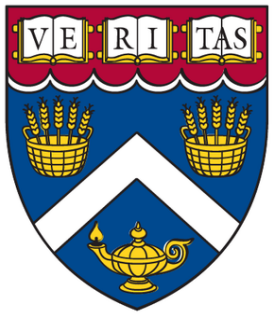
AI-Powered FDA Approval Prediction



João P. Diogo[1], Shaoyi Li[2], Shekhar Gupta[3], Mofei Wang[4], Gassan Yacteen[5]

A Capstone Project in the Field of Data Science

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

December 2024

[1] jod9730@g.harvard.edu
[2] lishaoyi89@gmail.com
[3] shg429@g.harvard.edu
[4] mow479@g.harvard.edu
[5] gus767@g.harvard.edu

Abstract

Evergrowth BioHealthcare Capital, a hedge fund specializing in small-cap biotech stocks, seeks to enhance its investment decision-making process by accurately predicting the likelihood of U.S. Food and Drug Administration (FDA) drug approvals. The current manual approach, reliant on subject matter experts sifting through vast amounts of unstructured textual data, is time-consuming, unscalable, and prone to human error. In this project we developed a data-driven predictive model that leverages advanced machine learning techniques and natural language processing (NLP) to forecast FDA drug approvals based on Phase 2 clinical trial data.

Our research addressed key challenges: integrating heterogeneous datasets, ensuring scalability for large data volumes, applying advanced analytical methods, mitigating data leakage and bias, and providing sufficient model interpretability. Publicly available data, primarily from Aggregate Analysis of ClinicalTrials.gov (AACT), and supplementary Cortellis Drug Discovery Intelligence Database records formed the basis of our analysis. We focused on clinical trial attributes and other relevant variables associated with approval outcomes.

Our methodology is structured into four stages: (1) Process the data so that we have datasets with acceptable quality for model fitting, (2) Use language models (LMs) to generate more features, (3) expand the features list and make the model more realistic by introducing the Trial Sequence concept, and (4) tune the model. Ethical considerations and data trustworthiness were prioritized throughout the project, ensuring adherence to legal guidelines and data integrity.

Our final models, when tested on held-out data, achieved up to 0.77 accuracy for a One-row-per-pair dataset and 0.71 accuracy for a more complex Trial Sequence dataset. These models identified the most useful features for predicting drug-development outcomes: patients enrolled, days since trial start, number of trial facilities, trial duration and number of secondary outcomes to measure. Although language model-driven features provided moderate improvements, their impact was constrained by available time and data quality. Future work will focus on integrating richer features, refining language model applications, and exploring advanced time-series modeling techniques. Ultimately, this research demonstrates that a systematic, data-driven approach can enhance the prediction of FDA drug approvals, improving investment decision-making and offering a more scalable, efficient process than traditional manual analyses.

Acknowledgements

We would like to extend our heartfelt gratitude to Dr. Bruce Huang, Director of the Harvard Extension School (HES) Master's Degree Program in Information Technology, for his invaluable guidance and structured approach. His wisdom on navigating datasets and shaping our analytical methodology greatly influenced the trajectory of this project.

We also want to thank Dr. Stephen F. Elston for his thoughtful suggestions on machine learning approaches and model sampling methods. His expertise and willingness to share knowledge significantly improved the rigor and quality of our analytical framework.

We also acknowledge Teaching Fellow Leonardo Neves for his guidance on large language models and natural language processing techniques. His technical insights helped us integrate cutting-edge tools and methods into our analysis, ultimately enhancing the project's depth and sophistication.

A special note of appreciation goes to Paulo Silva (ALM '23). His deep understanding of the biotech landscape, the hedge fund investment environment, and the nuances of the FDA approval process, coupled with his dedication and constructive feedback was instrumental. Our regular consultations with him provided not only technical insights but also strategic direction, consistently helping us refine our goals and strengthen our overall approach.

We also thank Grammarly and ChatGPT for their assistance with grammar and phrasing.

Finally, we acknowledge the entire Data Science Capstone cohort, whose constructive feedback, encouragement, and collaborative spirit helped us refine our work at every stage. We are immensely grateful for the collective support and expertise that made this project possible.

Chapter 1: Introduction

1.1 | Background

Evergrowth BioHealthcare Capital is a start-up hedge fund that invests in small-cap biotech stocks. They have a record of outperforming market benchmarks such as the Dow Jones Industrial Average, Nasdaq Biotechnology Index (NBI), and S&P 500. One important criterion of Evergrowth's stock investment process is predicting the likelihood of a company's new drug being approved by the U.S. Food and Drug Administration (FDA) . This represents a binary event pivotal for a biotech company, as it leads to significant fluctuations in its stock value. We aim to enhance their investment decision-making process by developing a predictive model that leverages advanced data analysis techniques to forecast FDA drug approvals.

1.2 | Problem Statement

The current process for determining the likelihood of drug approval or rejection relies heavily on subject matter experts (SMEs) manually sifting through vast amounts of unstructured textual data. This manual approach is time-consuming, unscalable, and prone to human error, potentially leading to gaps and delays in the assessment analysis and less informed investment decisions. There is a need for a systematic, data-driven approach that can efficiently analyze both structured and unstructured data to predict FDA drug approvals with higher accuracy.

1.3 | Sub-problems

Scalability presents a challenge, especially when working with large and diverse datasets from the biotech domain. Additionally, we must navigate the complexities of these fields while also incorporating expertise in finance and investment, which requires focused efforts to improve domain knowledge and skills. Furthermore, investing in small-cap biotech companies is inherently risky, influenced by volatile macroeconomic conditions, sector-specific challenges, and microeconomic variables. These factors show the importance of developing a robust risk management framework to ensure the final investment model can effectively mitigate potential losses and support sound decision-making.

1.3.1 | Challenges with Data Collection

One of the key challenges in this project was the extensive effort required to collect suitable data sources. Obtaining drug intelligence data and clinical trial data presented challenges, as manually

downloading data for thousands of trials from ClinicalTrials.gov was impractical. The commercial data source options such as Citeline Trialtrove and Cortellis Clinical Trials Intelligence were prohibitively expensive and beyond the budget constraints of this academic project. The publicly available AACT (Aggregate Analysis of ClinicalTrials.gov), a widely used industry-standard database recommended by industry experts we consulted, is substantial, comprising over 51 structured data tables with a total size exceeding 400 GB, and includes multiple schemas derived from ClinicalTrials.gov. A significant amount of time was spent reviewing, cleaning, and processing the AACT data, as well as integrating it with the Cortellis Drug dataset to ensure compatibility and usability for our analysis.

1.3.2 | Challenges with Data Preprocessing for Predictive Models

Integrating heterogeneous sources presented significant challenges due to inconsistencies, missing values, and varying levels of granularity. Ensuring quality and completeness was critical for building a reliable model. The volume of information demanded efficient processing techniques to maintain scalability and performance. Additionally, handling complex inputs, including unstructured textual content from clinical trial descriptions, required advanced language models and sophisticated machine learning algorithms. Furthermore, the model's predictions were designed to be interpretable and actionable, enabling their practical applications in investment decision-making.

Finally efforts were also made to prevent data leakage. In this project, data leakage is defined as the unrealistic situation where we have engineered features derived from the target variable and the situation where a single data point appears simultaneously in training, validation and testing sets. Data leakage will make the model performance over-optimistically high which can not be realised when the model is used in practice out-of-sample.

1.4 | Conceptual Framework

1.4.1 | The Food and Drug Administration

The U.S. Food and Drug Administration (FDA) ensures the safety and effectiveness of human and veterinary medicines, biologics, and medical devices. With over 18,000 employees, the FDA regulates various products, including food and cosmetics[1]. A key function of the FDA is the drug approval process, which rigorously evaluates new drugs for safety and efficacy, significantly impacting the biotech industry [1]. The pharmaceutical drug development process, from discovery to FDA approval, takes 12 to 15 years and involves many steps and variables [2].

1.4.2 | Binary Event Forecasting

Evergrowth focuses on binary event forecasting. This involves predicting outcomes that have two possible states, we will represent these as success or failure. In the biotech industry, this refers to predicting the approval or rejection of a drug by regulatory authorities (FDA). Accurate binary event forecasting can impact investment decisions, as these outcomes directly affect the financial prospects of biotech companies [3].

1.4.3 | FDA's 5-Step Drug Development Process

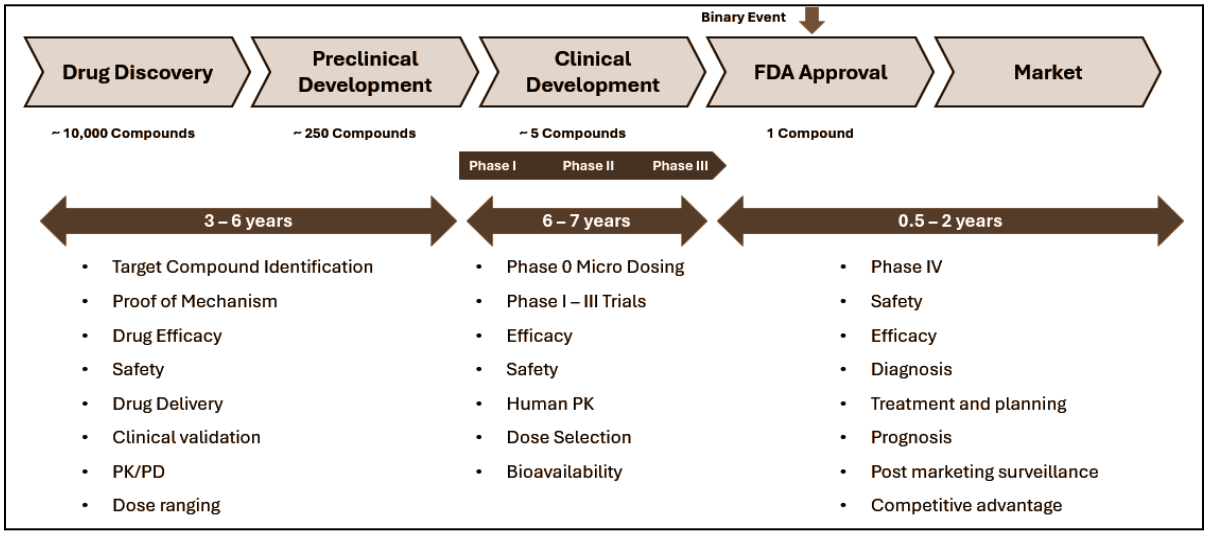


Figure 1: Drug discovery approval process diagram [4, 5].

1. Discover new drugs by understanding diseases and testing compounds through experiments.
2. Preclinical research involves assessing a drug's toxicity and includes animal testing before human trials[5].
3. In Clinical Research, the IND (Investigational New Drug) application is filed to ensure clinical trial safety (human testing). After approval, a clinical trial starts with 4 phases. Phase 1 involves 20 to 100 volunteers, with about 70% of drugs progressing. Phase 2 includes several hundred patients, with 33% advancing. Phase 3 tests 300 to 3,000 patients, with 25-30% moving forward. Phase 4 involves several thousand patients[5].
4. In FDA drug review, the New Drug Application (NDA) demonstrates a drug's safety and effectiveness, encompassing preclinical data through Phase 3 trial results. Upon receiving an NDA, the FDA has 60 days to assess completeness; incomplete submissions may be refused. A complete NDA undergoes 6 to 10 months of review. Once safety and effectiveness are

confirmed, the FDA collaborates with the applicant to finalize prescribing information, known as "labeling", detailing approval basis and usage guidelines.

5. FDA continually monitors drugs post-approval, addressing issues through cautions, usage updates, and more serious measures [5].

1.5 | Priori Hypotheses

The quantitative and qualitative information from clinical trials was leveraged to predict the approval or rejection of new drugs by the FDA. Our primary hypothesis was that utilizing advanced data analysis techniques, including AI-driven models such as machine learning and natural language processing (NLP), to analyze this information would enhance the accuracy of predicting FDA drug approvals. We further hypothesized that Phase 2 clinical trial data, augmented with insights from advanced language models, would provide the most relevant features for biotech hedge funds' stock selection processes and biotech stock investments [5, 6].

1.6 | Data Sources, Variables, and Key Concepts

We collected data regarding key clinical results and milestones from various research reports and reputable databases to build a predictive model for FDA drug approvals. The data was intended to be sourced from new drug applications filed with the FDA, ClinicalTrials.gov, research reports, and papers presented at healthcare industry conferences. Additionally, we attempted to integrate data from structured databases such as the AACT Clinical Trials Database, Cortellis Drug Discovery Intelligence Database, and DrugBank.

1.6.1 | Data Sources

1. **AACT (Aggregate Analysis of ClinicalTrials.gov):** AACT is a publicly available relational database that contains all information (protocol and result data elements) about every study registered in ClinicalTrials.gov. This database provided comprehensive trial-specific attributes, including trial design, patient demographics, inclusion and exclusion criteria, trial outcomes, and investigator experience. It offered a detailed structure for analyzing clinical trials by consolidating data on enrollment, trial status, interventions, and conditions studied. AACT's daily updates ensure that users accessed the most recent trial information, making it valuable for monitoring trial progress and identifying trends over time. For this study, we downloaded a static copy of the AACT database and pipe-delimited filesets as on Sep 23, 2024. Furthermore,

the database supported advanced queries for specific trial attributes, enabling tailored analyses for unique research needs. Its role extended beyond individual trial analysis to allow for meta-analyses and comparative studies across multiple trials, enhancing its utility as a critical resource for analyzing clinical trial trends and outcomes [7].

2. **Cortellis Drug Discovery Intelligence database:** This database is commercially available from Clarivate. It has detailed drug-specific information, such as therapeutic class, mechanism of action, development status, and biological targets, enhancing the contextual relevance of clinical trial data [8]. We used the Cortellis Drug Discovery data available from Baker Library at Harvard Business School.

3. **DrugBank database:** DrugBank was reviewed as a potential candidate but due to limitations in funding was not utilized to aid in our final dataset [9].

1.6.2 | Variables

The independent variables considered for this study spanned a broad range of drug- and trial-level characteristics derived from publicly available databases and literature. Financial variables were intentionally excluded to maintain focus on clinical and biological indicators. On the drug side, information from the Cortellis database included key attributes such as the drug's highest development status (e.g., preclinical, phase completion, or approval), therapy area, active indications, mechanism of action, and underlying technologies. Additional drug-level factors included molecular structure, any partnership or licensing deals, and the first launch date, each potentially influencing the drug's regulatory trajectory and ultimate approval likelihood.

From the AACT clinical trial database, we integrated variables capturing trial-level complexity and patient characteristics. These encompassed factors like accrual (number of enrolled participants), disease type, trial duration, inclusion and exclusion criteria, and patient demographics (e.g., gender and age distributions). We also considered more nuanced trial features such as investigator experience, sponsor identity and track record, trial design, and location (including number of sites and country). Further, disease segmentation, trial endpoints, adverse event profiles, and any FDA-granted special statuses (e.g., novel therapy designation or RMAT) were included, along with classifications like mono or combination therapy and orphan drug status. In sum, this diverse array of variables was intended to capture the multifaceted nature of the drug development and approval process, allowing for a more comprehensive and informative predictive model.

Financial data was not included in this study. Future work in this area can include financial independent variables such as (i) the market cap of the publicly listed company ($250MM - $2.0B) and (ii) macroeconomic variables.

The dependent variable was defined as the approval of a therapy or the occurrence of a significant FDA regulatory event.

1.7 | Assumptions, Delimitations, and Limitations

1.7.1 | Assumptions

- Publicly available data sources provided sufficient and high-quality data for building a robust predictive model.
- Advanced language models effectively extracted relevant features from unstructured textual data.

1.7.2 | Delimitations

- The analysis focused exclusively on Phase 2 clinical trial data, which provided pivotal information on drug efficacy and was critical for investment decision-making.
- Data was sourced solely from publicly accessible databases, given accessibility constraints.

1.7.3 | Limitations

- Data quality issues, such as missing values or inconsistencies, adversely affected the performance and reliability of the predictive model.
- The generalizability of the model was constrained by the scope of the data and the specific therapeutic areas represented.
- Ethical considerations, including potential biases in pre-trained language models and responsible data usage, were carefully managed.
- Limited computing power and time constraints restricted the complexity of models that could be implemented.

This framework acknowledged the constraints and assumptions while leveraging publicly available data to build a predictive model that focused on high-value insights from Phase 2 clinical trials [6].

1.8 | Importance of the Study

Binary event forecasting in the biotech industry is fundamental for Evergrowth BioHealthcare Capital to maintain its competitive edge and deliver superior returns. The current manual process is time-consuming, unscalable, and prone to errors. However, by implementing a data-driven approach, Evergrowth can enhance investment opportunity evaluation, increasing the chances of outperforming market benchmarks. With this approach we hoped to reduce the risk of uninformed decisions through systematic analysis while increasing efficiency and scalability by automating the extraction and analysis of unstructured data. Additionally, this advanced use of technology and data analytics can provide a competitive advantage. This study is important for maintaining leadership in biotech investment and delivering profitable returns on investment to investors. By the use of a data-driven approach, Evergrowth can improve its investment decision-making process, reduce risk, and increase efficiency, all while staying ahead of the competition in the biotech industry.

Furthermore, the study underscored the broader implications of adopting predictive analytics in healthcare investment. By integrating advanced machine learning models and leveraging structured and unstructured data sources, Evergrowth set a precedent for innovation in investment practices. This initiative aligned with the growing demand for transparency and accountability in the biotech sector, fostering trust among stakeholders and reinforcing Evergrowth's reputation as a forward-thinking leader in the industry. Finally, these advancements enabled the identification of novel investment opportunities, such as emerging therapeutic areas or underrepresented biotech startups, broadening the scope of Evergrowth's portfolio diversification.

Chapter 2: Literature Review

In this study, we examined the multifaceted processes involved in clinical trials and drug approvals, identified high-quality data sources critical for predictive modeling, and explored machine learning methodologies that enhanced the accuracy of drug approval predictions. Through this review, we analyzed the clinical trial phases and processes to better understand how types of data were generated at each stage and their relevance to regulatory decisions. We also identified structured and unstructured datasets from public and proprietary sources, offering insights into features and tools used in prior studies. Additionally, we analyzed recent advancements in ML models, including their application to trial outcomes and regulatory predictions, as well as research in using language models to assess the potential of data-driven approaches in improving decision-making for drug development and approval. Finally, we reflected on the broader implications of these advancements for the pharmaceutical industry and investors, recognizing the transformative role of AI and ML in accelerating and refining the drug approval process.

2.1 | The Drug Approval Process and Clinical Trial Phases

Clinical trials are fundamental in drug development, comprising multiple phases that progressively evaluate a drug's safety, efficacy, and overall suitability for FDA approval. Each phase generates critical data, shaping predictions for approval outcomes [10].

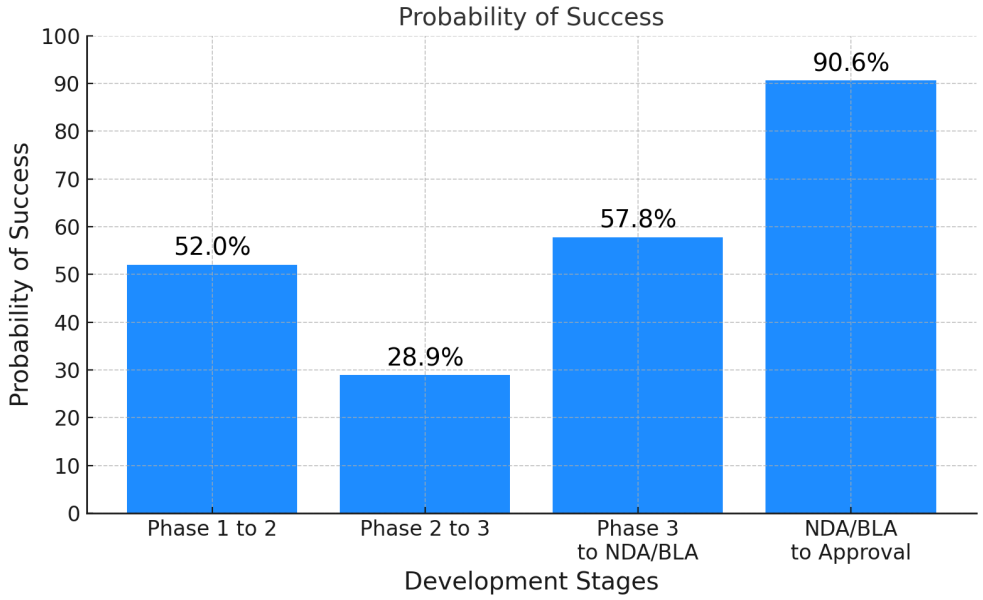


Figure 2: Phase transition success rates from Phase 1 to FDA approval for all diseases and modalities, showing the probability of progression at each stage of clinical trials [11].

- **Phase 1**: This phase evaluates the drug's safety, tolerability, pharmacokinetics, and pharmacodynamics in a small cohort of healthy volunteers or patients. Approximately 52% of drugs successfully transition from Phase 1 to Phase 2 [10, 11].
- **Phase 2**: Focused on efficacy and identifying side effects, Phase 2 involves a larger patient population and provides the first substantial evidence of therapeutic potential. Data from this phase is a strong predictor of FDA approval outcomes, with 28.9% of drugs progressing to Phase 3 [11].
- **Phase 3**: This phase confirms the drug's efficacy, monitors adverse reactions, and compares it to existing treatments in large-scale trials. About 57.8% of drugs advance from Phase 3 to the NDA/BLA submission stage, emphasizing the importance of robust evidence for regulatory evaluation [11].
- **NDA/BLA to Approval**: After Phase 3, a New Drug Application (NDA) or Biologics License Application (BLA) is submitted to the FDA for approval. At this stage, the drug undergoes comprehensive evaluation, with 90.6% of submissions receiving final approval[11].

2.2 | Data Sources for Predictive Modeling

The effectiveness of ML models in predicting FDA approvals was highly contingent on the quality and type of data utilized. Given the specificity and sparsity of the available data, we used methods similar to those of Lo et al., which emphasized the importance of statistical imputation for handling missing data, thereby improving the robustness of our predictive models [12]. Clinical trial data, sponsor track records, and prior approval histories were among the most predictive features [12].

We reviewed the literature for all the data sources that could help with our study. We identified many structured and unstructured data sources. However, many of these data sources were inaccessible to us due to being commercial third-party data sources. The study "A Tool for Predicting Regulatory Approval After Phase 2 Testing of New Oncology Compounds" provided a valuable reference for clinical data sources, collecting data from public sources such as clinicaltrials.gov, drugs@fda, and commercial pipeline databases like IMS Health (now IQVIA) R&D Focus, Thomson Reuters Partnering, and Citeline Trialtrove, as well as abstracts at scientific conferences and publications in peer-reviewed journals [13]. Although we could not access the exact dataset utilized in Andrew W. Lo's paper [14], it provided insights into relevant features, as summarized in Table (1).

While searching the Harvard Library Catalog, we identified clinical data available at Baker Library at Harvard Business School. The clinical data, referred to as the 'Cortellis Drug Discovery Intelligence'

(CDDI) database, included granular target and method of action (MOA) information as well as extensive pre-clinical and clinical biomarker data. The CDDI database contained trial data for over 600,000 drugs and biologics.

2.3 | Machine Learning Techniques for Predicting FDA Approvals

	Description	**Type**
Drug Features		
Route	Route of administration of the drug, the path by which the drug is taken into the body.	Multi-label
Origin	Origin of the active ingredient in the drug.	Multi-label
Medium	Medium of the drug.	Multi-label
Biological target family	Family of proteins in the body whose activity is modified by the drug, resulting in a specific effect.	Multi-label
Pharmacological target family	Mechanism of action of the drug, the biochemical interaction through which the drug produces its pharmacological effect.	Multi-label
Drug-indication development status	Current phase of development of the drug for the indication.	Binary
Prior approval of drug for another indication	Approval of the drug for another indication prior to the indication under consideration (specific to drug-indication pair).	Binary
Trial Features		
Duration	Duration of the trial (from reported start date to end date) in days.	Continuous
Study design	Design of the trial (keywords).	Multi-label
Sponsor type	Sponsors of the trial grouped by types.	Multi-label
Therapeutic area	Therapeutic areas targeted by the trial.	Multi-label
Trial status	Status of the trial.	Binary
Trial outcome	Results of the trial.	Multi-label
Target accrual	Target patient accrual of the trial.	Continuous
Actual accrual	Actual patient accrual of the trial.	Continuous
Locations	Locations of the trial by country.	Multi-label
Number of identified sites	Number of sites where the trial was conducted.	Continuous
Biomarker involvement	Type of biomarker involvement in the trial.	Multi-label
Sponsor track record	Sponsor's success in developing other drugs prior to the drug-indication pair under consideration.	Continuous
Investigator experience	Primary investigator's success in developing other drugs prior to the drug-indication pair under consideration.	Continuous

Table 1: Features used in "Machine Learning With Statistical Imputation for Predicting Drug Approvals" [12].

We explored which models were utilized for similar kinds of research projects. The application of ML techniques, such as ensemble models that combine various algorithms to improve prediction accuracy, was featured in recent studies. In "The Novartis Data Science and Artificial Intelligence Challenge," data from Phase 2 trials was utilized, where the top-performing ML model included a combination of extreme gradient boosting and Bayesian logistic regression, along with additional feature engineering [14]. These techniques demonstrated efficacy in enhancing predictive performance, achieving a testing AUC value of 0.88 (95% CI [0.85, 0.90]) for predicting drug approval in 2021 [14]. This model showed noticeable improvement compared to the 5NN-RandomForest model constructed in 2019 [12] on the same testing set. Similarly, the RESOLVED2 model employed ML to predict outcomes post-Phase 1 clinical trials in oncology, indicating high predictive accuracy and utility in early decision-making [15].

2.4 | Language Models for FDA Drug Approval Predictions

The adoption of language models and machine learning frameworks is transforming the way FDA drug approvals are predicted. With recent advancements in NLP, unstructured data such as clinical trial documents, patient records, and scientific literature can now be utilized within predictive models. These advancements provide new opportunities for analyzing the complexities of drug approval processes. For instance, large language models (LLMs) allow for the extraction of critical data from unstructured sources, improving the efficiency and accuracy of information retrieval, expediting the approval pipeline while maintaining rigorous safety standards.

Transformer-based models like BioBERT significantly advanced biomedical text mining. BioBERT, trained on large-scale biomedical datasets, excelled in tasks such as information extraction and text classification [16]. Specifically, BioBERT demonstrated notable improvements in biomedical named entity recognition, with a 0.62% F1 score increase, relation extraction, had a 2.80% F1 score increase, and question answering, had an improvement of 12.24% for the mean reciprocal rank metric, over previous state-of-the-art models [16].

Meta AI's LLaMA (Large Language Model Meta AI) series represents a noteworthy advancement in the application of LLMs [17]. Although initially designed for general NLP tasks, there are emerging usages of LLaMa models that have been adapted for medical applications, such as Me-LLaMa, which specializes in medical text analysis and diagnostics some research suggestions that these models outperforms GPT-4 on 5 out of 8 datasets [18].

GPT-4 demonstrates potential in enhancing clinical trial workflows relevant to FDA drug approval processes. Its application in tools like TrialGPT shows its ability to match patients to clinical trials by evaluating eligibility criteria, a critical step in ensuring trial integrity and compliance with FDA requirements [19]. Additionally, its capability for automated report generation can speed up the preparation of regulatory documents, aligning to the FDA standards for submission [20]. These features position GPT-4 as a valuable asset in navigating the complex landscape and language of clinical trial design and execution, particularly in regulatory contexts.

The CTP-LLM model automates clinical trial outcome prediction by leveraging trial design documents to forecast phase transitions. This model represents an advancement in applying large LLMs to predict the progression of clinical trials through various phases. By accurately predicting phase transitions, CTP-LLM improves strategic planning and resource allocation for pharmaceutical companies and research institutions. The model achieves 67% accuracy in predicting trial phase transitions across all phases and 75% accuracy specifically in predicting the transition from Phase 3 to final approval. These results show the potential of utilizing LLMs to enhance decision-making in clinical trial management and streamline the drug development pipeline [21].

Tx-LLM, a large language model trained on a collection of 709 datasets and fine-tuned from PaLM-2, integrates unstructured textual data with structured molecular information to predict drug approval outcomes. In evaluations across 66 tasks from the Therapeutics Data Commons (TDC), Tx-LLM achieved performance competitive with state-of-the-art (SOTA) models on 43 tasks and surpassed SOTA on 22 tasks. Notably, it excelled in tasks that combined molecular SMILES representations with textual data, such as disease or cell line names. Traditional models often specialize in specific domains, focusing solely on either molecular data or textual information. The Tx-LLM's ability to process both data types simultaneously allows for a more comprehensive analysis of factors influencing drug approval. This comprehensive approach enables Tx-LLM to outperform traditional methods, especially in tasks involving small molecule drugs, by effectively capturing the complex interplay between molecular properties and clinical contexts [22].

The FDA has increasingly explored the potential of LLMs to enhance safety monitoring in the drug approval pipeline. For instance, AE-GPT, fine-tuned from GPT-3.5, achieved a micro F1 score of 0.704 (strict matches) and 0.816 (relaxed matches) for extracting adverse events from the Vaccine Adverse Event Reporting System [23]. The BERTox Initiative uses NLP to improve regulatory science at the FDA. It focuses on tasks like information retrieval, text classification, and named entity recognition to analyze FDA documents and public data, its goal is to improve efficiency and accuracy while ensuring safety and efficacy assessments remain robust [24].

Integrating LLMs with real-world data sources, such as electronic health records, has also enhanced post-market surveillance, allowing regulators to monitor adverse events and ensure compliance with safety standards. For example, a UCSF BERT model identified serious adverse events in clinical notes with a 5-10% accuracy improvement over previous models [25].

While LLMs offer much potential in the FDA drug approval pipeline, significant challenges remain. Current models often struggle with generalization across diverse datasets and require extensive fine-tuning for domain-specific tasks [26]. Additionally, concerns about data privacy, model interpretability, and biases in training data pose obstacles to broader adoption [26]. For instance, LLMs trained on diverse datasets often inadvertently incorporate biases, affecting fairness and equity in healthcare applications [26]. These shortcomings push the focus of new research to developing more robust domain-specific LLMs for regulatory applications while also addressing ethical concerns through transparent and interpretable modeling techniques [26]. It is also important to address populations that are underrepresented in the datasets used to train or fine-tune these models to mitigate biases. By addressing these gaps, researchers and industry can more comprehensively utilize the power of language models, enabling faster, safer, and more efficient drug approvals [26].

2.5 | Implications for Industry and Investors

Moreover, the discussion section of "The Novartis Data Science and Artificial Intelligence Challenge" shared additional information on the limitations of the models and studies, such as not utilizing enough information from unstructured text and not considering data availability at the time of decision-making. The discussion section also included feedback on the data science approach to drug approval prediction from a domain expert working as a Pharmaceutical Researcher [14].

As part of the data science capstone, we were interested in identifying which AI and ML methodologies had been used in previous studies. The integration of AI and ML into drug development was found to not only enhance predictive accuracy but also streamline the drug approval process. By predicting outcomes more accurately, these technologies helped pharmaceutical companies focus resources on the most promising drug candidates, potentially reducing the time and cost associated with drug development. Additionally, this benefit attracted investors to drug development projects, accelerating and improving the overall availability of new drugs, thus increasing sales and potentially generating higher profits. This, in turn, led investors to allocate more resources for greater returns. This research was referenced by Mullard, who discussed the role of AI in predicting trends and outcomes within the FDA approval process, noting the increasing importance of biologics and specialized drug categories [27].

Our main objective was to predict binary events of FDA approval. FDA approval went through a number of intermediate steps, each carrying a risk of approval or denial. In researching this business problem further, we reviewed a document titled "Scientific and Regulatory Reasons for Delay and Denial of FDA Approval of Initial Applications for New Drugs, 2000-2012 [28]," which investigated the factors contributing to the delay or denial of FDA marketing approval for new drugs from 2000 to 2012. It explicitly addressed new molecular entities (NMEs), active ingredients that had never been marketed in the US. Among the 302 NME applications reviewed, 151 (50%) were approved upon their first submission, while 222 (73.5%) ultimately gained approval after one or more resubmissions. The primary factors for initial rejection included can be seen in Table 2.

Type of Deficiency	First-Cycle Review Failures (n = 151)	Delayed Approvals Following Resubmission (n = 71)	Drugs Never Approved During Study (n = 80)	P Value
Efficacy deficiencies only	48 (31.8)	15 (21.1)	33 (41.3)	.01
Safety and efficacy deficiencies	41 (27.2)	13 (18.3)	28 (35.0)	.03
Safety deficiencies only	39 (25.8)	24 (33.8)	15 (18.8)	.04
CMC alone	17 (11.3)	13 (18.3)	4 (5.0)	.02
Labeling alone	4 (2.6)	4 (5.6)	0	.05
CMC and labeling	2 (1.3)	2 (2.8)	0	.22

Table 2: The table above shows the frequency of safety, efficacy, CMC (Chemistry, Manufacturing, and Controls), and labeling deficiencies in drugs that failed their first-cycle FDA review. It compares these deficiencies across three categories: drugs that failed initially but were approved upon resubmission, drugs never approved during the study, and their statistical significance (P Value) [28].

The report “Clinical Development Success Rates and Contributing Factors 2011–2020” analyzed 12,728 phase transitions in drug development, identifying disease indication, drug target, modality, and novelty as key success factors. These findings were vital for refining our predictive model, directly enhancing its ability to accurately forecast FDA approvals and informing smarter investment decisions in biotech stocks [11].

Phase Success	Phase I to II		Phase II to III		Phase III to NDA/BLA		NDA/BLA to Approval	
	n	Phase POS	n	Phase POS	n	Phase POS	n	Phase POS
Hematology	92	69.6%	106	48.1%	82	76.8%	72	93.1%
Metabolic	136	61.8%	149	45.0%	66	63.6%	48	87.5%
Infectious disease	403	57.8%	414	38.4%	197	64.0%	156	92.9%
Others	154	63.6%	228	38.6%	90	60.0%	69	88.4%
Ophthalmology	88	71.6%	200	35.5%	82	51.2%	45	91.1%
Autoimmune	413	55.2%	471	31.4%	219	65.3%	202	94.1%
Allergy	55	56.4%	92	28.3%	34	64.7%	20	100.0%
Gastroenterology	45	46.7%	73	34.2%	35	57.1%	33	90.9%
All indications	4414	52.0%	4933	28.9%	1928	57.8%	1453	90.6%
Respiratory	179	55.9%	215	21.9%	62	64.5%	45	95.6%
Psychiatry	150	52.7%	164	26.8%	71	56.3%	57	91.2%
Endocrine	319	43.3%	293	26.6%	151	66.2%	124	86.3%
Neurology	516	47.7%	504	26.8%	226	53.1%	165	86.7%
Oncology	1628	48.8%	1732	24.6%	495	47.7%	324	92.0%
Cardiovascular	214	50.0%	252	21.0%	105	55.2%	80	82.5%
Urology	22	40.9%	40	15.0%	13	69.2%	13	84.6%

Table 3: The table above presents phase transition probability of success (POS) by disease, showing the percentages of drugs advancing through each clinical trial phase and receiving FDA approval. The data is segmented into categories, showing a comparison from different phases to final approval. Note, that NDA and BLA stand for New Drug Application, and Biologics License Application respectively [11].

The document "Analysis of US Food and Drug Administration New Drug and Biologic Approvals, Regulatory Pathways, and Review Times, 1980–2022" reviewed FDA approvals from the past four decades, highlighting how regulatory changes shortened approval times. For example, median review times dropped from 26.6 months to 9.9 months following significant legislative changes in 1992 and 2012 [29]. These changes in FDA regulations affected the state of the data utilized, its long-term robustness, and the constructed model, impacting both its accuracy and long-term viability.

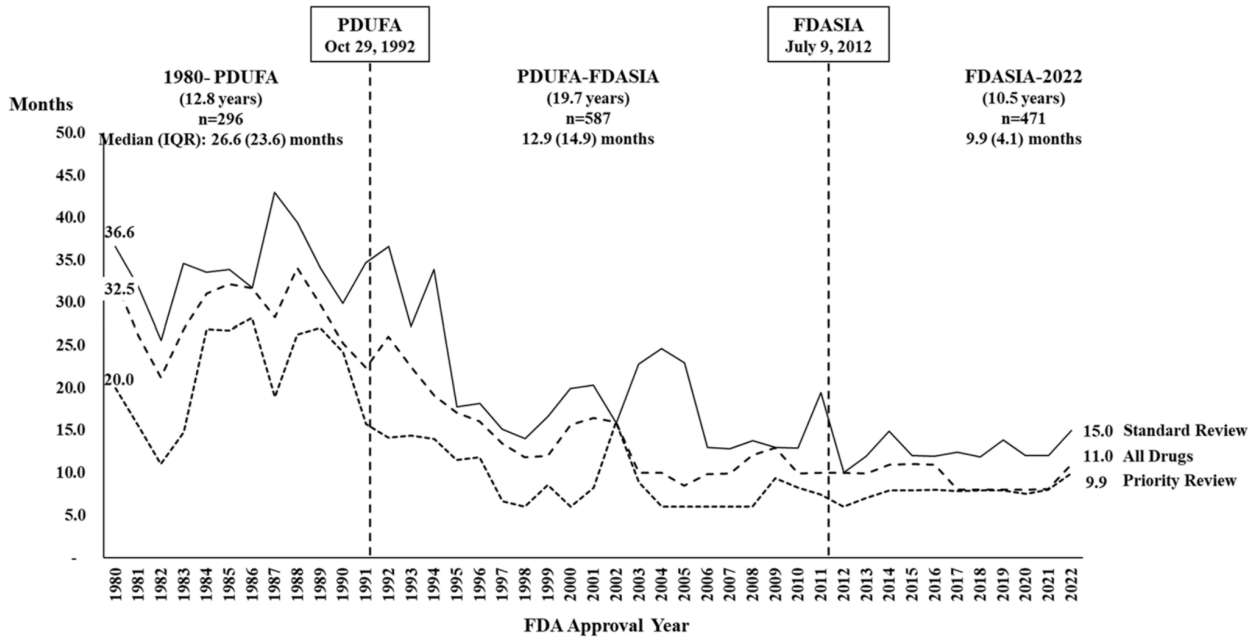


Figure 3: FDA New Drug Approvals Median Review Time, 1980–2022. Annual median and interquartile range (IQR) of FDA review time for new drugs approved between 1980 and 2022. The trends are presented for standard review, priority review, and the total for all drugs. The figure includes the number of drugs and the annual median and IQR of the FDA review time for new drugs during the periods defined by PDUFA and FDASIA [19].

The literature shows a growing dependence on AI and ML in drug approval forecasting, driven by the need for more accurate and efficient investment decision-making in the biotech sector. As AI technologies evolve, their integration into predictive models is expected to become more pronounced, leveraging complex datasets and advanced analytical techniques to forecast drug approval outcomes with ever-improving precision.

Chapter 3: Methodology

3.1 | Project Methodology

This project focused on developing a predictive model for FDA drug approvals using Phase 2 clinical trial data. The approach combined key features, language models, and an expanded feature set to enhance prediction accuracy. The methodology was divided into four stages:

1. **Data collection and processing:** Ensuring the datasets met acceptable quality standards for model fitting.
2. **Feature generation using language models:** Leveraging language models to create additional features.
3. **Feature expansion and Trial Sequence integration:** Making the model more realistic by introducing the Trial Sequence concept and broadening the feature set.
4. **Final models:** Refining the models to optimize performance.

Ethical considerations and data trustworthiness were central to the project, ensuring compliance with legal standards and maintaining data integrity. Each stage was designed with specific steps and a timeline to achieve the desired outcomes efficiently.

3.2 | Stage 1: Data Collection and Processing

In this stage, we gathered data from various sources, including Cortellis datasets accessed through the Harvard Baker Library, and online resources such as the AACT database, DrugBank, and ClinicalTrials.gov. The goal is to process this data to predict the FDA approval outcome for specific drug/indication pairs based on Phase 2 clinical trial results.

During data processing, we addressed duplicate entries and handled missing values to ensure data integrity. Stratified sampling was employed to maintain proportional representation of each outcome category (approved, not approved) within the dataset, preserving class balance across training, validation, and testing splits. This approach was critical for achieving robust model performance. Additionally, we eliminated some features that could potentially leak future information into the model, safeguarding against data leakage and enhancing prediction accuracy.

3.3 | Stage 2: Feature Generation Using Language Models

Hypothesis: Leveraging language models to enhance dataset completeness will lead to improvements in baseline model performance.

Steps:

1. **Initial information extraction using large language models:** LLMs were used to extract relevant information from long and complex clinical trial information text.
2. **Small language model:** A small language model was utilized for efficiency in automating the information extraction process.
3. **Model ensemble for information extraction:** An ensemble of language models was explored to improve the robustness of the information extraction process.
4. **Retraining and evaluation of the baseline model:** The newly extracted information was incorporated into the dataset. The baseline model was retrained, and its performance was evaluated to measure the impact of the enhanced dataset.

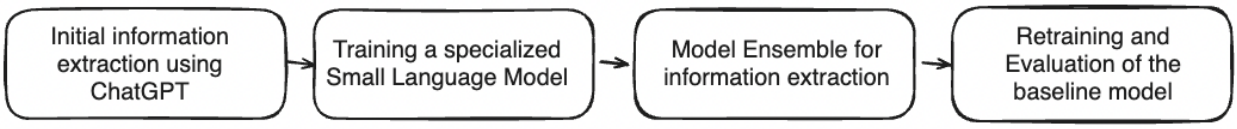


Figure 4: Overview of Stage 2 Methodology

In this stage, we are considering the following assumptions, risks, delimitations, limitations, and expected results:

Assumptions:

- Language models like ChatGPT accurately extracted relevant information from clinical trial documents.
- Improved data completeness enhanced model performance.

Risks:

- Language models may misinterpret or miss critical information.
- Dependency on external models could introduce variability in results.

Delimitations:

- The focus was on specific clinical trial documents and press releases for data extraction.
- ChatGPT and other language models were used for initial information extraction.

Limitations:

- The quality of extracted data is dependent on the language model's accuracy.
- Time and computational resources were required for utilizing language models.

Expected Results:

- The dataset was enhanced with additional extracted information from clinical trials.
- Model performance improved due to more comprehensive feature sets.

3.4 | Stage 3: Feature Expansion and Trial Sequence Integration

Based on our literature review, the following categories of variables were identified as critical for improving predictive accuracy in the final model. These features encompass both drug-specific attributes and clinical trial characteristics, offering a comprehensive view of factors that may influence FDA approval outcomes. While these features are essential for the final modeling phase, they were not fully utilized during the baseline model stage.

Drug-Specific Features:

- Biological Target: Identifies the class of the drug's target (e.g., cytokine/growth factor, enzyme, ion channel, receptor, transporter).
- Mechanism of Action: Describes the drug's mode of action (e.g., cell cycle inhibitor, DNA inhibitor, ion channel antagonist, protein kinase inhibitor).
- Origin: Indicates the drug's source or composition (e.g., biological protein antibody; biological protein recombinant; chemical synthetic).
- Therapeutic Class: Classifies the drug by its primary therapeutic use (e.g., anti-viral, anti-HIV; anti-cancer, immunological; anti-epileptic).

Clinical Trial Features:

- Disease Type: Specifies the condition or disease area under study (e.g., bladder, colorectal, ovarian).
- Duration of the Trial: Measures the length of the trial period.
- Number of Identified Sites: Counts how many clinical sites are participating in the trial.
- Patient Demographics: Includes relevant participant characteristics such as age and gender.
- Trial Design: Details the methodological framework of the study (e.g., cross-over, double-blind/blinded, efficacy-focused, multiple-arm, open-label, pharmacodynamics, pharmacokinetics, placebo-controlled, randomized, single-arm).

- Outcome: Reflects the trial's result (e.g., completed with positive outcome/primary endpoints met; completed but with negative or indeterminate outcome; terminated due to safety or adverse effects).
- Primary Endpoint and Exclusion Criteria: Defines the main objective(s) of the trial and the criteria used to exclude certain participants.
- Sponsor Type and Track Record: Includes the nature of the sponsoring entity and historical performance metrics (e.g., number of prior approved drug–indication pairs, number of prior failed pairs, total number of trials sponsored, number of trials with positive results, number of terminated trials).

In addition to the work done on the features, we also came up with a second processed dataset, with each data point representing one clinical trial in the trial sequence of a drug-indication pair, given that many drug-indication pairs had more than one clinical trial. This made the model prediction more realistic for hedge fund investors, as they usually observed one clinical trial at a time when making investment decisions. Essentially, we obtained more data points.

3.5 | Stage 4: Final Model

In this stage, we explored the tree-based models below and their ensemble techniques to develop robust predictive models.

- Random Forest
- XGBoost (Extreme Gradient Boosting)
- LightGBM (Light Gradient Boosting Machine)
- H2O's GBM (H2O Gradient Boosting Machine)

Note XGBoost, LightGBM, and H2O's GBM are implementation variations in Python for Gradient Boosting and we would like to take the opportunity to experiment these popular open-source variations on the same datasets.

We also studied stacked models that combined simple logistic regression and the above tree based models.

In this process, we did not use cross-validation due to dataset constraints. Instead, we adopted an iterative approach, tuning the core hyperparameters of each model to achieve the best validation AUC.

Additionally, we analyzed feature importance using these models to identify which features most significantly influenced FDA approval. This analysis provided insights into the key factors driving the predictions and helped refine the model further.

3.6 | Data Trustworthiness

Ensuring the trustworthiness of our datasets was fundamental for the robustness of our models. Key challenges included addressing missing entries and confirming the relevance of non-US data, as well as the relevance of data on firms not in Evergrowth's portfolio. The integrity of our datasets was also influenced by potential revisions to the datasets. Similarly, availability timestamps were essential to assess the relevance of each data attribute and the practical benefits to hedge fund investors. These timestamps were challenging to obtain, and input from subject matter experts was required to identify the attributes most likely available before the typical decision-making time, in this case for hedge fund investment.

3.7 | Ethical and Legal Considerations

Ethical and legal considerations played a critical role in our project, particularly in how we sourced, managed, and utilized data. All data for this project was obtained from AACT (Aggregate Analysis of ClinicalTrials.gov), a publicly available resource. While AACT provided open access to clinical trial data, it was essential to ensure that its use adhered to ethical and legal guidelines, particularly regarding data aggregation and analysis.

Additionally, gaps in the available data, such as information on biological targets, mechanisms of action, or sponsor track records, underscored the limitations of relying solely on public datasets. Any future attempt to fill these gaps would need to strictly comply with ethical standards, particularly in cases where data might involve sensitive clinical trial information. Safeguarding participant privacy and adhering to data usage policies remained paramount to maintaining the integrity and ethical grounding of this research.

3.8 | Project Timeline

The project was completed in four stages with the following timeline:

- **Stage 1:** Mid September - Data Collection & Review (obtained the primary datasets)

- **Stage 2:** Late October - Data Pre-processing & Integration (and attempted to acquire the alternative commercial datasets)
- **Stage 3:** Late November - Baseline Model & Language Model Development
- **Stage 4**: Early December - Final Model & Analysis
- **Final Report:** Mid December

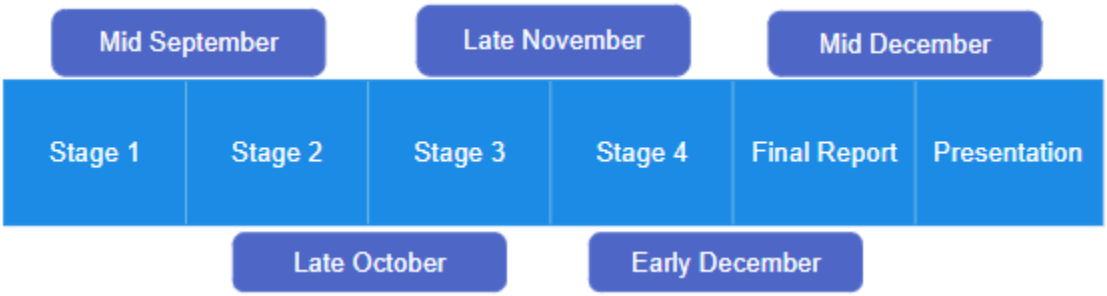


Figure 5: Project Timeline

The development of this project was iterative. We deployed and reassessed the next steps as we progressed, based on the learnings from previous stages. While some tasks from different stages could have been performed concurrently, we chose a sequential and structured approach to ensure a comprehensive analysis and development process. This approach leveraged traditional modeling techniques and advanced language models to achieve the best possible predictive performance.

Our methodology was structured to effectively predict FDA drug approvals. The project was organized into four stages to ensure clarity and focus, allowing for systematic progress toward our objectives. Each stage built on the previous one, enhancing the model's accuracy through iterative development and real-time adjustments. This approach enabled us to maintain flexibility while steadily advancing toward our ambitious goals.

Chapter 4: Analysis

4.1 | Description of Data and Analysis

4.1.1 | Data Sources and Relevance

The primary objective of this study was to develop a predictive model for FDA drug approvals based on Phase 2 clinical trial data. To achieve this goal, a dataset was constructed by integrating multiple reputable data sources, each contributing unique and critical information to the analysis. AACT Clinical Trials Database [7] provided trial-specific attributes such as trial design, patient demographics, inclusion and exclusion criteria, trial outcomes, and investigator experience. AACT, a rich repository of clinical trial data, enabled in-depth analysis of trial characteristics and outcomes, serving as our main data source.

Cortellis Drug Discovery Intelligence Database [2] supplied detailed drug-specific information, including therapeutic class, mechanism of action, development status, biological targets, and other attributes pertinent to drug development. This database, known for its extensive coverage of pharmaceutical data, is widely used in the industry for drug intelligence purposes. The DrugBank Database was explored but did not contribute to our baseline dataset derived from AACT [7, 9].

While the AACT database and Cortellis datasets provided valuable information about clinical trials and drug attributes, they lacked a critical component: the approval status at the drug-indication pair (DIP) level, which was central to our study. For example, Cortellis reported approval status only at the drug level. If a drug was tested for type 1 and type 2 diabetes and was marked as approved, it was unclear whether the approval applied to type 1, type 2, or both indications.

To address this gap, we explored two potential solutions. The first involved searching for new datasets that included DIP-level approval details. The second relied on manually enriching the existing datasets by making logical assumptions and conducting supplementary research. After evaluating both options, we determined that the latter approach was the only feasible solution, enabling us to create the necessary DIP-level approval data to support our predictive model.

4.1.2 | Outcome Variables

Our study focused on two primary binary outcome variables:

- **Market launch status of drugs**: This variable reflects whether a drug received FDA approval and was successfully launched into the market (coded as 1) or not (coded as 0). This outcome is the ultimate goal of drug development efforts, and predicting it accurately can significantly enhance strategic decision-making in pharmaceutical development.
- **Completion status of clinical trials**: This variable indicates whether a clinical trial was completed (coded as 1) or not completed (coded as 0). Understanding the factors contributing to trial completion is crucial, as incomplete trials represent significant financial and time losses in drug development.

These outcome variables were central to addressing the research problem of improving predictive accuracy in the drug development pipeline, particularly in the critical transition from Phase 2 trials to market approval. On the other hand, we also realised these outcome variables don't suffice to create the target classification variable for our clinical trial approval prediction purpose. The solutions are discussed in the later sections.

4.1.3 | The Option of Using Commercial Datasets

During our literature review, we identified several studies that used datasets from Citeline (formerly part of Informa's Pharma Intelligence division) to predict drug approvals. These datasets included DIP level approval status, a critical feature absent from both the AACT and Cortellis datasets.

Using these datasets would have presented several advantages. First, they included DIP-level approval status, directly addressing a key gap in our existing data sources. Second, leveraging datasets already utilized in published work could have streamlined the process and reused established data cleaning and preprocessing pipelines. Finally, employing the same datasets as previous studies would have facilitated a transparent evaluation of our methodology, enabling direct comparisons to existing approaches.

We initially contacted Professor Andrew Lo, author of [12], who kindly provided a point of contact at Citeline. Despite this, our efforts to engage with Citeline through email, phone, and text were unsuccessful. We then turned to Citeline's customer support, which ultimately led to productive meetings with their marketing and data engineering teams. After presenting our requirements, Citeline offered two academic-use product options. Unfortunately, the quoted prices were prohibitively expensive and far exceeded the budget allocated for our capstone project.

Given these constraints, we had to abandon efforts to acquire Citeline's datasets and instead focus on enriching the AACT and Cortellis datasets with DIP-level approval status, as detailed in Section 4.1.1.

4.1.4 | Data Preprocessing

Data integration challenges

In integrating data from multiple heterogeneous sources presented several challenges that required careful consideration:

- **Data alignment and matching**: In our datasets, each clinical trial and drug is associated with multiple conditions, represented as unstructured strings. These strings often lack consistency, exhibit significant variation, and may include spelling errors, even when referring to the same indication. Our goal was to establish **drug-indication pairs (DIPs)**—explained further down below in *"Drug-indication pair (DIP) as the primary data point entity"*—using disease types as proxies for standardized indication names. To address the variability in condition strings, we tested fuzzy matching algorithms but that did not do well in the data matching. We also experimented with both LLMs and small language models (SLMs) for condition-indication matching, but their matching accuracy remained below 80%, which did not meet our requirements. Consequently, we adopted a manual approach, systematically matching conditions to disease types by identifying relevant keywords through resources like Google and ClinicalTrials.gov. This method achieved over 98% accuracy, providing a highly reliable alignment between conditions and disease types.

- **Data quality and completeness**: The datasets contained inconsistencies, missing values, and varying levels of data granularity. An extensive data cleaning process was undertaken, involving:

    - **Removing duplicates**: Duplicate records were identified and removed to prevent skewed analyses and ensure that each observation was unique.

    - **Reducing features**: Features with more than 30 categories were excluded to reduce high-dimensional sparsity and computational inefficiency. This included features such as irrelevant dates, unhelpful description fields, and source or name-related features that added unnecessary complexity without contributing meaningful insights to the model. While reducing dimensionality by grouping some of the categories could be beneficial, it requires significant manual effort and careful stratification when splitting the dataset for training. Given the complexity of this process, it may be considered for future improvements to the model.

    - **Handling missing values**: Features with over 50% NaN values were removed to avoid limited information and complex imputation. This included features related to expanded access, FDA regulation, and device approval status, which were dropped to streamline the dataset and minimize unnecessary complexity in the analysis.Missing numerical data were imputed using statistical methods appropriate for the data distribution, while categorical data with insufficient representation were carefully reviewed. In cases where imputation was not feasible, records with critical missing information were excluded to maintain data integrity.

    - **Normalizing features**: Variables were normalized across datasets to ensure consistent magnitude for logistic regression's regularization. The other models we experimented didn't require normalization.

- **Eliminating data leakage**: Strict data partitioning was forced to prevent future or outcome-related data from inadvertently influencing the training data. This was critical after initial models exhibited unnaturally high accuracies due to data leakage. Data leakage can occur when information that would not be available at the time of prediction is included in the training data, leading to over-optimistic performance estimates.

Drug-indication pair (DIP) as the primary data point entity

In Section 4.1.1, we explain that two DIPs (e.g., Drug A for Condition X and Drug A for Condition Y) can have different approval statuses, even though they are linked to the same drug. This reflects the fact that approval is granted at the level of a specific drug-indication pair, not the drug as a whole. However, within a single DIP (e.g., Drug A for Condition X), all associated clinical trials should share the same approval status, as they collectively contribute evidence toward the regulatory decision for that DIP.

We treated each DIP as a single data point for our modeling dataset, consistent with prior studies [12 , 14]. This approach ensures that the predictive model aligns with the regulatory process, where decisions are made at the DIP level.

As explained in 4.1.1, two DIP can have different approval status even if they are linked to the same drug. Meanwhile, all the clinical trials under the same DIP should have the same approval status. It's conceptually natural to treat one DIP as one data point, for the processed dataset for modeling, which is also the setup of the existing work [12 , 14].

However, the AACT dataset does not include a specific indication field. The closest available field, "condition," is poorly standardized and inconsistent, making it challenging to identify DIPs directly. To address this, we mapped the "condition" values from the AACT dataset to disease types defined in [14]. Further details on this mapping process are provided in the "Data Alignment and Matching" section above.

We recognize that disease types are not equivalent to indications, but this mapping was the most practical solution given our time and resource constraints. It is important to note that we used disease types only to group clinical trials into data points for modeling purposes and did not treat them as categorical variables in the model.

Beyond drug-indication pairs: Trial Sequence datasets

While our initial approach grouped data points at the DIP level, this setup assumes that investors know when the final clinical trial for a given DIP has concluded. In reality, hedge fund investors must often make decisions based on incomplete information, evaluating ongoing clinical trials without knowing the eventual outcome. This constraint underscores the importance of designing a predictive model that aligns with decisions made in practice.

To address this challenge, we developed a second dataset tailored for real-world use cases. Unlike the DIP-level grouping, this dataset treats individual clinical trials as distinct data points. For each trial, we incorporate its specific attributes and aggregate information from all prior trials for the same DIP completed by that time. This approach maintains consistency in approval status across all trials for the same DIP while enabling predictions at any stage of the clinical trial process.

The **Trial Sequence** dataset complements the original **One-Row-Per-Pair** setup by simulating how investors access and interpret trial data over time. By introducing this temporal element, our work moves beyond static modeling to offer a more dynamic and actionable framework, ensuring practical and relevant insights aligned with the decision-making timelines faced by hedge fund investors.

Target variable value determination

As described in Sections 4.1.1 and 4.1.2, we manually determined the target variable, DIP-level approval status, due to its absence in the available datasets. The following approach balances data accuracy with the constraints of our project timeline.

We labeled all DIPs associated with drugs classified as disapproved in the Cortellis dataset as disapproved. For drugs classified as approved, we retained only the DIPs that had progressed to Phase 4 and labeled them as approved while dropping the remaining. Phase 4 trials, conducted after regulatory

approval, monitor a drug's performance in real-world use and assess its broader safety profile [30]. We excluded DIPs without sufficient evidence to confirm approval status in Phase 4, as we could not reliably determine their outcomes.

We also made significant efforts to verify the exact approved indications for drugs labeled as approved in the Cortellis dataset. This process involved cross-referencing information using online resources, including Google and ChatGPT. However, after realizing that completing this task would require substantial time beyond the project's scope, we decided to discontinue it.

Feature Engineering

To enhance the predictive power of the dataset, comprehensive feature engineering was performed, involving both structured and unstructured data.

- **Categorical variable encoding**: One-hot encoding was applied to categorical variables to convert them into a format suitable for machine learning algorithms. To mitigate the high dimensionality resulting from variables with many categories, similar categories were grouped based on domain knowledge, and dimensionality reduction techniques were considered. This step was important to avoid the curse of dimensionality, which can adversely affect model performance.

- **Natural language processing**: Advanced NLP techniques were used to extract useful features from unstructured text fields, such as trial descriptions and inclusion/exclusion criteria. This involved the use of state-of-the-art language models to process and derive meaningful features from textual data, as detailed in the next section.

Attempt to utilize the information from Principal investigators:

Since previous studies, including the [14], used investigator information as a feature, we explored ways to incorporate it into our model. During discussions with a senior clinical physician, we learned about the importance of principal investigators in clinical trials. However, we determined that using their names as categorical variables would not add value for model fitting due to the large number of unique principal investigators in the Phase 2 trial subset.

To address this, we proposed generating an **Experience Score** for each principal investigator. This score would increase with the number of trials they participated in. To align with hedge fund investors' needs, we ensured the Experience Score remained **point-in-time**, meaning we calculated it using only trials completed:

- Before the last clinical trial in the DIP for the **One-row-per-pair dataset**, or
- Before the current trial for the **Trial Sequence dataset**.

After cleaning and standardizing principal investigator names, we found that the maximum number of trials any investigator participated in was only nine across the full dataset. When restricting the score to point-in-time trials, the number decreased further, making it unlikely that the Experience Score would effectively differentiate between DIPs or clinical trials. Given the coding complexity required to implement a point-in-time Experience Score and its limited potential impact, we stopped this task after completing the name cleaning and exploratory data analysis (EDA).

The process of manually matching and expanding features takes significant time. To support parallel workstreams, we built a baseline model during this stage. For the baseline model, we tried Random Forest and XGBoost, as these models were selected for their robustness to handle high-dimensional tabular data, and effectiveness in capturing non-linear relationships between variables. This initial model provided a foundational framework for evaluation, enabling us to proceed with other analyses while continuing to expand features and data points DIP in subsequent stages.

4.1.5 | Rationale for Analytical Techniques

The complex nature of the dataset, characterized by high dimensionality and a mix of structured and unstructured data, necessitated the use of sophisticated analytical techniques. Each technique was carefully chosen to address specific aspects of the data and research objectives.

Exploratory Data Analysis (EDA) was conducted to understand the underlying structure of the data, identify patterns, detect anomalies, and formulate hypotheses. This step was necessary for informing subsequent modeling efforts and ensuring that the models were grounded in a solid understanding of the data.

Machine learning models

To tackle the predictive modeling tasks, we employed machine learning algorithms, including Random Forest (RF), XGBoost, LightGBM, and H2O GBM. These models were selected for their ability to handle complex datasets, robustness to overfitting, and complementary strengths in capturing non-linear relationships between variables. Each model is briefly described below:

- **Random Forest (RF):** An ensemble learning technique that builds multiple decision trees and aggregates their predictions to improve accuracy and reduce overfitting. It excels in handling large feature sets, mixed data types, and missing values while providing feature importance metrics.

- **XGBoost (Extreme Gradient Boosting):** A gradient-boosting framework optimized for speed and performance. XGBoost incorporates regularization to control overfitting and employs tree-based algorithms to capture non-linear relationships effectively.
- **LightGBM (Light Gradient Boosting Machine):** A gradient-boosting algorithm designed for efficiency and scalability. It uses a histogram-based approach for faster computation, making it suitable for large datasets. LightGBM is particularly effective in handling categorical variables and sparse data.
- **H2O GBM (H2O Gradient Boosting Machine):** A high-performance, distributed implementation of gradient boosting provided by the H2O platform. H2O GBM is known for its scalability and ability to work seamlessly with big data frameworks. It supports flexible configurations for hyperparameter tuning and handles missing data gracefully.
- **Ridge Logistic Regression:** A simple and efficient model for classification problems. Not expecting it to out-perform the tree-based models above, but it could add value by stacking with an existing high-performance model.

By using these models, we ensured that our analysis leveraged the strengths of each approach to address the challenges posed by the dataset, thus enhancing the robustness and reliability of the predictions.

4.2 | Initial Modeling Exploration

4.2.1 | Data Preparation for Modeling

Before training the models, the dataset underwent several pre-processing steps:

Train-test split: The data was partitioned into training (70%), validation (15%), and test (15%) sets to ensure that the models were evaluated on unseen data, reducing the risk of overfitting.

- **Splitting based on drugs:** The trials or DIPs associated with the same drug tend to have very similar data points in both features and target variables. Having different data points associated with the same drug separate in training, validation and testing sets can cause implicit data leakage and inflated performance metrics. Here we split based on the drugs first and then subsetting the data points based on the drug splitting.

- **Stratified by therapy areas:** The most important high-degree categorical variable in our dataset is the therapy area. We introduced stratification to ensure a similar distribution of different therapy areas in the training, validation and testing sets.

- **Split count:**

	Split count at Drug level	Split count One-row-per-pair	Split count Trial Sequence
Train Count	678	1490	3031
Test Count	145	288	652
Validation Count	146	261	608

Table 4: Various train-test split methods applied. Note that split count at the **drug level** was performed instead of drug-indication pair (DIP).

Feature Scaling: Numerical features were standardized to have zero mean and unit variance, which is essential for algorithms sensitive to the scale of data.

4.2.2 | Initial Results and Observations

The models were trained using the training set, with hyperparameters tuned based on performance on the validation set. Key evaluation metrics included:

- **Accuracy**: The proportion of correct predictions over the total predictions made.
- **Precision and recall**: Precision measures the accuracy of positive predictions, while recall measures the ability to identify all positive instances.
- **F1 Score**: The harmonic mean of precision and recall, providing a balance between the two.
- **Area Under the Receiver Operating Characteristic Curve (AUC-ROC)**: Assesses the model's ability to distinguish between classes across all thresholds.

The Random Forest model initially achieved a high accuracy of 98% on the validation dataset; however, data leakage was causing this unusually high performance. Outcome-related information was inadvertently included in the training data, leading to over-optimistic estimates. A similar issue was identified in the other tree based models, which also showed comparable results, emphasizing the need for careful data preprocessing to ensure reliable performance evaluation.

4.2.3 | Data Leakage and Updated Model Performance

Recognizing the critical issue of data leakage, corrective actions were implemented:

- **Data re-partitioning**: The dataset was re-examined to ensure that no information from the validation and test sets was present in the training set. Strict temporal separation was enforced, with trials initiated after a certain cutoff date reserved for validation and testing.
- **Feature review**: Features that were potential proxies for the outcome, such as 'Development Status' or any variables updated post-Phase 2, were excluded from the training data.
- **Model retraining**: The models were retrained using the corrected dataset. This resulted in more realistic performance metrics, reflecting the true predictive capabilities.

4.2.4 | Interpretation of Results

The revised models demonstrated strong predictive performance, with balanced accuracy and AUC-ROC scores. The following observations were made:

- **Class imbalance management**: The datasets used to fit the models have relatively balanced distribution between approval and disapproval classes. No action is necessary here.
- **Feature importance**: Significant predictors included the Investigator Experience Score, Trial Complexity Index, and features derived from NLP analysis of trial descriptions.
- **Model comparison**: XGBoost slightly outperformed Random Forest, possibly due to its ability to capture complex relationships through gradient boosting.

Despite the promising results, several limitations were identified:

- **Data quality**: Remaining issues with data quality and missing values may have impacted model performance.
- **Generalizability**: The model may not generalize well to new data, especially in therapeutic areas underrepresented in the training set.
- **Interpretability**: While ensemble models provide feature importance, they are still relatively black-box models compared to simpler algorithms.

The baseline model provided a solid foundation for predicting FDA drug approvals based on Phase 2 clinical trial data. It highlighted the critical importance of rigorous data preprocessing and the potential

pitfalls of data leakage. The understanding gained informed the subsequent development of more advanced models incorporating additional features and leveraging sophisticated language models.

4.3 | Language Model Analysis

To further improve the dataset and increase our model's predictive power, several advanced language models were utilized, each contributing unique features derived from the textual data. Models such as Llama3, GPT-4o, and NLP tools like SpaCy were used to extract meaningful features from unstructured text, enhancing our baseline model. These models and methods enabled us to leverage the subtle information contained in trial descriptions and criteria, which were difficult to analyze using traditional methods. The use of these models required careful consideration of potential data leakage and biases inherent in the models due to their training data.

4.3.1 | Small Language Models For Criteria Extraction

The utilization of the Llama3's small language models (SLMs) were introduced to evaluate the robustness of inclusion and exclusion criteria within clinical trial descriptions. Robust was theorized to be important for ensuring trial reliability, credibility, and replicability. Both the three billion and one billion mode parameters versions of Llama3's language models were tested. After review of the outputs, we utilized Llama3.2-Instruct, with 3 billion model parameters due to its better accuracy and its ability to capture nuances. The model, tokenizer, and prompt used can be seen in code 1.

```
# Initialize the pipeline with Llama3.2 3B Instruct model
pipe = pipeline("text-generation",
                model="/kaggle/input/llama-3.2/transformers/3b-instruct/1",
                tokenizer="/kaggle/input/llama-3.2/transformers/3b-instruct/1")

def clinical_trial_robustness(text):
   # chatbot role
   chatbot_role = "You are a clinical trial expert specializing in evaluating the
robustness of trial criteria."
   # Criteria Prompt Robustness
   prompt_template = f"""
   Evaluate the following clinical trial criteria text for robustness.
   Rate it as one of the following:
   - Strong: Clear, detailed, and minimizes bias.
   - Normal: Moderately clear but with some potential gaps.
   - Weak: Vague, incomplete, or prone to significant bias.
   - NA: Please contact site for information.

   Criteria Text:
   {text}
   """
   # Creating message for model
```

```
   messages = [ {"role": "system", "content": chatbot_role},
                {"role": "user", "content": prompt_template} ]

   # Generateresponse
   response = pipe(messages, max_new_tokens=20, temperature=0.7,
num_return_sequences=1)
   return response[0]['generated_text'][-1]['content']
```

Code 1: Llama 3.2 with 8 Billion parameters used had the following prompt above. Note that this is a SLM version of Llama3.

The LLama3_2_Criteria_Robustness feature was designed to quantify the clarity and detail of inclusion and exclusion criteria, as stronger documentation was expected to lead to better predictive accuracy by reducing uncertainty in the trial framework. To generate this feature, the clinical trial criteria were fed into the SLM and stored in a list for each national clinical trial identifier. The Llama 3.2 model classified the vast majority of criteria as strong, with only a fraction classified as normal or weak. This distribution, as can be seen in figure 6, suggests that many clinical trial descriptions were either well-defined or that the SLM lacked the complexity to capture these nuances.

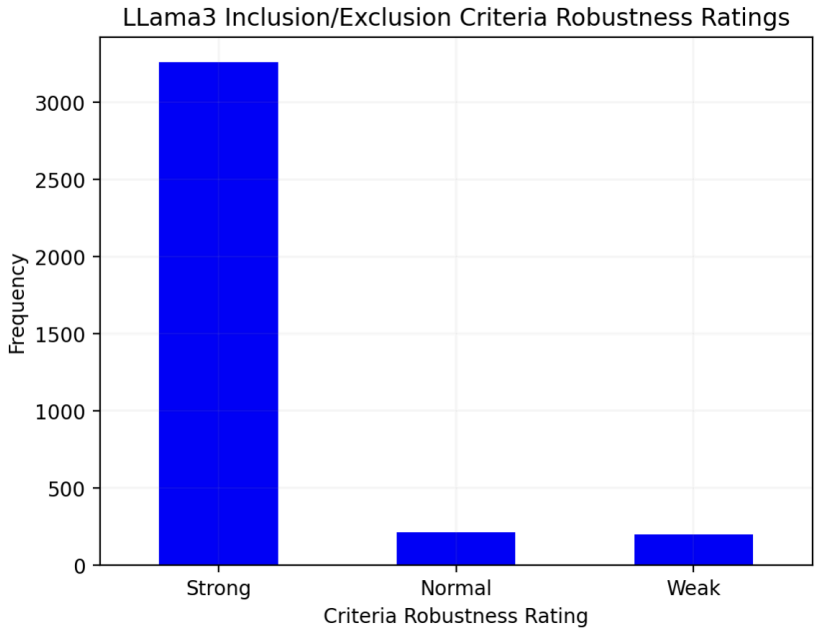


Figure 6: Bar chart showing the counts of inclusion and exclusion criteria robustness ratings as evaluated by the Llama 3.2 model

One challenge that arose was for the SLM to have a reference point for how these categories were defined. Introducing too much text, such as three examples of each category, into the prompt to aid the SLM caused the model to hallucinate.

4.3.2 | Large Language Models for Human Importance

GPT-4o's evaluation of trial descriptions for human interpretability aimed to address the qualitative aspects of trial documentation. The GPT4o_Human_Interpretability feature was intended to capture how important and relevant trial descriptions were with respect to humans. To derive this feature information with respect to trial descriptions were fed for each national clinical trial identifier. This was done with batches of size two hundred to help the large language model have relative context between the values. The prompt used to generate the responses can be seen in code 2, please note that no other parameters were not changed.

```
# GPT-4o Setup
prompt_template = """
You are an expert in clinical trial evaluation tasked with assessing and categorizing
trial descriptions based on their importance and relevance to public health and
medical innovation. Use the following scale to rate each trial:

- "Not important": Limited scope, low relevance, or lacking broader implications.
- "Slightly important": Moderately relevant but limited in scope or potential impact.
- "Moderately important": Provides useful information with practical implications but
not groundbreaking.
- "Very important": Significant potential for public health improvement or medical
innovation.
- "Extremely important": Transformative potential with wide-ranging public health or
scientific implications.

Respond in the following format only:
(index, Rating)

Descriptions:
{batch_descriptions}
"""
```

Code 2: GPT4o was used for its understanding and well roundedness in context and human relationship.

Clear and interpretable documentation and descriptions was hypothesized to improve model training by reducing ambiguities and enhancing the alignment of input data with clinical objectives. Trials with comprehensive and well-structured descriptions scored higher. The data reflects mapped importance levels due to insufficient representation in the categories "Not important" (33 instances) and "Extremely important" (6 instances), which were merged with "Slightly important" and "Very important," respectively, for analysis purposes with respect to enough representations for our ML model to learn from. The majority of evaluations fall under "Moderately important" (2245 instances), followed by "Slightly important" (840 instances) and "Very important" (588 instances).

This feature was shown to have more importance than the Llama 3.2's criteria robustness feature in the above section.

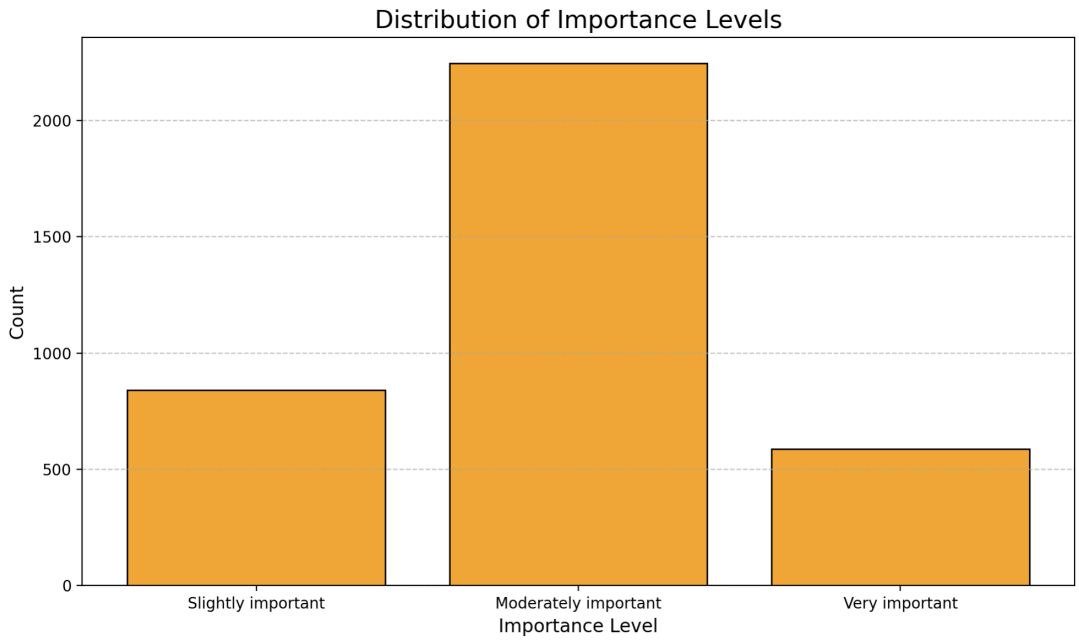


Figure 7: Bar chart showing the distribution of importance levels as outputted by the GPT-4o model.

4.3.3 | Advanced NLP Methods For Pregnancy Relationship

SpaCy's entity recognition pipeline was introduced to address aspects of clinical trial documentation, the identification of exclusion criteria, specifically for pregnant women. The Spacy_Pregnant_Women_Excluded feature was hypothesized to improve the model's ability to capture patterns in trial design that are often overlooked by generic NLP tools. Identifying such exclusions is important for understanding trial demographics and assessing the generalizability of trial outcomes. Approximately 46% of trials were identified as excluding pregnant participants. An interesting outcome as a A 2018 analysis of actively recruiting NIH-funded phase 3 and 4 clinical trials found that 68% explicitly excluded pregnant women. This number does have high variance with other research suggesting 95% [31, 32].

By merging these features with our dataset, we sought to improve the predictive accuracy of our baseline model, which we will discuss below. Due to the usage of such models, we noted potential pitfalls, such as data leakage and model biases, and took measures to maintain the integrity of the analysis.

4.3.4 | Data Leakage Considerations

The integration of advanced language models into our analysis introduced potential risks of data leakage and biases that could compromise the validity of our findings. We carefully evaluated these risks

for each model and implemented strategies to mitigate them, ensuring the integrity of our analysis throughout. SLM Llama Models and GPT-4o, trained on large-scale corpora that might include publicly available clinical trial data, required additional precautions.

We removed some models and their derived features from the analysis to eliminate any risk of data leakage. One of the models removed was ClinicalTrialBioBERT-NLI4CT, which had been fine-tuned on the NLI4CT. This dataset contained clinical trial statements for natural language inference, and its use in FDA drug approval models could have risked data leakage if overlapping trials caused memorization of specific trial details instead of generalizable patterns. Therefore, it was dropped and not utilized [33, 34]. Another model tested was Drug-BERT, which had been trained on drug reviews but was excluded for similar reasons. This model had been built on Google's BERT architecture and fine-tuned on the Drug Review Dataset from Drugs.com, likely overlapping with information in our dataset. Using its outputs would have introduced bias and undermined the reliability of our results.

Recognizing the limitations of models trained on specific datasets, we were cautious in interpreting results. Patterns were only considered meaningful if consistently supported by multiple models and analyses. These measures ensured that the contributions from language models enhanced predictive accuracy without compromising the integrity of the analysis.

4.4 | Final Data Table & Analysis

We constructed our dataset, by merging AACT clinical trials data with Cortellis Drug data, which are two separate databases. We extracted drug specific features from Cortellis and trial specific features from AACT. Our complete process is shown in the following flow chart:

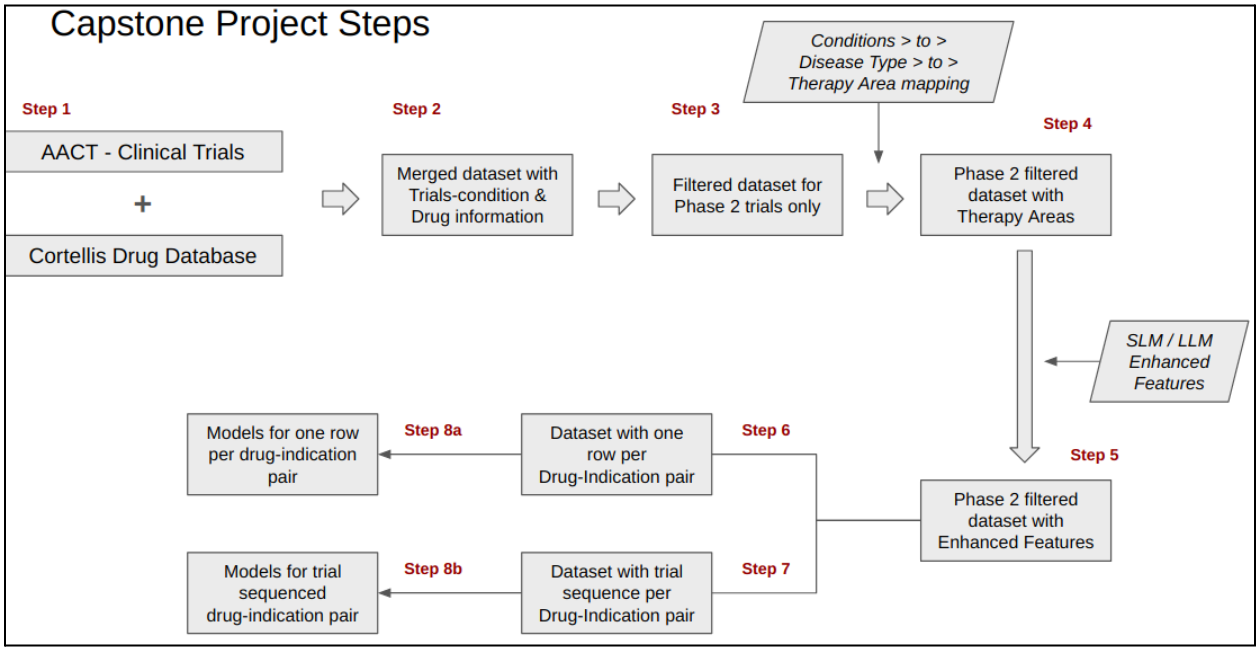


Figure 8: Capstone project Steps from data pre-processing to final models development

For data preparation, we utilized multiple data tables from the AACT database, such as Study, Intervention, Conditions, Patient Eligibilities, Trial Results, Study Calculated Values, etc. The data model for AACT database including all tables is available on the AACT website [7]. While merging AACT with Cortellis, we first created a dataset of drug-condition pairs, and then filtered on phase 2 trial data.

We took the conditions available in AACT to pair them with a drug. Since there were 2,274 indications present in AACT tables, we categorized them first into 147 disease types and then further categorized them into 26 therapy areas. Identification and categorization of conditions into disease types and therapy areas was done by reviewing previous studies, LLM models and online research.

Consequently the final datasets obtained have a fairly balanced distribution between approval and disapproval classes and further distribution insights in other variables can be found in Figure 11, 12, 13 and 14 as well as in Table 5 and Table 6.

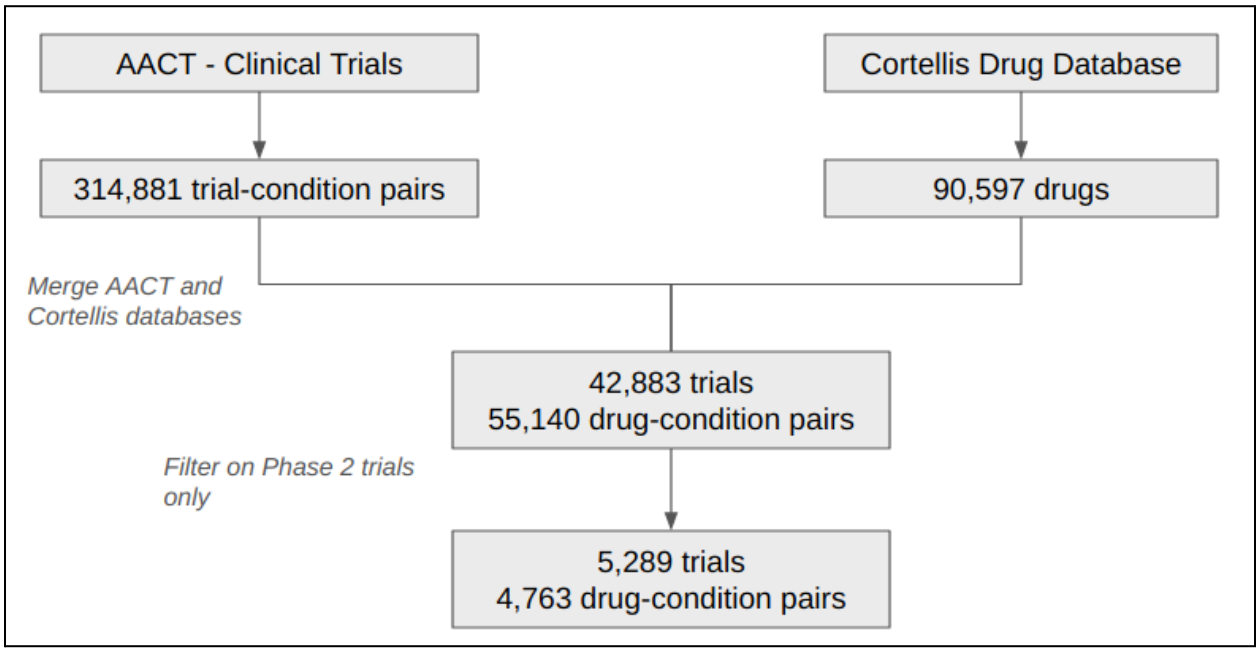


Figure 9: Data filtering and sample sizes at each step of data pre-processing

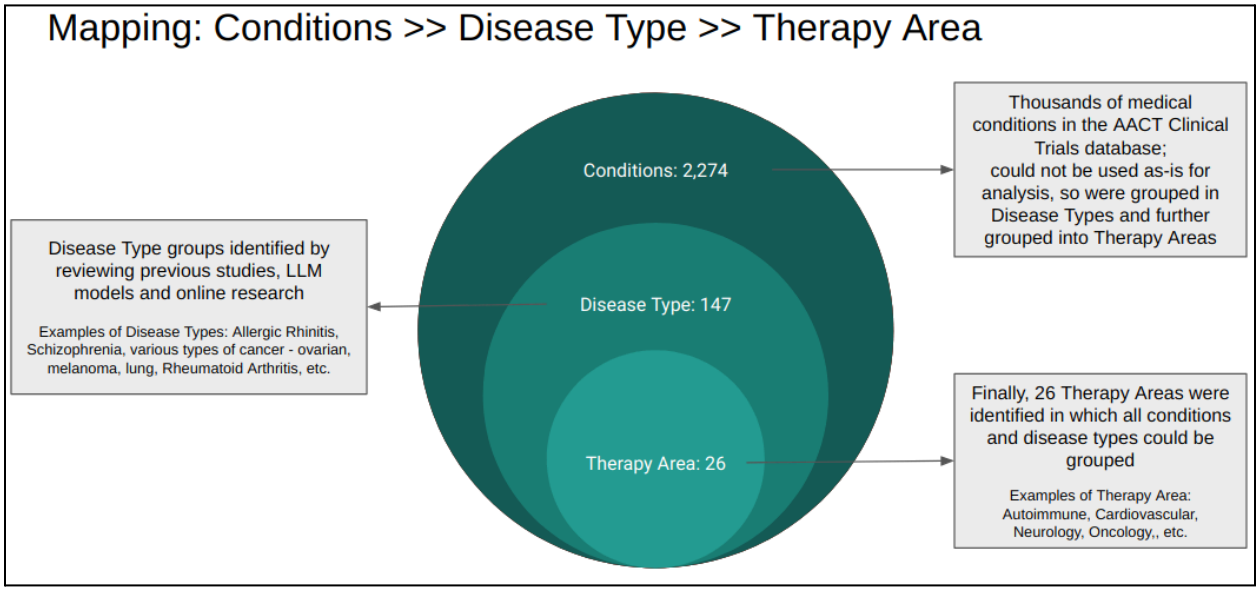


Figure 10: Categorization of AACT Conditions into Disease Types and Therapy Area

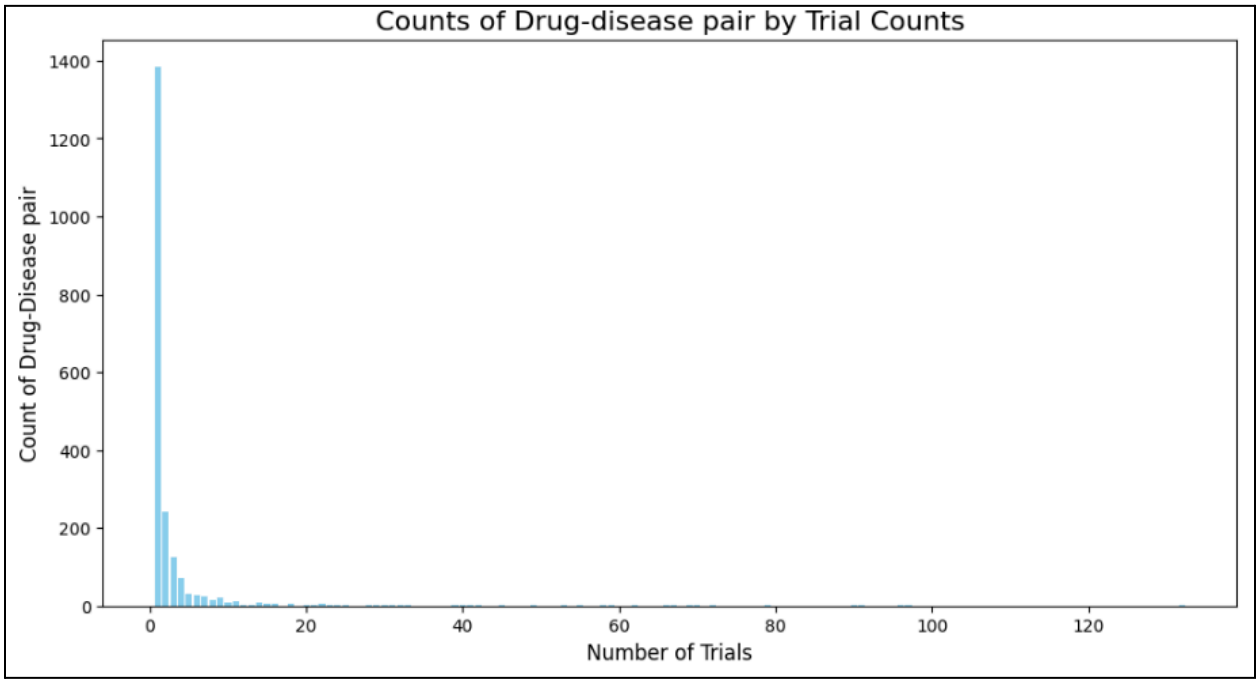


Figure 11: Most of the Drug-disease pairs had 1 trial

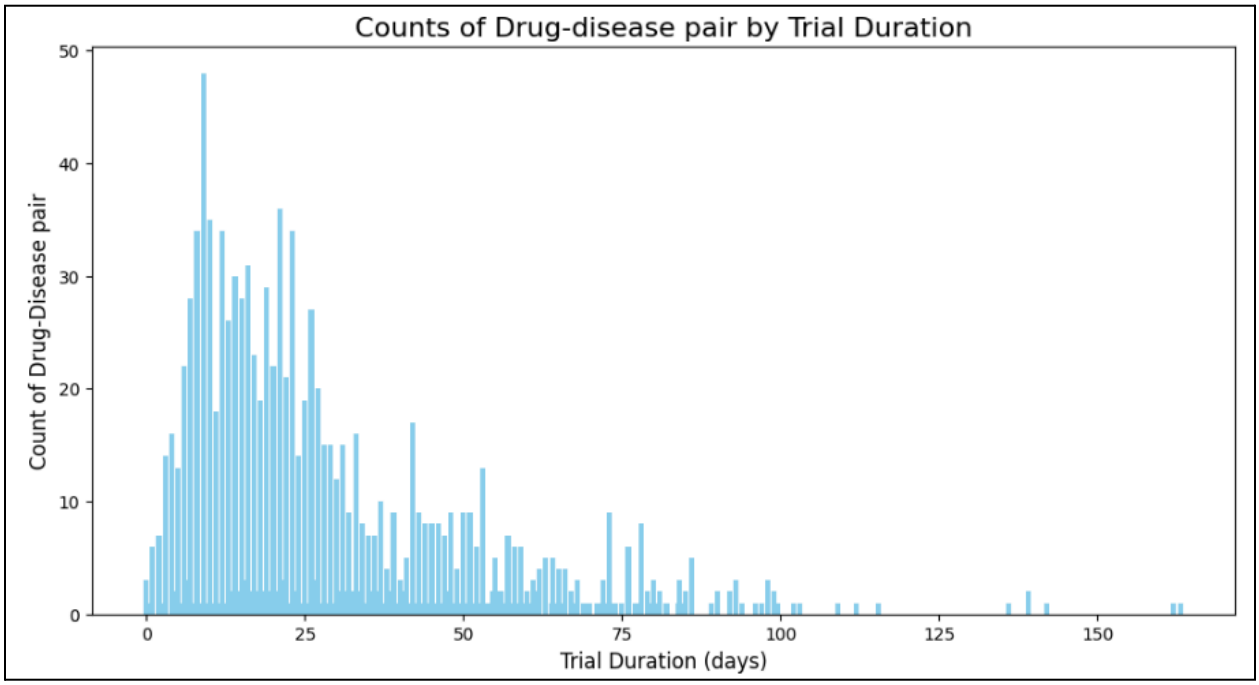


Figure 12: Most of the Drug-disease pairs had trials run for < 50 days

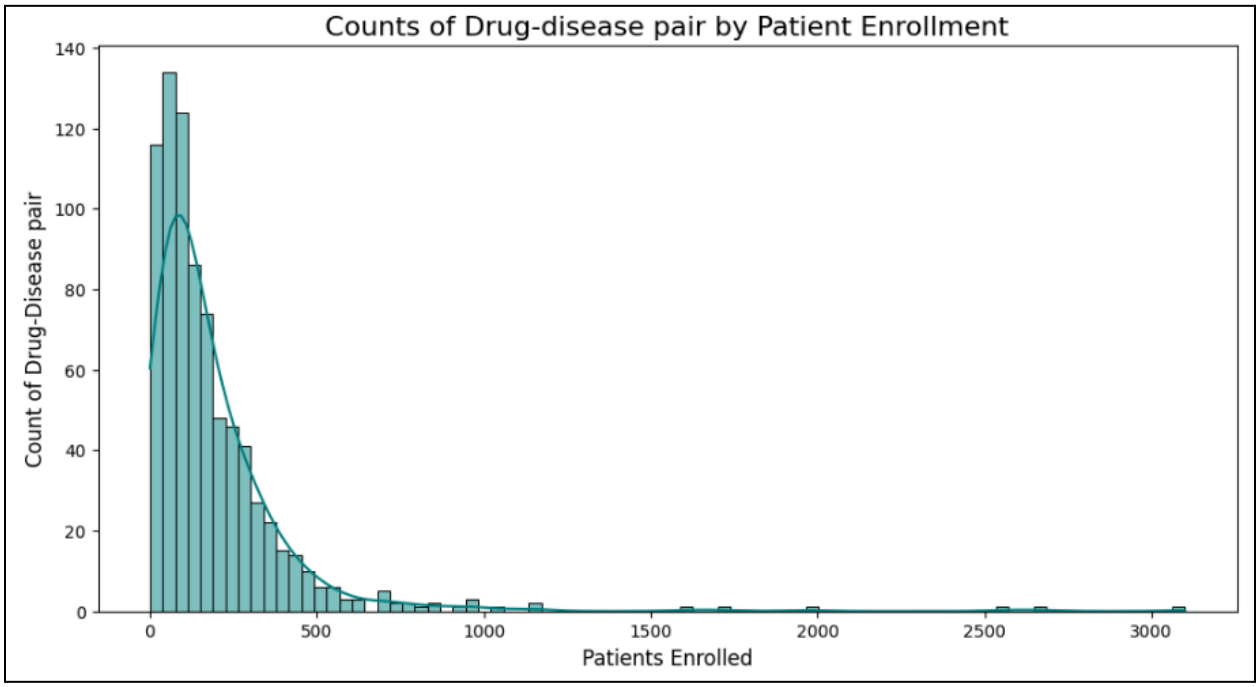


Figure 13: Most of the Drug-disease pairs had < 500 patients enrolled in the clinical trial

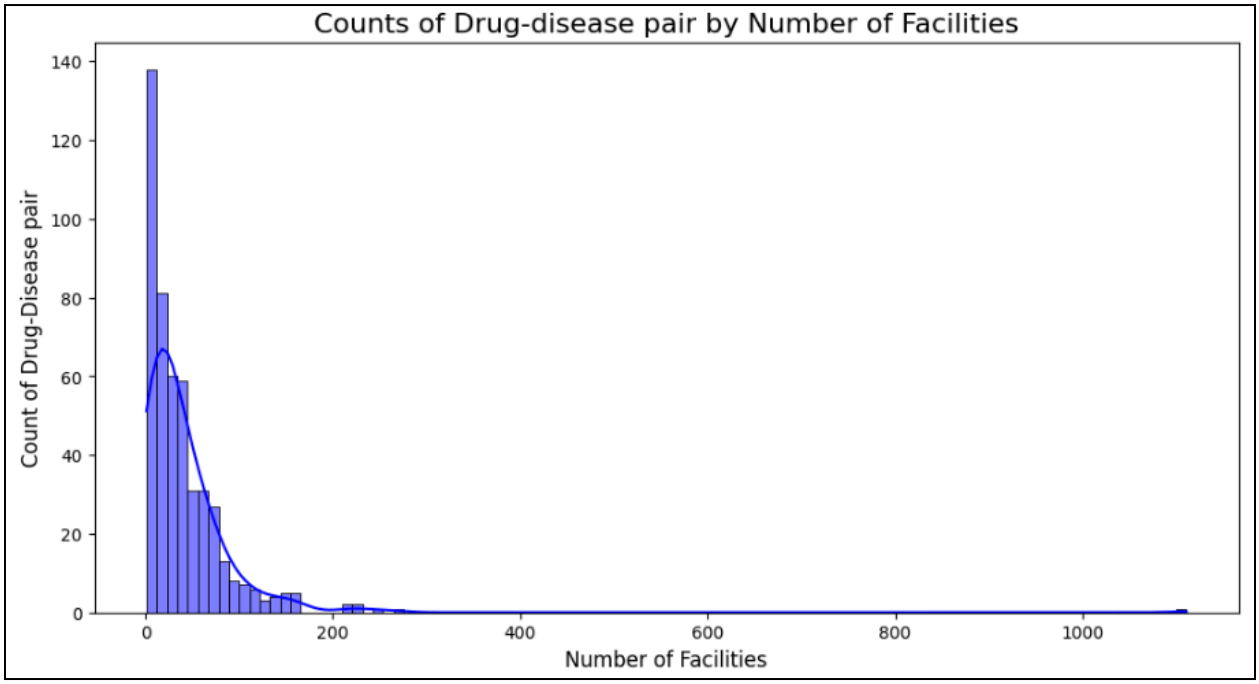


Figure 14: Most of the Drug-disease pairs had < 200 facilities for clinical trial

	Drug Outcome	
	Approved	**Failed**
Trials with adults included	97.5%	98.9%
Trials with children included	25.2%	13.2%
Trials with US facility	78.8%	73.2%
Trials with Single facility	65.4%	36.6%
Trials with healthy volunteers	16.5%	5.0%

Table 5: Percentage of Trials with different categories

Therapy Area	**Drug Outcome**	
	Approved	**Failed**
Autoimmune	20	15
Bone	1	4
Cardiovascular	68	49
Dermatology	9	37
Endocrinology	2	6
Gastrointestinal	9	29
Genetic and Rare Diseases	5	31
Hematology	5	28
Hepatic Diseases	0	2
Infections (bacterial)	30	34
Infections (viral)	39	36
Infectious Diseases	0	1
Metabolic Disorders	66	87
Neurology	22	26
Neuroscience (non-progressive)	13	28
Oncology	136	650
Ophthalmology	13	13
Other	78	93
Pain	11	27
Neuroscience (progressive)	21	63
Psychiatry & Mental Health	32	36
Respiratory	24	55
Rheumatology	24	46
Urology and Reproductive Health	14	16

Table 6: Drug-disease count by Therapy area and Trial Outcome

4.5 | Final Model & Analysis

4.5.1 | Final Models

Inspired by the work demonstrated in [14], we employed tree-based models (Random Forest, XGBoost, LightGBM, and H2O's GBM) to analyze our datasets. To ensure representative datasets, we used stratified sampling when splitting the datasets into training, validation, and test sets, as explained further in 4.2.1. The training process involved hyperparameter tuning on the training dataset to identify the configuration with the highest validation AUC. Subsequently, the final proposed model was retrained using the combined training and validation datasets with the best hyperparameters. The tree-based model optimization also includes model selection from the four mentioned above. Based on the validation AUC, XGBoost performs the best for One-row-per pair and LightGBM is the winner for Trial Sequence.

In addition, we selected the top 10 features, as listed in 4.5.5, based on the four tree-based models' feature importance as simple yet effective feature reduction. Using these selected features, we trained a basic L2 regularized logistic regression model with hyperparameter tuning on the training dataset. Unlike the tree-based models, which inherently handled missing values and did not require imputation, the logistic regression model used simple median imputation to address missing data. Note the median imputation actually makes the processed dataset not point-in-time which is not ideal for hedge fund investors because information available after the trials' completion date becomes part of the features. Similar to the tree-based models, we selected the best hyperparameters based on validation performance and retrained the logistic regression model on the combined training and validation datasets if we choose to include it as part of the final proposed model .

For the final stacking, we applied a soft voting approach. The stacking combined the optimized tree-based model (weighted at 2/3) and the logistic regression model (weighted at 1/3), in the probability space.

On top of the modeling choices, two processed datasets were utilized for fitting, testing and analysis, as discussed in 4.1.4:

1. One-row-per-pair dataset
2. Trial Sequence dataset

For Trial Sequence, the stacked model achieved a final test accuracy of **0.71**. For One-row-per-pair, which may be too optimistic for hedge fund investors, the stacked model achieved a test accuracy of

0.77. As a result, we decided to use the stacked model for One-row-per-pair and the lightGBM model alone for Trial Sequence. See 4.5.2 for further details.

Throughout the model tuning and other decision making process, the test dataset remained untouched and was only used to evaluate the final proposed model's performance

4.5.2 | Model Stacking Analysis

Below are the evaluation metrics comparison for the model stacking. Based on the validation metrics, that logistic regression added some value to the tree-based model for One-row-per-pair but failed to contribute much for Trial Sequence. This is not too difficult to understand as the Trial Sequence dataset is more complicated and is more likely to have non-linear relationships and interactions between features. Meanwhile the logistic regression is neither hurting too much the stacked model even for Trial Sequence in terms of validation performance, which confirms the merit of ensemble technique. Given the validation AUC and accuracy we chose the stacked model for One-row-per-pair and lightGBM alone for Trial Sequence.

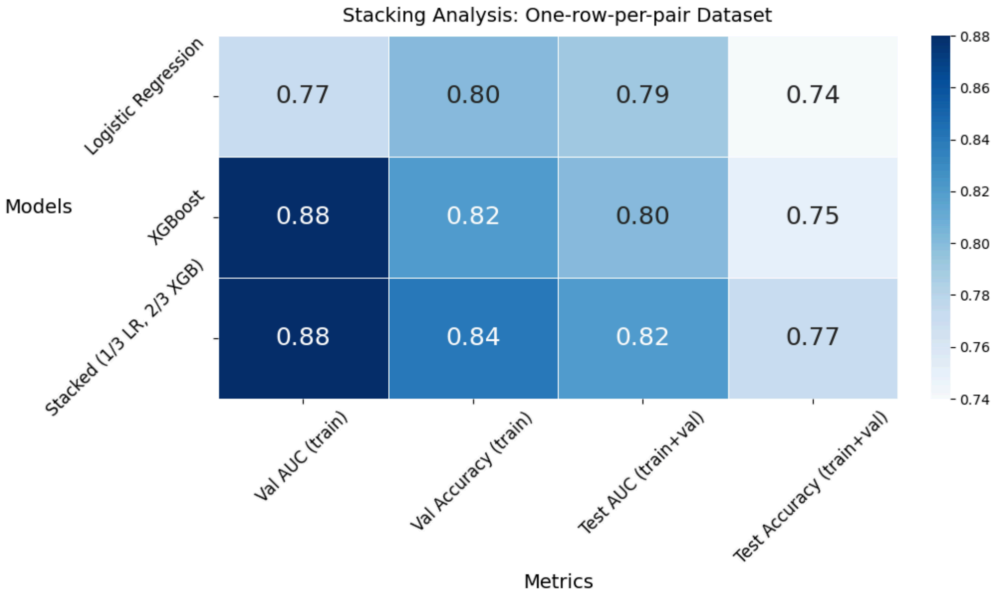


Figure 15: Stacking analysis heatmap of models used and trained on the One-row-per-pair dataset.

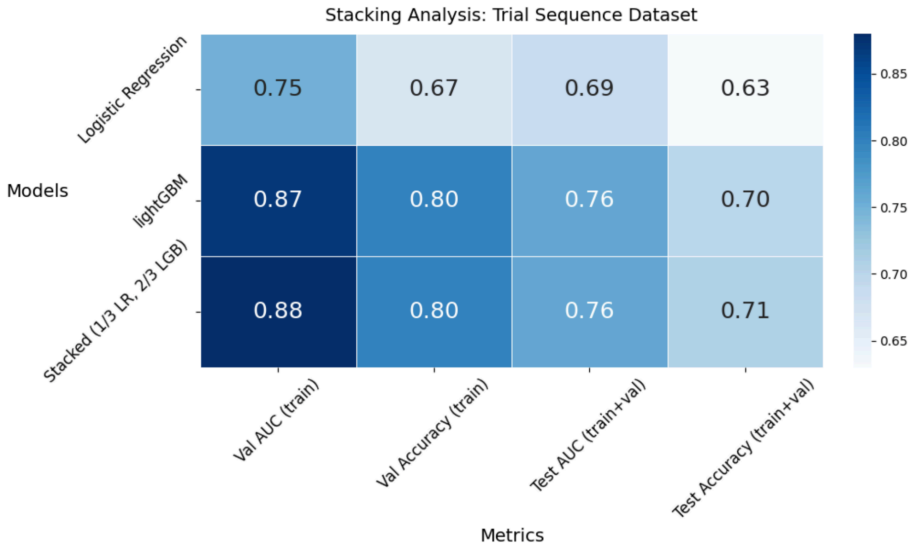


Figure 16: Stacking analysis heatmap of models used and trained on the Trial Sequence dataset.

4.5.3 | Features

For our final models and analysis, we included the following key features:

Feature	Description
Enrollment	Number of participants enrolled in the trial.
Actual Duration	Duration of the trial (in days).
Number of primary outcomes to measure	Number of primary outcomes being measured in the trial.
Number of secondary outcomes to measure	Number of secondary outcomes being measured in the trial.
Number of arms	Number of arms (groups) in the clinical trial.
Adult **Child**	Binary indicators of whether the trial included adults and/or children.
Has Single facility **Has US facility**	Binary indicators of trial facility location.
Healthy Volunteers	Indicates inclusion of healthy volunteers.
Gender	Binary indicators for gender representation in the trial.
Min Max Age	Minimum and maximum age of participants.
Number of facilities	Number of facilities conducting the trial.
GPT 4o Human Importance Rating	Large langue model metric for the clinical trials description and its relationship to humans.
LLama3.2 Criteria Robustness	Small language model metric generated for the robustness of the inclusion/exclusion criteria.
Spacy pregnant women excluded	Binary indicator if pregnant women were excluded from the trial.
Days since first start	Time, in days, elapsed since the start of the trial.

Feature	Description
Cumulative sum enrollment **Cumulative average enrollment**	Cumulative sum and average enrollment numbers across the trials until the latest trial's completion date.
Cumulative sum trial duration **Cumulative Average trial duration**	Cumulative sum and average trial durations.
Pair trial sequence	Sequence number of the trial in its associated DIP
New therapy area*	Binary encoding for therapeutic areas (e.g., Oncology, Rheumatology).
Drug outcome	The target variable representing trial success or failure.

Table 7: Features used in our final model

Further details in methodology

The columns Disease type and Trial drug were utilized in the methodology to establish a robust linkage between clinical trials and drug-indication pairs. The disease type column was critical for identifying the therapeutic area targeted by each clinical trial, ensuring alignment between the trial's purpose and the condition being addressed. Similarly, the trial drug column was used to standardize drug names, enabling consistent mapping across trials and facilitating the evaluation of drug-specific outcomes. These columns were essential in forming the foundational relationships required for subsequent analyses, such as assessing drug performance or trial alignment within specific disease categories.

Missing features

Despite the efforts to incorporate a comprehensive set of features, some key attributes identified in the literature review and outlined in section 3.2.1 could not be included due to the unavailability of relevant data. For drug-specific features, we were unable to attain details on the biological target, mechanism of action, drug origin, and therapeutic class, which could have provided deeper insights into the pharmacological basis of trial success. In terms of clinical trial features, we lacked granular information on trial design elements such as randomization, blinding, and placebo control, as well as detailed outcomes beyond success or failure, including primary endpoints and exclusion criteria. Additionally, sponsor-related metrics, such as track records and prior trial success rates, were not accessible, limiting our ability to analyze trial sponsorship as a factor. These omissions reflect gaps in the available data but highlight areas for potential expansion in future studies.

4.5.4 | Hyperparameter Tuning

We did not use cross-validation because our study requires very specific subset splitting. Instead, we employed a simple iterative process, adjusting hyperparameters and assessing performance on a predefined validation set as discussed in 4.2.1 (train-test split). This approach allowed us to systematically refine the models within the constraints of the dataset. Additionally, we set the random seed to 42 across all tuning processes to ensure consistency and reproducibility.

Here are the specific hyperparameters that that were believed to have higher priority and tuned for each model:

- **Random Forest**: We tuned the n_estimators and max_depth within a specified range, adjusting the depth with a step size. Additionally, we set the maximum depth based on the square root of the number of features, as it is a classification task.
- **XGBoost**: We tuned the max_depth, learning_rate, and min_child_weight for the model.
- **LightGBM**: We tuned the num_leaves, learning_rate, and max_depth for the model.
- **H2O GBM**: We tuned the ntrees, max_depth, and learn_rate for the model.
- **Logistic Regression**: We tuned the C parameter, which controls the regularization strength, and also set the penalty to Ridge (L2 regularization).

4.5.5 | Feature Importance and Coefficients

Since we proposed the stacked model for One-row-per-pair and lightGBM alone for Trial Sequence as the final models, we have one feature importance plot for each processed dataset and one coefficient plot for One-row-per-pair.

Based on the two feature importance plots, generally the positively cumulative fields like enrollment in One-row-per-pair and cumsum_enrollment in Trial Sequence tend to be the most important features. Meanwhile neither trial count nor pair trial sequence is the top one in their own plot. So generally the number of clinical trials done in the past is critical information to predict the approval decisions while value is limited by only looking at clinical trial counts. This is confirmed by non-cumulative fields like avg_days_since_first_start and cumavg_enrollment also being the top important features.

In addition, Oncology is the top specific therapy area feature and it adds moderate value to the ensemble trees' predictive power. The language model derived features also add moderate value.

The linear coefficient plot for One-row-per-pair is sorted by coefficient absolute values. Total days since first start has the largest influence in the model prediction and generally the more days since the first start the more likely the DIP is predicted to be approved. Enrollment also has a big positive coefficient. Meanwhile, a high number of total facilities and therapy area being Oncology make the DIP less likely to be predicted to be approved.

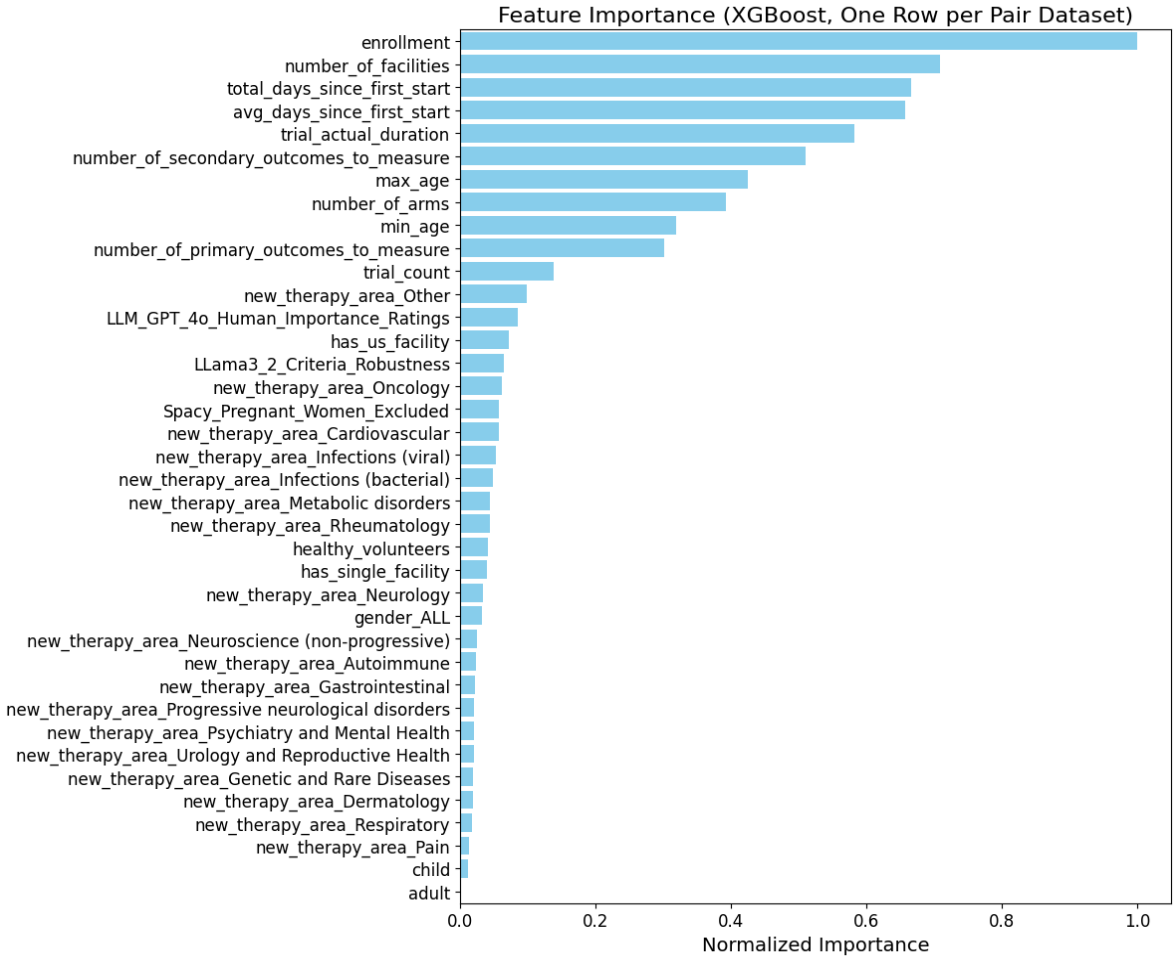


Figure 17: Feature importance ratings of XGboost model trained on one-row-per-pair dataset.

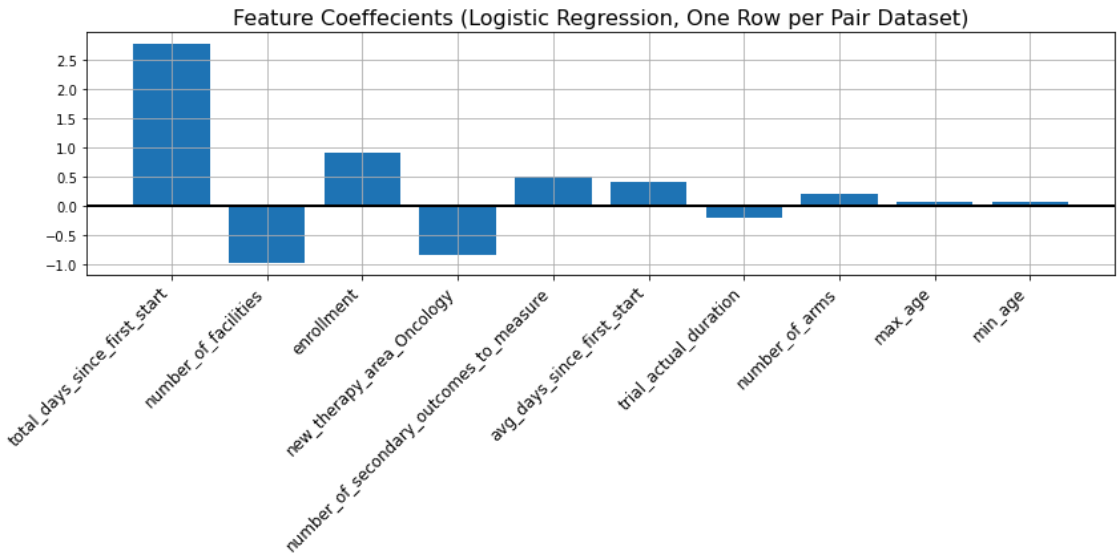


Figure 18: Feature coefficients of Logistic regression model trained on one-row-per-pair dataset.

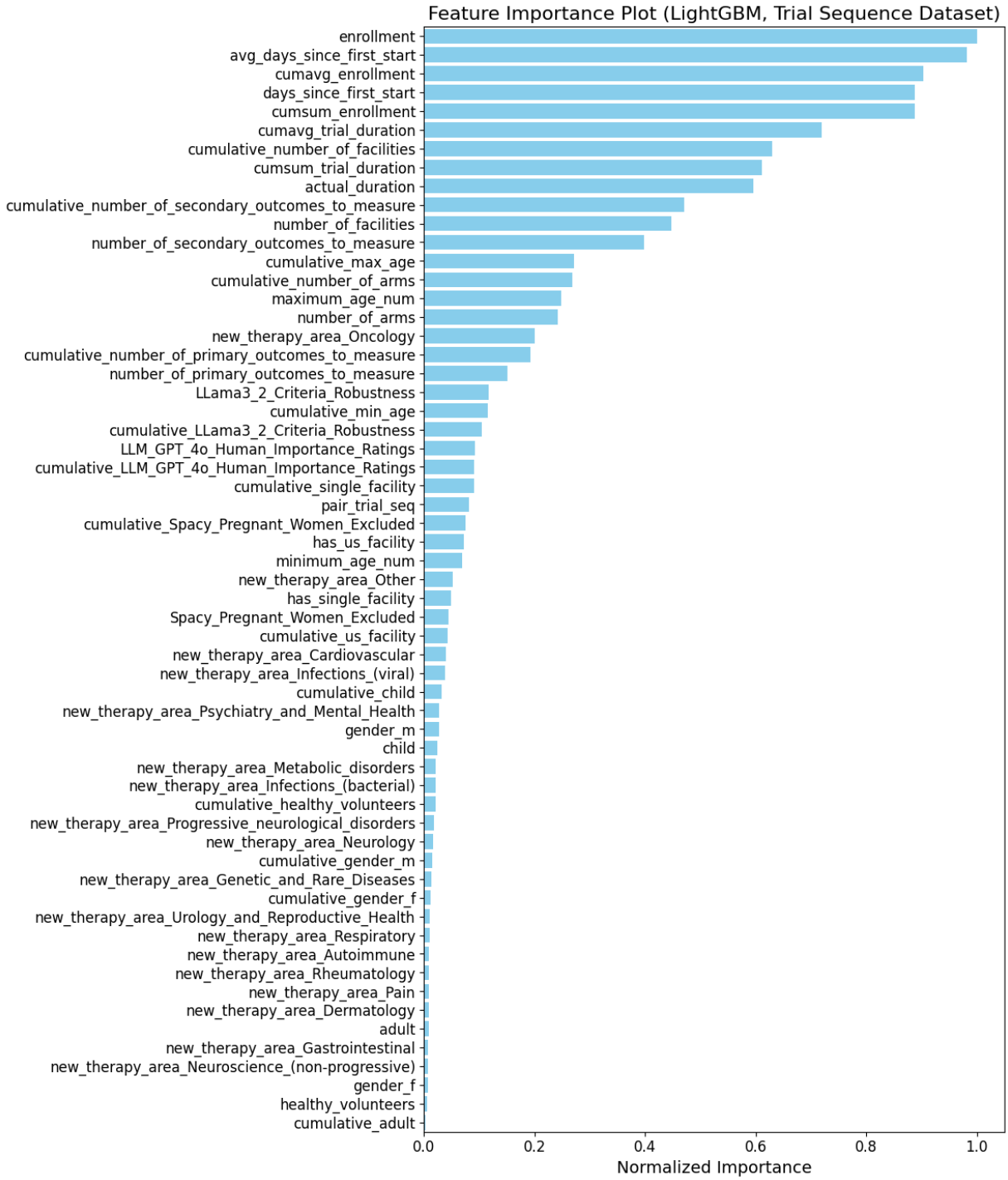


Figure 19: Feature importance ratings of LightGBM model trained on train sequence dataset.

4.5.6 | Performance on the Trial Sequence dataset at n-th clinical trial

Since we introduced the concept of Trial Sequence, it's interesting to evaluate the test performance of the final model at each sequence number. For example, does the model perform better at the first

completed clinical trials of the DIPs than the fourth trial in the sequence? We used the final lightGBM model for this specific analysis.

Note we had a small testing set so the first step was to look into the count of data points at each trial sequence number. We could see the count of data points decreased quickly after each sequence number which is intuitive. Meanwhile, it makes sense to group sequence number 5, 6, 7 and 8 together to reduce the noises in evaluation. In the two bottom plots of Figure 20, sequence number 5 means the aggregated group of 5, 6, 7 and 8.

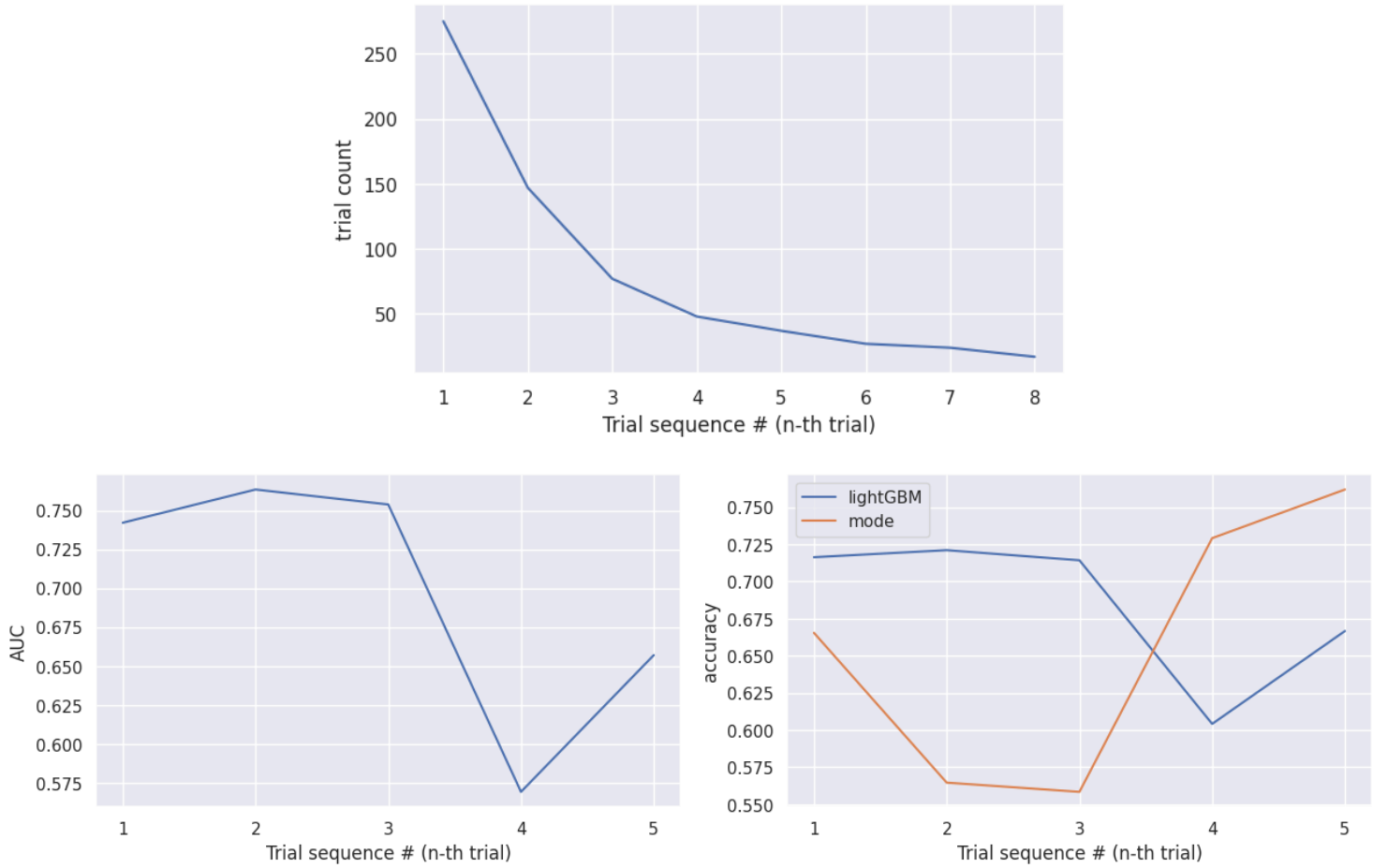


Figure 20: The top middle plot shows the trial count on the x-axis and the trial outcome (trial count) on the y-axis, showing that the trial outcome tends to decrease as more trials are completed. The bottom right plot displays the AUC over the trial sequence number. The bottom right plots focus on the model's accuracy metric, indicating that the reference mode model generally outperforms LightGBM as the trial sequence progresses.

The plot of AUC per sequence number reveals that the final model performs significantly better in the first three DIP trials compared to the fourth trial and those beyond the fourth.

Furthermore we also analyzed the accuracy metrics, and as contrast we introduced a reference mode model. The reference model uses the mode of the target variable in the combined training and validation dataset as the prediction result. For example if most of the first trials are failures in the combined training and validation dataset, then the predicted outcomes for the first trials in the test set are always disapproved. Again the final lightGBM model is beating the reference mode model for the first three trials of DIPs but not in the later trials. This in some degree also confirms the value of the preprocessing where we excluded the clinical trials later than the eighth of a DIP. We hope to have better performing models for larger trial sequence numbers by having higher quality datasets in the future.

Chapter 5: Conclusions & Future Research

This study demonstrates predictive modeling to aid in Evergrowth BioHealthcare Capital's needs for forecasting FDA drug approvals based on Phase 2 clinical trial data. By leveraging advanced machine learning techniques and natural language processing, the model addresses the inefficiencies and limitations of the existing manual approach. The integration of heterogeneous datasets from AACT and Cortellis, coupled with meticulous data preprocessing to mitigate issues such as data leakage and bias, ensured the robustness and reliability of the predictive outcomes. Although we made progress in extracting meaningful variables and found that certain language model-driven features contributed to feature importance rankings, their overall impact on prediction accuracy was more modest than anticipated.

5.1 | Final Model Results & Conclusions

The final models demonstrated commendable accuracy, achieving up to 77% on the One-row-per-pair dataset and 71% on the more intricate Trial Sequence dataset. These results underscore the models' potential to enhance investment decision-making by providing timely and scalable predictions. Key features contributing to the model's performance included the number of patients enrolled, days since trial initiation, the number of trial facilities, trial duration, and the number of secondary outcomes measured. These factors align with the critical aspects highlighted in the abstract, emphasizing their significance in predicting FDA drug approvals.

Additionally, the incorporation of language model-driven features, though offering only moderate improvements, contributed to a more nuanced understanding of trial descriptions and criteria. For example, one human-importance derived feature ranked 23rd and 24th cumulatively for the LightGBM trial sequence model, while a feature extracted using a small Llama3 variant ranked 20th and 22nd, and an NLP-based pregnancy exclusion feature ranked 32nd and 27th in cumulative importance for the same model. In the XGBoost with One-row-per-pair setup, the human importance feature improved to 13th, the small Llama3 criteria robustness feature reached 15th, and the NLP pregnancy feature also ranked 15th. While these rankings indicate that language model-derived features provided some value, they did not yield the substantial predictive improvements we had hoped for, due in part to the limited time for experimentation and optimization.

Overall, the research demonstrates that a systematic, data-driven approach can significantly improve the prediction of FDA drug approvals. This advancement not only streamlines the investment analysis

process but also provides a scalable and efficient alternative to traditional manual methods, thereby offering substantial value to Evergrowth BioHealthcare Capital.

5.2 | Future Research

Future research should first focus on enhancing data quality, breadth, and granularity. Obtaining more reliable DIP-level approval statuses, higher-quality indication fields, and additional attributes—such as biological targets, mechanisms of action, sponsor track records, and detailed trial methodologies—would yield a richer dataset. Increasing the quantity of data points through academic collaborations, commercial licenses, or tailored extraction scripts would further strengthen the modeling process. Efforts should also be directed toward refining domain-specific language models by providing more systematic training, using advanced architectures, and integrating textual information more carefully, all of which could result in greater predictive gains.

Beyond these data-driven improvements, integrating dynamic, real-time data and external sources—such as ongoing trial updates, regulatory changes, and financial or patent-related indicators—could enhance the model's timeliness and applicability. Additional benefits might emerge from exploring transfer learning and model ensembling strategies to improve generalizability and robustness across diverse therapeutic areas. In parallel, it will be important to increase model explainability. Applying interpretability techniques like SHAP or LIME and visualizing feature importance would make the model's predictions more transparent and thus more trustworthy for stakeholders such as investors, regulatory specialists, and industry partners.

In terms of methodological refinements, future work might include experimenting with alternative trial-level aggregation methods, explicit feature reduction techniques, and more comprehensive parameter tuning. Employing realistic point-in-time imputation methods and advanced sequence modeling techniques, such as recurrent neural networks or Temporal Fusion Transformers, would allow a more accurate capture of temporal dynamics. Finally, it may be necessary to revisit filtering rules, trial sequencing thresholds, and other data inclusion criteria. Engaging with biotech investors and clinical trial experts can guide such decisions, ensuring that the model reflects practical investment scenarios and domain-specific considerations, ultimately increasing its reliability and utility.

Project Code GitHub Repository: https://github.com/GusGitMath/AI_Powered_Drug_Approval

References

[1] Commissioner, O. of the. (n.d.). *About FDA*. U.S. Food and Drug Administration. Retrieved from https://www.fda.gov/about-fda

[2] Drugs.com. (n.d.). *FDA drug approval process*. Retrieved from https://www.drugs.com/fda-approval-process.html

[3] Lahiri, K., & Yang, L. (2013). Forecasting binary outcomes. In G. Elliott & A. Timmermann (Eds.), *Handbook of economic forecasting* (Vol. 2, Part B, pp. 1025–1106). Elsevier. https://doi.org/10.1016/B978-0-444-62731-5.00019-1

[4] Prader-Willi Syndrome Association (PWSA) USA. (2021). *Clinical trials presentation: March 27, 2021*. Retrieved from https://www.pwsausa.org/wp-content/uploads/2021/03/Clinical-Trials-Presentation-March-27-2021.pdf

[5] Center for Drug Evaluation and Research. (n.d.). Drug Development Process. U.S. Food and Drug Administration. https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process

[6] Seeking Alpha. (2011, April 14). *Why phase II trials are the best time for biotech investing*. Nasdaq. Retrieved from https://www.nasdaq.com/articles/why-phase-ii-trials-are-best-time-biotech-investing-2011-04-14

[7] Clinical Trials Transformation Initiative. (n.d.). *Improving public access to aggregate content of ClinicalTrials.gov: AACT database*. Retrieved from https://aact.ctti-clinicaltrials.org/

[8] Clarivate. (2024, October 30). *Cortellis Drug Discovery Intelligence Platform*. Life Sciences & Healthcare. Retrieved from https://clarivate.com/life-sciences-healthcare/research-development/discovery-development/cortellis-pre-clinical-intelligence/

[9] DrugBank Online. (n.d.). *Database for drug and drug target info*. Retrieved from https://go.drugbank.com/

[10] Lipsky, M. S., & Sharp, L. K. (2001). From Idea to Market: The Drug Approval Process. Journal of the American Board of Family Practice.

[11] BIO, QLS Advisors, Informa UK Ltd. (2021). Clinical Development Success Rates and Contributing Factors 2011–2020.

[12] Lo, A. W., Siah, K. W., & Wong, C. H. (2019). Machine Learning With Statistical Imputation for Predicting Drug Approvals. Harvard Data Science Review.

[13] DiMasi JA, Hermann JC, Twyman K, Kondru RK, Stergiopoulos S, Getz KA, Rackoff W. A Tool for Predicting Regulatory Approval After Phase II Testing of New Oncology Compounds. Clin Pharmacol Ther. 2015 Nov;98(5):506-13. doi: 10.1002/cpt.194. Epub 2015 Sep 24. PMID: 26239772.

[14] Siah KW, Kelley NW, Ballerstedt S, Holzhauer B, Lyu T, Mettler D, Sun S, Wandel S, Zhong Y, Zhou B, Pan S, Zhou Y, Lo AW. Predicting drug approvals: The Novartis data science and artificial intelligence challenge. Patterns (N Y). 2021 Jul 21;2(8):100312. doi: 10.1016/j.patter.2021.100312. PMID: 34430930; PMCID: PMC8369231.

[15] Beinse, G., et al. (2019). Prediction of Drug Approval After Phase I Clinical Trials in Oncology: RESOLVED2. JCO Clinical Cancer Informatics.

[16] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics*, Volume 36, Issue 4, February 2020, Pages 1234–1240, https://doi.org/10.1093/bioinformatics/btz682

[17] Dubey, Abhimanyu, et al. "The llama 3 herd of models." *arXiv preprint arXiv:2407.21783* (2024).

[18] Me LLaMA: Foundation Large Language Models for Medical Applications. (2024). *arXiv preprint*. Retrieved from https://arxiv.org/abs/2402.12749

[19] OpenAI. “TrialGPT: Matching Patients with Clinical Trials Using GPT-4.” arXiv, 2023. Available: https://arxiv.org/abs/2307.15051.

[20] OpenAI. “GPT-4 for Automated Report Generation in Clinical Trials.” medRxiv, 2024. Available: https://www.medrxiv.org/content/10.1101/2024.02.08.24302376v1.full.pdf.

[21] Zhang, Y., et al. (2024). *CTP-LLM: Predicting Clinical Trial Progression Using Language Models.* Retrieved from arXiv:2408.10995.

[22] Google Research. (2023). Tx-LLM: Supporting therapeutic development with large language models. Retrieved from https://research.google/blog/tx-llm-supporting-therapeutic-development-with-large-language-models

[23] AE-GPT Model for AE Detection: PLOS One article discussing the performance of the AE-GPT model in extracting adverse events from VAERS data. Retrieved from https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0300919

[24] FDA BERTox Initiative: FDA webpage highlighting BERTox's application of BERT-based NLP techniques for classifying drug-induced liver injury risks. Retrieved from https://www.fda.gov/about-fda/nctr-research-focus-areas/bertox-initiative

[25] UCSF BERT for Serious Adverse Events: MedRxiv study showcasing accuracy improvements using UCSF BERT for identifying serious adverse events. Retrieved from https://www.medrxiv.org/content/10.1101/2023.09.06.23295149v1

[26] Dandan Wang & Shiqing Zhang et al. (2024). Large language models in medical and healthcare fields: applications, advances, and challenges https://link.springer.com/article/10.1007/s10462-024-10921-0

[27] Mullard A. 2022 FDA approvals.Nature Reviews Drug Discovery.. 2023 Feb;22(2):83-88. doi: 10.1038/d41573-023-00001-3. PMID: 36596858.

[28] Sacks LV, Shamsuddin HH, Yasinskaya YI, Bouri K, Lanthier ML, Sherman RE. Scientific and Regulatory Reasons for Delay and Denial of FDA Approval of Initial Applications for New Drugs, 2000-2012. *JAMA*.2014;311(4):378–384. doi:10.1001/jama.2013.282542

[29] U.S. Department of Health and Human Services, Food and Drug Administration. (2024). Analysis of US Food and Drug Administration new drug and biologic approvals, regulatory pathways, and review times, 1980–2022.

[30] *NCI Dictionary of Cancer terms*. Comprehensive Cancer Information - NCI. (n.d.). https://www.cancer.gov/publications/dictionaries/cancer-terms/def/phase-4-clinical-trial

[31] Thiele LR, Spong CY. Inclusion of Pregnant and Lactating People in Clinical Research: Lessons Learned and Opportunities. Obstet Gynecol Clin North Am. 2023 Mar;50(1):17-25. doi: 10.1016/j.ogc.2022.10.001. PMID: 36822702.

[32] Shields, Kristine & Lyerly, Anne. (2013). Shields KE, Lyerly AD. Exclusion of pregnant women from industry-sponsored clinical trials. Obstetrics & Gynecology 2013; 122(5):1077-1081.. https://pubmed.ncbi.nlm.nih.gov/24104789/

[33] Rosati, D. (2023, January 31). ClinicalTrialBioBert-NLI4CT. *Hugging Face*. computer software. Retrieved 2024, from https://huggingface.co/domenicrosati/ClinicalTrialBioBert-NLI4CT/tree/main.

[34] Natural Language Inference for Clinical Trial (NLI4CT) Dataset. Available at: https://paperswithcode.com/dataset/nli4ct.. Accessed December 8, 2024.vation