} More recently, with the advent of generative LLMs, even more automated solutions were proposed to assist or improve CT design,³⁶ patient-to-trial matching,³⁷ and participant recruitment.³⁸

Recent work trained a BERT-like model with contrastive learning and rephrasing via generative LLMs to improve the quality of eligibility criterion embeddings.³⁹ Interestingly, clustering eligibility criteria was used to enhance the generation of negative sampling pairs. As semantically equivalent criteria can be described using different expressions, the current study focuses on evaluating the quality and relevance of the information available in semantic clusters. Relying on criterion clusters can identify common patterns and essential information across a large historical dataset of CTs while discarding irrelevant historical details. Moreover, clusters can be used to retrieve critical eligibility criteria which are more likely to be relevant and broadly applicable for a given set of phase(s), condition(s), and/or intervention(s).

To evaluate the amount and quality of the information available in criterion clusters, we performed 3 experiments.

- 1) First, we intrinsically evaluated the alignment between generated clusters and information relevant to CT-protocol design by measuring mutual information scores between clusters and CT labels. Moreover, we directly assessed the coherence of generated eligibility criterion clusters using a human expert analysis.
- 2) Second, we extrinsically evaluated information retained in clusters by using them as input features for ML models to classify CT-level outcomes and comparing the performance to using raw eligibility criterion embeddings.
- 3) Third, we analyzed the effectiveness of semantic eligibility clusters to assist the design of new CT protocols by generating eligibility sections and compared the results against those provided by a generative LLM prompted with CT information.

Materials and methods

Dataset

We extracted a dataset of eligibility criteria from the ClinicalTrials.gov registry,¹⁸ queried on June 17, 2024. Descriptive statistics for the dataset are available in [Table S1A](#). From the database, we selected only interventional CTs:

- 1) whose status was either completed or terminated,
- 2) whose phase section included phases 1, 2, 3, or 4, and
- 3) whose study started between January 1, 2000 and June 1, 2024.

Then, we split the criterion section of each CT into a list of individual criteria using a custom parsing algorithm based on sentence tokenization⁴⁰ and regular expression matching. Each criterion was associated with related CT-level information: phase(s), condition(s), intervention(s). Each condition and intervention were mapped to their corresponding MeSH⁴¹ tree ID. Moreover, regular expression matching was used to identify, whenever possible, whether a criterion was exclusive or inclusive. The final raw dataset was written to a csv file in which each row contains an inclusion or exclusion criterion, the associated CT ID, matching condition and intervention MeSH IDs, and phase(s). The full dataset creation pipeline is shown in [Figure 1A](#). The scripts to extract and build the dataset are available in our code repository.

Cluster generation

To cluster eligibility criteria based on LLM embeddings, we implemented a pipeline with 3 main steps—text embedding, dimensionality reduction, and clustering ([Figure 1B](#)). We adapted the BERTopic⁴² package, which integrates most required functionalities. BERTopic combines BERT-like embeddings with a clustering algorithm and a class-based TF-IDF (which we did not use in our analysis), making it particularly suitable for organizing and summarizing the complex information found in CTs.

To embed eligibility criterion texts, we compared 4 different pretrained language models: BERT,²⁹ Sentence-BERT,⁴³ a BERT model fine-tuned to produce sentence embeddings, PubMed-BERT,³¹ a BERT model further pretrained on biomedical literature, and PubMed-Sentence-BERT,⁴⁴ a PubMed-BERT model fine-tuned to produce sentence

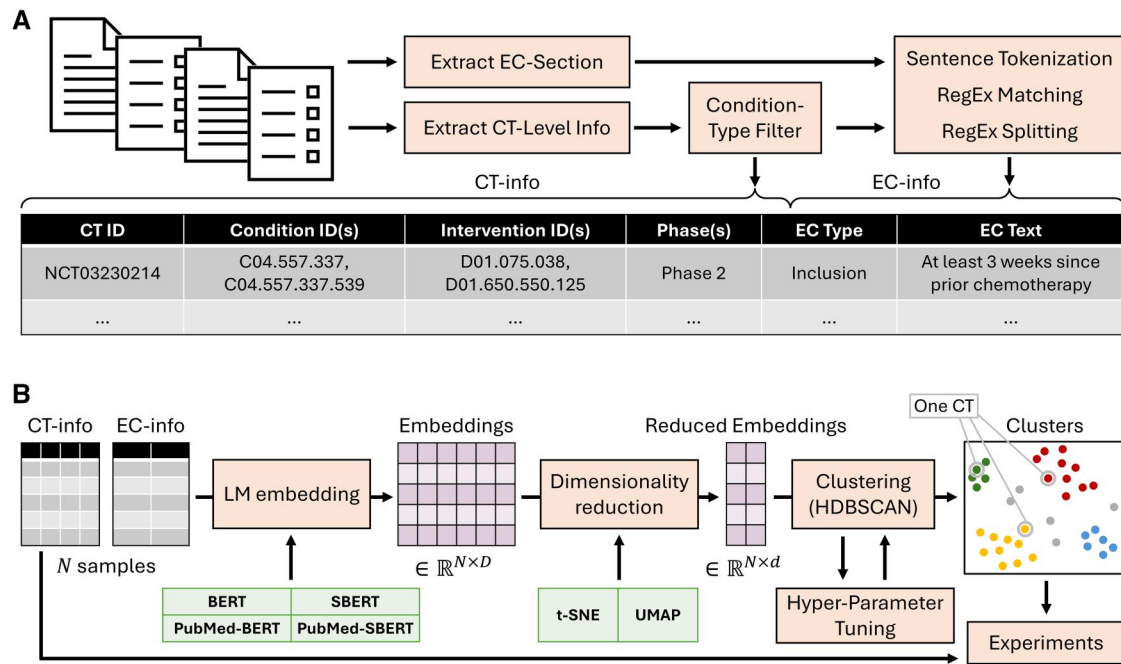


Figure 1. (A) Eligibility criterion dataset generation. (B) Cluster generation pipeline. One CT is composed of several criteria and can be represented by a set of cluster IDs. Abbreviations: CT, clinical trial; EC, eligibility criterion; SBERT, Sentence-BERT.

embeddings. Our goal was to determine the contribution of biomedical pretraining and the use of sentence embeddings in aligning criterion clusters with CT-protocol information. For the token-level models, we used the output embedding of the [CLS] token as the numerical representation of a criterion. For the sentence-level models, we used the average of all non-[PAD] tokens.

Based on previous results^{45,46} showing that reducing dimensionality improves the performance of clustering algorithms with biomedical data, we implemented a dimensionality reduction step in our pipeline. We empirically selected t-SNE⁴⁷ with dimension 2 as this configuration has proven effective in creating qualitative clusters in our prior experiments.^{45,46}

To improve the quality of the criteria clusters, we built a custom procedure on top of HDBSCAN⁴⁸ (BERTopic's default) for the clustering step. First, clusters were generated from the reduced embeddings using HDBSCAN. Then, reduced embeddings of each "primary" cluster were fed to a new instance of HDBSCAN to attempt further "secondary" clustering. Primary clusters were kept intact in case the algorithm did not converge. The hyperparameters of the clustering procedure were optimized using Optuna⁴⁹ with 100 trials and the Tree-structured Parzen Estimator (TPE) sampler.⁵⁰ The objective function was defined as follows:

$$O(\{C_i\}|\theta) = \text{Silhouette Score}(\{C_i\}|\theta) + \frac{N_{EC}}{N_{EC \in \{C_i\}}(\{C_i\}|\theta)},$$

where $\{C_i\}$ is the result of the evaluated clustering procedure, θ is the set of hyperparameters, N_{EC} is the total number of eligibility criteria, and $N_{EC \in \{C_i\}}$ is the number of criteria with an assigned cluster (HDBSCAN allows for unassigned samples). The set of hyperparameters that were optimized by Optuna and their value ranges are described in [Supplementary Information S2](#). To efficiently test many conditions and

explore large hyperparameter ranges, we used GPU-adapted versions of t-SNE, HDBSCAN, and Silhouette Score from the CuML suite.⁵¹

To improve the interpretability of criteria clusters, we created interactive HTML visualizations to explore the spatial distribution of clusters and their underlying structure. In these visualizations, the reader can navigate the eligibility criterion embedding space and hover over the different samples to reveal detailed information about individual criteria (eg, raw text and assigned cluster labels). In these visualizations, we generated descriptive cluster labels with the output of a generative LLM (GPT-3.5-Turbo). The LLM was tasked with summarizing the raw text extracted from a representative sample of 20 eligibility criteria from each cluster. The template used to create the prompts is shown in [Supplementary Information S3](#). The labels were not used for cluster evaluation, but only to help their visualization and interpretability. The HTML visualization files are available for download (https://minhaskamal.github.io/DownGit/#/home?url=https://github.com/ds4dh/eligibility_criterion_clustering/tree/master/visualizations) (link created using DownGit⁵²) and can be opened in a standard web browser.

We assessed cluster quality through 3 experiments summarized in [Figure 2](#) and detailed in the next subsections. All experiments were run for different MeSH condition types: we filtered CT data to include only those with at least 1 condition matching C01 (infections), C04 (neoplasms), C14 (cardiovascular diseases), and C20 (immune system diseases).

Experiment 1—intrinsic evaluation—alignment with CT-level information

To evaluate how eligibility clusters are aligned with the underlying structure of CTs, we measured the similarity of each cluster with CT-level information ([Figure 2A](#)). Each eligibility criterion was associated with a unique label, combining 1 CT phase, 1 condition, and 1 intervention. To reflect

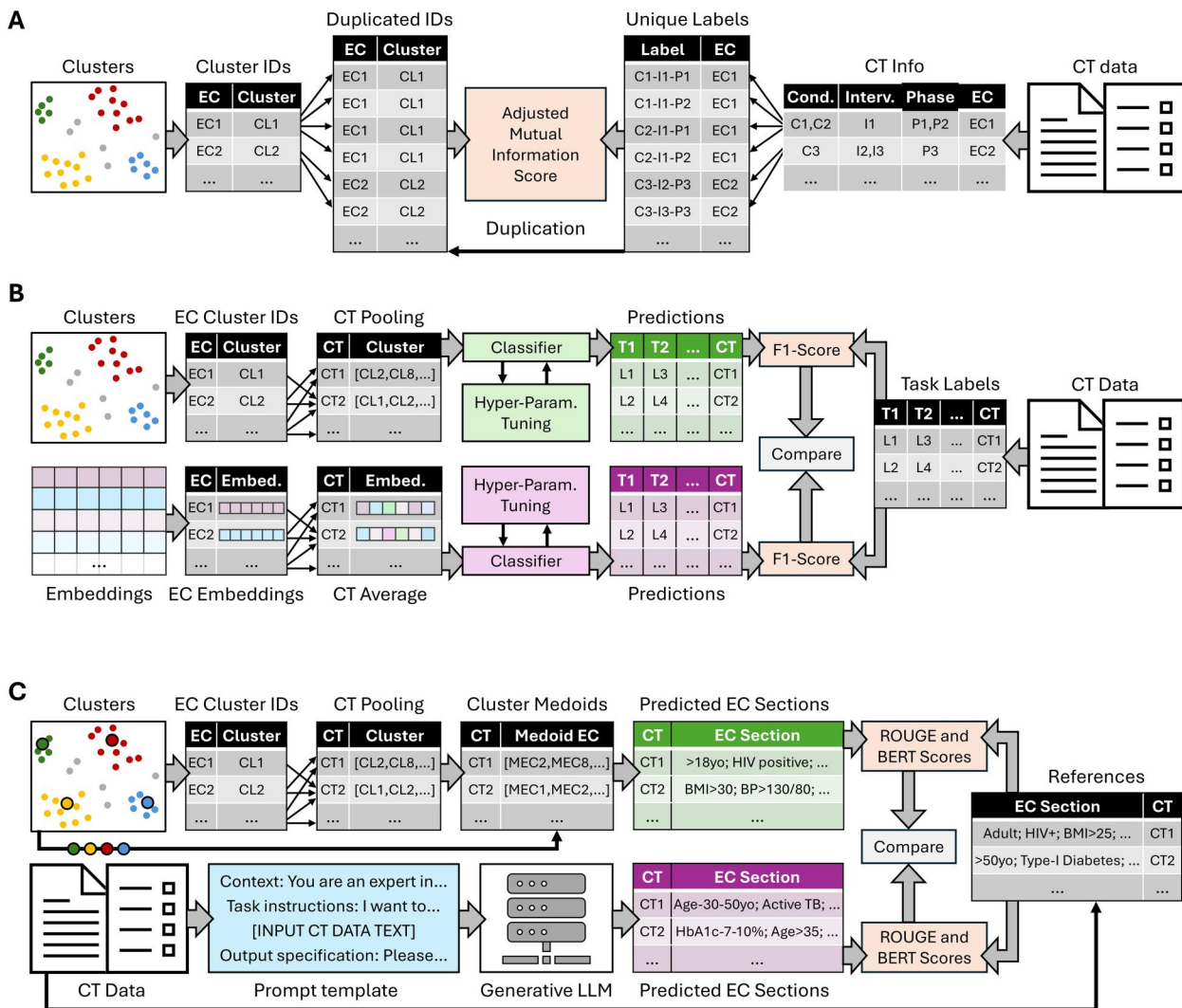


Figure 2. (A) Experiment 1—alignment with CT-level information. (B) Experiment 2—CT-level classification. (C) Experiment 3—eligibility section generation. Abbreviations: C, condition; CL, cluster; CT, clinical trial; EC, eligibility criterion; I, intervention; L, label; MEC, medoid EC; P, phase; T, task.

the required specificity of CTs, condition and intervention labels were derived from MeSH tree IDs at 3 different levels of granularity: fourth level for conditions and third level for intervention, third level and second level, and finally second level and first level. Eligibility criteria from CTs with several phases, conditions, or interventions were duplicated for evaluation using each possible unique label combination. For example, if 1 criterion belonged to a CT with phases P_i and P_j , intervention I_k , and conditions C_m and C_n , the corresponding duplicated samples would have the following labels: “ $P_i-I_k-C_m$,” “ $P_i-I_k-C_n$,” “ $P_j-I_k-C_m$,” and “ $P_j-I_k-C_n$.” We evaluated the alignment between clusters and CT-level information by computing the amount of mutual information between cluster IDs and the constructed labels. Given the high number of possible labels, we used the adjusted mutual information (AMI) score to correct the effect of random alignment between clusters and labels.⁵³ The resulting scores were compared across all models and condition types.

To contextualize model performance, we computed theoretical maximum AMI scores. Since eligibility criteria within a CT address distinct aspects of the eligibility section, each criterion is likely to fall in a different cluster. Hence, achieving an AMI score of 1.0 is practically impossible because labels are defined

at the CT level. Under the assumption that each eligibility criterion within an eligibility section falls in a different cluster, the best alignment between cluster labels and true CT-based labels corresponds to exactly 1 eligibility criterion per CT assigned to the correct cluster. Based on these considerations, we computed ceiling AMI scores for all condition types.

In addition, we asked a human expert to evaluate the quality of the clusters given a set of 1000 eligibility criteria. The clustering pipeline was applied to CTs with conditions matching “C04” (Neoplasms). From the 50 largest clusters, 20 eligibility criteria closest to their cluster’s medoid were selected for evaluation. The expert evaluated whether the raw text of each eligibility criterion was correctly assigned to its respective cluster, considering the other criteria within the same cluster. Possible answers were “correct,” “not correct,” and “unclear.” The performance was computed as the number of correctly assigned samples over the total number of samples.

Experiment 2—extrinsic evaluation—CT-level classification

To extrinsically evaluate the quality of criterion clusters, we selected the best embedding model from Experiment 1 and

used the generated clusters as input features for different CT-level classification tasks (Figure 2B). The goal of this experiment was to measure how much protocol information is retained by cluster IDs only, as compared to raw criterion embeddings. After splitting CTs into training (70%), validation (10%), and test (20%) sets, we trained regularized regression models (Lasso Regression,⁵⁴ Ridge Regression,⁵⁵ Elastic Net⁵⁶) to predict CT-level outcomes such as phase, study duration, and enrollment count. For continuous outcomes, we created binary labels by splitting the data evenly, with 50% of instances in each class. Each CT was represented by a vector of length N , where N is the number of identified clusters. Each element in the vector indicated the frequency of a particular criterion cluster within that CT. Then, we compared performance to using the raw embeddings from which the clusters were derived as input features. Each CT was represented by the average of its eligibility criterion embeddings, that is, by a vector of length D , where D is the dimension of the embedding model.

Classifier hyperparameters were optimized with Optuna, setting validation F1 score as the objective function. The choice of the classification algorithm was part of the hyperparameter set. Ranges for all hyper parameters are described in Supplementary Information S2. For each condition, the algorithm with the best selected parameters was retrained using the training and validation splits and evaluated with F1 score computed on the test set.

Experiment 3—utility for eligibility criterion section generation

To assess the utility of semantic-equivalent eligibility clusters in helping clinicians in the design of new CTs, we generated eligibility sections of study protocols only using cluster information (see Figure 2C). We first filtered for specific condition types (C01, C04, C14, C20) and then randomly selected, for each condition type, 100 CTs that included at least 1 matching condition. The process used to select evaluated CTs was as follows. For each potentially selected CT, we retrieved all CTs from ClinicalTrials.gov that shared at least 1 common phase, intervention, and condition, excluding the evaluated CT itself. Filters for condition and intervention were initially applied using fourth- and third-level MeSH tree IDs. If fewer than 5000 eligibility criteria were retrieved from the identified CTs, we relaxed the filters to third- and second-level MeSH IDs. If necessary, we further expanded to second- and first-level IDs. If insufficient criteria were still found, we moved to the next potentially selected CT. The final evaluation dataset included 400 CTs (ie, 100 for each condition type), from which the eligibility section was extracted as the target reference, and where each CT was mapped to a set of at least 2500 relevant eligibility criteria.

Using the best model from Experiment 1 and for each evaluated CT, we ran the clustering pipeline on the corresponding set of eligibility criteria to identify clusters expected to reflect relevant historical information. Then, we used the medoid and prevalence of the clusters to generate the eligibility section of the evaluated CT. For each cluster, we selected the criterion closest to the cluster medoid as a candidate. We then looped over selected criteria and added them to the generated section with a probability proportional to the corresponding cluster prevalence. We added criteria until the average number of eligibility criteria per CT, based on the condition type (21 for C01, 30 for C04, 21 for C14, and 24 for C20), was

reached. We measured the quality of all generated sections with ROUGE⁵⁷ and BERTScore⁵⁸ metrics, using the eligibility section of the corresponding CTs in the evaluation set as references.

Finally, we compared the performance of the clustering method to a more expensive approach using a generative LLM, GPT-3.5-Turbo. The LLM was prompted with task instructions and output specifications, followed by the whole text of the evaluated CT, from which only the eligibility section was removed. The exact prompt template is shown in Supplementary Information S3. Large language model-generated sections were evaluated with ROUGE and BERTScore metrics, using the same references as for the cluster method.

As a random baseline for both generation methods, we computed the ROUGE and BERTScore metrics against reference sections randomly sampled from the whole ClinicalTrials.gov dataset. This random baseline assesses each generation method's lexicosemantic similarity with CT protocols in general. The difference between a method's score and its random baseline measures how accurately the method captures the specific content of the target CT. We also generated a theoretical maximum of performance by prompting GPT-3.5-Turbo, tasked with reformulating the eligibility section of each evaluated CT 10 times, from which we computed average ROUGE and BERTScore metrics. The template used to create the reformulation prompts is shown in Supplementary Information S3.

Results

In this section, we present the results of our 3 experiments, each aimed at evaluating different aspects of the information contained within eligibility criterion clusters. All statistical comparisons were done using paired t -tests, with a significance threshold set at an alpha level of 0.05. To control for the increased likelihood of type I errors in multiple comparisons, we also report significance after Bonferroni correction, which divides alpha by the number of comparisons.

Experiment 1—intrinsic evaluation—alignment with CT-level information

We generated 4 sets of clusters based on eligibility criteria from all CTs matching conditions C01, C04, C14, and C20. Figure 3 shows the eligibility clusters extracted from CTs matching condition type C01. Figure S4A-C provides the visualizations for other condition types. Clusters produced with sentence embedding are more separated. Moreover, embeddings further pretrained with biomedical data produce more structured clusters and more samples find a cluster. Interactive HTML visualizations for all condition types are available for download (https://minhaskamal.github.io/DownGit/#/home?url=https://github.com/ds4dh/eligibility_criterion_clustering/tree/master/visualizations).

Figure 4 presents the results of experiment 1, which are in line with the qualitative assessment of clusters in Figure 3. Clusters were evaluated intrinsically by computing the alignment between cluster IDs and labels relevant to CT-protocol design, using AMI. Figure 4A shows the AMI obtained by all embedding models stratified by condition type, for different levels of granularity in the CT-based labels. We normalized AMI values using the ceiling performance inherent to each condition type (see “Methods” for detailed computation). Figure S4D shows the absolute AMI scores obtained with the

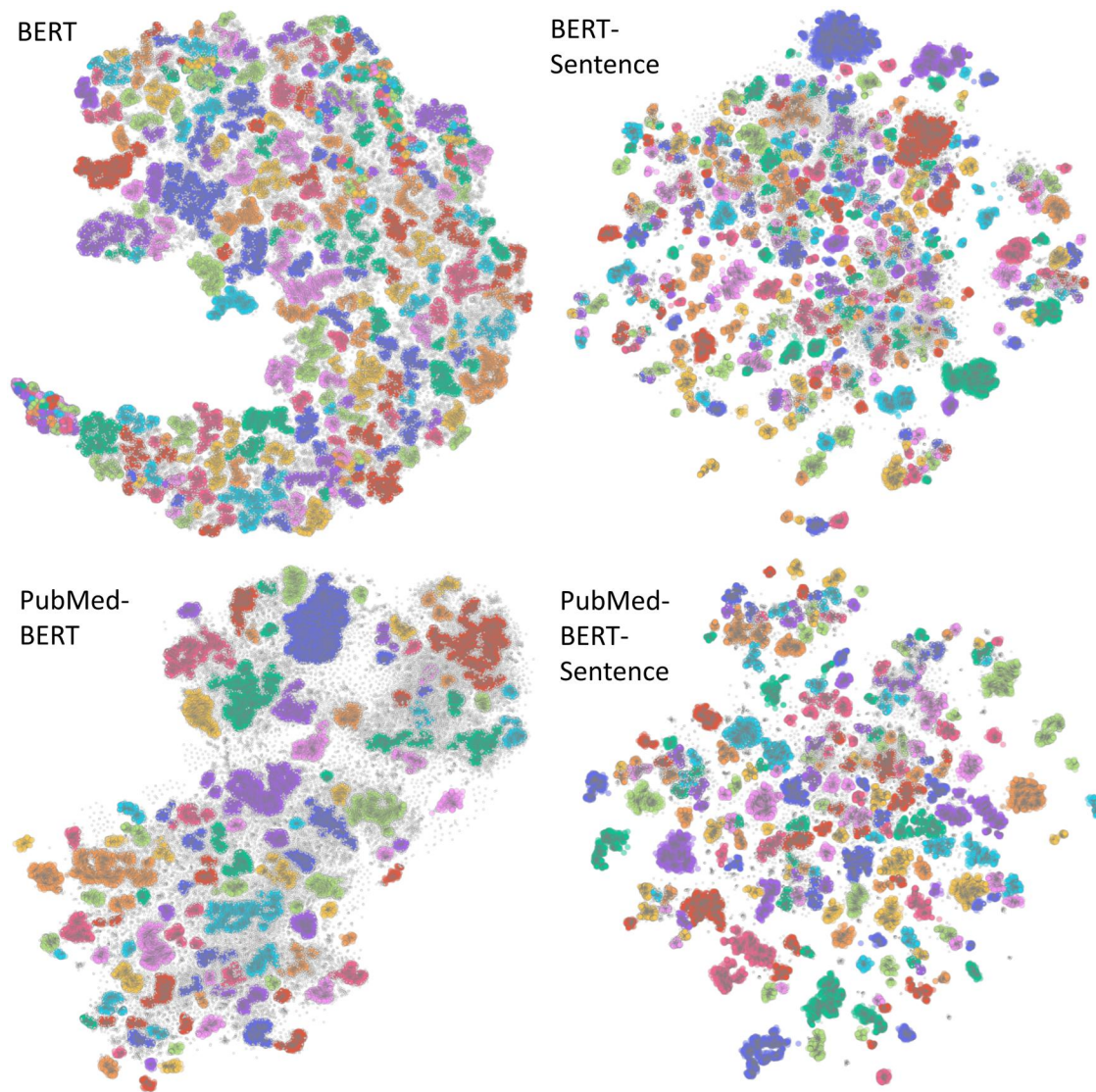


Figure 3. Eligibility clusters extracted from CT protocols matching at least 1 condition starting with C01. Clusters are represented by different colors and gray dots are samples that were not assigned any cluster by the HDBSCAN algorithm. N_{ECs} : 76 695; N_{CTs} : 4655; $N_{clusters}$: {BERT: 245, BERT-Sentence: 200, PubMed-Bert: 310, PubMed-BERT-Sentence: 328}. Abbreviation: CT, clinical trial.

clustered embeddings. Clusters generated by the PubMed-Sentence-BERT model exhibit higher alignment with protocol information compared to other models, across all tested condition types, and reach between 40% and 90% of the ceiling performance, depending on the condition type and label granularity. Moreover, both using sentence embeddings and using embeddings further pretrained with biomedical data consistently improves cluster alignment.

Figure 4B shows the results of the human expert cluster quality assessment for the best model identified in Figure 4A, that is, PubMed-Sentence-BERT. Using 1000 criteria, the proportion of criteria the human expert deemed correctly assigned to its cluster by the model is 84.3%. This suggests that sentence-level embeddings further pretrained with biomedical data are effective in producing coherent eligibility clusters.

Experiment 2—extrinsic evaluation—CT-level classification

Experiment 2 extrinsically evaluates the CT-level information encoded by the eligibility clusters. Using only the cluster IDs,

different CT-level information is predicted. The results are compared to using raw embeddings as the upper bound, as they encode the maximum amount of eligibility information available to the clustering algorithm, and using random vectors as the lower bound, which do not encode any CT-level information. The random condition used uniform random vectors of the same dimensionality as for the cluster IDs condition. As for experiment 1, this experiment was run separately for condition filters C01, C04, C14, and C20. For the phase classification task, all filtered samples were used. For the other classification tasks, the training, validation, and evaluation procedures were separated by phase. This means that there are 4 data points (each condition-type filter) reported for the phase classification task, and 16 data points (4 condition-type filters \times 4 phase filters) reported for all other classification tasks.

Figure 5 shows the macroaveraged F1 scores for different classification tasks and input features. For each classification task, we performed paired *t*-tests to compare the performance reached using cluster IDs to using raw embeddings, as well as

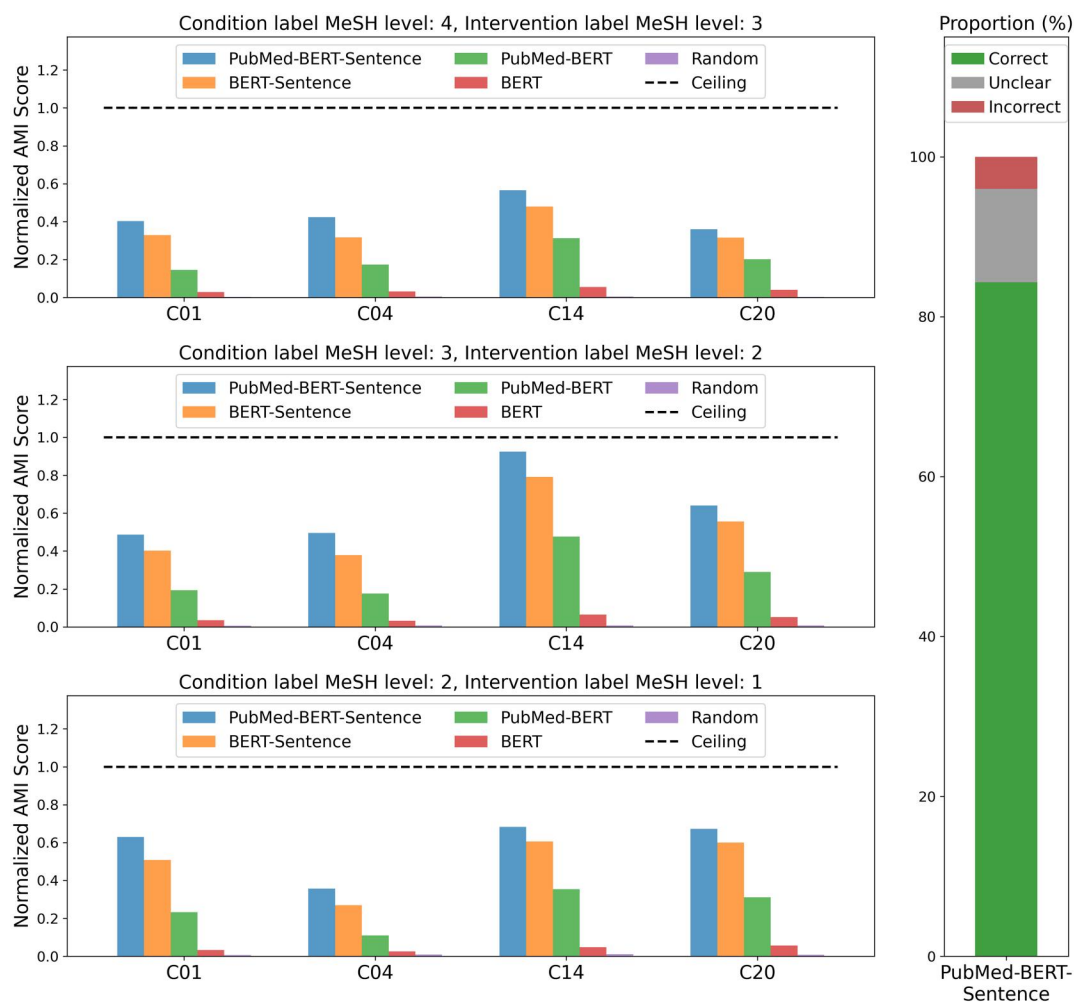


Figure 4. (A) Normalized AMI scores for alignment with CT-protocol labels across different levels of label granularity and condition types. C01— infections; C04—neoplasms; C14—cardiovascular diseases; C20—immune system diseases. (B) Human evaluation of cluster quality using PubMed-Sentence-BERT to embed eligibility criteria. Abbreviation: CT, clinical trial.

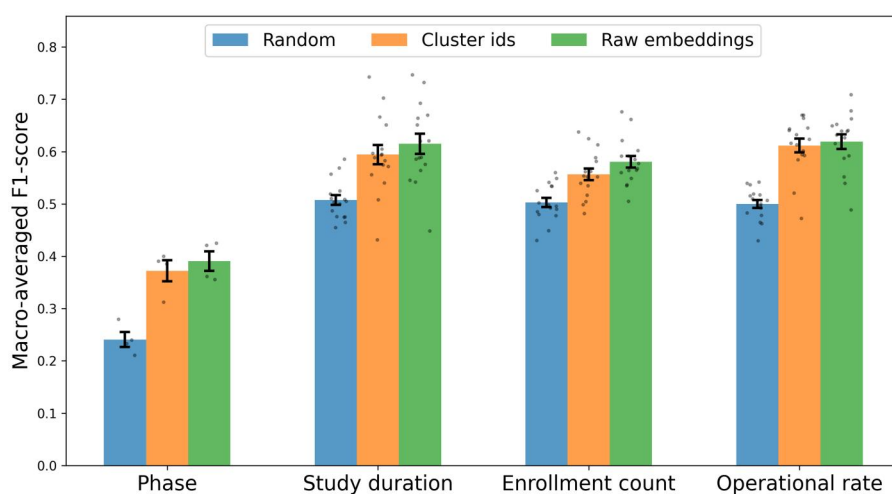


Figure 5. F1 scores for CT-level classification tasks comparing cluster-based features and raw embeddings. Error bars represent the SE of the mean. Abbreviation: CT, clinical trial.

using both input feature types to using random inputs (ie, 12 comparisons in total). Performance using both cluster IDs and raw embeddings significantly outperforms the random

baseline before and after Bonferroni correction (all task $P < .004$ for cluster IDs, and all task $P < .002$ for raw embeddings). Moreover, comparing performance using cluster IDs

to using raw embeddings did not show any significant difference, before or after Bonferroni correction (P -values ranging from .135 for enrollment count classification to .626 for operational rate classification). Still, scores obtained with raw embedding are slightly higher than with cluster IDs. In terms of mean values, cluster IDs achieve on average scores that reach 97% of the scores obtained with raw embeddings, while the difference with the random baseline corresponds to 84% of the difference obtained between raw embeddings and the random baseline. These results suggest that, despite significant compression compared to the raw embeddings, the clusters effectively retain the majority of essential information contained in CTs.

Experiment 3—utility for eligibility criterion section generation

Experiment 3 uses a generation task to assess the extent to which cluster information is useful to assisting clinicians in the creation of eligibility sections for new CT protocols. For each condition type (C01, C04, C14, C20), we randomly selected 100 CT protocols as the out-of-sample evaluation dataset, which we excluded from the analysis. We compared sections generated using eligibility clusters extracted from similar CT protocols to sections generated by prompting GPT-3.5-Turbo with CT-protocol details. Figure 6A shows the examples of cluster sets from which eligibility criteria were sampled using the cluster method. Interactive visualizations for criteria clusters extracted from similar CTs to different example target trials are available for download (https://minhaskamal.github.io/DownGit/#/home?url=https://github.com/ds4dh/eligibility_criterion_clustering/tree/master/visualizations). We also analyze key details of the output of clustering pipeline in Figure S6J-L.

Figure 6B shows the average of the ROUGE-1, ROUGE-2, and ROUGE-L F1 scores, as well as BERTScore F_{BERT} (harmonic mean between recall R_{BERT} and precision P_{BERT} , using SciBERT), computed over all samples of the evaluation dataset, for both methods and their associated random baselines. First, it should be noted that all scores are significantly lower than ceiling performance, which was computed by reformulating reference eligibility sections. Then, for both metrics, we performed paired t -tests to compare the cluster and generative LLM methods, as well as each method against its random baseline (ie, 8 comparisons in total). Both the clustering and LLM methods significantly outperform their random baselines before and after Bonferroni correction, both in terms of average ROUGE F1 score ($P < .001$) and in terms of BERTScore F_{BERT} ($P < .001$). Moreover, the scores obtained with the clustering and LLM methods are significantly different from each other, both before and after Bonferroni correction ($P < .001$ for both metrics). This means that, when considering the 400 CTs from the evaluation dataset, the LLM method outperforms the clustering method. Still, in terms of mean values, the clustering method reaches 95% of the ROUGE F1 score and 98% of the BERTScore obtained with the generative LLM method (despite relying on an embedding model with $O(10^3)$ fewer parameters). Moreover, the difference between the average ROUGE F1 score obtained using the cluster method and its random baseline reaches 58% of the difference between the generative LLM method and its random baseline (and 65% for the BERTScore F_{BERT}). Finally, the random baseline of the clustering method significantly outperforms the random baseline of the LLM method

for the ROUGE F1 score, both before and after Bonferroni correction ($P < .001$).

Figure S5E presents the same comparisons, stratifying by condition type, and provides details for ROUGE-1, ROUGE-2, and ROUGE-L metrics as well as BERTScore, computed with different BERT models (Figure S5G for F1 scores, Figure S5H for recalls, Figure S5I for precisions). Finally, Figure S5F, shows a small, yet significant correlation between cluster quality, measured by Silhouette Score, and the metrics used in Figure 6B ($r = 0.18$ for average ROUGE F1 score, $r = 0.23$ for F_{BERT} ; $P < .01$ for both metrics). In contrast, no significant correlation is observed when using the random baseline scores ($P = .23$ for average ROUGE F1 score, $P = .58$ for F_{BERT}).

Discussion

This study explores how eligibility criterion clusters extracted from CTs using encoder-based LLM embeddings align with CT-protocol information and their effectiveness in summarizing complex CT data. The results obtained in our experiments show that clusters encode lexicosemantic eligibility criteria information offers insights that can be useful to improve the efficiency of CT design.

Alignment of clusters with protocol-design information

The intrinsic evaluation of the clusters (experiment 1) aims to directly evaluate their alignment with information relevant to protocol design. The AMI scores achieved by the different embedding models suggest that sentence-level embeddings fine-tuned on biomedical literature are the most effective in capturing relationships between different eligibility criteria and their corresponding CT characteristics. This highlights the importance of using tailored embeddings for extracting meaningful patterns from complex, unstructured clinical text data. Moreover, human expert evaluation confirms the coherence of the clusters generated with our best embedding model. The accuracy score (84.3%) achieved by the best model shows that clusters generated with sentence-level embeddings tailored for the biomedical domain align well with expert judgments.

Retention of CT-level information in clusters

The extrinsic evaluation of the clusters through CT-level classification tasks (experiment 2) aims to assess how much relevant information is retained in compressed cluster representations. Although consistently lower, the cluster-based features achieve 97% of the performance obtained using raw embeddings. This suggests that semantic-equivalent eligibility criterion clusters, despite drastically compressing information (sparse, interpretable, 200- to 300-dimensional integer vectors instead of dense, noninterpretable, 768-dimensional floating-point vectors), retain most of the essential features present in raw embeddings. Cluster information offers a balance between data compression and the preservation of information which is crucial for scalable and efficient CT design. For example, when designing a new CT, screening all eligibility criteria from similar trials would be time-consuming and impractical, whereas cluster representations allow for rapid identification of relevant criteria while retrieving the most pertinent information. It should be noted that while raw embeddings are pooled using an average before being used as

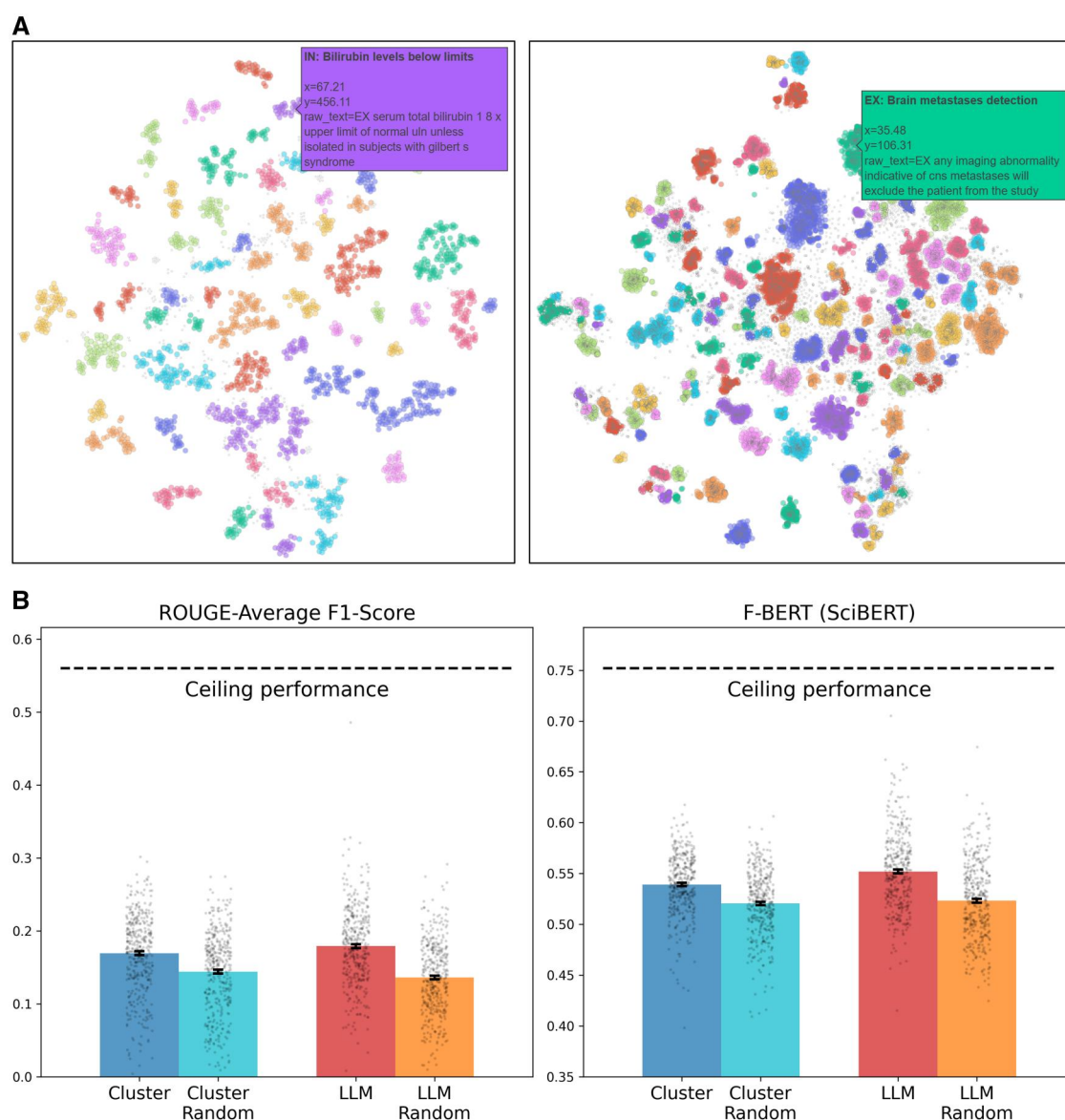


Figure 6. (A) Example visualizations of the embedding space from which eligibility criteria are sampled with the cluster method. (Left) 2817 eligibility criteria for a phase 3 CT evaluating an antiviral treatment for hepatitis C. (Right) 23 647 eligibility criteria for a phase 2 CT investigating a cell signaling inhibitor for metastatic kidney cancer. (B) Average ROUGE F1 score and F_{BERT} (using SciBERT) for eligibility section generation using cluster medoids vs generative LLM prompting, and corresponding random baselines. Error bars represent the SE of the mean. Abbreviations: CT, clinical trial; LLM, large language model.

features for classification, their high dimensionality reduces the likelihood of significant information loss.

Comparison with LLM outputs

To assess the utility of cluster information for CT-protocol design, we generated entire eligibility sections for CT protocols (experiment 3). Eligibility clusters derived from PubMed-Sentence-BERT embeddings perform reasonably well compared to outputs from a generative LLM (GPT-3.5-Turbo) prompted with nearly all information contained in the CT. Using only the semantic eligibility cluster information achieves 95% of the ROUGE scores and 98% of the BERT scores reached by the generative LLM.

Still, all obtained scores are significantly lower than the ceiling performance, computed by reformulating reference eligibility sections. While clusters and LLM-generated

eligibility sections show comparable alignment with reference ECs, both methods do not achieve clinically meaningful equivalence. Importantly, the metrics we used—ROUGE and BERTScore—only serve as proxies for textual similarity and information retention rather than definitive measures of clinical validity. While using only eligibility clusters to draft entire eligibility sections of CT protocols is clearly not sufficient, the results of experiment 3 suggest that eligibility clusters contain information that is helpful for CT design. Still, our analysis was limited to cases where sufficient data could be extracted from ClinicalTrials.gov, given the set of phase(s), condition(s), and intervention(s) of the evaluated CT (ie, at least 2500 eligibility criteria). Expanding to larger or more diverse databases could mitigate this limitation.

It should be noted that the cluster random baseline, which assesses lexicosemantic similarity with CT protocols in

general, significantly outperforms the LLM random baseline. This suggests that eligibility sections generated using cluster information are more aligned with the actual vocabulary of CTs, as our method directly pulls eligibility criteria from relevant trials based on cluster relevance. This is consistent with the findings shown in [Figure S5H and I](#), where the cluster method generally outperforms the LLM method in terms of recall (sensitivity), while the LLM tends to show better precision.

Importantly, cluster quality, as measured by Silhouette Score, demonstrates a significant, yet small, correlation with the quality of the generated eligibility sections, as measured by both ROUGE and BERTScore (see [Figure S5F](#)). This suggests that when well-defined eligibility clusters are identified for a given CT, they are likely to provide valuable information for generating accurate eligibility sections. In contrast, when using random CT references to compute these metrics, no significant correlation is observed. This means that the improved performance of the cluster method over its random baseline is mostly due to the fact that cluster information helps generate sections that are more tailored to the specific requirements of the evaluated CT.

Implications for CT design

The findings of this study have significant implications for CT-protocol design. Clustering eligibility criteria using encoder-based LLMs trained on biomedical corpora retains critical CT-level information while reducing the complexity of the data that clinical researchers need to screen. With a sufficiently large and diverse CT database and given that specific filtering (eg, based on phase, conditions, and interventions) is provided, semantic-equivalent eligibility clusters could provide a compressed yet comprehensive overview of historically relevant eligibility criteria. A practical example is our interactive visualizations, which provide a means to navigate the clustered embedding space of eligibility criteria extracted from similar CTs and review cluster labels alongside the corresponding raw criteria texts. This added information can help clinical researchers identify trends and patterns across various CTs, leading to better decision-making in the initial design phases. The key idea is that, even though language model embeddings can be rather similar in terms of, for example, cosine similarity, assigning cluster labels adds interpretable information that can help differentiate between criteria addressing distinct aspects of the eligibility section. A clustering approach could ensure that no important information is overlooked during protocol design and that eligibility criteria are characterized by a wide range of pertinent historical data. Integrating such clustering methods into real-time data-driven tools could help the design process, ultimately improving the efficiency and success of CTs.

Limitations and future research

While our study presents interesting insights for CT-protocol design, the metrics presented in our experiments remain relatively low. For experiments 1 and 2, this is primarily due to the inherent challenge of comparing model outputs based on individual eligibility criteria to labels that are defined at the CT level. For experiment 3, future work could explore the development of more tailored metrics to better capture the quality and relevance of generated eligibility sections, for example, involving human expertise to assess clinical utility. Another concern is reproducibility, as methods like t-SNE

and HDBSCAN introduce inherent randomness even after fixing a seed throughout the pipeline, meaning that running the pipeline twice may yield slightly different cluster assignment patterns. Still, we observed that the outcome of the clustering algorithm led to results that were stable across runs (see, for example, [Figure S6J and K](#)). Moreover, further research is needed to explore the scalability of the clustering approach across a broader range of CTs, conditions, and interventions. Future work could investigate the integration of clustering methods in data-driven tools, such as real-time updating of eligibility criteria based on new trial data becoming available. The potential and benefit of cluster information for CT design based on historical information remains to be demonstrated in a real use-case scenario. Finally, an avenue for future research could extend the clustering approach to electronic health record (EHR) data. Mapping clusters of patients in EHRs to semantic clusters of criteria could facilitate feasibility studies and improve the efficiency and accuracy of patient-to-protocol matching.

Conclusion

In conclusion, summarizing eligibility criterion information using cluster analysis based on LLMs provides a balance between information compression and CT-protocol relevance, particularly when using sentence-level embeddings from medically tailored language models. Our findings offer insights that could be used to help the design of CT protocols.

Author contributions

Alban Bornet, Douglas Teodoro, and Poorya Amini conceived and designed the study. Alban Bornet was responsible for writing the manuscript, designing, implementing, and performing the experiments, as well as analyzing and visualizing the data. Quentin Haas contributed as the human expert for Experiment 1. Philipp Khlebnikov, Florian Meer, Anthony Yazdani, Boya Zhang, Poorya Amini, and Douglas Teodoro provided critical feedback on the manuscript. All authors reviewed and approved the final version of the manuscript.

Supplementary material

[Supplementary material](#) is available at *Journal of the American Medical Informatics Association* online.

Funding

This work was supported by the Swiss Innovation Agency Innosuisse under the project with funding number 101.466 IP-ICT “CTxAI: quality by design of clinical studies using explainable AI”.

Conflicts of interest

The authors declare the following competing financial interest(s): P.K., F.M., Q.H., and P.A. work for Risklick AG. All other authors declare no competing financial interest.

Data availability

The source data for this study is publicly available from ClinicalTrials.gov. The processed dataset generated and analyzed

during the current study will be made available upon manuscript acceptance.

References

- Friedman LM, Furberg C, DeMets DL, et al. *Fundamentals of Clinical Trials*. Springer; 2010.
- Guyatt GH, Haynes RB, Jaeschke RZ, et al.; Evidence-Based Medicine Working Group. Users' guides to the medical literature: XXV. Evidence-based medicine: principles for applying the users' guides to patient care. *JAMA*. 2000;284:1290-1296.
- Van Spall HGC, Toren A, Kiss A, et al. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. *JAMA*. 2007;297:1233-1240.
- Gross CP, Mallory R, Heiat A, et al. Reporting the recruitment process in clinical trials: who are these patients and how did they get there? *Ann Intern Med*. 2002;137:10-16.
- Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet*. 2005;365:82-93.
- Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 2005;294:218-228.
- Su Q, Cheng G, Huang J. A review of research on eligibility criteria for clinical trials. *Clin Exp Med*. 2023;23:1867-1879.
- Hutson M. How AI is being used to accelerate clinical trials. *Nature*. 2024;627:S2-S5.
- Desai M. Recruitment and retention of participants in clinical studies: critical issues and challenges. *Perspect Clin Res*. 2020;11:51-53.
- Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. *Contemp Clin Trials Commun*. 2018;11:156-164.
- Harrison RK. Phase II and phase III failures: 2013–2015. *Nat Rev Drug Discov*. 2016;15:817-818.
- Williams RJ, Tse T, DiPiazza K, et al. Terminated trials in the ClinicalTrials.gov results database: evaluation of availability of primary outcome data and reasons for termination. *PLoS One*. 2015;10:e0127242.
- Sokka T, Pincus T. Most patients receiving routine care for rheumatoid arthritis in 2001 did not meet inclusion criteria for most recent clinical trials or American College of Rheumatology criteria for remission. *J Rheumatol*. 2003;30:1138-1146.
- Britton A, McKee M, Black N, et al. Threats to applicability of randomised trials: exclusions and selective participation. *J Health Serv Res Policy*. 1999;4:112-121.
- Masoudi FA, Havranek EP, Wolfe P, et al. Most hospitalized older persons do not meet the enrollment criteria for clinical trials in heart failure. *Am Heart J*. 2003;146:250-257.
- Heiat A, Gross CP, Krumholz HM. Representation of the elderly, women, and minorities in heart failure clinical trials. *Arch Intern Med*. 2002;162:1682-1688.
- Woo M. An AI boost for clinical trials. *Nature*. 2019;573:S100-S102.
- Zarin DA, Tse T, Williams RJ, et al. The ClinicalTrials.gov results database—update and key issues. *N Engl J Med*. 2011;364:852-860.
- Miller MI, Shih LC, Kolachalama VB. Machine learning in clinical trials: a primer with applications to neurology. *Neurotherapeutics*. 2023;20:1066-1080.
- Kang T, Zhang S, Tang Y, et al. EliIE: an open-source information extraction system for clinical trial eligibility criteria. *J Am Med Inform Assoc*. 2017;24:1062-1071.
- Yuan C, Ryan PB, Ta C, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. *J Am Med Inform Assoc*. 2019;26:294-305.
- Weng C, Wu X, Luo Z, et al. EliXR: an approach to eligibility criteria extraction and representation. *J Am Med Inform Assoc*. 2011;18:i116-i124.
- Bieganeck C, Aliferis C, Ma S. Prediction of clinical trial enrollment rates. *PLoS One*. 2022;17:e0263193.
- Liu J, Allen PJ, Benz L, et al. A machine learning approach for recruitment prediction in clinical trial design. arXiv preprint arXiv:2111.07407. 2021.
- Lan Y, Tang G, Heitjan DF. Statistical modeling and prediction of clinical trial recruitment. *Stat Med*. 2019;38:945-955.
- Wu K, Wu E, DAndrea M, et al. Machine learning prediction of clinical trial operational efficiency. *AAPS J*. 2022;24:57.
- Bustos A, Pertusa A. Learning eligibility in cancer clinical trials using deep neural networks. *Appl Sci*. 2018;8:1206.
- Chuan C-H. Classifying eligibility criteria in clinical trials using active deep learning. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Orlando, FL. IEEE; 2018:305-310.
- Devlin J. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018.
- Zhang X, Xiao C, Glass LM, et al. DeepEnroll: patient-trial matching with deep embedding and entailment prediction. In: *Proceedings of the Web Conference 2020*, Taipei Taiwan. ACM; 2020:1029-1037.
- Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthc*. 2022;3:1-23.
- Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36:1234-1240.
- Wang Z, Sun J. Trial2vec: zero-shot clinical trial document similarity search using self-supervision. arXiv preprint arXiv:2206.14719. 2022.
- Ferdowsi S, Knafou J, Borissov N, et al. Deep learning-based risk prediction for interventional clinical trials based on protocol design: a retrospective study. *Patterns*. 2023;4:100689.
- Ferdowsi S, Copara J, Gouareb R, et al. On graph construction for classification of clinical trials protocols using graph neural networks. In: *International Conference on Artificial Intelligence in Medicine*. Springer; 2022:249-259.
- Wang Z, Xiao C, Sun J. AutoTrial: prompting language models for clinical trial design. arXiv preprint arXiv:2305.11366. 2023.
- Jin Q, Wang Z, Floudas CS, et al. Matching patients to clinical trials with large language models. *Nat Commun*. 2024;15:9074.
- Guan Z, Wu Z, Liu Z, et al. Cohortgpt: an enhanced gpt for participant recruitment in clinical study. arXiv preprint arXiv:2307.11346. 2023.
- Kim S, Won JH, Lee D, et al. CReSE: benchmark data and automatic evaluation framework for recommending eligibility criteria from clinical trial information. In: *Findings of the Association for Computational Linguistics: EACL 2024*. ACL; 2024:2243-2273.
- Kiss T, Strunk J. Unsupervised multilingual sentence boundary detection. *Comput Linguist*. 2006;32:485-525.
- Lipscomb CE. Medical subject headings (MeSH). *Bull Med Lib Assoc*. 2000;88:265-266.
- Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794. 2022.
- Reimers M. Sentence-BERT: sentence embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084. 2019.
- Deka P, Jurek-Loughrey ANNA, Padmanabhan D. Improved methods to aid unsupervised evidence-based fact checking for online health news. *JDI*. 2022;3:474-504.
- Jaume-Santero F, Zhang B, Proios D, et al. Cluster analysis of low-dimensional medical concept representations from electronic health records. In: *Health Information Science: 11th International Conference, HIS 2022, Virtual Event, October 28-30, 2022*. Springer; 2022:313-324.
- Bornet A, Proios D, Yazdani A, et al. Comparing neural language models for medical concept representation and patient trajectory prediction. medRxiv. 2023:2023-06.
- Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9:2579-2605.

48. McInnes L, Healy J, Astels S. HDBSCAN: hierarchical density based clustering. *JOSS*. 2017;2:205.
49. Akiba T, Sano S, Yanase T, et al. Optuna: a next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019. ACM; 2019:2623-2631.
50. Bergstra J, Bardenet R, Bengio Y, et al. Algorithms for hyperparameter optimization. *Adv Neural Inf Process Syst*. 2011;24:2546-2554.
51. Raschka S, Patterson J, Nolet C. Machine learning in python: main developments and technology trends in data science, machine learning, and artificial intelligence. *Information*. 2020;11:193.
52. Minhas KM. DownGit [Internet]. Accessed November 25, 2024. <https://github.com/MinhasKamal/DownGit>
53. Xuan Vinh N, Epps J, Bailey J. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *J Mach Learn Res*. 2010;11:2837-2854.
54. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol*. 1996;58:267-288.
55. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 1970;12:55-67.
56. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*. 2005;67:301-320.
57. Lin C-Y. ROUGE: a package for automatic evaluation of summaries. In: *Text Summarization Branches Out*. ACL; 2004:74-81.
58. Zhang T, Kishore V, Wu F, et al. Bertscore: evaluating text generation with bert. arXiv preprint arXiv:1904.09675. 2019.