# Predicting the Probability of Clinical Trial Succes

## Background

Drug development is costly and uncertain, with success rates varying widely across therapeutic areas and phases. Predicting the **Probability of Trial Success (PTS)** can guide better R&D investment, pipeline prioritization, and business development decisions.

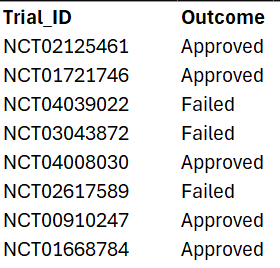
In this hackathon, your challenge is to **develop a machine learning model that predicts the PTS for ongoing (active) Phase-3 clinical trials**, based on learnings from historical trials.

## Objective

Create a model that **predicts the probability of success** for each trial. A successful solution will combine **domain-aware feature engineering**, **modeling sophistication**, and **transparent explainability**.

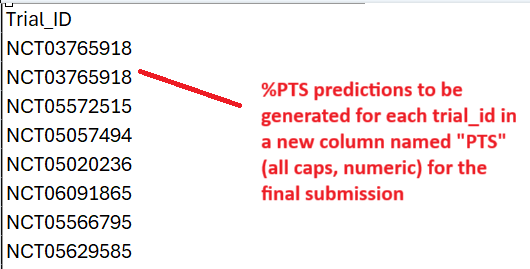
## Data Provided

1. **Historical Trials Dataset (Training Set)**:
   * **1273** completed trial IDs (<https://clinicaltrials.gov/>) with known outcomes *(Failed/Approved)*



* + You will be given a Google colab starter code (Refer ***Trial\_Data\_Download\_Starter.ipynb****)* to download the data from ct.gov website given a clinical trial ID with some default features. Feel free to explore the API guide at the official ct.gov website & modify the starter code for additional data fields or feel free to explore any other additional options for feature collation.
  + You will be given a Google colab starter code (Refer ***Hackathon\_LocalLLM\_Starter.ipynb***) to use a local LLM for data processing if you want to as the data is largely unstructured.
  + You are free to use traditional models for unstructured data processing like BERT, if you wish to not use Gen AI or use commercial APIs like Gemini and Open AI if you have access to it.

1. **Active Trials Dataset (Test Set)**:
   * **340** active trials (Trial IDs only)
   * Your model must generate **%PTS predictions for each trial.**

****

1. **Optional External Data**:
   * **You are encouraged to enrich your solution using publicly available datasets, provided the process is well-documented.**

## Evaluation Criteria

| **Area** | **Weight** | **Details** |
| --- | --- | --- |
| **Feature Engineering** | 30% | How well does the solution reflect trial complexity? Creativity in incorporating trial design, sponsor track record, mechanism novelty, timelines, or external data sources. |
| **Modeling Approach** | 30% | Quality of model setup: handling TA/phase-specific baselines, non-linearities, class imbalance, etc. |
| **Interpretability** | 20% | Transparent explanation of why trials get specific %PTS scores. |
| **Prediction Calibration & Fidelity** | 20% | Quality of predicted probabilities - Are %PTS values well-calibrated and discriminative? Metrics like RMSE, MAE, and rank correlation are used for quantifying the quality of your submission |

*Bonus points* for:

* Providing interpretable dashboards or interactive summary visuals.
* Calibrated uncertainty estimation (e.g., confidence intervals around %PTS).

## Deliverables

* **Codebase + notebooks** showing full pipeline: data prep, feature engineering, model training, prediction.
* **Predicted %PTS** scores (0–100%) for each trial in the test set (CSV).
* **Explainability assets** (e.g., SHAP plots or insights summary).
* **One-page executive summary + video walkthrough** of your approach, key assumptions, and insights.

## Timeline

* Problem Statement release: June 13th, 2025, 5.30 PM IST
* Submission deadline: **June 22rd, 2025, 11.59 PM IST** *(No request for delays will be entertained)*
* Judging & results: To be announced later

## Guidelines

1. Treat this as a real-world use case; noisy metadata, missing values, and TA-specific patterns are all part of the challenge.
2. Interpretability and feature reasoning matter as much as raw model performance.
3. Use appropriate model validation strategies.

## Understanding Clinical Trials: Quick Primer for Participants

1. **What is a Clinical Trial?**
   * Clinical trials are research studies conducted to evaluate the safety and efficacy of new drugs, treatments, or medical devices.
   * They are conducted in phases (I, II, III, IV) and are registered with key details like sponsor, indication, design, and primary endpoints.
2. **Navigating ClinicalTrials.gov (CT.GOV) and Trial IDs**
   * https://clinicaltrials.gov is a public registry of global clinical trials.
   * Each trial is identified by a unique NCT ID (e.g., NCT01234567)
   * You can search trials by condition, sponsor, phase, or status and many other fields mentioned in the official website.
   * Some key fields include: phase, intervention, status (Completed, Recruiting, etc.), primary outcome measures, start/end dates.
   * There is also an option for bulk download/API export. Explore the website for more details on the search terms and supported parameters.
3. **Trial Outcome: What Does "Success" or "Failure" Mean?**
   * In this problem, a trial is considered:
     + **Approved** if it met its primary endpoint(s) and progressed (or contributed) to a regulatory submission or approval
     + **Failed** if it did not meet endpoints, was terminated early, or was not continued for strategic/efficacy reasons
4. **How to Think About Predicting %PTS**
   * A trial’s success depends on many signals: trial design quality, sponsor track record, novelty of the drug, etc.
   * Your task is to quantify the likelihood (0–100%) that an active trial will be successful, using insights learned from historical patterns in the training data