NOVARTIS

NEST

Nurturing Excellence,
Strengthening **Talent.**

Let’s reimagine medicine, together!

Problem Statements from Process Innovation & AI, Development

Overall Theme: Accelerating clinical development by authoring a complete protocol that minimizes avoidable amendments and maximizes effectiveness

NOVARTIS

Contents

NOTE: PS is “Problem Statement”

Knowledge ramp up on clinical trials & protocol

A clinical trial is a research study conducted to evaluate the safety, and efficacy of medical interventions such as drugs, devices, or treatments on humans.

These trials are often conducted in phases to progressively assess the intervention's safety, dosage, and therapeutic effect.

The clinical trial protocol is a detailed plan that outlines the objectives, design, methodology, statistical considerations, and organization of a clinical trial. It ensures the trial is conducted consistently and ethically, providing a framework for how the study is carried out and how data is collected and analyzed. Following the protocol is crucial for the validity and integrity of the trial results.

The two major reasons a clinical protocol gets delayed or causes delays in the clinical trial timeline are:

- Patient Recruitment and Enrolment Challenges: Stringent patient inclusion & exclusion criteria can lead to difficulties in identifying and enrolling the right patients can significantly slow down the trial process.
- Protocol Amendments: Unanticipated amendments to the protocol, often due to changes in trial design or new regulatory requirements, can lead to considerable delays.

Important documents to develop understanding

(Participants are free to use any other free, non-proprietary sources for secondary research to develop a clearer business & process understanding)

AI in Clinical Trials Adobe Acrobat Document

Introduction to Clinical Trials Adobe Acrobat Document

CFR21: Code of Federal Regulations Title 21
ICH: International Council for Harmonisation
USFDA: United States Federal Drug Administration
GCP: Good Clinical Practice

Problem Statement - 1

Semantic grouping of clinical studies for retrieval & strategic insights

NOVARTIS

PS- 1: Semantic* grouping of clinical studies for retrieval & strategic insights

Business context & challenge:

While designing and writing a clinical trial protocol document, multiple empirical, scientific and medical references are used. Like any other hypothesis testing, it is helpful in gaining insights from similar historical research / experiments / clinical studies from the past to minimize chances of failure and improve predictability, quality & execution. The researcher / author's interest would be focused on finding a subset of trials as relevant as possible, categorized by drug(s) and/or disease(s) and other factors, given past clinical trials data. These insights can help avoid errors and improve the speed and efficiency of the clinical trial design, which remains a challenge for pharma sector. Delays in designing, errors etc. have a cascading impact on timelines for clinical trials.

Problem statement from technical lens:

Obtaining a matched subset of clinical trials for only a combination of 4 unstructured features (textual data: Study Title, Primary outcome measure, Secondary outcome measure and Criteria) and using the patterns found in them after applying human logical thought process. Expectation is to design the approach, model, method and/or algorithm to enable this.

Dataset:

A smaller subset of 450,000 clinical trials data will be provided from clinicaltrials.gov (publicly available)

Technical solution requirement:

Utilizing AI & deep learning methods, obtain a solution such that there is a mechanism to retrieve the relevant clinical trials based on a query.

*Connected with the meaning of words and sentences

PS- 1: Semantic grouping of clinical studies for retrieval & strategic insights

Input (for example):

Study Title, Primary outcome measure, Secondary outcome measure and Criteria

Output (for example):

For each unique clinical trial present in dataset, output must be 10 unique similar* clinical trials from the dataset.

Metrics for Evaluation:

Using 3 different trials NCT00385736 , NCT00386607 and NCT03518073 we would evaluate 10 similar trials coming from each trial respectively using our internal algorithm.

Additional evaluation:

Exploratory data analysis, Data preparation/cleaning, feature engineering, AI Framework, evaluation using appropriate metrics and the model explainability (It is crucial that the model's explainability aligns with domain understanding of clinical trials); weighted more on model explainability.

[Note: "Criteria" column is found in eligibilities.txt file]

*similarity based on the match on: Phases, Conditions, Interventions, Criteria, etc.

Problem Statement - 2

Predicting Actual Enrollment Duration of clinical studies with explainability

NOVARTIS

PS- 2: Predicting actual enrolment duration for clinical studies with explainability

Business context:
While designing and writing a clinical trial protocol document, multiple empirical, scientific and medical references are used. Like any other hypothesis testing, it is helpful in gaining insights from similar historical research / experiments / clinical studies from the past to minimize chances of failure and improve predictability, quality & execution. Authors of protocol documents would benefit from accurately identifying the right criteria for patients and accelerating patient recruitment for the clinical trial. Time taken for actual enrollment ~~prediction~~ during protocol development will provide valuable insights into whether the criteria should be more restrictive or broad, given past clinical trials data (e.g., avoid overly restrictive criteria that hinder recruitment, but also avoid criteria so broad that they fail to adequately test drug efficacy). These insights can help avoid errors and improve the speed and efficiency of the clinical trial design, which remains a challenge for pharma sector. Delays in designing, errors etc. have a cascading impact on timelines for clinical trials.

Problem statement from technical lens:
Predicting the actual enrolment duration (in months) of Completed "interventional" studies based on features (structured and unstructured) such as enrollment, study design, criteria, facility, country etc. Additionally, providing explainability that answers the reason for the prediction.

Dataset:
A smaller subset of 450,000 clinical trials data will be provided from clinicaltrials.gov (publicly available)

Technical solution requirement:
Utilizing AI & deep learning methods, obtain a solution such that the prediction of enrollment time taken based on Disease/Condition is backed by explainability (positively or negatively) affecting the magnitude. Causal inference approach will be given more points.

Explainability: easily interpretable outcome with human / understandable logic

NOVARTIS

PS- 2: Predicting actual enrolment duration for clinical studies with explainability

Input (for example):

Condition, Phases, Facility location, Enrollment, Criteria, Study design, Study title, Intervention, etc. (various other features can be included)

Output (for example):

30 months (predicted) with explainability – which features that positively and negatively influence the prediction.[Explainability: show the weightage of each feature on impact to make prediction]

Metrics for Evaluation:

Response variable: `Time taken for Enrollment`
RMSE (root mean squared error), R squared and adjusted R squared.

Symmetric Mean Absolute Percent Error (SMAPE)

Additional evaluation:

Exploratory data analysis, Data preparation/cleaning, feature engineering, modeling, evaluation using mentioned metrics and the model explainability (It is crucial that the model's explainability aligns with domain understanding of clinical trials); weighted more on model explainability. Important to identify the features used by the team.

[Note: You are free to choose as many or few features from the dataset and not just study design and criteria]

[Note: "Criteria" is found in eligibilities.txt file]

Problem Statement - 3

Predicting Completion of clinical studies with explainability

NOVARTIS

PS- 3: Predicting completion of clinical studies with explainability

Business context:
While designing and writing a clinical trial protocol document, multiple empirical, scientific and medical references are used. Like any other hypothesis testing, it is helpful in gaining insights from similar historical research / experiments / clinical studies from the past to minimize chances of failure and improve predictability, quality & execution.
Scientists who write clinical trial protocol documents would benefit from being able to predict whether the trial being designed would be completed or not due to some factors, given past clinical trials data. These factors may include (but not be limited to) trial amendments, criteria, disease indication etc.
These insights can also help determine the probability of success of clinical trials, allocation of investments in R&D, also help avoid errors and improve the speed and efficiency of the clinical trial design, which remains a challenge for Pharma sector. Delays in designing, errors etc. have a cascading impact on timelines for clinical trials.

Problem statement:
Prediction of the trial completion status ("Completed" or "Not Completed") based on features (unstructured and structured data) such as study design, criteria, adverse events, etc. Explainability is optional but adds more value to the solution. Here, "Not Completed" is "Suspended", "Withdrawn" or "Terminated".

Dataset:
A smaller subset of 450,000 clinical trials data will be provided from clinicaltrials.gov (publicly available)

Technical solution requirement:
Using AI & deep learning methods, obtain a solution such that the prediction of trial status is backed by explainability (positively or negatively) affecting the status. Causal inference approach would be given more points.

PS- 3: Predicting completion of clinical studies with explainability

Input (for example):

Study design, Criteria, Study title, Condition, Intervention, etc. (various other features can be included)

Output (for example):

"Not Completed" (trial status predicted) with explainability of such prediction. [Explainability: show the weightage of each feature on impact to make prediction]

Metrics for Evaluation:

Response variable: `Study Status`
Precision, Recall, F1, Confusion matrix and AUC-ROC

Additional evaluation:

Exploratory data analysis, Data preparation/cleaning, feature engineering, modeling, evaluation using mentioned metrics and the model explainability (It is crucial that the model's explainability aligns with domain understanding of clinical trials); weighted more on model explainability.

[Note: You are free to choose as many or few features from the dataset and not just criteria]

Problem Statement - 4

Utilizing data to predict recruitment rate (RR) in clinical trial for benchmarking

NOVARTIS

PS- 4: Utilizing data to predict recruitment rate (RR) in clinical trial for benchmarking

Business context

Effective recruitment is one of the most significant challenges in conducting clinical trials. The success of a trial often hinges on the ability to enrol the right **number of participants** within a **specified timeframe**. Understanding recruitment rates and identifying appropriate benchmarks are critical for optimizing this process. At present, the method for benchmarking historical clinical studies to determine the Recruitment Rate (RR) relies on factors like study population similarity, sample size, study phase, sponsor type, geographic scope, etc. However, this approach fails to account for dynamics such as external competition, availability of standard of care (SOC), specific populations or sub-populations, and niche or rare diseases lacking trial history.

- **Definition:** Recruitment Rate (RR) refers to speed at which patients are enrolled in a clinical trial, typically measured as **patients per site per month** (p/s/m)
- **Process:** Benchmarking involves comparing recruitment rates against established standards or historical data from similar trials. This practice helps identify realistic targets and informs strategies to improve resource allocation (e.g., #sites needed), time allocation (how long to plan the trial), and adjust operations

Problem statement

A predictive model / framework which calculates with a degree of confidence the estimated Recruitment Rate of a clinical study based on various internal (e.g., study parameters) & external (e.g., competition) factors

Dataset

A smaller subset of 450,000 clinical trials data will be provided from clinicaltrials.gov (publicly available)

Technical solution requirement

Data driven scalable solution that can be incorporated in internal planning system

Adobe Acrobat Document

Evaluation of factors associated with recruitment rates in early phase clinical trials based on the European Clinical Trials Register data

Adobe Acrobat Document

Clinical Trial Recruitment & Retention

NOVARTIS

PS- 4: Utilizing data to predict recruitment rate (RR) in clinical trial for benchmarking

Inputs (illustrative)

Study indication, phase, design, endpoints, sponsor, competition, enrolment target, etc. (non-exhaustive)

Output (for example)

- RR
- Key predictors impacting RR
- Weightage each predictor will carry at a trial level

Metrics for Evaluation:

Approach/Methodology (scenarios, no of parameters used, relevant factors selected, type/relevance of algorithm)

- **Data Processing and Analysis** (collection, cleaning, exploratory data analysis, etc.)
- **Model Selection and Designing** (selection, justification, design consideration etc.)
- **Model Training and Evaluation** (training process, evaluation, RMSE, Accuracy, Precision, Recall, Mean Absolute Error, R-squared (R^2) score, Overfitting and underfitting analysis etc.)
- **Overall Analysis** (summary of findings, model interpretability, visualization of results, comparison with baseline models, practical implementations and next steps etc.)

Submission Guidelines

NOVARTIS

Submission Guidelines Template

Innovation Challenge Submission Guidelines

Participants must submit their final solutions in a specified format to ensure consistency and ease of evaluation. The solution should be packaged into a single zip file containing a detailed report and all the code necessary to reproduce the results. Below are the detailed guidelines for preparing and submitting your solution.

Contents of the File:

1.Report (pdf format): [Not more than 5 pages, else will be disqualified]

- **File Name:** report.pdf
- **Content:**
  - **Title Page:**
    - Team Name
    - Use-Case Number
    - Team Members' Names and Contact Information
  - **Abstract:**
    - A brief summary of the problem, key assumptions, approach, and results (also highlight and key limitation identified in model)
  - **Introduction:**
    - Background information and problem statement.
  - **Methodology:**
    - Detailed explanation of the approach, including data preprocessing, feature engineering, model development, and evaluation.
  - **Results:**
    - Presentation of results using appropriate metrics.
    - Diagrams and visualizations to support the findings (e.g., charts, graphs, confusion matrices).
    - Analysis of the results and insights gained.

2.Code: (must be reproducible)

Data for all 4 Problem Statements

PS- 1: similar trials - only studies that have funder as industry; additionally add **eligibilities.txt** using *nct_id* column to join.

PS- 2: Time taken for enrollment (in months) - only interventional and completed studies; additionally add **drop_withdrawls.txt**, **eligibilities.txt**, **facilities.txt** using *nct_id* column to join.

PS- 3: trial status - completed vs "not completed" (i.e. Withdrawn + Terminated + Suspended); additionally add **drop_withdrawls.txt**, **eligibilities.txt**, **facilities.txt**, **reported_events.txt** using *nct_id* column to join.

PS- 4: File **usecase_4_.xlsx** contains completed industry sponsored studies from last 10 years. Study Recruitment rate (column AC) based on patients per site per month. **Additional parameters to be explored from this dataset for model building.**

NOTE: PS 1,2 and 3 have different base datasets, and we expect that you will join the suggested txt files to then execute the use-case.

All the best!

NOVARTIS