Clinical Trial Recommendation System Using NLP

Introduction

Clinical trials are essential for advancing medical research, yet patient recruitment remains a major bottleneck. Eligibility criteria for trials are typically written in long, unstructured text, which makes it difficult for clinicians to quickly identify suitable trials. As a result, many eligible patients remain un-enrolled, delaying research progress and leading to underrepresentation of certain populations. This project addresses this challenge by building a decision-support system that semantically matches patients to trials using natural language processing (NLP), ensures compliance with eligibility criteria, and provides explainable, ranked recommendations.

Exploring the Data

To simulate a realistic scenario while maintaining privacy, we combined synthetic patient profiles with a COVID-19 clinical trials dataset from Kaggle, which contains over 3,500 trials with metadata including age and sex criteria, trial phase, and inclusion and exclusion criteria. During initial exploration, we observed that most trials targeted adults aged 18 to 65 and included both sexes, with a smaller number focusing on pediatric or elderly populations. The inclusion and exclusion criteria were often long paragraphs, medically dense, and varied widely in terminology. For instance, the same condition might be referred to as “COVID-19 pneumonia,” “SARS-CoV-2 infection,” or “viral respiratory illness,” illustrating the high variability of clinical language. This observation indicated that simple keyword matching would be insufficient for accurate patient-trial matching, motivating the use of semantic embeddings. Visual exploration further supported these findings: histograms of trial age ranges confirmed that most studies targeted adults, bar charts of trial phases revealed that Phase 2 and 3 trials were predominant, and word clouds of eligibility text highlighted a diversity of keywords such as “covid,” “infection,” “hospitalization,” and “pneumonia.”

Data Preprocessing

Before modeling, the data required substantial cleaning and standardization. Inclusion and exclusion criteria were combined into a single `eligibility_text` column. The text was lowercased, punctuation and extra whitespace were removed, and age and sex columns were standardized to allow consistent eligibility filtering. Extreme values and missing entries were addressed to ensure accurate evaluation. This preprocessing step ensured that both the synthetic patient profiles and the clinical trial text were suitable for NLP-based modeling.

Modeling Approach

We implemented two approaches to patient-trial matching. The baseline model uses TF-IDF to vectorize trial text and compute cosine similarity with patient profiles, applying hard eligibility filters for age and sex. While transparent and interpretable, this method relies entirely on exact keyword overlap, which can miss relevant trials when terminology varies.

To overcome this limitation, we implemented a semantic model using Sentence-BERT. This model encodes both trial and patient text into dense vectors, capturing the semantic meaning beyond keywords, while still applying the same eligibility rules. For example, Sentence-BERT can recognize that “viral respiratory infection” is relevant for a patient with “COVID-19 pneumonia,” a connection that TF-IDF would likely miss. Both models rank trials according to similarity, allowing for top-K recommendations for each patient.

Evaluation

Because ground-truth labels for trial-patient matches are unavailable, we defined proxy relevance based on eligibility criteria compliance and condition-specific keywords. We evaluated both models using Precision@5, which measures how many of the top five recommended trials are relevant. The TF-IDF baseline achieved a Precision@5 of 0.8, whereas the Sentence-BERT model achieved a perfect 1.0, demonstrating the effectiveness of semantic embeddings in retrieving clinically relevant trials. Visualization of the results further emphasized these findings. A bar chart comparing Precision@5 showed the clear advantage of the semantic model, while a line chart of Precision@K across values from one to ten demonstrated that Sentence-BERT consistently outperforms TF-IDF, highlighting its robustness in ranking trials.

Explainability and Dashboard

To ensure transparency, each recommendation includes age and sex eligibility checks, matched condition keywords, and a snippet from the inclusion or exclusion text. Additionally, an interactive dashboard was developed using IPyWidgets, allowing users to select a patient, choose a model (TF-IDF or Sentence-BERT), adjust the number of top recommendations, and view both trial rankings and detailed explanations alongside dynamic Precision@K plots. This interactive system transforms the model from a static tool into a practical decision-support system that clinicians and research coordinators can trust and explore in real time.

Findings

The analysis revealed several important insights. Semantic embeddings substantially improve retrieval quality, capturing trials that keyword-based models miss due to variability in medical terminology. Eligibility filtering ensures that only trials suitable for a patient’s age and sex are recommended. Moreover, interactive explanations enhance user trust by making the reasoning behind recommendations transparent. The combination of semantic matching and eligibility filtering results in a highly accurate and interpretable recommendation system.

Recommendations

Based on these findings, three actionable recommendations can help maximize the impact of this system. First, integrating semantic trial matching into internal clinical dashboards can significantly accelerate patient recruitment. Second, providing explanation snippets for each recommendation allows clinicians to understand why a trial was suggested, increasing trust and usability. Third, regularly updating patient profiles and trial datasets will maintain the relevance and accuracy of recommendations over time.

Future Work

There are several avenues for future research. Incorporating additional patient attributes, such as comorbidities or lab results, could improve the richness of recommendations. Expanding the system to handle multi-lingual and multi-center trial datasets would enhance generalizability. Employing active learning or human-in-the-loop methods could refine proxy relevance labels and improve evaluation accuracy. Finally, exploring graph-based models to capture complex patient-trial relationships may lead to even more effective recommendations.

Conclusion

This project demonstrates a privacy-compliant, NLP-driven clinical trial recommendation system that semantically matches patients to trials, applies eligibility rules, and provides explainable, ranked recommendations. Compared to a keyword baseline, the semantic model shows measurable improvement, confirming that semantic understanding is crucial for navigating unstructured clinical text. By combining semantic retrieval, eligibility filtering, and transparent explanations, this system has the potential to streamline patient recruitment, support clinical decision-making, and accelerate medical research, all while maintaining ethical and practical compliance.