BIOLOGITS

USE-CASE 4

TEAM MEMBERS:

Name	Contact	
Paidi Geetesh Chandra	6309264223	paidigeetesh@gmail.com
Udit Mohan Pradhan	6370820572	uditraj0707@gmail.com
Lalit Karthikeyan M A	9965170922	karthikeyanlalit05@gmail.com
Pratyush Dash	6372682656	pratyush1dash@gmail.com
Raghuveer V	7845274630	vraghuveer382@gmail.com

ABSTRACT

Clinical trial recruitment is a critical yet challenging aspect of medical research, often hindered by the inability to account for external factors that affect recruitment rates fully. Traditional methods typically rely on internal study parameters, overlooking essential elements such as disease prevalence and competition. This study presents a predictive model integrating internal study data and external factors from open-sourced data and APIs. Extensive feature engineering is used to create meaningful features, and advanced feature selection techniques have been employed to refine the feature set. The solution accurately predicts recruitment rates using robust regression models, offering valuable insights that optimise trial planning, resource allocation, and decision-making.

PROBLEM STATEMENT

Effective recruitment is one of the most significant challenges in conducting clinical trials. The success of a trial often depends on the ability to enrol the correct number of participants within a specified timeframe. Understanding recruitment rates and identifying appropriate benchmarks are critical for optimising this process. To achieve this, a predictive model is required to calculate the estimated Recruitment Rate of a clinical trial with a degree of confidence. This model must account for internal and external factors that influence recruitment. The goal is to provide accurate predictions that will assist in better planning and managing clinical trials. To build and validate this model, a smaller subset of 450,000 clinical trial data will be provided from clinicaltrials.gov, a publicly available resource. The technical solution will be a data-driven, scalable model that can be integrated into internal planning systems, enabling clinical trial teams to make more informed, strategic decisions.

EXPLORATORY DATA ANALYSIS

The exploratory data analysis phase was conducted to gain insights into the dataset, identify patterns, and understand the distribution of key features. This involved summarising the data statistically and visualising it through a series of plots to uncover trends, outliers, and relationships between variables.

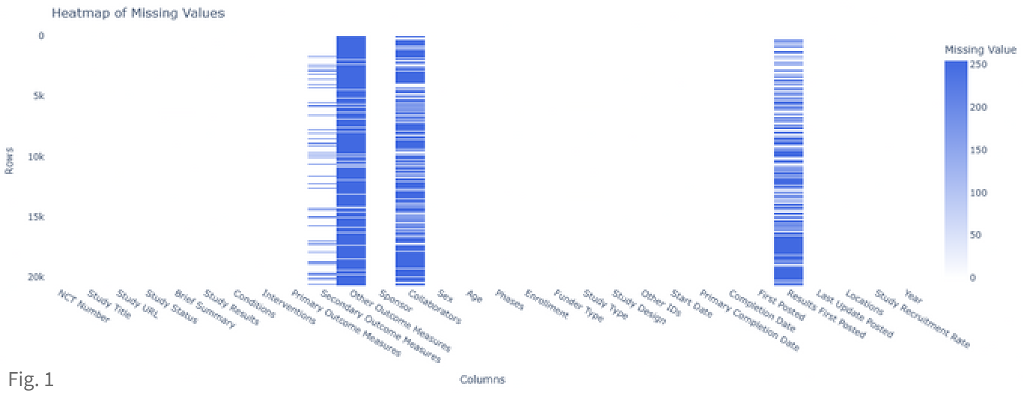


Fig. 1

Through EDA, NaN (missing) values can be identified (Fig. 1), helping to assess redundant features for potential elimination. Additionally, EDA helps in detecting outliers/anomalies (Fig. 2) and identifying dominating values across features (Fig. 3), providing critical insights into data quality and distribution.

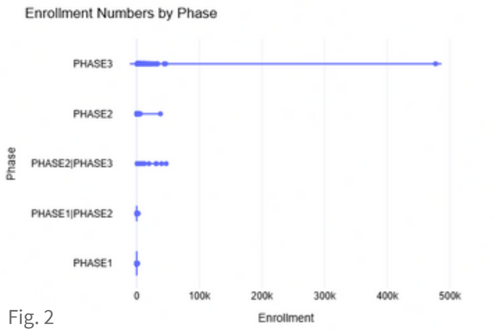


Fig. 2

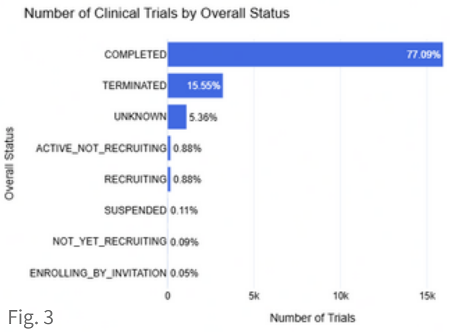


Fig. 3

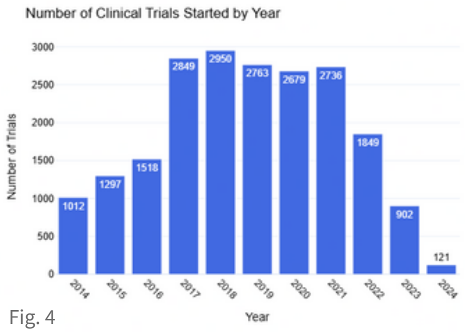


Fig. 4

The EDA process was extensive, involving numerous plots and visualisations to understand the dataset thoroughly. Due to the volume of insights generated, the complete EDA and all associated plots can be found in the **code** and **saved plots**.

FEATURE ENGINEERING

Geographic data was created by extracting city and country names from the location column in the dataset. This was achieved using **clinicaltrials.gov API**, which was sufficient for the task. The extracted city names were then matched with **population** data sourced from SimpleMaps' World Cities Database, which contains demographic information for approximately 47,000 major cities worldwide. The population of each city was included as a feature, as it can influence the number of potential participants available for recruitment in a clinical trial. The data on the country and number of trials can be interactively viewed on your desktop from **here**.

Two key **external factors** were incorporated into the analysis:

Competition: This feature represents the number of concurrent clinical trials occurring(overlap) in the same location (city) because if multiple trials were recruiting for participants in the same area, the availability for a given trial might be reduced.

Disease Prevalence: This factor was included to represent the condition under study in a specific location. Higher disease prevalence in a region was expected to correlate with a larger pool of eligible participants. **Disease burden** was quantified using **Disability-Adjusted Life Years** (DALYs), which were sourced from the WHO's Burden of Disease Database. Since the disease data source contained conditions organised in four hierarchical levels, **semantic matching** was performed to align the dataset's condition column with the most relevant hierarchy. The **medembed-small** model, an open-source embedding model fine-tuned for clinical data, was used to calculate semantic similarity scores. The hierarchy with the best match (highest similarity score) was selected for each condition, ensuring accurate and contextually relevant representation.

The **study design** column in the dataset contained four components, which were separated into individual columns:
Allocation: Described how participants were assigned to different groups in the trial.
Intervention Model: Outlined how participants were grouped or treated during the study.
Masking: Indicated the level of blinding (e.g., single-blind, double-blind) used to minimise bias in the trial.
Primary Purpose: Explained the main objective of the clinical trial (e.g., treatment, prevention, diagnostic).

Conditions listed in the dataset were mapped to **Medical Subject Headings (MeSH)** terms, reducing the total number of unique condition values from **7860** to **1328**, represented by **106** MeSH IDs. This generalisation improved the usability of the data by consolidating similar conditions into standardised categories. To achieve this, the MeSH dataset **medembed v1.0** was utilised to create embeddings for the conditions column. **Semantic matching** was performed to identify the top 5 MeSH terms for each condition, and their corresponding descriptor UIs (unique identifiers) were extracted. These descriptor UIs were then used to access the MeSH API, which provided tree IDs (e.g., C04, D03) in JSON format. These tree IDs enabled the creation of the feature column, adding a hierarchical and structured representation of the conditions to the dataset. This approach improved the interpretability of the data and **enhanced its utility for machine learning models**.

I**nclusion and exclusion criteria** were scraped from the study URLs provided in the dataset. To quantify the complexity of these criteria, **Shannon entropy** (both first-order and second-order) was calculated. Shannon entropy measures the unpredictability or variability in the criteria, with higher entropy indicating greater complexity or diversity in the requirements. This metric was useful in assessing the variability and specificity of the criteria, which could influence recruitment rates by reflecting how **restrictive or broad the eligibility conditions** were.

A feature representing the **sponsor's historical success** rate was created by calculating the ratio of successful trials to the total number of trials sponsored by each entity. This feature was included to account for the potential influence of sponsor experience on trial outcomes. This data was obtained from the HINT paper(see references) and a dataset from their Github.

The **duration of each trial** was calculated as the difference between the start and end dates, measured in days. This feature provided insight into the time commitment required for each study, which could impact participant recruitment.

DATA PREPROCESSING

The data processing phase focused on consolidating and refining the dataset to ensure it was suitable for model training. All engineered features, such as geographic data, disease prevalence, trial competition, sponsor success rates, study duration, and condition embeddings derived from MeSH terms, were **merged with the original dataset**, enriching it with additional contextual and predictive information. Several columns were dropped to streamline the dataset. The NCT Number was kept as an identifier but excluded from model training. Columns like Study Title, Study URL, Brief Summary, and date-related fields (e.g., Primary Completion Date, First Posted) were removed, as they were either **descriptive or unrelated to recruitment dynamics**. Features such as Interventions and outcome measures (Primary Outcome Measures, Secondary Outcome Measures, and Other Outcome Measures) were also dropped, as they **represented post-recruitment processes** and had no influence on recruitment outcomes. The Study Type column was removed due to its single value (INTERVENTIONAL). Columns like Start Date, Completion Date, Study Design, Locations, and Eligibility Criteria were dropped, as their useful information had **already been extracted into new features**. Similarly, the Sponsor and Collaborators columns were removed, as their information was generalised in the Funder Type feature and supplemented by the sponsor success rate feature.

One-hot encoding was applied to **polytomous categorical variables**, and NaN values in these columns were replaced with 0 to indicate absence. For numerical columns, NaN values were filled with their **mean** values, and one outlier identified in enrollment during EDA was removed to prevent skewing the model's predictions. MeSH IDs were encoded using a frequency encoder, preserving the relative importance of common conditions while reducing dimensionality. Finally, all numerical values were **scaled to a 0-1 range** to ensure uniformity and improve model convergence.

The correlation between all the features can be seen quantitatively through an Excel sheet we've created. **Excel Sheet Link**

FEATURE SELECTION

After data preprocessing, the dataset comprised **56 columns**. To identify the most relevant features for recruitment prediction, feature importance scores were calculated using a Random Forest (RF) model, which evaluates features based on their contribution to reducing impurity across decision trees. The **top 40** features with the highest importance scores were selected for model training. This approach not only **reduces dimensionality** but also focuses on the most impactful features, enhancing model efficiency, reducing overfitting risks, and improving the model's ability to generalize to unseen data while maintaining high predictive accuracy. Click for **Larger Plot**.

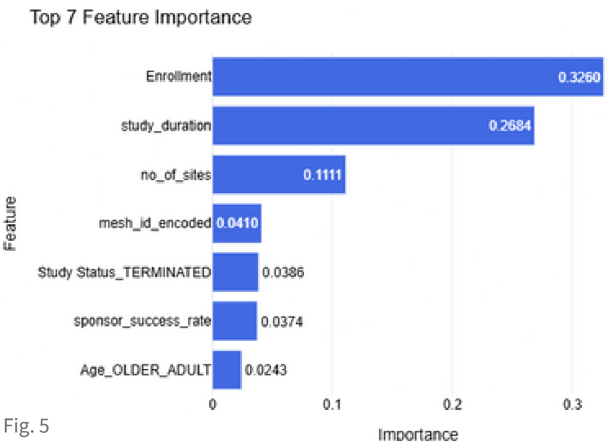
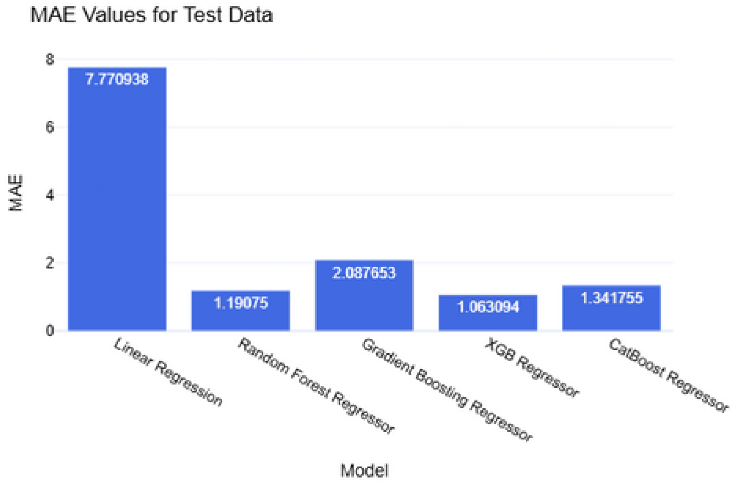


Fig. 5



MODEL TRAINING

In the baseline testing phase, several models were evaluated, including Linear Regression, Extreme Gradient Boosting (XGBoost), Gradient Boosting, Random Forest, and CatBoost. The best performance was achieved by Gradient Boosting, with an R^2 score of 0.840811 and an MAE of 2.087653. XGBoost also performed well, achieving an R^2 score of 0.809940 and an MAE of 1.063094. But this might be an overfit too.

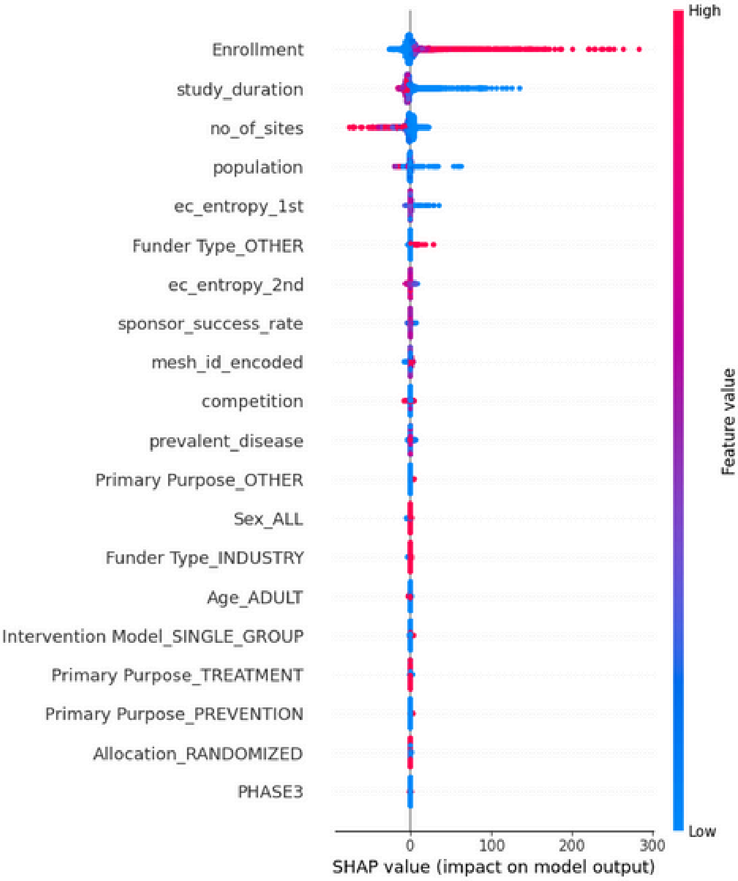
MODEL TUNING AND RESULTS

Following baseline testing, hyperparameter tuning was performed using Optuna, a Bayesian optimization framework, to fine-tune the models. Optuna was run for 25 trials to identify the best hyperparameters. After optimization, the best results achieved were an R^2 score of **0.7626** and an MAE of **2.098**, demonstrating improved model performance.

MODEL EXPLAINABILITY

The features were ranked by their mean absolute SHAP values, reflecting their overall impact on the model's predictions. The top features included Enrollment, study_duration, no_of_sites, population, and ec_entropy_1st/2nd. Enrollment had the highest impact, with larger trials (high enrollment) often correlating with lower predicted outcomes, possibly due to logistical challenges.

Longer study durations and a higher number of sites were also linked to reduced predictions, likely due to increased complexity. Population demographics and eligibility criteria entropy (ec_entropy_1st/2nd) further influenced outcomes, with higher entropy indicating greater variability and potential complications. Categorical features like Funder Type_INDUSTRY and Primary Purpose_TREATMENT also played a significant role. Industry-funded trials and those focused on treatment were associated with positive SHAP values, indicating higher success probabilities. Gold-standard trial design features, such as Allocation_RANDOMIZED and PHASE3, further enhanced predictions, aligning with their established reliability. Feature interactions revealed that higher entropy in eligibility criteria (ec_entropy_1st/2nd) and trials for prevalent diseases could reduce predictability, leading to negative SHAP values. These insights highlight the importance of risk mitigation (e.g., careful planning for large or long trials), design optimization (e.g., prioritizing randomized, Phase 3 trials with experienced sponsors), and targeted recruitment (e.g., smaller, focused trials) to improve outcomes.



CHALLENGES AND FUTURE WORK

The dataset’s heavy bias toward trials from the United States may limit the model’s generalizability to other regions. Future work could involve incorporating trials from more diverse countries. Exploring deep learning models and advanced embedding techniques could improve feature representation. Dynamically updating the competition factor through real-time API calls and developing a scalable product could enhance usability. Additionally, sourcing a larger and more detailed disease prevalence dataset would address the computational challenges of semantic search with open-source LLMs.

REFERENCES

ClinicalAgent: Clinical Trial Multi-Agent System with Large Language Model-based Reasoning link

TrialDura: Hierarchical Attention Transformer for Interpretable Clinical Trial Duration Prediction link

HINT: Hierarchical Interaction Network for Clinical Trial Outcome Prediction link

Uncovering key clinical trial features influencing recruitment link

Prediction of clinical trial enrollment rates link