SVKM'S NMIMS Deemed to be UNIVERSITY

SCHOOL OF MATHEMATICS, APPLIED STATISTICS & ANALYTICS

IMPROVISING CLINICAL TRIAL RECRUITMENT PROCESS

A DISSERTATION SUBMITTED TO
SVKM'S NMIMS (DEEMED TO BE UNIVERSITY)
IN PARTIAL FULFILMENT FOR THE DEGREE OF
MASTERS OF SCIENCE
IN
STATISTICS AND DATA SCIENCE

BY

RISHABH PANDEY

KEYUR PARKHI

DARSHANA PARLE

GAUTAMI PATANKAR

NIRAJ PATIL

UNDER THE SUPERVISION OF

Mr. Parag Jadhav

NILKAMAL SCHOOL OF MATHEMATICS, APPLIED STATISTICS AND ANALYTICS
SVKM's Narsee Monjee Institute of Management Studies
(Deemed-To-Be-University)
V.L. Mehta Rd, Vile Parle (West), Mumbai – 400056

March 2025

Group Information

Group 9

Name	SAP ID	Program	Year
Rishabh Pandey	86062400077	M.Sc. Statistics and Data Science	2024-25
Keyur Parkhi	86062400048	M.Sc. Statistics and Data Science	2024-25
Darshana Parle	86062400032	M.Sc. Statistics and Data Science	2024-25
Gautami Patankar	86062400037	M.Sc. Statistics and Data Science	2024-25
Niraj Patil	86062400004	M.Sc. Statistics and Data Science	2024-25

Acknowledgement

We would like to extend our heartfelt gratitude to our mentor Mr. Parag Jadhav for his unwavering determination, guidance and support throughout this project. His expertise, encouragement, and insightful feedback have been instrumental in shaping our work and fostering our growth as learners. His motivation encouraged us to keep learning something new at every corner of this project.

We are deeply appreciative of our Dean, Dr. Sushil Kulkarni, for giving us an opportunity to work on a project in the field of our interest and for permitting us to access and utilise all the facilities available in the college. This support has played a crucial role in the successful execution of our project.

Furthermore, we would like to extend our sincere gratitude to our professor Dr. Pradnya Khandekarpar whose constant support, supervision and guidance have been invaluable. Their dedication to our academic endeavours has not only enriched our learning experience but has also been and will also be a source of inspiration for several more projects yet to come.

Last but definitely not the least, we acknowledge and appreciate the collaborative efforts of all those who have played a role in the success of this project. Your support has been integral, and we are thankful for the opportunities and knowledge you have provided. Moreover, we would thank the readers too for giving us an opportunity to intrigue you on the topic of improvising the clinical trial recruitment process.

Table of Contents

1.Abstract

Our study emphasizes the critical need to revolutionize the clinical trial recruitment process by harnessing advanced machine learning and natural language processing (NLP) techniques. We undertook comprehensive data cleaning and rigorous exploratory data analysis to uncover hidden patterns affecting participant enrolment. By deploying robust ensemble models, we successfully predicted recruitment outcomes and identified high-impact features driving participant engagement. Furthermore, NLP methodologies were applied to process and analyse unstructured data, enabling deeper understanding of participant sentiments and common barriers to enrolment. Our integrated approach combined predictive modelling with text analytics to offer actionable insights that enhance recruitment efficiency, improve participant targeting, and reduce trial timelines. The findings demonstrate that leveraging data science innovations can significantly address recruitment challenges, laying the groundwork for scalable, intelligent recruitment frameworks that can adapt dynamically to evolving clinical trial requirements.

Keywords: Natural Language Processing (NLP), Ensemble Learning, Unstructured text data, Chia-with-scope (Clinical Health Information Annotation), Patient EHR Data Generation, Synthea, Bio-BERT NER Model, Sorensen-Dice Index (SDI)

2.Introduction

Clinical trials represent a cornerstone of modern medicine, enabling researchers and healthcare professionals to assess the safety, efficacy, and applicability of new treatments, diagnostics, and preventive strategies across a wide range of diseases and conditions. From the development of advanced imaging technologies to life-saving therapeutics, clinical trials have played a pivotal role in shaping medical progress. Despite their critical importance, one of the persistent challenges facing clinical trials is the difficulty of recruiting and enrolling eligible patients. Recruitment delays not only increase the operational costs of clinical trials but can also result in underpowered studies, regulatory setbacks, and in some cases, the complete termination of promising research endeavors.

A major bottleneck in patient recruitment lies in the ability to effectively identify and match eligible participants. This process is often hindered by a lack of awareness among both patients and healthcare providers about ongoing clinical trials. Additionally, eligibility criteria are typically written in complex medical language and often require time-consuming manual review of patient health records. As a result, trial recruitment has traditionally depended on clinicians manually screening patients against eligibility criteria—an error-prone, inconsistent, and labor-intensive process. Streamlining this process could help increase access to potentially life-saving treatments, improve diversity and representation in clinical trials, and accelerate the pace of discovery in biomedical research.

With the growing digitization of healthcare, electronic health records (EHRs) have emerged as a rich source of patient information that can be leveraged for clinical trial matching. EHRs store a wide array of patient data, including demographics, diagnoses, lab results, medications, procedures, and allergies, all of which are critical in determining trial eligibility. The use of standardized medical vocabularies and ontologies—such as ICD-10 for diagnoses, CPT for procedures, LOINC for lab tests, and ATC for drugs—has further improved the consistency and interoperability of health data across institutions. These advancements offer a valuable opportunity to develop automated systems that can parse and interpret structured and unstructured EHR data to identify eligible patients with high precision.

However, despite the potential, the integration of EHR data into clinical trial workflows remains limited. Existing clinical trial registries, like ClinicalTrials.gov, offer search functionalities for users to find trials by condition, location, and status, but they fall short in automating the matching process. Patients and providers must still manually interpret complex eligibility criteria, often written in dense clinical language, and enter relevant data manually. These inefficiencies can lead to mismatches, missed opportunities, and increased burden on all stakeholders. Therefore, there is a pressing need for intelligent systems that can automatically parse eligibility criteria and match them to patient profiles with minimal human intervention.

Natural Language Processing (NLP) offers a promising solution to this challenge. Eligibility criteria and EHR data both contain substantial amounts of free-text information—physician notes, lab interpretations, condition descriptions, and more—that are rich in detail but difficult to structure. NLP techniques, particularly those enhanced by deep learning and biomedical language models, can be harnessed to extract relevant clinical concepts, normalize them using standard ontologies, and perform semantic comparisons to match patients with appropriate clinical trials. An NLP-based trial matcher, capable of leveraging structured and unstructured EHR data, has the potential to revolutionize the clinical trial recruitment landscape by making it faster, more accurate, and more equitable.

3. Review of Literature

1. **Patient Enrollment Patterns in Clinical Trials**
   The study by Haidich and Ioannidis (2001) provides a comprehensive analysis of patient enrollment patterns in randomized controlled trials (RCTs), revealing significant demographic and geographic biases. Their research demonstrates that clinical trials systematically underrepresented women, elderly populations, and ethnic minorities while disproportionately recruiting participants from North America and Western Europe. This selective enrollment creates trial populations that often differ substantially from real-world patient demographics, potentially compromising the external validity of research findings. The authors identify strict eligibility criteria as a major contributing factor, with approximately 60% of potentially eligible patients being excluded from participation. These exclusion patterns may lead to inflated efficacy estimates and incomplete safety profiles, as trial participants tend to be healthier and have fewer comorbidities than the general patient population. The findings underscore the importance of developing more inclusive trial designs that better reflect the diversity of patients who will ultimately use the medications in clinical practice.

2. **Economic Challenges in Drug Development**
   The economic landscape of pharmaceutical development has been extensively documented by DiMasi et al. (2003), Morgan et al. (2011), and Dickson and Gagnon (2004). These studies collectively demonstrate a dramatic escalation in drug development costs over time, with DiMasi's seminal work estimating an average cost of $802 million (in 2000 dollars) to bring a new drug to market. Morgan's systematic review corroborates these findings while identifying clinical trial operations as the largest cost component, accounting for nearly 60% of total expenditures. Dickson's analysis further elucidates the structural factors driving these rising costs, including increasingly complex scientific targets, expanding regulatory requirements, and the growing size and duration of clinical trials. The authors note that the average New Drug Application submission grew from 21,000 pages in 1985 to over 100,000 pages by 2003, reflecting both increased regulatory scrutiny and more comprehensive data requirements. These economic pressures have significant implications for the pharmaceutical industry, potentially discouraging innovation in areas with smaller market potential or greater scientific uncertainty.

3. **Fred Brauer-** In the Research Paper titled, “Mathematical epidemiology: Past, present, and future”. The author gives us a thorough outline of the history, present state and probable future of Epidemiology. The author includes several Mathematical and Statistical tools, models and techniques used to analyse, predict and assess the spread of infectious

outbreaks. The main objective of the paper is to shed light on the development and evolution of mathematical computing in Epidemiology. The author focuses on various models used in classic epidemiology and reduction number which is used to decide if a disease will sustain or die out in a given period of time. The use of stochastic models and heterogeneity of disease spread are also researched in the paper with several mentions of emerging issues in the field like drug resistant microorganisms and slower than exponential growth in epidemics which not only makes it difficult to treat the disease and stop the spread of epidemics but also causes some serious issues in predicting the next stage of the epidemic outbreak. Thus, the predictions of the time stamp where the epidemic reaches a crest or a trough become off. Early works like John Graunt's study of death records and Daniel Bernoulli's model for smallpox were instrumental in defining modern epidemiology. As time passed, new models like the SIR (Susceptible, Infectious, Recovered) model were created to track how diseases spread in different groups of people and the Kermack-McKendrick model which studies infection patterns were born. The paper discusses how stochastic models account for randomness in the spread of the disease, especially during the outbreaks and how network-based models can replicate transmission in social settings. Next-generation matrices for meta-population scenarios to handle real world problems are studied too. The paper concludes that interventions and vaccinations are effective in stopping the spread of the epidemic and deterministic models may fail for a smaller population where stochastic and network-based models are more effective. Newer drugs are used to treat drug resistant microorganisms and new models are in development to model slower than exponential outbreaks, like Ebola. Advancements in computation power and better data quality are important for developing Epidemiology.

4. Integration of Findings and Contemporary Relevance

The reviewed literature reveals fundamental tensions between scientific rigor, demographic representation, and economic feasibility in clinical research. The enrollment biases identified by Haidich and Ioannidis persist despite growing recognition of their importance, suggesting that current incentives and practices in clinical research may inadvertently perpetuate health disparities. Meanwhile, the escalating costs documented by DiMasi, Morgan, and Dickson create substantial barriers to innovation, particularly for treatments targeting rare diseases or underserved populations. These challenges are compounded by the methodological complexities outlined in clinical trial design literature, as researchers attempt to balance scientific validity with practical and financial constraints. Recent advances in adaptive trial designs, real-world evidence generation, and decentralized clinical trials may offer potential solutions to some of these longstanding issues. However, the literature suggests that meaningful progress will require coordinated efforts across multiple stakeholders, including regulatory agencies, pharmaceutical companies, academic researchers, and patient advocacy groups, to develop more efficient and inclusive approaches to clinical research.

5. **Future Directions in Clinical Research**

The collective findings from these studies point to several critical areas for future research and policy development. First, there is a need for more systematic investigation of strategies to improve patient representation in clinical trials without substantially increasing costs or compromising scientific validity. Second, the development of alternative funding models and regulatory pathways may help mitigate the economic barriers to innovation, particularly for treatments addressing unmet medical needs. Third, the integration of novel technologies and data sources, such as electronic health records and digital health tools, could potentially enhance both the efficiency and generalizability of clinical research. Finally, the literature underscores the importance of ongoing monitoring and evaluation of trial enrollment patterns and development costs to identify emerging trends and assess the impact of policy interventions. These efforts will be essential for creating a more sustainable and equitable clinical research ecosystem that can deliver meaningful medical advances to diverse patient populations.

6. **Recent Advances in Hypertension Clinical Trials**

Götzinger et al. (2022) provide a focused review of hypertension-related clinical trials, summarizing key developments in blood pressure management. The authors highlight renal denervation therapy, a minimally invasive procedure that targets overactive renal nerves to reduce blood pressure in treatment-resistant patients. Recent trials (e.g., SPYRAL HTN-OFF MED and RADIANCE-HTN SOLO) suggest that this approach may offer sustained benefits, though long-term data are still needed. The paper also examines novel antihypertensive drugs, such as dual endothelin receptor antagonists and aldosterone synthase inhibitors, which show promise for specific patient subgroups.

Beyond therapeutics, the authors discuss innovative trial designs, including pragmatic trials that evaluate interventions in real-world settings rather than controlled environments. They emphasize the importance of patient-centered outcomes (e.g., quality of life, cardiovascular event reduction) over surrogate markers (e.g., blood pressure readings alone). The review concludes by identifying gaps in current research, such as underrepresentation of elderly and comorbid populations, which may limit generalizability**.**

7. **Ethical Challenges in Clinical Trial Representation**

Alhalel et al. (2022) address a critical but often overlooked issue in clinical research: the underrepresentation of individuals with Limited English Proficiency (LEP). The authors argue that language barriers contribute to health disparities, as LEP populations are frequently excluded from trials due to inadequate recruitment strategies, lack of

multilingual consent forms, and insufficient cultural competency among researchers. This exclusion not only raises ethical concerns about justice and equity but also undermines the external validity of trial results, as findings may not apply to diverse populations.

To promote inclusivity, the authors propose practical solutions, such as:

Community-engaged recruitment (e.g., partnering with trusted local organizations),

Professional interpreter services (rather than ad-hoc translations),

Regulatory incentives (e.g., FDA requirements for diversity in trial enrollment).
They also call for structural changes, including training researchers in cultural humility and revising IRB policies to mandate LEP-inclusive practices. The paper aligns with broader movements toward diversity, equity, and inclusion (DEI) in clinical research, emphasizing that equitable representation is both a moral imperative and a scientific necessity.

8. **Radiomics Harmonization for Multicenter Trials-**Soliman et al. (2022) tackle a pressing challenge in quantitative imaging research: the lack of standardization in radiomics data across institutions. Radiomics—the extraction of high-dimensional features from medical images—has potential applications in cancer diagnosis, prognosis, and treatment response prediction. However, the authors note that technical variability (e.g., differences in scanners, acquisition protocols, and segmentation methods) can lead to irreproducible results, limiting clinical utility.

   To address this, the paper proposes a harmonization framework with three key components:

   Pre-acquisition standardization: Consensus on imaging parameters (e.g., resolution, contrast timing).

   Phantom-based calibration: Using synthetic test objects to correct inter-scanner differences.

   Post-processing computational tools: Algorithms like ComBat to remove batch effects in feature extraction.
   The authors highlight successful applications in multicenter cancer trials, where harmonized radiomics improved predictive accuracy for outcomes like tumor progression. They also discuss regulatory challenges, noting that FDA approval of radiomics

biomarkers will require rigorous validation across diverse datasets. The framework bridges the gap between technical innovation and clinical translation, ensuring that radiomics can reliably inform personalized medicine.

4. Data Generation and Description

4.1 Chia_with_scope -

The CHIA (Clinical Health Information Annotation) dataset is a specialized corpus curated for advancing natural language processing in the biomedical domain, particularly for clinical trial eligibility extraction. It contains a collection of clinical trial protocols from ClinicalTrials.gov, where the **inclusion** and **exclusion criteria** for each trial are available as separate plain text files. These criteria outline the eligibility requirements for patient enrollment and typically include conditions, treatments, demographic requirements, measurements, and more.

The CHIA with Scope dataset is a structured and annotated corpus curated to support natural language processing tasks in the biomedical domain, particularly those involving clinical trial eligibility criteria. It is derived from real-world clinical trial protocols available on ClinicalTrials.gov, with each trial identified by a unique NCT number. For each trial, the dataset contains two primary text files: one detailing the inclusion criteria (NCTxxxxxxx_inc.txt) and the other detailing the exclusion criteria (NCTxxxxxxx_exc.txt). Alongside each of these text files, there is a corresponding annotation file (.ann) in the BRAT standoff format, which provides manually labeled spans of named entities relevant to clinical decision-making.

The annotations identify several key biomedical entity types, including CONDITION (e.g., "Type 2 Diabetes"), TREATMENT (e.g., "Metformin"), MEASUREMENT (e.g., "HbA1c > 7%"), DEMOGRAPHIC (e.g., "Male, aged 40–60"), INCLUSION, and EXCLUSION. Each annotated entity is associated with its character-level position in the text, enabling precise training of named entity recognition (NER) models. What distinguishes the "with Scope" version of the CHIA dataset is its emphasis on contextual labeling of entities based on whether they occur within inclusion or exclusion criteria. This allows for a more nuanced understanding of clinical trial eligibility logic, which is crucial for tasks like patient-trial matching.

Overall, the chia_with_scope dataset serves as a valuable resource for developing transformer-based models like BioBERT for clinical NLP. It supports fine-grained extraction of structured information from unstructured clinical trial texts and lays the groundwork for intelligent systems capable of identifying suitable clinical trials for individual patients based on their medical profiles.

```
Patients with biopsy-proven metastatic carcinoid tumors or other neuroendocrine tumors (Islet cell, Gastrinomas and VIPomas) with at least
one measurable lesion (other than bone) that has either not been previously irradiated or if previously irradiated has demonstrated
progression since the radiation therapy
The patient has no major impairment of renal or hepatic function, as defined by the following laboratory parameters: total bilirubin <1.5 X
ULN; AST, ALT<2.5X ULN (<5 X ULN if liver metastases are present)
Patients on Sandostatin Lar (long acting somatostatin analogue) must be on a stable dose for 30 days prior to study entry and short acting
somatostatin analogues must be judged to be on a clinically stable dose by the investigator prior to study entry
Must have a life expectancy of greater than three (3) months
Karnofsky Performance Status > 60
Female patients must have a negative serum pregnancy test at screening. (Not applicable to patients with bilateral oophorectomy and/or
hysterectomy or to those patients who are postmenopausal.)
```

Fig.4.1.1. Snapshot of the inclusion criterion data

```
T1      Condition 28 55 metastatic carcinoid tumors
T2      Value 21 27     proven
T3      Procedure 14 20 biopsy
R1      Has_value Arg1:T3 Arg2:T2
R2      AND Arg1:T1 Arg2:T3
T4      Condition 59 86 other neuroendocrine tumors
T5      Condition 88 98 Islet cell
T6      Condition 100 111       Gastrinomas
T7      Condition 116 123       VIPomas
T8      Scope 88 123    Islet cell, Gastrinomas and VIPomas
*       OR T5 T6 T7
R3      Subsumes Arg1:T4 Arg2:T8
*       OR T1 T4
T9      Condition 143 160       measurable lesion
T10     Qualifier 173 177       bone
T11     Scope 14 124    biopsy-proven metastatic carcinoid tumors or other neuroendocrine tumors (Islet cell, Gastrinomas and VIPomas)
T12     Procedure 215 225       irradiated
T13     Procedure 243 253       irradiated
T14     Observation 271 282     progression
T15     Procedure 293 310       radiation therapy
T16     Temporal 283 310        since the radiation therapy
R6      Has_temporal Arg1:T14 Arg2:T16
```

Fig.4.1.2. Snapshot of the inclusion data annotations

```
Patients with symptomatic CNS metastases or leptomeningeal involvement
Patients with known brain metastases, unless these metastases have been treated and/or have been stable for at least six months prior to
study start. Subjects with a history of brain metastases must have a head CT with contrast to document either response or progression.
Patients with bone metastases as the only site(s) of measurable disease
Patients with hepatic artery chemoembolization within the last 6 months (one month if there are other sites of measurable disease)
Patients who have been previously treated with radioactive directed therapies
Patients who have been previously treated with epothilone
Patients with any peripheral neuropathy or unresolved diarrhea greater than Grade 1
Patients with severe cardiac insufficiency patients taking Coumadin or other warfarin-containing agents with the exception of low dose
warfarin (1 mg or less) for the maintenance of in-dwelling lines or ports
Patients taking any experimental therapies history of another malignancy within 5 years prior to study entry except curatively treated non
melanoma skin cancer, prostate cancer, or cervical cancer in situ
Patients with active or suspected acute or chronic uncontrolled infection including abcesses or fistulae
Patients with a medical or psychiatric illness that would preclude study or informed consent and/or history of noncompliance to medical
regimens or inability or unwillingness to return for all scheduled visits
HIV+ patients
Pregnant or lactating females.
```

Fig.4.1.3. Snapshot of the exclusion criterion data

```
T1      Condition 26 40 CNS metastases
T2      Condition 44 70 leptomeningeal involvement
*       OR T1 T2
T5      Procedure 144 151       treated
T6      Qualifier 164 179       been stable for
T7      Temporal 180 220        at least six months prior to study start
R1      Has_temporal Arg1:T6 Arg2:T7
*       OR T5 T6
T4      Condition 92 108        brain metastases
T8      Observation 238 248     history of
T9      Condition 249 265       brain metastases
R3      AND Arg1:T8 Arg2:T9
T10     Procedure 278 299       head CT with contrast
T11     Scope 238 265   history of brain metastases
A1      Optional T11
R4      AND Arg1:T11 Arg2:T10
T12     Condition 359 374       bone metastases
T13     Procedure 432 464       hepatic artery chemoembolization
T14     Temporal 465 489        within the last 6 months
T15     Temporal 491 500        one month
```

Fig.4.1.4. Snapshot of the exclusion data annotations

4.2. Generation of Patients EHR health records from Synthea

Synthea (Synthetic Patient Population Simulator) is an open-source tool developed by The MITRE Corporation to generate synthetic yet realistically modeled electronic health records (EHRs). It simulates the medical histories of entire patient populations, using clinically accurate disease progression models, treatment pathways, and healthcare workflows. Each synthetic patient is created with a unique identity comprising demographic information (such as age, gender, race, and location) along with social determinants of health, genetic predispositions, and lifestyle factors. As the simulation progresses, the patients encounter clinical events such as disease onset, physician visits, hospital admissions, prescriptions, and lab tests based on guideline-driven modules built using real epidemiological and medical data. These events are recorded over time to form a longitudinal health record for each individual, reflecting a complete, life-like healthcare journey.

One of the major strengths of Synthea is its ability to generate privacy-preserving, high-fidelity data that mirrors the complexity and variability of real-world clinical data without any risk of identifying real patients. The generated data is output in multiple formats including CSV, JSON, HL7 FHIR, and CCDA, allowing for easy ingestion into data science pipelines or electronic health systems. In this project, Synthea-generated records were used to simulate a patient cohort for downstream clinical trial eligibility matching. These records provided detailed fields such as conditions (ICD codes), medications (RxNorm codes), procedures, and lab results, which were used to extract structured representations of patient health. This synthetic dataset served as a valuable and ethical substitute for real patient data, facilitating the development and evaluation of NLP and matching algorithms without compromising patient privacy.

5. Data preparation and Preprocessing

Preprocessing

The preprocessing of patient data plays a pivotal role in any natural language processing (NLP) pipeline, especially in the healthcare domain, where clinical data is rich in unstructured text. In our project, we used the CHIA with Scope dataset to build a model that extracts relevant medical entities from clinical trial eligibility criteria. To match this with patient data effectively, it was essential to design a preprocessing pipeline that standardizes, cleans, and converts free-text data into a structured format suitable for NLP model input. This preprocessing was applied to both the CHIA dataset (for model training) and the synthetic EHR patient data (for trial matching).

Initially, each patient's profile was represented in a row-wise manner across multiple files, with important health attributes like conditions, medications, procedures, allergies, and demographics stored in separate tables. These tables were merged using a unique patient identifier to create a consolidated "wide format" dataset. This transformation ensured that each patient's profile was represented as a single row, where attributes like medications or conditions were stored as lists or concatenated strings. This format was crucial for converting each patient's profile into a human-readable clinical summary that could be input to an NLP model.

Once the full textual profile of a patient was assembled, it was passed through a preprocessing pipeline developed using spaCy, an advanced NLP library. The pipeline included several components tailored for clinical and biomedical text. The first stage was tokenization, where the raw text was split into individual tokens or words. Tokenization in the clinical domain is non-trivial due to the presence of medical abbreviations, dosage forms, and measurement units. Therefore, we used scispaCy's models, which are trained on biomedical literature and offer more accurate token splitting for this domain.

Following tokenization, we applied Part-of-Speech (POS) tagging and dependency parsing, which allowed us to understand the grammatical structure and relationships between tokens. POS tagging was particularly helpful for distinguishing between nouns (e.g., "diabetes") and verbs (e.g., "treated"), which plays an important role in entity recognition. The dependency parser helped in understanding complex sentence structures often found in medical documentation. This was important when parsing multi-clause conditions like "Patients with chronic kidney disease who are not on dialysis."

In the next step, we used lemmatization to convert each word to its base or dictionary form. This ensured consistency in entity recognition, as variations of the same term (e.g., "treating," "treated," "treatment") were reduced to a single root ("treat"). In medical text, lemmatization helps unify various surface forms of the same concept, increasing the robustness of entity matching. We also incorporated abbreviation detection using `AbbreviationDetector` from the scispaCy library. Since medical text is rife with abbreviations like "HTN" (Hypertension) or

"T2DM" (Type 2 Diabetes Mellitus), expanding these abbreviations was essential for accurate entity extraction and matching with clinical trial criteria.

Lastly, the cleaned and processed text was passed into a fine-tuned BioBERT-based Named Entity Recognition (NER) model. This model, trained on the CHIA dataset, was capable of identifying and classifying entities into categories such as CONDITION, TREATMENT, MEASUREMENT, and DEMOGRAPHIC. These structured outputs were then used to compare against clinical trial eligibility criteria, thereby enabling automated patient-trial matching. This comprehensive preprocessing pipeline ensured high-quality inputs to the model and significantly improved its performance in identifying relevant medical entities.

Preparation

To evaluate our clinical trial matching system in a controlled and reproducible environment, we constructed a series of synthetic patient cohorts using trial criteria extracted from publicly available NCT protocols. These synthetic datasets were carefully curated to reflect trial-level characteristics such as medical condition categories, complexity of criteria, and diversity of patient attributes. The simulation followed principles similar to synthetic data generators like Synthea, which emulate realistic electronic health records (EHRs) based on publicly available health models. The goal was to mirror actual trial eligibility requirements while ensuring no real patient information was used.

Each trial was associated with a set of inclusion and exclusion criteria, and patients were programmatically generated based on these logical conditions. Depending on the disease category such as infectious diseases (ID), substance use disorders (SUD), or cardiovascular disease (CVD), Drug, Anxiety, etc.. and most of them were unknown synthetic patients were assigned medical histories, medications, and demographic profiles that either matched or conflicted with the eligibility definitions. The difficulty level (e.g., EASY) was determined based on the number of criteria and diversity of attribute types used in the trial logic, such as discrete (e.g., gender), continuous (e.g., age), or categorical variables (e.g., diagnosis types).

NCTID	Category	Subcategory	Difficulty	Sample Size	Attributes
NCT00050349	ID	infection	EASY	140	14
NCT00050349	Unknown	Unknown	EASY	60	6
NCT00061308	ID	infection	EASY	110	11
NCT00061308	Unknown	Unknown	EASY	40	4
NCT00094861	ID	virus	EASY	180	18
NCT00094861	Unknown	Unknown	EASY	160	16
NCT00122070	SUD	drug	EASY	90	9
NCT00122070	Unknown	Unknown	EASY	40	4
NCT00182520	SUD	substance	EASY	70	7
NCT00182520	Unknown	Unknown	EASY	30	3
NCT00183885	ID	infection	EASY	50	5
NCT00183885	Unknown	Unknown	EASY	90	9
NCT00198913	Unknown	Unknown	EASY	20	2
NCT00198913	Unknown	Unknown	EASY	10	1
NCT00235170	CVD	heart	EASY	110	11
NCT00235170	CVD	cardiovascular	EASY	80	8
NCT00236340	Unknown	Unknown	EASY	60	6
NCT00236340	Unknown	Unknown	EASY	20	2
NCT00250640	Unknown	Unknown	EASY	20	2
NCT00250640	CVD	heart	EASY	30	3
NCT00279552	Unknown	Unknown	EASY	20	2
NCT00279552	Unknown	Unknown	EASY	10	1
NCT00305097	CVD	heart	EASY	40	4
NCT00305097	Unknown	Unknown	EASY	40	4
NCT00312429	Unknown	Unknown	EASY	40	4

Fig 5.1.1. Basic properties of initial set of 25 clinical trials used to build synthetic patient cohorts.

6. What we are trying to Achieve

The general workflow for the TrialMatcher algorithm is diagrammed in Following Figure

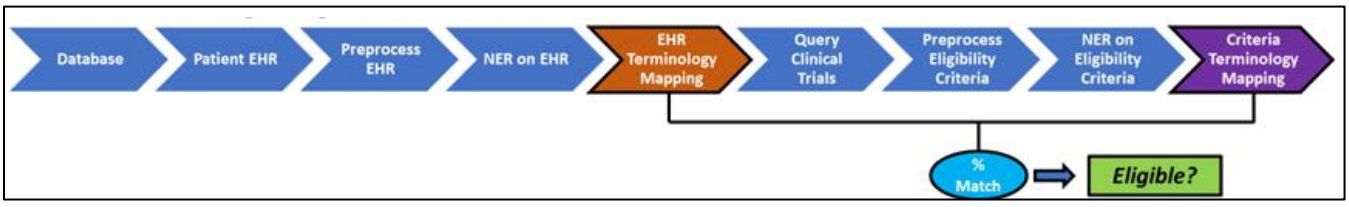

Fig 6.1.1. Process diagram for TrialMatcher

7. Challenges and Cons of Dealing with Unstructured Data for Clinical Trial Matching

1. Ambiguity in Language and Terminology

Clinical trial eligibility criteria and patient health records are often written in **free-text form**, using varying terminologies, abbreviations, and synonyms. For instance, “heart attack,” “myocardial infarction,” and “MI” could all mean the same thing, but may not be uniformly recognized by the system without extensive synonym mapping. Similarly, negations or double negations (e.g., “no signs of no infection”) can lead to misinterpretation of eligibility requirements or patient conditions, affecting the accuracy of trial matches

2. Lack of Standardization

Unstructured text lacks a standard schema, which makes it difficult to apply consistent logic or rules. Clinical trial criteria written by different sponsors or organizations vary widely in structure, phrasing, and level of specificity. This inconsistency complicates parsing and semantic interpretation. Unlike structured data (like ICD-10 codes or lab values), unstructured text doesn’t follow predictable patterns, requiring more sophisticated natural language processing (NLP) techniques that are computationally expensive and less accurate.

3. Entity Recognition Errors

Even advanced NER models like BioBERT or transformer-based spaCy pipelines can struggle with contextual understanding of medical entities. These models may extract the correct entity but miss modifiers, qualifiers, or numerical thresholds that are essential for accurate matching (e.g., “Hemoglobin < 10” vs. “Hemoglobin > 10”). They may also confuse between patient and caregiver eligibility or fail to distinguish exclusion criteria from inclusion statements. These misclassifications reduce the trustworthiness of the system.

4. Data Noise and Incompleteness

Patient data in EHRs often contains typos, incomplete records, outdated information, or irrelevant notes that add noise to the processing pipeline. Free-text notes may mention past conditions, suspected diagnoses, or family history that could be incorrectly treated as active conditions. This introduces false positives or false negatives in the match scoring, lowering the reliability of automated matching compared to manual review by experts.

5. Difficulty in Extracting Numeric or Logical Constraints

Clinical trial criteria often include complex numerical ranges, logical combinations, or temporal constraints, like “systolic blood pressure between 120-140 mmHg for at least 3 months.” Extracting these conditions from raw text using current NLP tools is extremely challenging. Regular expressions can only partially help and often require manual rule engineering. The inability to accurately parse these constraints can result in incorrect eligibility determinations.

6. Computational and Annotation Overhead

Processing large volumes of unstructured text requires high computational power and extensive data annotation for model training and evaluation. Annotating clinical text is expensive, time-consuming, and requires domain experts due to the sensitive and technical nature of the content. Moreover, fine-tuning language models like BioBERT or RoBERTa on custom annotations consumes resources and may still fall short of human-level understanding.

Hence, we used the approach of converting our unstructured data to structured data for further analysis. While doing this we faced certain challenges.

Real-World Difficulties We Faced

1. Identifying Medical Entities Accurately:

○ Extracting allergies, conditions, medications, and observations from raw text involved manual inspection and scripting.

○ No consistent format.

Example:

■ “Patient has a history of hypertension and asthma.”

■ “BP controlled with Amlodipine. No diabetes.”

2. Negation Handling:

○ Terms like "no diabetes", "denies fever" require negation detection, which was not initially implemented. This led to false positives in structured output.

3. Inconsistent Terminologies:

○ Same condition mentioned in various forms:

■ “Heart attack”, “Myocardial infarction”, “MI” — all refer to the same diagnosis.

○ Required mapping to a common ontology like SNOMED-CT or UMLS (which was not integrated).

4. Ambiguous Phrasing:

○ "Family history of cancer" ≠ "patient has cancer".

○ Rule-based models failed to distinguish between such cases.

5. Nested Conditions and Exceptions:

○ Example criterion: "Diabetic patients not on insulin."

■ Needs conditional logic: Diabetes == True and Insulin == False

6. Incomplete and Missing Data:

○ Key lab values like BMI, BP, Creatinine often missing or mentioned in varying unit

○ Example: "Weight = 80kg, Height = 170cm" vs "BMI is 28.7"

7. Manual Structuring Burden:

○ We had to manually or semi-automatically separate extracted data into 4 sheets:

■ Allergies

■ Medications

■ Conditions

■ Observations

○ Then merge them using patient IDs.

○ Data loss and misalignment occurred due to inconsistent entries or missing values.

Finally, we have structured chia_with_scope data as follows:

NCT ID	riteria Typ	Category	Subcategory
NCT00050	Exclusion	Condition	CNS metastases
NCT00050	Exclusion	Condition	leptomeningeal involvement
NCT00050	Exclusion	Procedure	treated
NCT00050	Exclusion	Qualifier	been stable for
NCT00050	Exclusion	Temporal	at least six months prior to study start
NCT00050	Exclusion	Condition	brain metastases
NCT00050	Exclusion	Observation	history of
NCT00050	Exclusion	Condition	brain metastases
NCT00050	Exclusion	Procedure	head CT with contrast
NCT00050	Exclusion	Scope	history of brain metastases
NCT00050	Exclusion	Condition	bone metastases
NCT00050	Exclusion	Procedure	hepatic artery chemoembolization
NCT00050	Exclusion	Temporal	within the last 6 months
NCT00050	Exclusion	Temporal	one month
NCT00050	Exclusion	Observation	other sites of measurable disease
NCT00050	Exclusion	Scope	within the last 6 months (one month if there are other sites of measurable disease
NCT00050	Exclusion	Procedure	radioactive directed therapies
NCT00050	Exclusion	Drug	epothilone
NCT00050	Exclusion	Condition	peripheral neuropathy
NCT00050	Exclusion	Condition	unresolved diarrhea
NCT00050	Exclusion	Measurement	Grade
NCT00050	Exclusion	Value	greater than 1

Finally, we have structured Patient EHR data as follows:

Patient ID	Total Cholesterol	Low Density Lipopr	High Density Lipop	Diastolic Blood	Systolic Blood P	Tobacco smokin	Hemoglobin	Body Mass In	Body Weig	Body He	medications.DESCRIPTION	conditions.Column2
3fe783d9-b8f8-4d21-b320	177.2	0	0	0	0	0	0	0	0	0	Mirena 52 MG Intrauterine System	Prediabetes
1f8da3b0-dd67-4925-ae6d	177.2	0	0	0	0	0	0	0	0	0	Mirena 52 MG Intrauterine System	Anemia (disorder)
a001f3e9-5955-4b52-bb83	177.2	0	0	0	0	0	0	0	0	0	Mirena 52 MG Intrauterine System	Miscarriage in first trimester
f64742bf-8cac-4780-ae38	177.2	0	0	0	0	0	0	0	0	0	Mirena 52 MG Intrauterine System	Normal pregnancy
e3bba570-83a5-4105-b76	177.2	0	0	0	0	0	0	0	0	0	Mirena 52 MG Intrauterine System	Sprain of wrist
47392cc2-4fb3-4431-b56f	183.6	0	0	0	0	0	0	0	0	0	Errin 28 Day Pack	Chronic sinusitis (disorder)
04630e85-e9f5-4a9b-be75	183.6	0	0	0	0	0	0	0	0	0	Errin 28 Day Pack	Acute bronchitis (disorder)
f4946dee-9d49-456b-bdf5	183.6	0	0	0	0	0	0	0	0	0	Errin 28 Day Pack	Normal pregnancy
3e234e30-d471-4c3e-b43	183.6	0	0	0	0	0	0	0	0	0	Errin 28 Day Pack	Viral sinusitis (disorder)
851275c6-bc90-4e95-a9a1	183.6	0	0	0	0	0	0	0	0	0	Errin 28 Day Pack	Preeclampsia
e901ce7f-cce0-46dc-b04f-	183.6	0	0	0	0	0	0	0	0	0	Errin 28 Day Pack	Contact dermatitis
6c7ce9c1-78a9-4b72-beff-	183.6	0	0	0	0	0	0	0	0	0	Errin 28 Day Pack	Miscarriage in first trimester
5ce924c7-ec4c-4b5f-93ff-2	183.6	0	0	0	0	0	0	0	0	0	Errin 28 Day Pack	Fetus with unknown complication
1c2d500e-c583-40f8-83a8	199.8	0	0	0	0	0	0	0	0	0	Nitroglycerin 0.4 MG/ACTUAT Muco	Coronary Heart Disease
7fef7a7d-0198-457f-8aad-	199.8	0	0	0	0	0	0	0	0	0	Nitroglycerin 0.4 MG/ACTUAT Muco	Appendicitis
9394cb52-7d92-4577-976	199.8	0	0	0	0	0	0	0	0	0	Nitroglycerin 0.4 MG/ACTUAT Muco	History of appendectomy
9c69f2d5-b7de-4634-84c5	199.8	0	0	0	0	0	0	0	0	0	Nitroglycerin 0.4 MG/ACTUAT Muco	Body mass index 30+ - obesity (finding)
b36658dd-cb89-41cf-b2e9	199.8	0	0	0	0	0	0	0	0	0	Nitroglycerin 0.4 MG/ACTUAT Muco	Stroke
3ef4c7dc-28c2-4436-b0cc	199.8	0	0	0	0	0	0	0	0	0	Nitroglycerin 0.4 MG/ACTUAT Muco	Acute viral pharyngitis (disorder)
09cbb3f1-c2ed-426c-8c53	199.8	0	0	0	0	0	0	0	0	0	Nitroglycerin 0.4 MG/ACTUAT Muco	Viral sinusitis (disorder)
61417eb5-1c42-4207-afb3	166.6	0	0	0	0	0	0	0	0	0	Terfenadine 60 MG Oral Tablet	Body mass index 30+ - obesity (finding)
e0503bf7-6b9e-4c70-9dc8	166.6	0	0	0	0	0	0	0	0	0	Terfenadine 60 MG Oral Tablet	Hypertension
e598a677-2882-4849-ae1	166.6	0	0	0	0	0	0	0	0	0	Terfenadine 60 MG Oral Tablet	Acute viral pharyngitis (disorder)
9dfbf6ed-2378-4c27-9ff7-7	166.6	0	0	0	0	0	0	0	0	0	Terfenadine 60 MG Oral Tablet	Acute bacterial sinusitis (disorder)
5ad3dd08-6e9e-4feb-98ff-	166.6	0	0	0	0	0	0	0	0	0	Terfenadine 60 MG Oral Tablet	Chronic sinusitis (disorder)
c2a6847e-c5d2-4158-bee5	186.2	0	0	0	0	0	0	0	0	0	Etonogestrel 68 MG Drug Implant	Body mass index 30+ - obesity (finding)
202b3372-a1d4-42bd-b4b	186.2	0	0	0	0	0	0	0	0	0	Etonogestrel 68 MG Drug Implant	Sprain of ankle
8bf92fdc-0cff-4945-b5a9-7	186.2	0	0	0	0	0	0	0	0	0	Etonogestrel 68 MG Drug Implant	Normal pregnancy
354aefa8-0bd5-4943-8f93	186.2	0	0	0	0	0	0	0	0	0	Etonogestrel 68 MG Drug Implant	Viral sinusitis (disorder)
f77ee33d-ba6e-40f6-a0b5	186.2	0	0	0	0	0	0	0	0	0	Etonogestrel 68 MG Drug Implant	Antepartum eclampsia

Let's understand the reasons behind the process being so challenging

- Natural language in medicine is nuanced and domain-specific.
- Requires not just Named Entity Recognition (NER), but also semantic understanding, context awareness, and medical knowledge.
- Most off-the-shelf NLP tools are not tuned for medical text unless specifically customized.

What if Structured Data Is Available ?

- With cleaned and well-structured datasets, the next step is matching patients to trials using ensemble learning models.
- Ensemble models combine the strengths of multiple base models (e.g., decision trees, logistic regression, SVMs) to produce a more robust, stable, and generalizable output.

- They can better capture complex relationships and decision boundaries that exist in medical data, especially when eligibility criteria involve multiple interrelated clinical parameters.
- Feature importance and confidence estimation in ensemble methods (like Random Forests or Gradient Boosting) help prioritize patients who are more likely to be eligible, even when data is noisy or partially missing.
- Ensemble learning can reduce overfitting compared to individual models, which is especially useful in a medical context where the number of features may be large, but labeled examples of "eligible" vs "ineligible" patients are limited.

We faced major issues applying ensemble models:
- Lack of labeled data (eligible vs ineligible patients).
- Structured data did not fully align with inclusion/exclusion logic.
- Many extracted features were noisy or misclassified.

While machine learning techniques like ensemble learning are powerful, they rely fundamentally on structured, consistent, and high-quality data. The core of the problem lies not in model performance but in the complexity and ambiguity of converting unstructured clinical text into machine-readable format. Each word, context, and phrasing in clinical notes can drastically change the meaning, and current methods (without domain-tuned NLP) struggle to handle this at scale. The transition from unstructured clinical text to structured datasets is not a simple data-cleaning task — it's a deep information extraction challenge requiring both computational and medical insight. When structured data is reliably available, ensemble models offer a compelling approach to accurately and efficiently identify eligible patients, thereby improving the speed and effectiveness of clinical trial recruitment.

8. Why use NLP over other methods such as ensemble learning ?

1. Entity Extraction from Free Text

One of the core applications of NLP is Named Entity Recognition (NER), which helps identify key medical concepts such as *conditions*, *medications*, *demographics*, *measurements*, and *procedures* from unstructured patient records and clinical trial texts. NLP models like BioBERT, spaCy with transformer pipelines, or SciSpacy are trained to recognize biomedical entities and annotate them accurately, enabling the transformation of raw free text into structured representations.

2. Handling Synonyms, Abbreviations, and Context

NLP helps resolve semantic variation by mapping different phrases or abbreviations (e.g., "HTN," "hypertension," or "high blood pressure") to a unified concept. This is crucial in healthcare, where the same condition may be described in multiple ways. Through contextual embeddings (e.g., from transformers), NLP models capture the meaning of terms based on surrounding text, allowing for more accurate interpretation than traditional keyword-based methods.

3. Sentence Segmentation and Dependency Parsing

Clinical trial eligibility criteria and provider notes often include complex, multi-clause sentences. NLP pipelines segment this unstructured text into meaningful units using sentence segmentation, and further analyze grammatical relationships through dependency parsing. This enables better understanding of which entities are associated with inclusion vs. exclusion criteria, what values are being compared, and whether a condition is negated or affirmed.

4. Negation Detection and Logical Structures

In eligibility criteria, it is vital to know whether a condition is *present* or *absent*. NLP tools such as negspacy apply negation detection using predefined linguistic rules or models (e.g., NegEx) to identify when a medical condition or factor is negated in text. Similarly, logical constructs like "AND," "OR," or "NOT" can be interpreted to help determine whether multiple conditions are required or optional, improving the logic of match algorithms.

5. Information Structuring and Standardization

Once entities are extracted and disambiguated, NLP helps organize them into structured formats like dictionaries, JSON objects, or annotated token sequences. These structured representations can be directly compared using similarity metrics (e.g., Sørensen–Dice Index) or used as input for rule-based filtering and machine learning models. NLP also facilitates mapping to standard medical vocabularies like ICD-10, SNOMED, or LOINC for interoperability.

6. Enhancing Automation and Scalability

With NLP, tasks like parsing patient records, extracting trial criteria, or identifying matching entities can be automated at scale. This drastically **reduces manual review**, speeds up patient-trial matching, and allows real-time decision support. NLP pipelines can be retrained or fine-tuned on domain-specific datasets (like CHIA) to adapt to new clinical trials or population groups, making the system robust and continuously improving.

9. Models

9.1 BioBERT: Biomedical Bidirectional Encoder Representations from Transformers

1. What is BioBERT?

BioBERT is a domain-adapted version of BERT (Bidirectional Encoder Representations from Transformers) specifically pre-trained on large biomedical corpora. Developed by researchers at the DMIS Lab (Korea University), BioBERT extends BERT by continuing the pre-training phase using millions of abstracts from PubMed and full-text articles from PMC (PubMed Central). The idea is to retain BERT's contextual language understanding while adapting it to biomedical terminology and syntax. This makes BioBERT especially effective for tasks like NER, question answering, and relation extraction in the medical domain.

2. Architecture

BioBERT shares the same architecture as the original BERT-base model: it uses a 12-layer transformer encoder, each with 768 hidden units, 12 attention heads, and 110 million parameters. The model is trained using masked language modeling (MLM) and next sentence prediction (NSP) objectives. During fine-tuning, a task-specific classification head (such as a linear layer or CRF for NER) is added on top of the final hidden states. The pre-training corpus makes a significant difference: while BERT was trained on Wikipedia and BooksCorpus, BioBERT is fine-tuned on specialized biomedical text, improving performance in this niche.

3. How It Works

In a typical NER task using BioBERT, the model receives a tokenized input sentence with special tokens like `[CLS]` and `[SEP]`. It generates contextual embeddings for each token. These embeddings are passed through a classification head, which predicts an entity label for each token. Because of its bidirectional nature, BioBERT can take into account the full left and right context of each word, allowing it to disambiguate similar terms and understand complex clinical sentences. For example, it can differentiate between *"The patient has no signs of MI"* and *"MI was confirmed through ECG"*, assigning correct labels in both.

4. Use Cases in Clinical NER

BioBERT has outperformed several existing models in biomedical NER benchmarks such as NCBI Disease, BC5CDR, and CHEMDNER. In your use case, fine-tuning BioBERT on the CHIA dataset helps extract entities relevant to clinical trial eligibility criteria. Each token is tagged with entities like `INCLUSION`, `EXCLUSION`, `CONDITION`, or `MEASUREMENT`, and then

used for patient-trial matching via string similarity or the Sørensen–Dice index. Compared to general-purpose models, BioBERT exhibits a deeper understanding of medical context, abbreviations, and rare terms.

5. Where to Access and How to Use

BioBERT is available through Hugging Face's Model Hub (`dmis-lab/biobert-v1.1`) and can be directly loaded using the `transformers` library in Python. It supports both PyTorch and TensorFlow backends. Fine-tuning can be done with standard token classification scripts by preparing your training data in the BIO format. A simplified training loop includes loading the model, preparing tokenized datasets, attaching the classification layer, defining loss (usually cross-entropy), and training with AdamW optimizer.

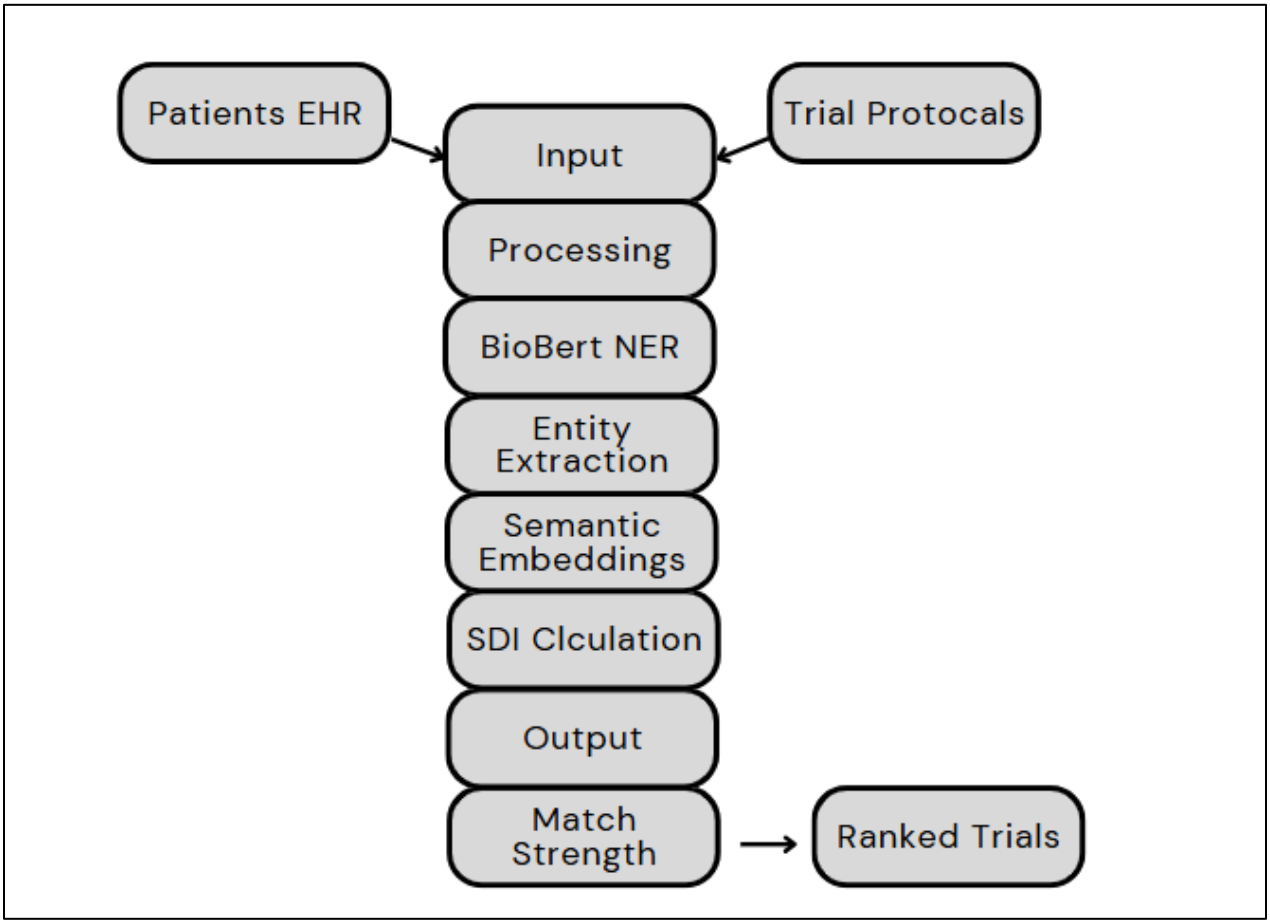

Fig 9.1.1. Flowchart of Model

10. Evaluation Using Similarity Metrics

1. Sorensen Dice Index

In our clinical trial matching system, evaluating the alignment between patient profiles and clinical trial eligibility criteria is a core component. To measure this alignment, we utilized the Sørensen–Dice Index (SDI)—a set similarity metric that quantifies the degree of overlap between two sets. In our context, one set represents the medical attributes of a patient, while the other represents the concepts extracted from a clinical trial's inclusion and exclusion criteria. SDI offers an interpretable similarity score ranging from 0 (no overlap) to 1 (perfect overlap), which can be used to rank patients based on how well they match a given trial.

Each clinical trial is defined by a standardized set of inclusion and exclusion criteria, which we parsed using regular expressions. The inclusion and exclusion texts were processed through the same spaCy + BioBERT NER pipeline used for patient data, extracting entities such as conditions, treatments, measurements, and demographics. For similarity comparison, these extracted entities are converted into structured profiles. The trial profile was encoded as a dictionary, where keys are entity names and values are booleans: `True` for inclusion and `False` for exclusion.

Patient profiles, originally sourced from structured or semi-structured electronic health record (EHR) formats, were similarly converted to dictionaries. Each key in the dictionary corresponds to a health attribute (like "hypertension" or "insulin"), and the value is `True` if that attribute is present and `False` otherwise. This structure allowed for fast and interpretable comparison of concept-level alignment between patients and trials. Importantly, we assumed that missing entries in patient data implied the absence of the attribute, enabling a uniform binary representation across all profiles.

One challenge with this binary matching is negation handling, where statements like "Unable to provide informed consent" in exclusion criteria are conceptually the inverse of "Able to provide informed consent" in inclusion criteria. While our initial system treats inclusion and exclusion concepts in isolation, future improvements can include negation detection using libraries like `negspacy`, which implements modified NegEx patterns to capture these subtleties. Similarly, handling numerical criteria (e.g., "Hemoglobin > 12") remains difficult with NER alone. Ongoing work involves integrating regex-based extractors like extractacy to link quantities with entities.

To compute the similarity between a patient and a clinical trial, we used the basic Sørensen–Dice Index formula:

$$\text{SDI}(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|}$$

Here, A is the set of attributes from the patient profile and B is the set of entities extracted from the clinical trial criteria. The index returns a score between 0 and 1, indicating the proportion of shared features. To extend this further, we calculated an aggregated SDI over dictionary-based profiles:

$$\text{SDI}_{agg} = \frac{1}{N} \sum_{i=1}^{N} \frac{2 \times |A_i \cap B_i|}{|A_i| + |B_i|}$$

In this formula, N is the total number of shared keys in both dictionaries, and Ai and Birepresent the sets (True/False values) for a particular entity from the patient and clinical trial profile, respectively. Since values are binary, each mini-SDI is either 0 or 1, and averaging them gives the final match score. The result reflects how well the patient matches the specific needs of a trial without being penalized for unrelated attributes.

This similarity scoring mechanism allowed us to simulate and evaluate the quality of trial matches across synthetic cohorts. Because the synthetic data was controllable, we could create test cases with predefined similarity levels and confirm that the SDI responded accordingly. This iterative feedback loop helped refine the preprocessing, entity recognition, and matching logic. While not perfect, SDI-based matching provides a strong baseline for scalable, interpretable, and clinically relevant patient-trial matching.

11. Results and discussion

We worked on an example “Patient has Type 2 Diabetes and is taking Metformin” , here each word/token gets one of the tags. Using the biobert model we provide bio tags to each of these tokens.

BIO stands for:

B:Beginning of an entity

I:Inside of an entity

O:Outside of an entity

The output generated for the above example is as follows.

Token	BIO Tag
Patient	O
has	O
Type	B-Disease
2	I-Disease
Diabetes	I-Disease
and	O
is	O
taking	O
Metformin	B-Drug

The table below displays the training progress of a model over three epochs, showing improvements in performance metrics across each epoch. The **training and validation loss** steadily decrease, indicating better model fit and generalization—training loss drops from 1.6523 to 1.1247, and validation loss from 1.4827 to 1.2133. Simultaneously, **precision**, **recall**, and **F1 score** improve with each epoch. Precision increases from 0.6471 to 0.7032, recall from 0.6011 to 0.6731, and the F1 score—a balance between precision and recall—rises from 0.6234 to **0.6882** by the third epoch. These trends suggest that the model is learning effectively and becoming more accurate in identifying relevant entities or matches.

Epoch	Training Loss	Validation Loss	Precision	Recall	F1
1	1.652300	1.482739	0.647136	0.601108	0.623411
2	1.334700	1.333021	0.683438	0.653379	0.668007
3	1.124700	1.213295	0.703208	0.673137	**0.688225**

The table shows the results of matching a patient's profile against various clinical trials using the Sørensen–Dice Index (SDI) for both inclusion and exclusion criteria. Each row corresponds to a trial ID, with SDI scores indicating the similarity between the patient's data and the trial's inclusion and exclusion criteria. Higher inclusion SDI and lower exclusion SDI suggest a better match. Based on these scores, an overall match is determined: "Strong Match" for high inclusion and zero exclusion similarity, "Partial Match" for moderate inclusion and some exclusion similarity, and "Weak Match" for low or no inclusion similarity and/or non-zero exclusion similarity.

```
Clinical Trial Matching Results:

   Trial ID  Inclusion SDI  Exclusion SDI Overall Match
NCT01991743           0.57            0.5 Partial Match
NCT03472508           0.00            0.6    Weak Match
NCT02573597           0.73            0.0  Strong Match
NCT03029078           0.00            0.0    Weak Match
NCT02552459           0.57            0.0  Strong Match
NCT02541955           0.67            0.0  Strong Match
NCT03062358           0.62            0.0  Strong Match
```

12. Business Impact of Clinical Trial Matcher in Healthcare

1. AI-powered solutions like TrialMatcher are set to revolutionize clinical trial recruitment, tackling the critical issue of low enrollment rates that delay medical innovation and increase costs. Manual screening of patients is time-consuming and resource-intensive, often causing trials to miss enrollment targets and extending development timelines by months. TrialMatcher, using advanced NLP, automates the process by matching patient data from electronic health records (EHRs) to trial eligibility criteria with speed and accuracy.

2. For healthcare organizations and research sponsors, this translates to significant cost savings and faster time-to-market for new therapies. Reduced manual effort allows clinical staff to prioritize patient care while improving recruitment efficiency, which can save millions in R&D costs and operational overhead. Moreover, better matching improves patient diversity in trials, aligning with regulatory expectations and enhancing the quality of research outcomes.

3. By integrating AI-driven matching tools, hospitals and health systems position themselves as leaders in digital health innovation, opening opportunities for strategic partnerships and additional revenue streams through sponsored trials. Most importantly, it accelerates patient access to life-saving treatments, creating a meaningful impact on patient care while delivering measurable business value.

13. Conclusion

We conclude that our research presents a transformative framework for improving clinical trial recruitment through the strategic application of machine learning and natural language processing. Our methodology involved meticulous data preparation, insightful exploratory analysis, and the deployment of powerful ensemble models to accurately predict recruitment success and optimize participant targeting. By integrating NLP techniques, we unlocked valuable insights from unstructured data sources, shedding light on participant perceptions and key barriers to trial enrollment. These advanced analytics empowered us to develop tailored recruitment strategies that enhance efficiency, diversity, and overall trial representativeness. Our study underscores the potential of data-driven decision-making in modernizing clinical trial operations, enabling stakeholders to implement adaptive, evidence-based recruitment processes. Ultimately, this work offers a scalable solution that not only accelerates recruitment timelines but also strengthens the integrity and inclusiveness of clinical research, paving the way for future advancements in trial management and patient engagement.

14. Limitations

1. AI-driven trial matching tools face several challenges.

2. Data privacy, interoperability, and inconsistent data quality across EHR systems affect accuracy.

3. Incomplete patient records or missing clinical details can lower matching precision.

4. The reliance on confidence thresholds for eligibility scoring means there's a risk of false positives or negatives if thresholds aren't properly calibrated.

5. Clinician trust is another barrier, requiring transparent, explainable AI outputs. Initial implementation demands high investment in infrastructure and training.

6. Furthermore, potential algorithmic bias could marginalize certain patient groups, and evolving regulatory standards require continuous adaptation. Overcoming these challenges is critical for long-term success.

15. Future Work

1. The future scope of AI-driven clinical trial matching is highly promising, with opportunities to further integrate real-time patient monitoring, genomic data, and wearable health devices for even more precise matching.

2. As healthcare data ecosystems expand, future models can leverage longitudinal patient records to predict trial eligibility proactively, even before a clinician considers recruitment.

3. Additionally, integrating multilingual capabilities will enable global scalability, improving access in diverse healthcare systems and underrepresented regions.

4. Advanced explainable AI techniques will also strengthen trust among clinicians and patients, encouraging widespread adoption.

5. Beyond recruitment, these systems can evolve to support patient retention strategies by predicting drop-out risks and offering personalized engagement.

6. Furthermore, collaboration with regulatory bodies can streamline approval processes, making trial matching tools an essential component of clinical research infrastructure. Ultimately, this evolution will contribute to faster drug development cycles, reduced healthcare costs, and improved patient outcomes worldwide.

16. References

1. Harrer S, Shah P, Antony B, Hu J. (2019). Artificial Intelligence for Clinical Trial Design https://doi.org/10.1016/j.tips.2019.06.003

2. Pulley JM et al. (2018). Connecting the public with clinical trial options: The ResearchMatch Trials Today tool. https://doi.org/10.1017/cts.2018.316

3. Victor M. Murcia, Vinod Aggarwal, Nikhil Pesaladinne, Ram Thammineni, Nhan Do, Gil Alterovitz, Rafael B. Fricks.Improving Clinical Trial Cohort Recruitment Using Natural Language Processing https://github.com/victormurcia/Clinical-Trial-Matcher

4. Sun Y et al. (2021) *The COVID-19 Trial Finder* https://doi.org/10.1093/jamia/ocaa306 .

5. Raza S et al. (2022) Large-scale application of named entity recognition to biomedicine and epidemiology https://doi.org/10.1371/journal.pdig.0000152

6. Improving clinical trial design using interpretable machine learning,. https://www.researchgate.net/publication/363198900_Improving_Clinical_Trial_Design_Using_Interpretable_Machine_Learning_Based_Approach

7. improving the Efficiency of Clinical Trial Recruitment Using an Ensemble Machine Learning to Assist With Eligibility Screening" by Tianrun Cai et al. was published in *ACR Open Rheumatology* on July 23, 2021 https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr2.11289

8. Penberthy LT, Dahman BA, Petkov VI, DeShazo JP. "Effort Required in Eligibility Screening for Clinical Trials." J Oncol Pract. 2012;8(6):365–70 https://ascopubs.org/doi/abs/10.1200/JOP.2012.000646

9. Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T, et al. "Automated Clinical Trial Eligibility Prescreening: Increasing the Efficiency of Patient Identification for Clinical Trials in the Emergency Department." J Am Med Inform Assoc. 2015;22(1):166–78.https://pubmed.ncbi.nlm.nih.gov/25030032/

10. Haidich AB, Ioannidis JP. "Patterns of Patient Enrollment in Randomized Controlled Trials." J Clin Epidemiol. 2001;54(9):877–83.https://pubmed.ncbi.nlm.nih.gov/11520646/