

# Leveraging Machine Learning Models to Predict Probability of Technical and Regulatory Success

**Mason Kim**

mason.kim@biogen.com

Biogen, Inc

**Michael Doane**

Iambic Therapeutics

**James Melican**

Biogen, Inc

**Kevin Wang**

**Haewon Jung**

UPMC Enterprises

**Cedric Liu**

Biogen, Inc

---

## Research Article

**Keywords:** Probability of technical and regulatory success (pTRS), clinical trial outcome prediction, ClinicalTrials.gov, interpretable machine learning, model calibration, Brier score

**Posted Date:** July 8th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-9416583/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.  
[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Title

Leveraging Machine Learning Models to Predict Probability of Technical and Regulatory Success

## Author Information

Mason Kim<sup>1</sup>, [mason.kim@biogen.com](mailto:mason.kim@biogen.com)

Michael Doane<sup>2</sup>, [mdoane@alumni.stanford.edu](mailto:mdoane@alumni.stanford.edu)

James Melican<sup>1</sup>, [james.melican@biogen.com](mailto:james.melican@biogen.com)

Kevin Wang, [kevin.wang7701@gmail.com](mailto:kevin.wang7701@gmail.com)

Haewon Jung<sup>3</sup>, [haewonj710@gmail.com](mailto:haewonj710@gmail.com)

Cedric Liu<sup>1</sup>, [hangyu.liu@biogen.com](mailto:hangyu.liu@biogen.com)

<sup>1</sup> Biogen, Inc., Cambridge, MA, USA

<sup>2</sup> Iambic Therapeutics, San Diego, CA, USA

<sup>3</sup> UPMC Enterprises, Pittsburgh, PA, USA

# Abstract

## Purpose

Probability of technical and regulatory success (pTRS) is a key driver of risk-adjusted valuation and portfolio decisions, yet many assessment methods of pTRS depend on proprietary data or complex, hard-to-replicate architectures. We aimed to develop a reproducible, interpretable pTRS assessment method using only public trial features and open labels.

## Methods

We extracted protocol-section features from ClinicalTrials.gov and paired them with binary outcome labels from the HINT clinical-trial-outcome repository (IQVIA-derived; small-molecule trials through 2021). Structured protocol variables were cleaned and one-hot encoded. Trials were mapped into 21 therapeutic areas using LLM-assisted classification of conditions/keywords. Data were split temporally: training included trials with primary completion dates before January 1, 2017 (n=10,224) and testing included trials with start dates after January 1, 2017 (n=478). To assess the impact of temporal separation on model performance, models were additionally trained and evaluated using datasets constructed with varying year gaps (0, 1, 3, and 5 years) between the training and test periods. We evaluated logistic regression, random forest, and XGBoost using ROC-AUC, PR-AUC, and Brier score, plus benchmark-relative comparison. Interpretability was assessed with SHAP.

## Results

Random forest produced the strongest aggregate performance (ROC-AUC 0.788; PR-AUC 0.794; Brier 0.196), improving upon baseline model and the benchmark comparator across most settings. Sponsor type, phase, therapeutic area, and key design variables were consistently influential.

## Conclusion

Interpretable models trained on open ClinicalTrials.gov protocol features can generate calibrated, decision-useful pTRS estimates that outperform standard benchmarks, offering a practical prior to complement expert elicitation in valuation and portfolio workflows.

# Summary

We developed a reproducible, interpretable pTRS prediction approach using only ClinicalTrials.gov protocol features and open outcome labels (small-molecule trials through 2021). After cleaning and one-hot encoding structured fields and mapping trials to 21 therapeutic areas via LLM-assisted categorization, we trained models with a temporal split (training: pre-2017; test: 2017 onward). A random forest achieved the best overall performance on held-out data (ROC-AUC 0.788; PR-AUC 0.794; Brier 0.196), outperforming the baseline logistic-regression model and the benchmark pTRS estimates. SHAP indicated sponsor type, phase, therapeutic area, and trial design features were the dominant drivers.

## Teaser

An interpretable, fully reproducible model using only ClinicalTrials.gov protocol data improves pTRS prediction versus benchmarks, highlighting sponsor, phase, therapeutic area, and design features to support valuation decisions.

## Keywords

Probability of technical and regulatory success (pTRS), clinical trial outcome prediction, ClinicalTrials.gov, interpretable machine learning, model calibration, Brier score

## Statements and Declarations

### **Funding**

This work was supported by Biogen, Inc.

### **Competing Interests**

The authors declare that they have no competing interests.

### **Ethics Approval**

Not applicable. This study used publicly available and/or de-identified secondary data and did not involve human participants as defined by applicable IRB oversight.

### **Consent to Participate**

Not applicable.

## **Consent to Publish**

Not applicable.

## **Data Availability**

The ClinicalTrials.gov source data used in this study are publicly available. The clinical trial outcome labels used for training and evaluation are available from the HINT clinical-trial-outcome-prediction repository. The code used to preprocess data, train models, and generate results is available at [GitHub - futianfan/clinical-trial-outcome-prediction: benchmark dataset and Deep learning method \(Hierarchical Interaction Network, HINT\) for clinical trial approval probability prediction, published in Cell Patterns 2022. · GitHub.](#)

## **Code Availability**

If the repository is not public: The code is available from the corresponding author upon reasonable request.

## **Author Contributions**

Not included.

## **Acknowledgments**

The authors acknowledge Sean Xu and Natalie Yang for their contributions to data cleaning and model development, and Professor Holly Wiberg (Carnegie Mellon University) for her guidance and valuable insights throughout the project.

# Introduction

The pharmaceutical industry operates within a uniquely challenging paradigm characterized by extraordinary capital requirements, extended development timelines, and exceptionally high failure rates that necessitate sophisticated risk assessment methodologies. Accurately assessing the present value of pharmaceutical products during the various stages of research and development poses a challenge for drug developers and investors alike because of the low probability of a new drug successfully completing clinical trials and becoming approved. After discovery, an investigational new drug (IND) has an average probability of technical and regulatory success (pTRS) of 9.6%, but some therapeutic areas experience average success rates as low as 5.1%. [1] This reality underscores the critical importance of developing and implementing appropriate pTRS assessment methodologies that can effectively inform valuation models, strategic business decisions, and investment allocation across the pharmaceutical development landscape. However, beyond methodological rigor, the practical use of pTRS also depends on cultural acceptance within organizations, as stakeholders may question its ability to reflect broader scientific merit, innovation potential, or long-term research value.

## **pTRS Background**

Companies use “probability of technical and regulatory success” (pTRS) to capture and quantify the technical and regulatory risks that can potentially derail a drug’s candidacy. This value can represent the probability of technical success (pTS) for a trial phase, probability of regulatory success (pRS), or the overall probability of technical and regulatory success (pTRS), which incorporates both. This industry-standard formulation of these risks permits program and portfolio analysis that accurately expresses the risk-adjusted cost, revenue, and expected value of a program, predicated on the probability of actually transitioning to each downstream phase.

## **Valuation Imperatives and Business Applications**

The estimation of pTRS serves as a cornerstone of pharmaceutical asset valuation, directly influencing how companies, investors, and strategic partners assess the economic potential of clinical programs. The pTRS or probability of success, is the likelihood that a biopharmaceutical asset can successfully progress through pivotal clinical trials and ultimately gain marketing approval from regulatory bodies such as the FDA or EMA. Clinical trials' probability of success affects the valuation of biotech assets and is one of the most critical drivers of value, and this probability is incorporated in the risk-adjusted NPV-based valuation. [2]

The application of appropriate pTRS estimates extends far beyond simple valuation exercises, serving multiple critical business functions including licensing negotiations, merger and acquisition decisions, and internal resource allocation strategies. Johnson and Johnson purchased Covagen AG in 2014 primarily for their lead product, COVA 322, a bispecific anti-

tumor necrosis factor in Phase 1b clinical trials for psoriasis with the potential to treat a broad range of inflammatory diseases. The historical probability of technical and regulatory success for investigational compounds in the Autoimmune/Inflammation therapeutic area from Phase 1 to Approval was found to be 15.1%, but Johnson and Johnson utilized a higher 26% probability of success factor when determining the present value of COVA 322, demonstrating how sophisticated organizations adjust benchmarks pTRS estimates based on asset-specific characteristics and strategic considerations. [1]

Net present value (rNPV) models represents the most common method for pharmaceutical asset valuation, with pTRS serving as the primary risk adjustment mechanism. rNPV enhances standard discounted cash flow (DCF) analysis by adjusting cash flow projections for the probability of success, i.e., adjusting for the pTRS or conversely, attrition rate. As a result, this method is also referred to as the expected net present value (eNPV) method [2]. This methodology enables organizations to incorporate the inherent uncertainty of drug development into their financial models, providing a more realistic assessment of asset value compared to traditional NPV approaches that rely solely on discount rate adjustments.

### **Industry Struggles with Risk Quantification**

The pharmaceutical industry has historically struggled with rigorous, standardized risk quantification, often relying on human judgment, narrow proxy metrics, or incompletely structured approaches that do not fully capture uncertainty and portfolio-level effects. [3-5] The published literature still lacks a standardized, widely adopted framework for assessing manufacturer pricing and market-access risk across products, stakeholders, and geographies. [6-8] This gap in comprehensive risk assessment contributes to inefficient resource allocation, weaker portfolio decisions, and lower R&D productivity, with very large economic consequences for the industry. [5, 9-11]

The consequences of inadequate risk assessment extend beyond individual companies to affect the entire innovation ecosystem. Imprecise, biased or otherwise incorrect pTRS assessments can lead to overvaluation of development programs, resulting in excessive capital deployment in high-risk, low-probability ventures while simultaneously underinvesting in more promising opportunities. This misallocation of resources ultimately slows the pace of innovation and delays the delivery of life-saving therapies to patients.

### **Current state of pTRS prediction**

Various methods exist for drug developers to assess their programs' pTRS values, including benchmarks, expert elicitation, and data inference.

#### ***Benchmarks***

Drug developers may elect to use an average probability of success to represent the pTRS for their own candidates. [12] These estimates can be improved by using the mean success rate within a given clinical development phase (e.g., Phase 2) and/or therapeutic area (e.g., cardiovascular). [13] Often these averages will come from subscription databases that compile industry-wide success rates, but those datasets may be incomplete or biased by geography and company size. Large companies may also generate internal benchmarks from their development history.

Benchmarks have the advantage of providing a rapid and objective snapshot of a drug candidate's pTRS. But they also fail to adjust for information gleaned from the individual clinical history or success factors of a given drug.

### ***Expert Elicitation***

Drug developers who wish to customize their programs' pTRS values based on individual properties of those candidates may utilize a process called expert elicitation. In this method, a team of internal experts is convened, often including representatives from Clinical Development, Biostatistics, Medical, Regulatory, Safety, and other functions. Following discussion, the experts can render an assessment of the probability of success of a specific program, usually phase-by-phase, or generate a Bayesian prior for statistical analysis. [14]

This technique has the benefit of being able to take into consideration all of the specific positive and/or negative attributes of a drug candidate as well as its previous clinical history. However, this method is also quite time consuming, and vulnerable to the subjectivity of experts who may work closely on the program in question, and misalignment with commercial forecast assumptions can lead to overly-optimistic eNPV calculations. [15]

### ***Data Inference***

Inferential techniques aim to be the best of both worlds, combining the objectivity of large datasets with the potential to make bespoke adjustments based on drug or target properties. Several research teams have attempted to apply this approach to prediction of success rates for drug candidates.

The ProCTOR methodology, developed by Gayvert, et al., combined drug “likeness” approach (which compares candidates with existing drugs of known safety profile) with structural and target features to predict the likelihood of toxicological concerns. [16] Evaluated on 784 drugs, this method demonstrated a ROC-AUC (receiver operator curve area under curve – a measure of performance for binary classification models) of 0.826, albeit on a relatively small sample. [17] A year later, Lo, et al., published work on a predictive model based on two large databases, *Pharmaprojects* and *Trialtrove*. [18] Using 31 drug features and 113 clinical trial attributes, they used a range of ML methods to generate success predictions with an average ROC-AUC of 0.78.

Subsequent attempts to use ML on text-based drug and trial data include the deep neural network (HINT) trained by Fu, et al. (mean ROC-AUC of 0.648) and work by Reinisch, et al., who used a GPT-3.5 enabled model with a ROC-AUC of 0.667, as well as Zheng, et al. (mean ROC-AUC of 0.678). [19][20][21] Aliper, et al.'s multimodal inClinico prediction platform stands out, with a ROC-AUC of 0.88). [22] This model leveraged a manually-curated ChEMBL database for molecular representation with a state-of-the-art Drug and Disease Interpretation Learning with Biomedical Entity Representation Transformer (DILBERT).

Despite the progress made in this space, many of these models are quite architecturally complex, and there has been no work which relied solely on the use of free, publicly available clinical trial databases (such as ClinicalTrials.gov) for training and test data. In this work, we describe the development of such an approach. We explored several popular, interpretable machine learning models including logistic regression, random forest, and XG boost to determine the pTRS and outcomes of clinical trials. After modeling, we compared each model across different metrics, emphasizing Brier score, ROC-AUC, PR-AUC, and our own outperforming metric. Additionally, we highlighted the most impactful features for our models for further analysis.

## Methods

### Data

After obtaining these two data sources, we preprocessed and curated the data for model training. Beginning with the clinicaltrials.gov dataset, we retrieved the full dataset from the API endpoint. The data consists of five sections: protocol, results, annotation, document, and derived. Our feature space consisted of only features from the protocol section. Within the protocol section, the data is further split into 8 modules describing the study details:

- Identification: Provides information about the identification of a clinical trial including IDs, titles, study types
- Sponsor collaborators: Provides information about primary individual, party, or collaborator sponsoring the clinical study
- Description: Provides context and narrative to clinical trial including summaries and more technical details
- Conditions: List of diseases, conditions, and other key words related to the clinical study
- Design: Provides information about how the clinical trials are set up, conducted, and evaluated
- Arm Interventions: Provides information on study arms, predefined participant groups, and interventions associated with each group
- Outcomes: Provides the outcomes or results of the clinical study through descriptions
- Eligibility: Defines the criteria that participant must meet

Across these 8 modules, there are categorical, numeric, boolean, and text features; not every feature across the 8 modules was incorporated for modeling. The specific features utilized in our analysis are detailed in Table 1.

## **Data Acquisition**

The dataset was obtained from two sources: clinicaltrials.gov for the feature space and IQVIA for the outcome labels.

Clinicaltrials.gov is a publicly accessible database containing information on clinical trials and results. The data can be retrieved on <https://clinicaltrials.gov/data-api/api>, using the API endpoint <https://clinicaltrials.gov/api/v2/studies/> and has a total of 438897 clinical trials as of January 19th, 2026.

The outcome labels were retrieved from the clinical-trial-outcome-prediction GitHub repository, associated with the research paper “HINT: Hierarchical Interaction Network for Clinical Trial Outcome Predictions” [19]. As noted in the paper, the labels were provided by an internal IQVIA team. The outcomes are binary, with 0 and 1 indicating failure and success, respectively. The data can be retrieved using a publicly hosted CSV file ([https://github.com/futianfan/clinical-trial-outcome-prediction/blob/main/data/raw\\_data.csv](https://github.com/futianfan/clinical-trial-outcome-prediction/blob/main/data/raw_data.csv)). The clinical trial labels included only small-molecule drugs and only contains labels/outcomes up to 2021. There is a total of 17615 benchmarks within this dataset.

## **Cleaning**

Cleaning was necessary because the clinicaltrials.gov data is heterogeneous and inconsistent, making it difficult to extract reliable features and accurately map disease categories. Standardization ensures that data can be effectively used for our modeling and classification tasks.

### ***Non-text Features***

After retrieving the clinicaltrials.gov data through the API, the protocol section of each trial gets extracted and appended to a list of JSONs. The list of JSONs (dictionaries) are converted to a data frame using the pandas package. This was followed by standard data cleaning practices including removal of duplicate trials, dropping rows where features are N/A, and setting the NCT IDs as unique indices.

Further cleaning was performed on individual columns as needed. The following contains details of individual cleaning:

- Conversion of string values to numerical values, as seen with the minimumAge column. For example, the string “18 years” becomes 18.
- In the interventionModel column, flattens the list to the first intervention type when possible.

- For multi-categorical columns, such as conditions, each contains a list of categories or labels. The lists were converted to a string with each element of the list being separated by '|'.

For the labels, we followed a similar procedure above with duplications and resetting the indices. The two datasets were merged on NCT IDs and were saved as a csv file called **base\_features.csv**. **base\_feature.csv** has the following features:

| Feature Type      | Feature Name                                  | Description                                                                                     |
|-------------------|-----------------------------------------------|-------------------------------------------------------------------------------------------------|
| Date              | statusModule_startDateStruct_date             | The actual date the first participant enrolled or the estimated date when enrollment will begin |
|                   | statusModule_startDateStruct_type             | Actual or estimated start date                                                                  |
|                   | statusModule_primaryCompletionDateStruct_date | The actual date where the last participant completed the trial or the estimated end date        |
|                   | statusModule_primaryCompletionDateStruct_type | Actual or estimated completion date                                                             |
| Numeric           | eligibilityModule_minimumAge                  | The minimum age a participant must be to be eligible                                            |
| Boolean           | eligibilityModule_healthyVolunteers           | Participants who do not have the condition are eligible to participate in the trial             |
| Categorical       | sponsorCollaboratorsModule_leadSponsor_class  | Type of sponsor that supported this trial (ex. NIH, industry)                                   |
|                   | designModule_designInfo_allocation            | Method of assigning arms (ex. randomized, non-randomized)                                       |
|                   | designModule_designInfo_interventionModel     | Method of assigning interventions (ex. single group, parallel)                                  |
|                   | designModule_designInfo_primaryPurpose        | Purpose of intervention (ex. treatment, prevention)                                             |
|                   | eligibilityModule_sex                         | The sex of the participant to be eligible to participate                                        |
|                   | armsInterventionsModule_interventions         | Intervention for each arm                                                                       |
| Multi-categorical | designModule_phases                           | The study phase for the clinical trial                                                          |
|                   | designModule_designInfo_maskingInfo_whoMasked | Groups of a clinical trial that are not allowed to have information regarding interventions     |
|                   | eligibilityModule_stdAges                     | Age group of the study                                                                          |

**Table 1.** Detailed information about feature space

### **Text Features**

There are two text features we primarily focused on: conditionsModule\_conditions and conditionsModule\_keywords. The columns contain a list of conditions/keywords (ex. Congenital

Adrenal Hyperplasia, Ocular Hypertension, etc.) for each clinical study. These two columns were combined into a single column with one list of all the descriptions.

We were also given a list of the top 21 most common disease categories; these categories are from the benchmark pTRS table that Biogen references. These categories are as follows: 'Allergy', 'Autoimmune/immunology', 'Cardiovascular', 'Dermatology', 'Endocrine', 'ENT/Dental', 'Gastroenterology (Non Inflammatory Bowel Disease)', 'Hematology', 'Infectious Disease', 'Metabolic', 'Neurology', 'Obstetrics/Gynecology', 'Oncology', 'Ophthalmology', 'Orthopedics', 'Psychiatry', 'Renal', 'Respiratory', 'Rheumatology (Non Autoimmune)', 'Urology'.

Given our new preprocessed column and disease category list, we used a GPT-4o API to classify each clinical study into one or more disease categories. We used (few-shot prompting) prompt engineering and batch processing to complete this task.

For each clinical trial, we combined the list of disease categories, the conditions from two text columns, and an example output to develop the prompt. The example output followed this structure: {"NCT111": ["Cardiovascular", "Metabolic"], "NCT112": ["Respiratory"], "NCT113": ["Neurology", "Autoimmune/immunology"]}.

Using multi-threading, the conditions were classified in batches (n = 50), allowing us to utilize the maximum token capacity and accelerate the process of mapping. This information was saved into **mapped\_categories.json**.

### ***One-hot encodings***

For categorical, multi-categorical, and conditions columns, we performed one-hot encodings to convert categories into numeric values. These features have been further split into new columns; each new column represents a category from the feature. These columns contain either 0 or 1. For example, sponsorCollaboratorsModule\_leadSponsor\_class feature has these unique values: 'NIH', 'OTHER', 'INDUSTRY', 'NETWORK', 'FED', 'OTHER\_GOV', 'INDIV', and 'UNKNOWN'. For each value, a column is created each unique value with a title such as 'sponsorCollaboratorsModule\_leadSponsor\_class\_nih'.

### ***Data Cleaning Workflow***

The data cleaning and transformation workflow is summarized in Figure 1, outlining the sequential reduction and processing of clinical trial records from raw download through cleaning, encoding, harmonization, and final dataset preparation for analysis.

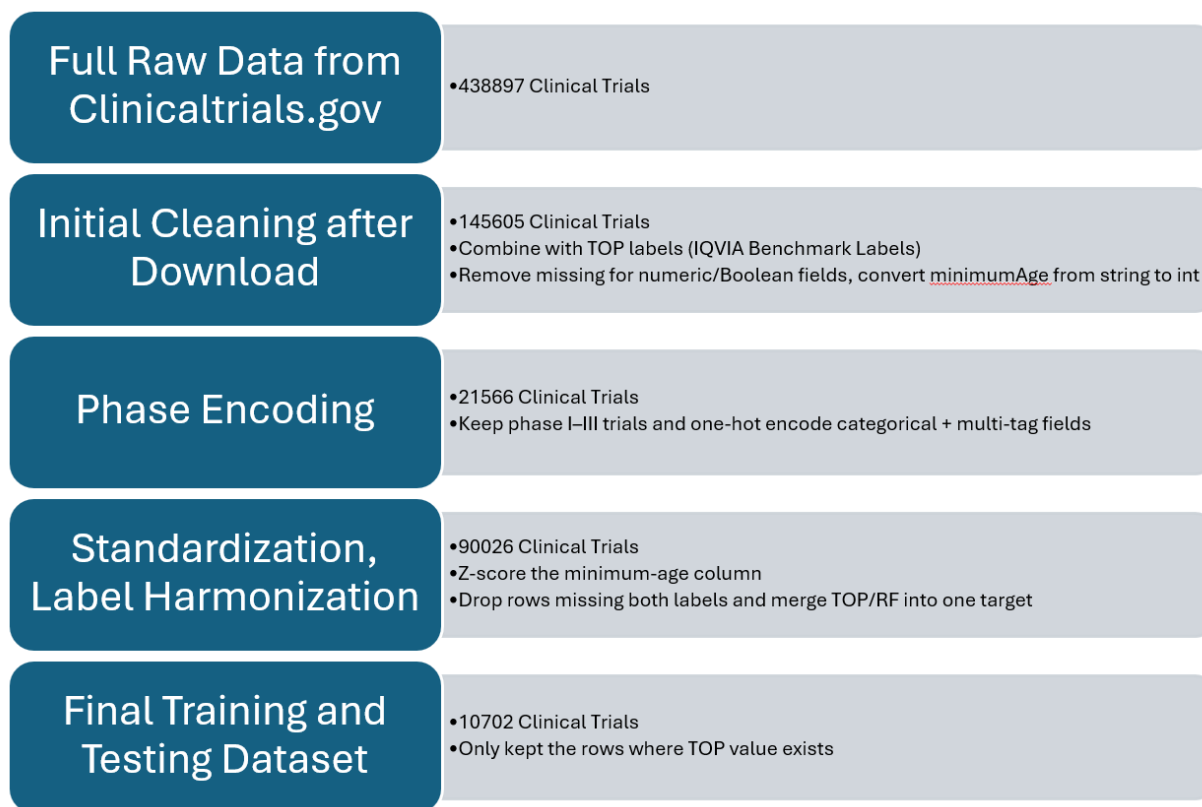


**Figure 1.** Clinical Trials Data Preprocessing and Selection Pipeline

### **Descriptive Statistics and Exploratory Data Analysis**

The complete merged dataset comprises 6,262 successful clinical trials and 4,440 unsuccessful trials, based on the label distribution.

Breaking down the dataset by clinical trial phase reveals distinct success patterns. Phase 1 trials achieve a 61.4% success rate (712 successes, 448 failures), Phase 2 trials are nearly evenly split at 49.2% success (2,577 successes, 2,662 failures), and Phase 3 trials demonstrate the highest success rate at 69.1% (2,973 successes, 1,330 failures).

Breaking down the dataset by training and testing splits, the training set comprises 6,007 successes (58.8%) and 4,217 failures (41.2%), while the testing set contains 255 successes (53.3%) and 223 failures (46.7%), as shown below.

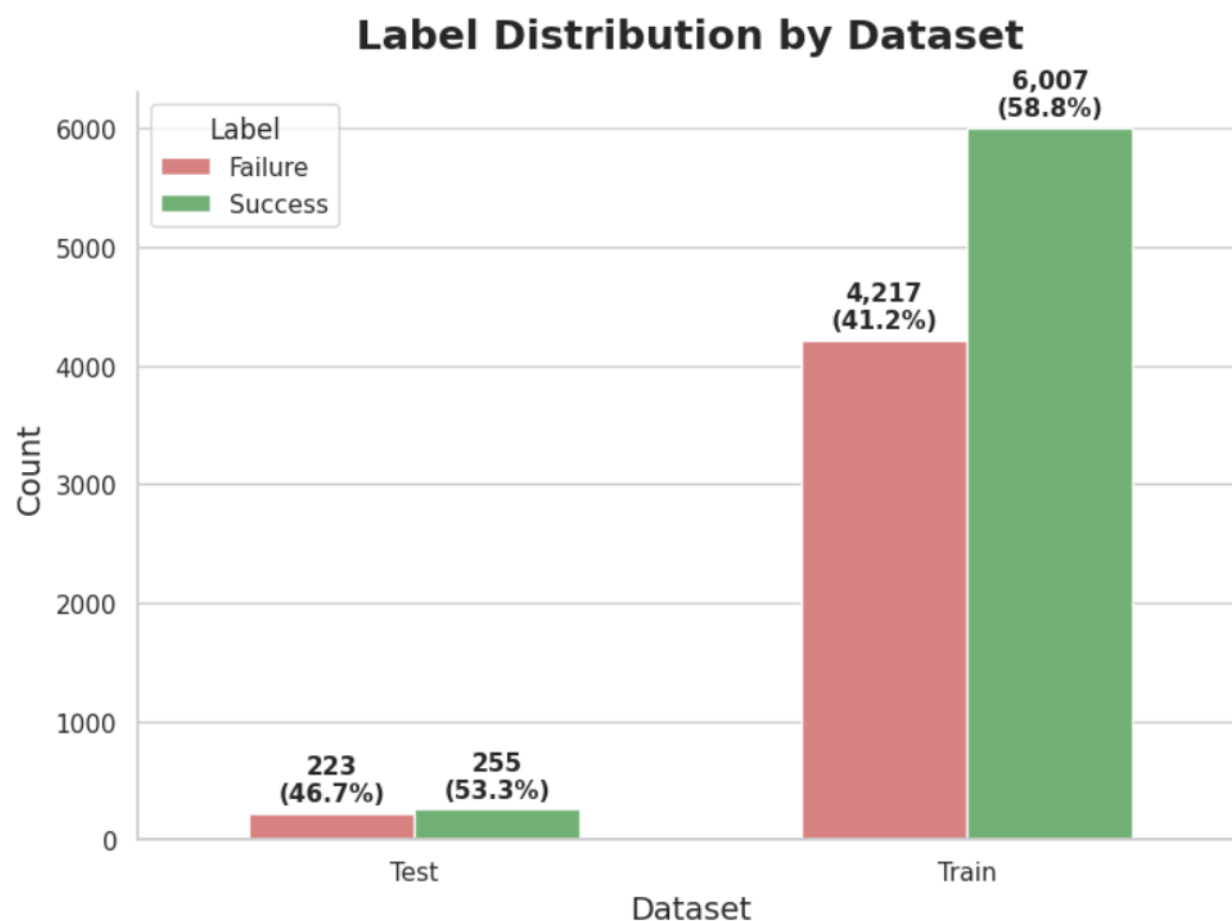


**Figure 2.** Distribution of success and failure for training and testing sets

## **Model/Training/GitHub**

### ***Training/Testing sets***

The training and testing were split based on the start and completion dates of the clinical trials. The training set included any clinical trial that had a completion date prior to January 1<sup>st</sup>, 2017. The testing set included any clinical trial that had a start date after January 1<sup>st</sup>, 2017. After splitting, there were 10224 clinical trials in training and 478 clinical trials in testing.

### ***Models***

We selected logistic regression as our baseline model because it provides an interpretable and widely used reference for binary classification. We also evaluated the random forest and XGBoost model, which returned predicted probabilities and used 0.5 as the threshold for classification.

For the latter two models, we performed hyperparameter tuning through random search. Random Forest and XGBoost models were implemented using the scikit-learn and XGBoost Python libraries, respectively. [23] [24]

During modeling, 20 different hyperparameter combinations were randomly sampled for each algorithm, and for each combination, the training data was split into 5-folds for cross validation, resulting in a total of  $20 \times 5 = 100$  model fits per algorithm. The scoring metric for the folds was ROC AUC.

The training and testing datasets were saved as pickle files once modeling was completed.

### ***Class weighting***

As shown in our exploratory data analysis, the classes are imbalanced, with more successes than failures. To ensure that the models pay equal attention to both classes and penalize misclassifying failures more severely, we used class weights. The weights were computed inversely proportional to the class frequencies using `compute_class_weight`. For the case of XGBoost, a similar parameter was computed using the ratio between failure and success classes.

### ***GitHub***

The code for this analysis is available upon request.

## **Evaluation**

Model performance was assessed using a combination of discrimination, calibration, benchmark comparison, and interpretability metrics.

### ***Discrimination Metrics***

Discriminative ability was measured using the area under the receiver operating characteristic curve (ROC-AUC) and the area under the precision-recall curve (PR-AUC). ROC-AUC quantified the model's capability to distinguish between successful and failed trials across thresholds, while PR-AUC provided a complementary assessment in the presence of class imbalance. Higher values on both metrics correspond to strong discriminative performance.

### ***Brier Score***

Calibration and probabilistic accuracy were evaluated using the Brier score, which measures the squared difference between what the model predicted (a probability) and what actually happened (0 or 1). The score was decomposed into reliability (calibration), resolution (discrimination across outcome groups), and uncertainty (outcome variability). Scores were computed for both model predictions and benchmark PTRS values, with lower values indicating better calibration.

$$\text{Brier Score} = \text{Reliability} - \text{Resolution} + \text{Uncertainty}$$

This decomposition provides insight into whether improvements in overall accuracy stem from better calibration, stronger discrimination, or differences in outcome variability.

### ***Collecting benchmark data***

The outperformance metric requires the predicted values to be compared against the benchmark. However, when using the external industry benchmarks for each disease category, our model always outperforms successful trials, while always underperforms failed trials. Therefore, we created our own benchmarks using the training data to derive meaningful insights.

1. The data was looped over each disease category
2. For each disease category, compute the mean of the outcome labels (0, 1) for each phase.
3. Store into dictionary as {disease: {phase1: X, phase2: Y, phase3: Z}}

### ***Outperformance Metric***

To quantify performance relative to benchmark pTRS values, an **Outperformance rate** metric was defined to capture trial-level deviations between the model's predicted probability of success and an established benchmark.

$$\text{Outperform}_i = \begin{cases} 1, & \text{if } |\hat{p}_i - y_i| < |p_i^{(\text{bench})} - y_i|, \\ 0, & \text{otherwise,} \end{cases}$$

where  $y_i \in \{0,1\}$  is the true outcome (0=failure, 1=success),  $\hat{p}_i$  is our predicted success probability, and  $p_i^{(\text{bench})}$  is the benchmark probability.

We then take the overall rate across  $n$  cases:

$$\text{Outperformance Rate} = \frac{1}{n} \sum_{i=1}^n \text{Outperform}_i .$$

### ***Interpretability***

Model transparency was evaluated using SHapley Additive exPlanations (SHAP). SHAP values attribute deviations from the model's mean prediction to individual features, allowing for consistent decomposition of feature contributions. Global interpretability was obtained by ranking predictors according to mean absolute SHAP values.

## Results

Random Forest was the training method that was ultimately selected for the model. When comparing ROC-AUC, PR-AUC, and Brier Score, the Random Forest model outperformed the baseline model on all three metrics for almost all phases. The only metric that the baseline model outperformed the Random Forest model on was the PR-AUC score for Phase 3 trials.

| Phase      | Random Forest |              |              | Baseline |              |             |
|------------|---------------|--------------|--------------|----------|--------------|-------------|
|            | ROC-AUC       | PR-AUC       | Brier Score  | ROC-AUC  | PR-AUC       | Brier Score |
| Phase 1    | <b>0.771</b>  | <b>0.627</b> | <b>0.244</b> | 0.719    | 0.593        | 0.279       |
| Phase 2    | <b>0.782</b>  | <b>0.755</b> | <b>0.202</b> | 0.710    | 0.594        | 0.229       |
| Phase 3    | <b>0.729</b>  | 0.867        | <b>0.165</b> | 0.707    | <b>0.873</b> | 0.177       |
| All Phases | <b>0.788</b>  | <b>0.794</b> | <b>0.196</b> | 0.726    | 0.748        | 0.220       |

**Table 2.** Random Forest and Baseline Comparison

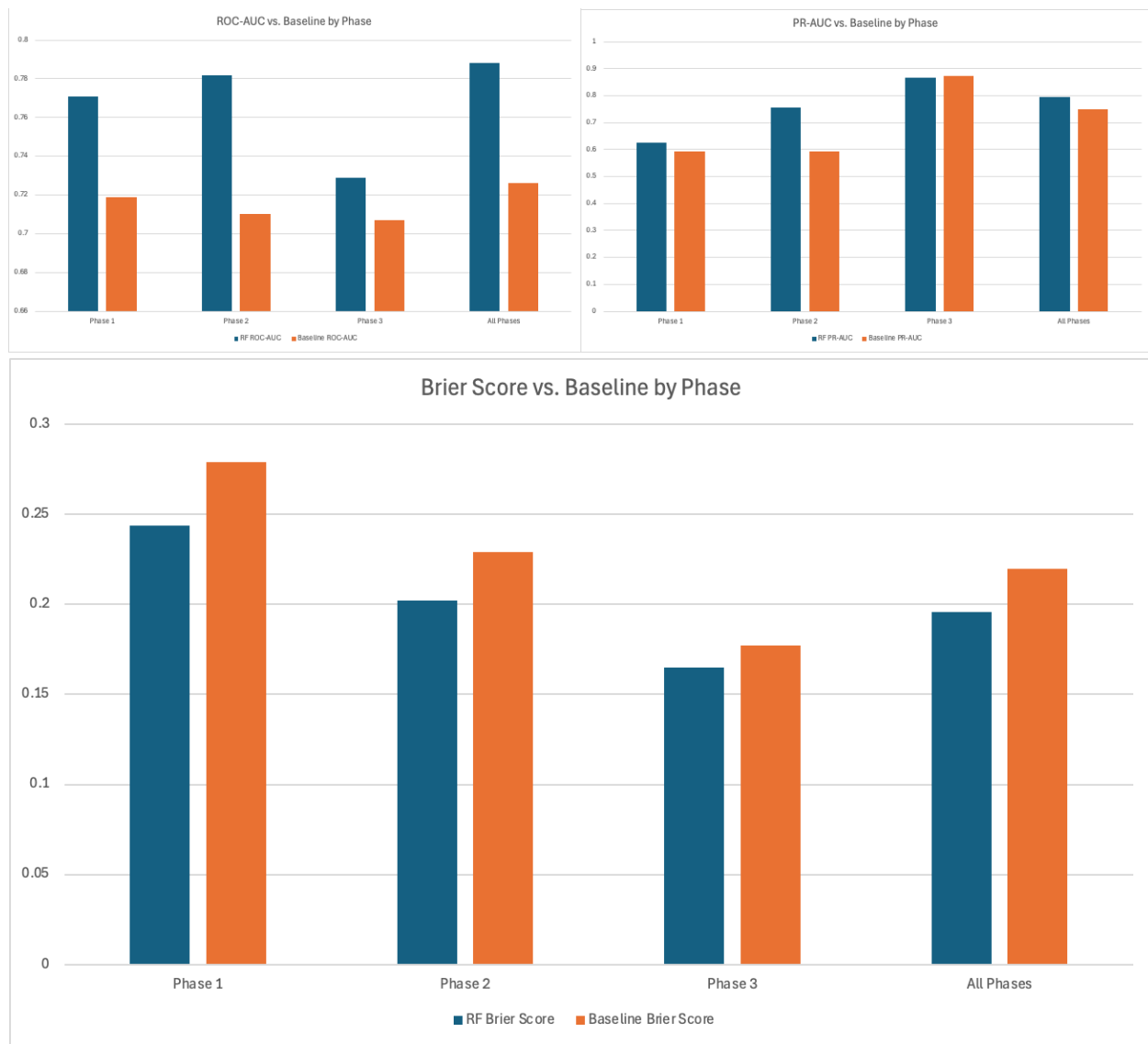


**Figure 3.** Model performance metrics (ROC-AUC, PR-AUC, and Brier score) by clinical trial phase for RF and baseline models

Year gaps were also implemented in training the Random Forest model. These modified datasets would feature a specified gap value between the years used for training and test data, with gap values of 0, 1, 3, and 5 years. While performance was relatively stable across models of different gap values, small yet consistent declines in performance were seen as the value of the time gap was increased, as displayed in the table below.

| Model         | Grouped Gap (Year) | ROC-AUC      | PR-AUC       | Brier Score (Predicted) | Brier Score (Benchmark) |
|---------------|--------------------|--------------|--------------|-------------------------|-------------------------|
| Random Forest | 0                  | <b>0.755</b> | <b>0.804</b> | <b>0.199</b>            | 0.215                   |
|               | 1                  | 0.753        | 0.802        | <b>0.199</b>            | 0.215                   |
|               | 3                  | 0.746        | 0.793        | 0.201                   | 0.215                   |
|               | 5                  | 0.748        | 0.795        | 0.201                   | 0.215                   |

**Table 3.** Model performance by grouped gap (years between training and test data). Brier Score (Benchmark) represents the training-derived benchmark from the no-gap split.

Ultimately, the model without gaps produced the strongest results, with a ROC-AUC score of 0.788, PR-AUC score of 0.794, and Brier Score of 0.196. These metrics as well as those of other models are displayed below.

### **Model Interpretability via SHAP Analysis**

As shown in Figure 4, the SHAP summary plot presents the top 15 most influential features driving model predictions across all 478 clinical trials. Each point represents a single trial, with color indicating the magnitude of the corresponding feature value (red = high, blue = low).

Industry sponsorship clearly dominates as the strongest positive driver of predicted trial success, while oncology and non-industry sponsorship contribute negatively. Phase 3 and metabolic indications are associated with higher predicted success, whereas Phase 2 trials exhibit modest negative contribution—consistent with the known attrition typical of mid-stage development.

Design-related factors such as randomized allocation, participant masking, and single-group assignments show moderate yet directionally consistent effects, suggesting that increased trial complexity may temper predicted outcomes. Clinical domain variables including endocrine, infectious disease, and metabolic indications contribute positively, reflecting historically higher success rates in these therapeutic areas.

Overall, these interpretability results underscore the central role of sponsor type, therapeutic domain, trial phase, and study design in shaping model behavior. While these associations align with established clinical-development trends, they should be interpreted as correlational rather than causal.

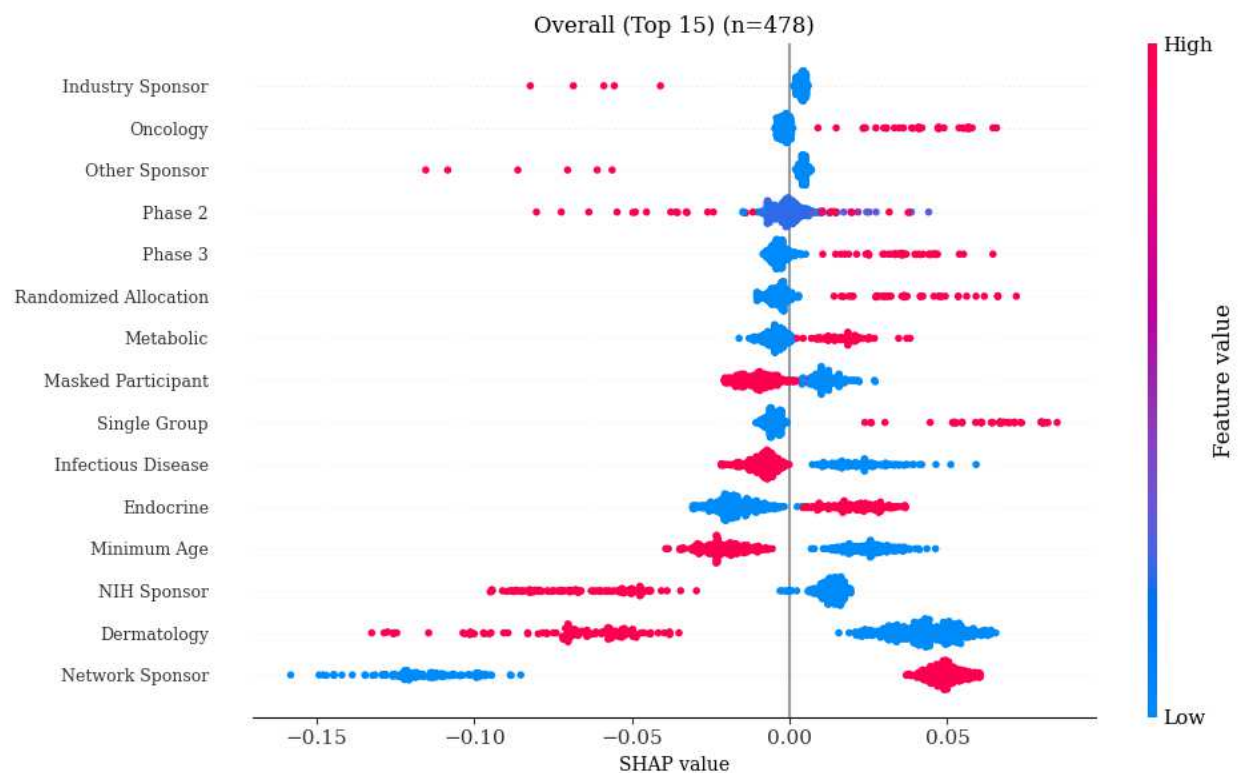


**Figure 4.** Random Forest SHAP Overall Feature Importance.  
*The plot depicts SHAP values for individual trials, colored by feature magnitude (red = high, blue = low) for top 15 features (n=478).*

# Discussion

This study demonstrates that decision-grade predictions of clinical-trial success can be produced using interpretable models trained solely on publicly available features from ClinicalTrials.gov and open labels. A Random Forest (RF) classifier improved discrimination and calibration relative to a baseline across most settings (aggregate ROC–AUC 0.788 vs. 0.726; PR–AUC 0.794 vs. 0.748; Brier 0.196 vs. 0.220), with only a slight PR–AUC shortfall in Phase 3. Performance decayed modestly as the temporal gap between training and testing widened, yet performance metrics remained above benchmark values even at five-year gaps. In the pharmaceutical setting, these findings may enable risk assessment for clinical trials based on historical data, supporting a more objective and data-driven foundation for decision-making.

These findings demonstrate that well-calibrated, competitive pTRS estimation is feasible using only public features. Our contribution is primarily practical: we show that calibrated, competitive pTRS estimates can be obtained without proprietary datasets or opaque architectures. Many recent approaches [19, 22] rely on restricted data or complex multimodal pipelines that impede replication. In contrast, we pair open features with training-derived phase-by-therapeutic-area benchmarks to create a fair and reproducible evaluation.

Interpretation of the model outputs focuses on identifying sources of incremental signal beyond what is already encoded in the benchmark, with an emphasis on understanding why certain features meaningfully influence predicted trial success rates. Both the feature importance analysis and the SHAP (SHapley Additive exPlanations) results indicate that sponsor characteristics—particularly whether a trial is industry- or non-industry-sponsored—and study-design variables such as allocation method, masking, and intervention model contribute substantially to predictions. In the model, industry-sponsored trials tend to show higher predicted probabilities of technical and regulatory success (pTRS) across development phases. This pattern should be interpreted as an association learned from the data rather than as evidence that sponsorship itself causally determines trial success. One possible explanation is that sponsor type may proxy for broader differences in program characteristics, study objectives, or development context that are also relevant to pTRS. In particular, non-industry-sponsored trials may be conducted for purposes other than advancing a program toward regulatory approval or commercial de-risking, such as exploratory research, methodological investigation, or public-health-oriented questions. As a result, lower predicted pTRS for these trials may reflect differences in trial purpose and composition within the sample, rather than an inherent disadvantage associated with non-industry sponsorship itself.

Therapeutic area also emerged as a prominent driver of model predictions. In particular, oncology programs were associated with a pTRS lower than average, while infectious disease and metabolic disorder trials tended to have higher pTRS. This finding aligns with historical

success rates reported in the literature; for instance, oncology drugs have the lowest likelihood of approval starting from Phase 1, with an estimated success rate of only 6.7%. [25] The biological complexity of cancer, coupled with high heterogeneity in disease subtypes, tumor microenvironments, and patient responses, makes demonstrating efficacy especially challenging. In addition, oncology programs often contend with stringent regulatory evidentiary standards and rapidly evolving competitive landscapes, factors that may be associated with a higher risk of failure. Conversely, infectious disease and metabolic disorder programs may benefit from more established mechanistic understanding and clearer clinical endpoints, consistent with the higher predicted and observed success rates seen in our analysis.

Phase of development was another influential predictor. The model reflected the expected pattern in which Phase 2 status programs have lower pTRS relative to other stages, while Phase 3 status programs tend to have higher pTRS. This is consistent with the role of Phase 2 as a proof-of-concept stage, where uncertainty remains high and many programs fail due to insufficient efficacy signals or emerging safety concerns. In contrast, Phase 3 trials have already passed these critical efficacy and safety gates, meaning the programs have been partially de-risked and are more likely to achieve their primary endpoints, provided operational execution and regulatory strategy remain sound.

Taken together, these findings suggest that the model is capturing meaningful and interpretable relationships between trial characteristics and success probabilities. Features such as sponsor type, therapeutic area, and phase not only contribute to predictive accuracy but also align with domain knowledge and historical trends in drug development, reinforcing confidence that the model's incremental signal beyond the benchmark is grounded in real-world development dynamics.

The model's stability across temporal splits suggests that recency limitations—while present—do not erase decision value over multi-year horizons. With the average lifecycle of a drug from its initial stages to regulatory approval taking anywhere from 10-15 years, the 'gap' analysis provides a realistic forecasting scenario where the model must be trained on historical data to predict future trial outcomes. In clinical development, the 'gap' reflects a challenge that leadership faces regularly when making decisions about investments for the portfolio. Performance remained broadly consistent across all temporal gap settings, indicating that the predictive relationships captured by the model are stable over multi-year horizons. However, the improvement over benchmarks in Brier score (mean squared error) declined by approximately 12.5% when extending from a 0 to 5 year gap, suggesting that the model's performance degradation is modest but measurable over longer timeframes. This modest decline may be attributable, at least in part, to the reduced volume of training data available in those scenarios. Despite this decline, the overall stability enhances confidence that the model is capturing signal that persists over multi-year forecasting horizons.

The approach is most applicable to Phase 1–3 interventional trials represented on ClinicalTrials.gov and to mixed-sponsor portfolios. In practice, organizations can use these probabilities as priors in pTRS selection, refine them with program-specific evidence during expert elicitation, and surface them directly in valuation and resource-allocation workflows. Routine monitoring of calibration within the strata used for benchmarking, plus light recalibration to local base rates, should be sufficient for operational use.

Several limitations merit emphasis. First, label coverage and recency are constrained (~17k labeled trials, latest through ~2021); shifts in trial conduct or regulatory expectations since then may degrade forward performance. Second, we did not exploit rich free text (e.g., detailed descriptions, endpoints), molecular properties, or PK/PD/tox information—omissions that likely cap performance, particularly in oncology and first-in-class settings. Within the ClinicalTrials.gov dataset, fields such as “Primary Outcomes” and “Description” have the potential to enrich the feature set by providing additional clinical context and details regarding trial conduct and outcomes. Natural language processing (NLP) techniques, including word embedding models and information extraction methods, could be applied to free-text fields to derive structured representations and clinically relevant features. Third, we did not isolate industry-sponsored trials for evaluation; if sponsor-mix differs materially from a given portfolio, subgroup calibration may vary. Fourth, therapeutic-area mapping relied on LLM-assisted categorization, which can introduce structured misclassification that affects the benchmark strata. Finally, because benchmarks were derived from training data, future work should complement our “outperforming” metric with additional external benchmarks where feasible.

Our group was aware of tradeoffs due to the limitations previously mentioned. However, in the context of portfolio analytics and decision-making for leadership, interpretability was a key component. For example, for the HINT model, features such as SMILES (simplified molecular input line entry system) were incorporated into their models, which could potentially have better performances. [19] Neural networks and variants of neural networks are popular approaches in prior research. However, in pTRS selection sessions with non-technical stakeholders, being able to explain the models and results in plain language expedites the decision-making process with more confidence. Ultimately, this facilitates more informed decision-making, with our results being another piece of information alongside benchmarks and expert elicitation. Even though our results are more interpretable, the other limitations could raise concerns about performance. Our model outperformed benchmarks around 60% of the time, with predicted probabilities 4 percentage points lower than benchmarks for failures and 1.3 percentage points higher for successes on average. Without context, this result does not look satisfying. However, in the context of clinical trials and decision-making, even small improvements to accuracy are meaningful. Clinical trials are influenced by numerous complex and out-of-control variables. Although the absolute improvements in probabilistic accuracy appear modest, they matter at portfolio scale. Small gains in calibration and rank-ordering propagate through value calculations and Go/No-Go decisions, improving capital deployment.

Strengths of the approach described here include full reproducibility with public inputs, emphasis on calibration (Brier and its decomposition) alongside discrimination, stakeholder-transparent explanations, and a benchmarking strategy designed to symmetrically assess both successes and failures. The reproducibility of the model allows easier integration into current decision frameworks and allows rapid assessment of a particular asset's risk profile using readily available public data. When benchmarks are generated from proprietary or organization-specific data, the resulting model is more closely aligned with the asset portfolio under evaluation. This alignment enables focused performance tracking within priority therapeutic areas and supports strategic decisions on where to deepen investment or reduce exposure.

In sum, interpretable models trained on open clinical-trial features can yield calibrated, decision-relevant pTRS estimates that outperform standard benchmarks, offering a practical foundation for valuation and portfolio decisions while leaving clear headroom for improvement through fresher labels and richer features.

## **Conclusions and Future Directions**

We have shown that interpretable machine learning models trained on publicly available ClinicalTrials.gov data can produce calibrated pTRS assessments that outperform traditional phase-by-therapeutic-area benchmarks. Our Random Forest classifier achieved an ROC-AUC of 0.788 and Brier score of 0.196 on held-out test data, representing meaningful improvements over benchmark performance across most evaluation settings. The model's stability across temporal gaps of up to five years and its reliance solely on accessible features make it practical for deployment in real-world portfolio decision-making. Feature importance analysis revealed that sponsor type, therapeutic area, and trial phase drive predictions in ways that align with established domain knowledge, providing stakeholders with interpretable risk assessments that can complement expert judgment in pTRS estimation.

The primary value of this work lies not in achieving state-of-the-art performance but in demonstrating that decision-useful risk quantification is achievable without proprietary data or architectural complexity. Performance improvements of 3-4 percentage points in probabilistic accuracy may appear modest in isolation, yet they compound meaningfully across portfolio-scale asset evaluations where capital deployment decisions hinge on marginal differences in expected value.

Future work could build upon the methodology described here by identifying additional labeled datasets that could be used to expand the training data for the model. Similar approaches could also be pursued using large language models to utilize the context of trial summaries and glean more context than can be had from categorizing the input fields. LLMs such as BioBERT are trained on biomedical text and can significantly enhance the predictive power of probabilistic models. [26] Finally, there are non-subscription catalogs of chemical structures that could be

used to augment this model with structural-analysis based inference to predict toxicological or pharmacokinetic properties that can affect drug candidate success.

# References

- [1] Alacrita Consulting. (2023). Valuing pharmaceutical assets: When to use NPV vs rNPV.
- [2] BiopharmaVantage. (2025). 2025 ultimate pharma & biotech valuation guide  
<https://www.biopharmavantage.com/pharma-biotech-valuation-best-practices>
- [3] Jekunen A. Decision-making in product portfolios of pharmaceutical research and development—managing streams of innovation in highly regulated markets. *Drug Des Devel Ther.* 2014;8:2009–2016. <https://doi.org/10.2147/DDDT.S68579>
- [4] Gouglas D, Marsh K. Prioritizing investments in rapid response vaccine technologies for emerging infections: a portfolio decision analysis. *PLoS One.* 2021;16(2):e0246235.  
<https://doi.org/10.1371/journal.pone.0246235>
- [5] Vergetis V, Skaltsas D, Gorgoulis VG, Tsirigos A. Assessing drug development risk using big data and machine learning. *Cancer Res.* 2021;81(4):816–819. <https://doi.org/10.1158/0008-5472.CAN-20-0866>
- [6] Sendyona S, Odeyemi I, Maman K. Perceptions and factors affecting pharmaceutical market access: results from a literature review and survey of stakeholders in different settings. *J Mark Access Health Policy.* 2016;4:31660. <https://doi.org/10.3402/jmahp.v4.31660>
- [7] Tirrell Z, Lybrand S, Soderlund N, Parkinson B. Accessing the risks: a mixed methods study of risk management in Australian pharmaceutical market access. *Health Policy.* 2026;168:105608. <https://doi.org/10.1016/j.healthpol.2026.105608>
- [8] Manders EA, van den Berg S, de Visser SJ, Hollak CEM. Drug pricing models, no ‘one-size-fits-all’ approach: a systematic review and critical evaluation of pricing models in an evolving pharmaceutical landscape. *Eur J Health Econ.* 2025;26:683–696. <https://doi.org/10.1007/s10198-024-01731-w>
- [9] Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat Rev Drug Discov.* 2012;11(3):191–200.  
<https://doi.org/10.1038/nrd3681>
- [10] Fernald KDS, Förster PC, Claassen E, van de Burgwal LHM. The pharmaceutical productivity gap – incremental decline in R&D efficiency despite transient improvements. *Drug Discov Today.* 2024;29(11):104160. <https://doi.org/10.1016/j.drudis.2024.104160>
- [11] Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL. How to improve R&D productivity: the pharmaceutical industry’s grand challenge. *Nat Rev Drug Discov.* 2010;9(3):203–214. <https://doi.org/10.1038/nrd3078>

- [12] Oreskovich A-M, Gittins J. The evaluation of risk in pharmaceutical research: a study of current models and techniques. *R&D Manag.* 2016;46(5):900–913.  
<https://doi.org/10.1111/radm.12200>
- [13] Schuhmacher A, Hinder M, Brief E, Gassmann O, Hartl D. Benchmarking R&D success rates of leading pharmaceutical companies: an empirical analysis of FDA approvals (2006–2022). *Drug Discov Today.* 2025;30(2):104291. <https://doi.org/10.1016/j.drudis.2025.104291>
- [14] Dallow N, Best N, Montague TH. Better decision making in drug development through adoption of formal prior elicitation. *Pharm Stat.* 2018;17(4):301–316.  
<https://doi.org/10.1002/pst.1854>
- [15] Hampson LV, Holzhauer B, Bornkamp B, Kahn J, Lange MR, Luo WL, Singh P, Ballerstedt S, Della Cioppa G. A new comprehensive approach to assess the probability of success of development programs before pivotal trials. *Clin Pharmacol Ther.* 2022;111(5):1050–1060.  
<https://doi.org/10.1002/cpt.2488>
- [16] Gayvert KM, Madhukar NS, Elemento O. A data-driven approach to predicting successes and failures of clinical trials. *Cell Chem Biol.* 2016;23(10):1294–1301.  
<https://doi.org/10.1016/j.chembiol.2016.07.023>
- [17] Nahm FS. Receiver operating characteristic curve: overview and practical use for clinicians. *Korean J Anesthesiol.* 2022;75(1):25–36. <https://doi.org/10.4097/kja.21209>
- [18] Lo AW, Siah KW, Wong CH. Machine learning with statistical imputation for predicting drug approvals. *SSRN Electron. J.* 2018. <https://doi.org/10.2139/ssrn.2973611>
- [19] Fu T, Huang K, Xiao C, Glass LM, Sun J. HINT: Hierarchical interaction network for clinical-trial-outcome predictions. *Patterns (N Y).* 2022;3(4):100445.  
<https://doi.org/10.1016/j.patter.2022.100445>
- [20] Reinisch M, He J, Liao C, Siddiqui SA, Xiao B. CTP-LLM: Clinical trial phase transition prediction using large language models. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2024.  
<https://doi.org/10.1109/BIBM62325.2024.10822746>
- [21] Zheng W, Peng D, Xu H, Li Y, Zhu H, Fu T, Yao H. LIFTED: multimodal mixture-of-experts for clinical trial outcome prediction. In: ICML 2024 FM-Wild Workshop. 2024. Available from: <https://openreview.net/forum?id=F1B1zTEAn5>
- [22] Aliper A, Kudrin R, Polykovskiy D, Kamya P, Tutubalina E, Chen S, Ren F, Zhavoronkov A. Prediction of clinical trials outcomes based on target choice and clinical trial design with multi-modal artificial intelligence. *Clin Pharmacol Ther.* 2023;114(5):972–980.  
<https://doi.org/10.1002/cpt.3008>

- [23] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://scikit-learn.org/>
- [24] XGBoost contributors. (2024). Get Started with XGBoost. XGBoost documentation. [https://xgboost.readthedocs.io/en/stable/get\\_started.html](https://xgboost.readthedocs.io/en/stable/get_started.html)
- [25] Hay, M., Thomas, D., Craighead, J. et al. Clinical development success rates for investigational drugs. *Nat Biotechnol* 32, 40–51 (2014). <https://doi.org/10.1038/nbt.2786>
- [26] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.1810.04805>