An Unsupervised Approach For Identifying Patient Trial Eligibility

Meet Gandhi, Department of Computer and Information Science[1]
[1]**University of Pennsylvania**

Abstract

This study presents an approach for finding eligible patients for clinical trials using patient records and trial inclusion and exclusion criteria in a completely unsupervised manner. Our algorithm uses a combination of fast string similarity functions and deep learning models to match patient records with inclusion and exclusion criteria to analyze 100k patient records and find the top 100 trial candidates from a database in just 15 minutes on a machine without any GPU. It also provides a similarity score for match confidence and highlights parts of the text in the patient record most relevant to the trial inclusion. Finally, we discuss limitations and future work, including improving sentence tokenization, deep learning models' performance, and incorporating user feedback to enhance model performances iteratively. The approach presented has the potential to significantly speed up the patient recruitment process for clinical trials, reducing the time and cost involved in conducting such studies.

Problem Statement/Motivation

Clinical trials play a crucial role in the development of new treatments for various medical conditions. However, identifying eligible patients for these trials can be a daunting task, and missing potential participants can cause delays in research progress and limit the effectiveness of new treatments. As such, there is a need for an efficient and effective approach to identifying eligible patients for clinical trials.

In this paper, we propose an unsupervised approach for identifying patient trial eligibility that can be used across different indications. By automatically analyzing large amounts of patient data, this approach has the potential to save time and resources while expanding the pool of potential participants for clinical trials. Ultimately, this can lead to more comprehensive research and faster development of new treatments, benefiting both patients and healthcare providers alike. The following sections describe the proposed approach in detail and provide results from the experiments.

Solution

The previous methods for automating the patient eligibility identification process for clinical trials are mainly rule-based, supervised, or information extraction. Rule-based approaches involve defining a set of rules that specify the eligibility criteria for a clinical trial, which are then applied to patient medical records to determine eligibility. This method is time-consuming and challenging to develop an exhaustive set of rules across indications. Supervised model development trains machine learning algorithms on labeled datasets to identify patients who meet the eligibility criteria for a clinical trial by analyzing their medical records. However, this method requires labeled data, making it indication and trial-dependent. Information extraction involves training named entity recognition models to extract medical entities from patient records and identify eligibility, which requires developing an exhaustive set of concepts that can change with indications and trials.

In contrast, the proposed solution in this paper is unsupervised and uses different string similarity measures to match patient records with trial inclusion/exclusion criteria and identify eligibility. It offers several benefits. Firstly, it is a general unsupervised framework that can work on different string similarity measures, so we can easily port it to any indication. Additionally, it highlights relevant parts of the text found most relevant to the trial inclusion, making it easier to validate the output quickly and train supervised models in the future. Furthermore, it saves time and resources as it does not require a labeled dataset or an exhaustive set of rules. However, it also has some limitations. It struggles to handle complex sentences with under-defined context and abbreviations, which is common in medical records. Also, the solution requires specialized deep-learning models trained on medical text to achieve good performance.

Data Sources

In this study, we utilize two main data sources to develop and evaluate our proposed approach for patient eligibility identification for clinical trials.

The first data source is the Medical Information Mart for Intensive Care III (MIMIC-III[3]) database, which is accessible through the PhysioNet website after obtaining appropriate permissions and completing the required data use agreement. MIMIC-III is one of the largest publicly available critical care datasets, containing data from over 60,000 ICU patients spanning over a decade. The database includes a wide range of data, such as clinical notes, laboratory results, vital signs, medications, procedures, diagnoses, and demographics. For this study, we utilize the gender, age, primary diagnosis, and clinical notes information from the MIMIC-III database.

The second data source is the ClinicalTrials.gov website data. This website serves as a comprehensive registry and database of clinical trials conducted worldwide, providing information about trial design, intervention details, study objectives, and inclusion and exclusion criteria. We utilized the publicly available API provided by the website to fetch trial data based on specific filters like indication, etc. For this study, we use the inclusion and exclusion criteria with structured information like min/max age and indication of actively recruiting trials. We then identify eligible patients in the MIMIC-III database based on these criteria and evaluate the algorithm.

By leveraging these two complementary data sources, we can develop a more comprehensive and accurate approach for patient eligibility identification for clinical trials, ultimately benefiting both patients and healthcare providers.

Data Model

The data model for this study involves joining multiple tables from the MIMIC-III database and the ClinicalTrials.gov API to create two final tables. The first intermediate table is created by joining the patients and admissions tables from the MIMIC-III database on the SUBJECT_ID field. The resulting intermediate table includes fields like SUBJECT_ID, HADM_ID, ADMITTIME, DIAGNOSIS, AGE (which is ADMITTIME - DOB), DOB, and GENDER. This intermediate table is then joined with the noteevents table from the MIMIC-III database on the SUBJECT_ID and HADM_ID fields, resulting in the final_table_mimic with fields such as SUBJECT_ID, HADM_ID, CHARTDATE, CATEGORY, TEXT, DIAGNOSIS, AGE, and GENDER. On the other hand, the final_table_ctgov is obtained by using the ClinicalTrials.gov API to fetch trial data based on specific filters like indication, etc. The final_table_ctgov includes fields like NCTId, MinimumAge, MaximumAge, Gender, Condition, InclusionCriteria, and ExclusionCriteria. By combining the data from these different sources, we can develop and evaluate our proposed approach for identifying patient eligibility for clinical trials.

Column Name	Description
SUBJECT_ID	Unique identifier for each patient
HADM_ID	Unique identifier for each hospital admission
CHARTDATE	Date and time of the clinical note
CATEGORY	Category of the clinical note
TEXT	Text of the clinical note
DIAGNOSIS	Primary diagnosis for the hospital admission
AGE	Age of the patient at the time of admission
GENDER	Gender of the patient

Table 1: Description of columns in MIMIC-III Final Table

Column Name	Description
NCTId	Unique identifier for each clinical trial
MinimumAge	Minimum age of patients for the trial
MaximumAge	Maximum age of patients for the trial
Gender	Gender eligibility for the trial
Condition	Medical condition being studied in the trial
InclusionCriteria	Criteria for inclusion in the trial
ExclusionCriteria	Criteria for exclusion from the trial

Table 2: Description of columns in ClinicalTrials.gov Final Table

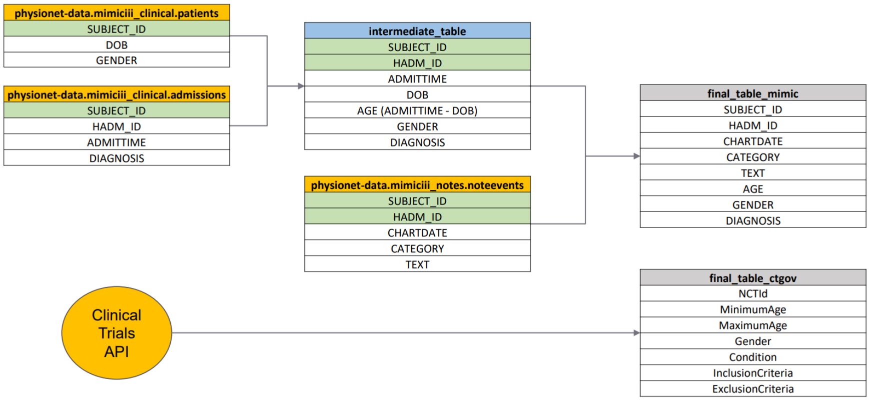


Figure 1: Data Model

Method

The proposed method for patient-trial eligibility determination involves a step-by-step approach. Initially, the age and gender of the patient are compared with the minimum and maximum age and gender criteria specified for the trial. If the patient's age and gender do not match the specified criteria, then the patient is deemed ineligible. If the patient meets the age and gender criteria, the patient record is sentence-tokenized, and the inclusion and exclusion criteria of the trial are also sentence-tokenized.

After tokenization, each sentence in the patient record is compared with all the sentences in the inclusion criteria, and the overall inclusion match is obtained using five string similarity functions. Similarly, an overall exclusion match is obtained by comparing each sentence in the patient record with all the sentences in the exclusion criteria. The exclusion match is compared with the inclusion match, and if the exclusion match is greater than the inclusion match, then the patient is deemed ineligible, else the patient is considered eligible for the trial.

To ensure that domain, contextual, and lexical similarities are captured between the text, five different string similarity functions are employed. These include Clinical Bert[2] Similarity, TFIDF Cosine Similarity, Fuzzy String Match, UMLS[4] Similarity, and Sentence Transformer[1] Embedding Cosine Similarity. The Clinical Bert model is fine-tuned for string similarity on the MEDSTS dataset. The TFIDF Cosine Similarity is calculated by measuring the cosine similarity between the TFIDF encodings of the patient record and the trial inclusion and exclusion criteria. Fuzzy String Match uses token set ratio to calculate similarity between strings. UMLS Similarity is calculated based on the intersection over union of extracted UMLS concepts from the patient record and inclusion and exclusion criteria. Finally, the Sentence Transformer embedding cosine similarity is used to measure general English and medical text similarity.

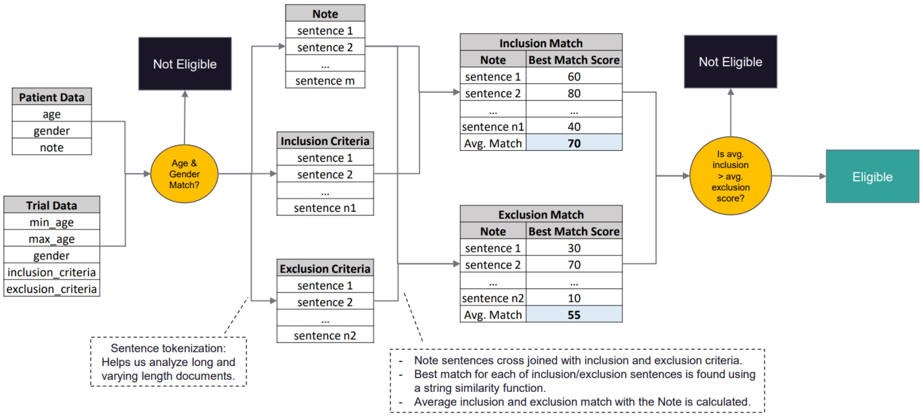
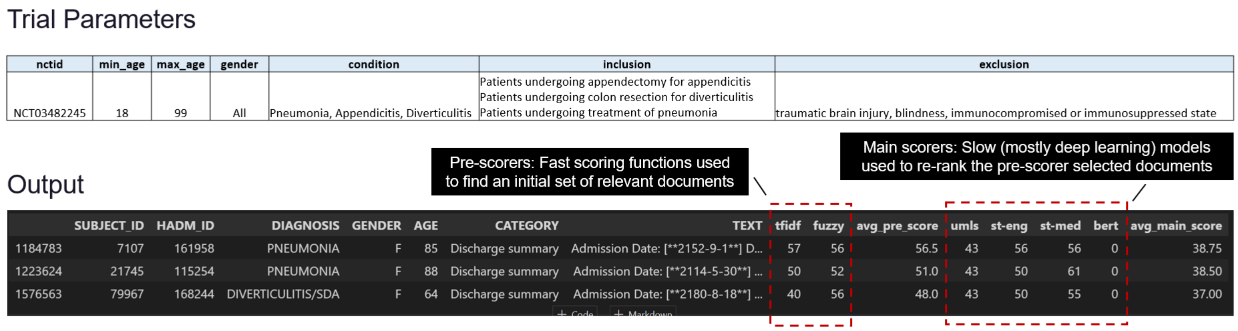


Figure 2: Atomic Example For One Similarity Function

Further, to improve the computational efficiency of the algorithm on large databases, we first run pre-scorers between patient records and the clinical trial. These pre-scorers are fast string similarity functions used to get an average pre-score for each patient in the database. Based on the pre-score, we select a list of top n patients, or all patients with a non-zero pre-score, to run the main scoring functions, which are majorly slow deep learning models. Finally, patients having their main score above a threshold are selected and sorted in descending as output. This approach helps in efficiently analyzing a large number of patient records in a time-efficient manner.



Trial Parameters

nctid	min_age	max_age	gender	condition	inclusion	exclusion
NCT03482245	18	99	All	Pneumonia, Appendicitis, Diverticulitis	Patients undergoing appendectomy for appendicitis Patients undergoing colon resection for diverticulitis Patients undergoing treatment of pneumonia	traumatic brain injury, blindness, immunocompromised or immunosuppressed state

Pre-scorers: Fast scoring functions used to find an initial set of relevant documents

Main scorers: Slow (mostly deep learning) models used to re-rank the pre-scorer selected documents

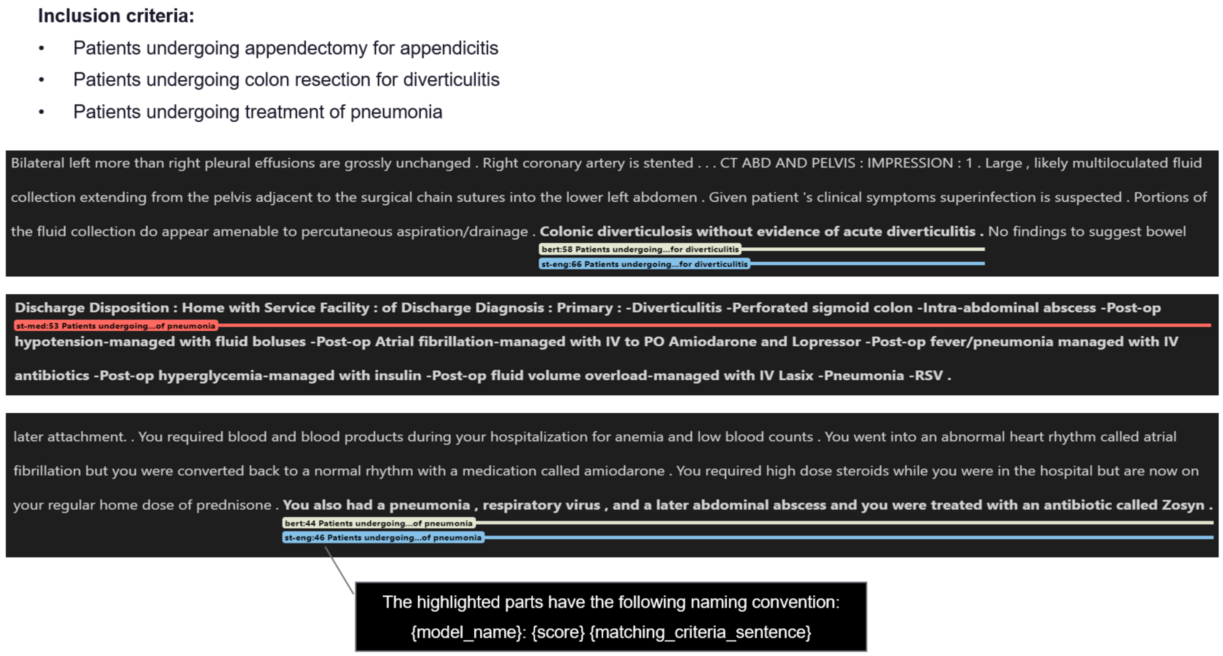
Output

	SUBJECT_ID	HADM_ID	DIAGNOSIS	GENDER	AGE	CATEGORY	TEXT	tfidf	fuzzy	avg_pre_score	umls	st-eng	st-med	bert	avg_main_score
1184783	7107	161958	PNEUMONIA	F	85	Discharge summary	Admission Date: [**2152-9-1**] D...	57	56	56.5	43	56	56	0	38.75
1223624	21745	115254	PNEUMONIA	F	88	Discharge summary	Admission Date: [**2114-5-30**] ...	50	52	51.0	43	50	61	0	38.50
1576563	79967	168244	DIVERTICULITIS/SDA	F	64	Discharge summary	Admission Date: [**2180-8-18**] ...	40	56	48.0	43	50	55	0	37.00

Figure 3: Output from algorithm, when run on a database

Results/Discussion

Although evaluating ranking algorithms without labels is a challenging task, we can measure its performance based on the efficiency benefits it provides. Our proposed algorithm can analyze a large database of 100k patient records and identify the top 100 trial candidates in just 15 minutes. Furthermore, the algorithm provides a match confidence score based on similarity measures, which helps in assessing the reliability of the output. Additionally, the algorithm highlights the most relevant parts of the text that it found to be significant for the trial inclusion, which could be used to quickly validate the output. These benefits can significantly improve the efficiency of the clinical trial matching process and reduce the burden on clinicians and researchers.



Inclusion criteria:

- Patients undergoing appendectomy for appendicitis
- Patients undergoing colon resection for diverticulitis
- Patients undergoing treatment of pneumonia

Bilateral left more than right pleural effusions are grossly unchanged . Right coronary artery is stented . . . CT ABD AND PELVIS : IMPRESSION : 1 . Large , likely multiloculated fluid collection extending from the pelvis adjacent to the surgical chain sutures into the lower left abdomen . Given patient 's clinical symptoms superinfection is suspected . Portions of the fluid collection do appear amenable to percutaneous aspiration/drainage . **Colonic diverticulosis without evidence of acute diverticulitis .** No findings to suggest bowel

bert:58 Patients undergoing...for diverticulitis
st-eng:66 Patients undergoing...for diverticulitis

Discharge Disposition : Home with Service Facility : of Discharge Diagnosis : Primary : -Diverticulitis -Perforated sigmoid colon -Intra-abdominal abscess -Post-op hypotension-managed with fluid boluses -Post-op Atrial fibrillation-managed with IV to PO Amiodarone and Lopressor -Post-op fever/pneumonia managed with IV antibiotics -Post-op hyperglycemia-managed with insulin -Post-op fluid volume overload-managed with IV Lasix -Pneumonia -RSV .

later attachment. . You required blood and blood products during your hospitalization for anemia and low blood counts . You went into an abnormal heart rhythm called atrial fibrillation but you were converted back to a normal rhythm with a medication called amiodarone . You required high dose steroids while you were in the hospital but are now on your regular home dose of prednisone . **You also had a pneumonia , respiratory virus , and a later abdominal abscess and you were treated with an antibiotic called Zosyn .**

bert:44 Patients undergoing...of pneumonia
st-eng:46 Patients undergoing...of pneumonia

The highlighted parts have the following naming convention:
{model_name}: {score} {matching_criteria_sentence}

Figure 4: Visual Explanation: For a 52% similar patient

Limitations & Future Work

We can address several limitations to this approach in future work. Firstly, we can improve sentence tokenization to handle complex sentence structures common in inclusion/exclusion criteria. The current method struggles to capture the meaning of conjunction compounded sentences. Further, contextual sentence similarity models work poorly with inadequate context, which is often the case with medical notes. Therefore, we must improve the performance of deep learning models used on clinical notes. Lastly, we can also focus on incorporating user feedback to iteratively improve model performances, which will help us make the algorithm more effective in finding eligible patients for clinical trials.

Conclusion

In conclusion, our proposed algorithm provides a novel approach to identifying eligible patients for clinical trials in a completely unsupervised manner. By combining pre-scorers and deep learning models, we are able to analyze large patient datasets and provide efficient trial candidate recommendations. While the current algorithm has limitations in accurately capturing complex sentence structures and inadequate context, it should not miss out on potentially eligible patients. We have identified areas for improvement in future work. Additionally, incorporating user feedback will allow us to iteratively improve our model performance and increase the effectiveness of our algorithm. Overall, this approach shows promise in addressing the challenge of identifying eligible patients for clinical trials and accelerating the drug development process.

References

1. Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084. 2019 Aug 27.
2. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, McDermott M. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323. 2019 Apr 6.
3. Johnson, A., Pollard, T., Shen, L. et al. MIMIC-III, a freely accessible critical care database. Sci Data 3, 160035 (2016). https://doi.org/10.1038/sdata.2016.35
4. Olivier Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Research, Volume 32, Issue suppl_1, 1 January 2004, Pages D267–D270, https://doi.org/10.1093/nar/gkh061
5. Kang T, Zhang S, Tang Y, Hruby GW, Rusanov A, Elhadad N, Weng C. EliIE: An open-source information extraction system for clinical trial eligibility criteria. Journal of the American Medical Informatics Association. 2017 Nov 1;24(6):1062-71.