ID Cover Page

Summary of WP Student Team

Predictive Modeling for Clinical Trial Completion:

Assessing the Phase Success

Group constitution:

Student Name	Program	Individual Title
Erica Francalanci	Business Analytics	Factors Driving Predictive Outcomes
Ginevra Rossi	Business Analytics	Enhancing Trial Predictions through Uncertainty Quantification
Jasmin Hamrouni	Business Analytics	Incorporating RAG Techniques for Predictive Modeling of Clinical Trial Outcomes
Júlia Zhou Cheng	Business Analytics	Incorporating Missing Value Imputation
Sara Sofia Almeida Favita	Business Analytics	A What-if Scenario Approach on the Enrollment

Work project carried out under the supervision of:

Advisor: Qiwei Han

Reading observations for jury: If applicable

Note: this template is a support document for the jury members before the defense, only.

A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

Predictive Modeling for Clinical Trial Completion: Assessing the Phase Success

Erica Francalanci, Ginevra Rossi, Jasmin Hamrouni, Júlia Zhou Cheng, Sara Sofia Almeida Favita

Work project carried out under the supervision of:

Qiwei Han

17/12/2024

Abstract (100 words maximum)

This study investigates predictive modeling of clinical trial completion using the *HINTBasic* and *HINTPlus* models. By integrating multimodal datasets, the models predict clinical trial phase success. It provides interpretability insights into the *HINTPlus* model's decision-making process. Imputation methods are incorporated to improve data consistency and model performance. To enhance reliability, a selective classification technique addresses uncertainty quantification, and a what-if analysis evaluates the impact of increased enrollment on trial outcomes. Retrieval-Augmented-Generation techniques were used to contextualize results. Our findings support informed decision-making, optimize resource allocation, and accelerate drug development in clinical trials.

Keywords (Clinical Trials, Health Care, Artificial Intelligence, Machine Learning Methods, Predictive Modeling, Missing Value Imputation, Model Interpretability, Selective Classification, What-if Analysis, Retrieval-Augmented-Generation)

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22

Index

1. Introduction

20 million new cases and 9.7 million deaths. These are the numbers reported by the World Health Organization (WHO)'s cancer agency (2024), describing the state of the growing worldwide burden of cancer in 2022. Their analysis also shows that 1 in 5 people develop cancer in the span of their lives. A look into the future suggests that over 35 million new cases are predicted in 2050, which is a 77% increase in comparison to the estimated 20 million cases in 2022. According to Dr. Adams, the Head of the Union for International Cancer Control, governments need to facilitate cancer care, so that all humans have access. Yet he also highlights the progress which has already been made in early detection of cancer as well as their treatment. One way to access the newest, most innovative treatments is through participation in clinical trials. For example, Singh et al. (2016) demonstrated the impact of clinical trials in cancer treatment, a type of blood cancer called acute lymphoblastic leukemia, which led to remission rates of 90%. To understand clinical trials in a broader way, we will now dive deeper inter their inner workings, as well as their impact on the market.

Clinical trials are a multi-phase process designed to assess the safety and efficacy of new drugs or treatment modalities before the official approval by the authorities for public use (Hay et al, 2014).

Clinical trials are an essential element of the drug development process, and their importance cannot be overstated, as they stand as the gold standard in medical research (Ghim & Ahn 2023). Despite their critical role, the process of conducting a trial carries a long list of inherent complexities, high costs, and time-consuming nature.

Indeed, the strict regulatory compliance (dictated by regulatory bodies such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA)) imposes rigorous standards to ensure both safety and reliability. The regulatory landscape additionally involves the adherence to Good Clinical Practice (GCP) guidelines and the observance of ethical considerations such as informed consent, patient privacy, and the use of placebos (National Institute of Health, World Health Organization).

Clinical trials are typically divided into Phases I, II, III, and IV. Phase I trials primarily focus on determining the safety of a new drug and identifying potential side effects, often involving a small group of healthy volunteers or patients. Around 20-80 patients are involved in this phase (National Institute of Health). However, only 70 % of these patients will progress to the next stage in 3-6 months. Phase II trials expand the focus to evaluate the efficacy of the drug and further assess its safety, usually involving a larger group of patients with the target condition. Phase III trials aim to confirm the drug's efficacy, monitor side effects, and compare it to existing standard treatments, involving a much larger patient population to ensure statistically significant results (Chopra et al 2024, 4212). In fact, Phase III is subject to 10 times more participants in comparison to Phase II, to establish efficacy, monitor adverse effects as well as compare the results to other treatments (National Institute of Health; Getz et al. 2016). This process has a duration of 1 to 4 years and only one fourth of participants will progress to the next phase (Chopra et al. 2024, 4212).

Phase IV trials are completed trials that study the side effects caused over time by a new treatment after it has been approved and is on the market. These could also be referred to as post marketing surveillance trials. The aim is to look for side effects that have not been detected in earlier phases of trials and may also assess the performance of a new treatment over a long period of time (National Cancer Institute). This phase might take a year or more to come to an end and it has a success rate of 70-90% (Chopra et al. 2024, 4212).

This essential process is also one of the most resource-intensive components of drug development in terms of both financial and time perspective. Indeed, the average time to bring a new drug to market exceeds a decade, while the total development is estimated to be $2.6 billion, with clinical trials accounting for a major part of these expense (DiMasi et al. 2016).

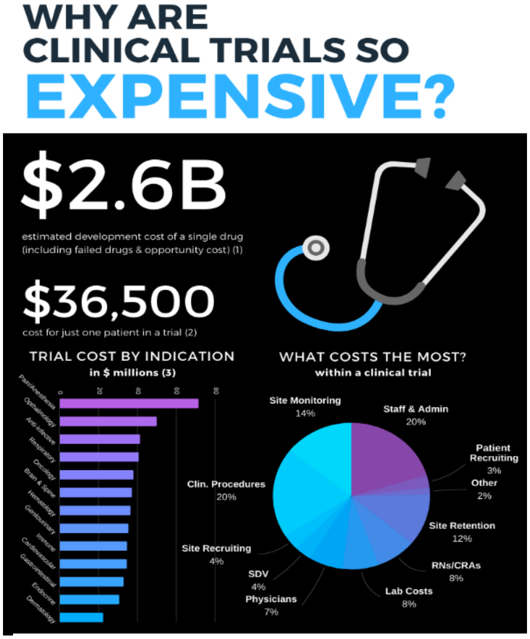


Figure 1 - Clinical Trials Cost distribution (Clinical Research IO 2018)

1.1 Recent Industry Insights

With regards to the market trends, it can be stated that the clinical trials field is expected to be thriving in the next decade, and major investments are being placed to favor the development of AI practices aimed at achieving improved efficacy and reducing expenditure. Recent industry reports provide valuable insights related to the current state and future projections of the industry.

An industry report by the Global Market Insights from 2024 has shown that the global clinical trials market has been valued at approximately $ 55.8 billion in 2023, with an annual growth rate (CAGR) of 5.4 %. This will lead to it reaching a value of around $ 89.8 billion by 2032. (2024) have estimated that the global clinical trial market has been valued at $ 57.76 billion in 2023, and it is expecting a CAGR of 7.1 %, therefore reaching $ 106.78 billion by 2032. (Faizullabhoy et al. 2024)

A closer look at clinical trials suggests that Phase III trials dominate the market, accounting for 53.3 % of the total revenue share in 2023. Phase II trials are also significant, with the segment projected to lead the market, accounting for the largest revenue of $ 23.4 billion in 2023 (Grand View Research 2024). A huge priority in treatments is oncology as it remains a primary focus, with a substantial number of trials initiated in this area. Other significant therapeutic areas include central nervous system disorders, cardiology, and infectious diseases (Grand View Research 2024).

The market also experiences a shifting focus when it comes to countries and regions in which these clinical trials are carried out. Europe's share of global commercial clinical drug trials shows a significant decrease, declining from 22 % in 2013 to 12 % in 2023. On the contrary, China has doubled its share, accounting for 18 % of the global total, while the US remains the leader despite a slight decrease. This might be explained by pharmaceutical companies favoring US and China for their more straightforward regulatory environments (Financial Times 2024).

1.2 Limits of Clinical Trials

While clinical trials are crucial for drug development, they are also limited by several constraints, primarily related to financial resources and time.

From an historical point of view, the cost of clinical trials has grown significantly over the past few decades. Specifically, it doubled every nine years since the 1950s, following a trend described by the Eroom’s Law, the reverse of Moore’s Law (Scannell et al. 2012).

Furthermore, Phase III clinical trials account for most of the expenditure, with cost spanning between $11 million and $52 million, depending on the therapeutic area and the study requirements (Mestre-Ferrandiz et al. 2012).

Another element that could enhance the overall financial burden is represented by the large-scale patient recruitment and the need for extensive data collection (Getz et al. 2016).

In terms of duration, clinical trials are notoriously lengthy. For instance, the average duration of each phase is approximately 6 or 7 years, even though later phases often exceed this estimation; hence, the total development process timeline exceeds the decade and could reach 15 years (Wouters et al. 2020). Likewise, the process of patient recruitment is another major challenge. Finding suitable candidates who meet all the eligibility criteria is extremely time-consuming and maintaining participant retention throughout the duration of a trial can be equally difficult. The process of patient recruitment often involves identifying individuals who meet strict inclusion and exclusion criteria, which may include factors such as age, gender, medical history, and current health status. This can be further hindered by a lack of awareness or understanding of clinical trials among potential participants, as well as logistical challenges such as travel requirements, and the time commitment involved in participating in a trial (Bieganek et al. 2022). This is enhanced in the case of trials focused on rare disease or demographic-specific studies, where the pool of eligible participants is already limited. Moreover, it is to be emphasized that the relevant failure rate of clinical trials. Failure root causes may include a broad variety of diverse factors, which range from the lack of efficacy or unforeseen safety issues to difficulties in managing the complexities of trial design (Getz et al. 2016; DiMasi et al. 2016).

1.3 Literature Review and Limitations of Existing Models

Artificial Intelligence has the power to be a real shift in the medical field as it could be able to accelerate the efficiency, economy and timeliness of drug development. In recent years, there have been notable advancements in artificial intelligence (AI) and machine learning, which have brought innovative techniques to the field of clinical trials, offering solutions to overcome challenges (Chopra et al. 2024).

AI-driven models have proven their ability to optimize various aspects of trials processes, such as patient recruitment, outcome prediction, and synthetic data generation. Specifically, the integration of Large Language Models (LLMs) has meaningfully facilitated the process by analyzing large datasets, thereby improving trial design and forecasting outcomes (McKinsey 2024).

The models developed in latest years, such as those developed by Kavalci & Hartshorn (2023), have delivered promising methodologies for predicting trial success and preventing early terminations. AI practices also suggested hypothetical resolutions to address the challenges of patient recruitment and retention. Indeed, through the examination of electronic health records (EHRs) and other real-world data, AI models can be leveraged to identify participants who fulfil eligibility criteria for a given trial, thus streamlining the recruitment process. Their predictive ability could also be employed to forecast patient drop-out rates and detect factors that may influence retention at an early stage, therefore letting trial organizers take proactive measures to preserve high engagement. Similarly, AI-driven approaches have been hired with the purpose of enhancing trial efficiency, by simulating diverse scenarios in which parameters such as dosage, sample size, and endpoint selection have been varied accordingly. These techniques evidenced their capability to shorten trial timelines and improve the overall efficiency of the drug development process (Lu et al. 2024). According to a study conducted by McKinsey in November 2023, the integration of AI into clinical studies might potentially reduce trials timeline by 15-30%.

Despite the enormous advancements, the models proposed thus far also have several limitations. A primary obstacle lies in the reliance on large amounts of high-quality data, which is often not available. For instance, Large Language Models, while highly effective, require extensive labelled datasets to accomplish accurate predictions, hence their applicability is only feasible in areas where data is consistent and reliable. Similarly, as many of the existing models are based on binary outcomes (trials are classified as either successful or unsuccessful) the complexity of clinical research is not fully captured, resulting in a more nuanced interpretation that does not underscore determinant factors for success or failure (Topol 2019; Yu et al. 2018). Another limitation of existing AI models might be found in the lack of transparency and interpretability, since numerous AI-driven models, especially those relying on deep learning techniques, act as hidden frameworks and do not clearly show how predictions are made. This might pose as a barrier, blocking these models from being adopted in clinical trials, as regulatory bodies and stakeholders require a clearer understanding of the factors affecting trial outcomes. Still, the models currently in use recurrently fail to achieve generalizability, thereby their performance firmly depends on the context (Topol 2019; Yu et al. 2018).

According to Chopra et al. (2024), understanding and validating seems to be demanding, guidelines and standards for AI-driven healthcare solutions are being developed by authorities such as the FDA. However, this takes a huge amount of time, since this requires a long testing and validation phase, while also requiring explanation of the underlying algorithms. Deep learning neural networks are often called “black box” models because of their high level of difficulty, making it harder to understand. This complexity as well as its unclear interpretability of its result can also be a reason why mainstream adoption slows down. As a result, it might also be more challenging to get patients’ permission to use their data, due to the lack of understanding of what the data is being used for. Consequently, it is vital that AI models are transparent and easy to understand. Upholding a high patients’ confidence and a high standard of ethical behavior also holds when it comes to data privacy and data ownership to avoid unjust treatment. This might be the case when trials are using face recognition software or other methods to track if participants are sticking to the rules of the study. For that reason, it is important that researchers using AI and ML algorithms provide their participants with an explicit explanation of the risks and advantages of their data gathering (Chopra et al. 2024, 4212-4216).

Another closely related limitation is the inherent bias in the data used. Trained Algorithms on a specific dataset are at risk of excluding large parts of the populations which have not been included. This is due to a serious lack of diversity in medical research, which can be traced back to research only including able bodied white people as the default. As a result, an AI trained model would not be able to have enough knowledge of underrepresented groups such as people of color and patients from lower socio-economic backgrounds, leading to biased findings that do not translate to the not included groups. One way to avoid this happening is if health equity consideration are implemented in AI and ML systems. Another way to treat this problem is the inclusion of domain experts such as medical professionals, which are consulted in the algorithm development process to help fill in missing values which are vital for providing context to the dataset (Chopra et al 2024, p.4218).

The limitations of traditional clinical trials, combined with the partial resolutions offered by existing AI models, emphasize the need for more sophisticated methodologies that can address these challenges comprehensively. There is a clear demand for multi-function approaches that could serve different purposes, such as integrating diverse data sources, providing interpretable insights, and generalizing adaptability across different trial settings, thereby proposing a more holistic solution to the issue faced by clinical trials.

1.4 Proposed Model: Hierarchical Interaction Network (*HINT*)

This thesis aims to address gaps identified in the current literature by implementing and furtherly refining the Hierarchical Interaction Network (*HINT*) model by Fu et al. (2022), which has demonstrated a strong potential in the prediction of clinical trial outcomes. *HINT* represents a novel approach that is characterized by the integration of a diverse dataset, including drug molecular structures, disease-specific information, trial eligibility criteria and pharmacokinetic interactions, to accurately predict the success or failure of clinical trials' outcomes across different phases. Unlike traditional models, *HINT* leverages a hierarchical interaction graph to display the complex relationships among these components, providing a better comprehensive perspective of the influential factors.

Fu et al.'s (2022) methodology employed comprises several key components. First, multimodal data embedding is used to encode different types of trial-related information. The embeddings obtained are then combined with external knowledge sources, including pharmacokinetic data, to reach a deeper understanding of drug interactions and their effects on trial outcomes. To advance the progress of ML models in the context of clinical trials, *HINT* introduces TOP (Trial Outcome Prediction), a benchmark dataset, containing all the data sources used in the development of the model, providing a comprehensive resource for training and evaluation purposes.

The objective of this thesis is to develop an effective model for predicting the factors that influence the success or failure of clinical trials by leveraging *HINT*. We are defining success as the clinical trial reaching the end of its phase. This is done by collecting data, that has been used by Fu et al. (2022) for their benchmark, to recreate an updated and enriched version of TOP, which will then be leveraged to replicate the *HINT* model.

The results are expected to provide valuable insights into how AI-driven models can be used to improve clinical trial efficiency, reduce costs, and ultimately accelerate the approval of new therapies. By providing accurate predictions of trial outcomes, *HINT* could help pharmaceutical companies make more informed decisions about which drug candidates to advance, thereby reducing the risk of costly late-phase failures. Furthermore, the ability to predict trial outcomes at an early stage could enable more efficient allocation of resources, such as patient recruitment efforts and financial investments, ultimately making the entire drug development process more cost-effective.

Our work is structured as follows: firstly, a literature review provides a deep dive into the *HINT* model, which is followed by an evaluation of existing research, methodology and their limitations, addressing related research questions. Secondly, the data collection and preparation encoding process is explained, which is followed by the performance of an Exploratory Data Analysis (EDA) containing insights as well as patterns and trends. This is followed by the model's learning phase. After, we explain the architecture of our model. In our results section, we present and evaluate our model's performance while also comparing it with other algorithms. Finally, we discuss the implications and limitations of our results and conclude key insights derived from our analysis, while also providing recommendations for future research.

In conclusion, this thesis will explore the potential of the *HINT* model to address some of the most pressing challenges in the field of clinical trials. The successful implementation of this model could pave the way for more widespread adoption of AI-driven approaches in drug development, ultimately benefiting patients by bringing new treatments to market faster and more efficiently. The implications of this research extend beyond the pharmaceutical industry, offering a framework that could be adapted to other areas of healthcare and medical research where complex, data-driven decision-making is required.

2. Literature review

Fu et al.'s (2022) study has a vital role in the field of clinical trial outcome predictions. By implementing the novel approach of the Hierarchical Interaction Network (*HINT*), which depicts the complex relationship between trial variables, predictive accuracy for clinical outcomes across phases is amplified. By capturing complex relationships among trial components, *HINT* can accurately predict the success or failure of clinical trials across different phases. Its ability to incorporate diverse data types and leverage external knowledge makes it one of the most comprehensive and effective tools in this domain. *HINT* functions as an end-to-end framework, which ultimately returns a success probability score for a clinical trial before it starts. By providing this insight, *HINT* could become a powerful tool for stakeholders to allocate resources in a more efficient way, while accelerating the clinical trials approval process.

Clinical trials are designed to assess the safety and efficacy of new treatments. A trial typically involves testing a treatment set, which includes one or more drug candidates, against a target disease set in a group of patients defined by eligibility criteria. The goal is to determine whether the treatment successfully meets its primary endpoints, such as reducing disease symptoms or achieving specific outcomes (Fu et al. 2022, 3-4).

In the *HINT* model, the outcome of a trial is indicated using a binary label: “1” indicates success, meaning that the trial met its primary endpoints, while “0” indicates failure.

HINT focuses on two key prediction tasks: the first is a phase-level prediction, which assesses the likelihood of success for a specific trial phase (e.g., Phase I, II, or III), while the second is an indication-level prediction, which judges whether a treatment will ultimately pass all three phases of the trial process (Fu et al. 2022, 4).

A carefully designed architecture that integrates multimodal data and external knowledge with modelling techniques achieves *HINT* predictive ability. The framework consists of several key components, featuring an input embedding module, a knowledge embedding module, a hierarchical interaction graph, and a dynamic attentive graph neural network (Fu et al. 2022, 4).

The first step consists in encoding multimodal data into embeddings, by using the input embedding module, which processes three primary data types: drug molecules, disease information, and trial protocols. Drug molecules are encoded through the SMILES (Simplified Molecular Input Line Entry System) strings and molecular graphs, which capture the structural and chemical properties of the drugs. To generate embeddings, these representations are processed through techniques such as Morgan fingerprints, SMILES encoders, and message-passing neural networks. The encoding of disease information is achieved by using hierarchical medical ontologies, such as ICD-10 codes and textual descriptions. *HINT* then leverages the Graph-based Attention Model to create embeddings that reflect the hierarchical relationships inherent in these datasets. Trial protocols (including inclusion and exclusion criteria) are then encoded through Bio-BERT, a domain-specific language model able to detect the semantic notes of trial descriptions (Fu et al. 2022, 4-6).

Ultimately, the output is a set of embeddings that represent the key components of the trial. These embeddings are further enhanced by using external knowledge sources. Indeed, to provide a greater understanding of drugs interactions, the pharmacokinetics data present in *ADMET* dataset are incorporated (pharmacokinetics include properties such as absorption, distribution, metabolism, excretion, and toxicity).

Lastly, historical trial data, including success rates for specific diseases, are also integrated to inform predictions. These knowledge embeddings are pretrained on external datasets, enabling *HINT* to leverage vast amounts of prior information (Fu et al. 2022, 6-7). The following step consists in the hierarchical interaction graph, which lies at the heart of *HINT* architecture. This graph aims to connect the embeddings and establish relationships between drugs, diseases, and trial protocols. Therefore, it features several types of nodes, such as input nodes (representing the trial components), external knowledge nodes (representing *ADMET* properties and disease risks), aggregation nodes (which summarize interactions between trial components), and prediction nodes (which generate the final trial outcome predictions). This structure is essential to enable the model to capture both direct and indirect relationships, providing a holistic view of the trial's dynamics (Fu et al. 2022, 8-9).

After processing the graph, *HINT* aggregates the information from the graph to make the final prediction. The prediction node summarizes the information gathered from the pharmacokinetics properties, disease risk, and the interactions between the drug, disease, and eligibility criteria. The output of this prediction node is the final predicted trial outcome, which is typically a binary success/failure label, though it can also provide probability scores indicating the likelihood of success (Fu et al 2022, 8).

To refine predictions, an attentive graph neural network is employed. This component uses graph convolutional layers to aggregate information from neighboring nodes, updating the embeddings to reflect their context within the graph. Consequently, attention mechanisms assign weights to the most critical interactions, emphasizing the relationships that are most likely to influence trial outcomes, to improve predictive accuracy (Fu et al 2022, 8-9).

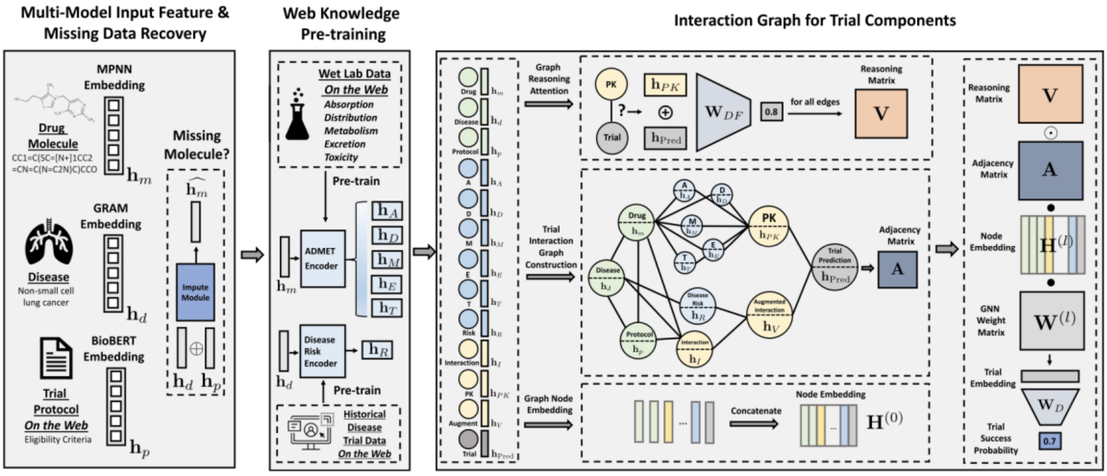


Figure 2 -HINT Framework, (Fu et al., 2022, Patterns 3, 100445) April 8, 2022, a 2022 The author(s)).

HINT was evaluated on the Trial Outcome Prediction (TOP) dataset, a comprehensive benchmark created to standardize clinical trial outcome prediction. The dataset includes 17,538 clinical trials, with 9,999 classified as successful and 7,539 as failures. For each trial, the dataset provides information on drug molecular structures, disease descriptions, trial protocols, and outcomes. Additional supplementary datasets include pharmacokinetics data (e.g., *ADMET* properties) and historical trial success rates. The TOP dataset spans a wide range of diseases and trial phases, providing a robust foundation for training and evaluation. By leveraging this dataset, *HINT* was able to demonstrate its ability to generalize across different types of trials and diseases (Fu et al. 2022 10-11).

HINT performance was validated through a series of experiments, showing significant improvements over existing models. For phase-level predictions, *HINT* achieved F1 scores of 0.665 for Phase I, 0.620 for Phase II, and 0.847 for Phase III. These results highlight *HINT* ability to handle the unique challenges of each trial phase, from toxicity testing in Phase I to large-scale efficacy studies in Phase III. *HINT* also demonstrated strong performance across different disease categories. For example, it achieved an F1 score of 0.867 for respiratory diseases, 0.786 for digestive diseases, and 0.585 for oncology. While oncology trials proved to be the most challenging, *HINT* performance in other categories underscores its versatility and robustness (Fu et al. 2022, 14).

Case studies further illustrated *HINT* practical utility. For instance, the model accurately predicted the failure of Entresto's $200 million Phase III trial, assigning it a low success probability of 0.476. Conversely, it successfully forecasted the outcome of Sitagliptin's diabetes trial, assigning it a high success probability of 0.742 (Fu et al. 2022, 14-15).

Despite its strengths, *HINT* is not without limitations. The model is currently limited to small-molecule drugs, excluding biologics and medical devices. It also struggles to work well in case of rare diseases due to the lack of sufficient training data. Additionally, *HINT* simplifies outcomes into binary success or failure labels, which may not fully capture the nuance of trial results. Ultimately, the complexity of the hierarchical interaction graph can make interpretation significantly more difficult.

Future work aims to address these limitations by expanding *HINT* to other trial types, integrating outcome labels with higher granularity, and improving interpretability through explainable AI techniques. These advancements would further enhance *HINT* utility and applicability. So, the Hierarchical Interaction Network (*HINT*) represents a paradigm shift in clinical trial modelling. By integrating multimodal data, leveraging external knowledge, and employing advanced graph-based techniques, *HINT* sets a new standard for accuracy and scalability in trial outcome prediction. As the pharmaceutical industry continues to embrace AI-driven solutions, *HINT* offers a powerful tool for optimizing drug development and improving global healthcare outcomes.

Even though *HINT* holds a fundamental role in this field, it still grapples with limitations, for instance challenges with handling missing data and fine-tuning predictions for phase-specific regulatory needs. Fu et al's. (2022) work has sparked subsequent studies expanding on *HINT* framework by including innovative machine learning techniques, robust data imputation methods, and large language models with the goal to improve predictive robustness and regulatory adaptability in clinical trial setting which will be explored in the following literature review.

Missing values, due to inconsistencies, can undermine the reliability of the success predictions, as they can result in potential biases and reduced modeling robustness. Lo et al. (2019) directly target the data completeness limitation described in Fu et al.'s (2022) work. By enabling accurate predictions even with incomplete data, this study enhances the foundational methods of *HINT*, allowing more robust trial predictions and continuity despite data gaps. Lo et al. (2019) study utilizes advanced machine learning techniques to impute the missing values in their dataset, for instance k-Nearest Neighbor (kNN) and Multiple Imputation by Chained Equations (MICE). Training machine learning models such as Random Forests and Support Vector Machines on the imputed datasets has shown a heightened predictive accuracy for drug approval outcomes. Lo et al. (2019) have allowed for more reliable trial predictions with their work.

While Lo et al. (2019) enhances prediction reliability by improving data completeness, Chen at al. (2024) improve clinical trial outcome predictions by integrating selective classification into the existing *HINT* framework to manage uncertainty. This approach permits the model to refrain from low-confidence predictions while boosting the accuracy of trials with increased prediction confidence. The authors demonstrate that including selective classification (SC) into clinical trial outcome models significantly enhances predictive reliability, which refines *HINT* framework by adding reliability in early, uncertainty-prone trial phases.

Aliper et al.'s (2023) further strengthens outcome predictions by developing a multi-modal AI framework to predict phase transitions based on trial design, target characteristics and omics through transformer-bases AI. The study demonstrates the effectiveness of the applied framework, by reaching a 79% accuracy rate. The model not merely refines the general outcome estimations of *HINT* but further offers phase-specific insights.

Following the advancements demonstrated by Aliper et al. (2023) in multi-modal AI, the potential of multi-modal approaches is further elevated by leveraging Large Language Models (LLMs). Zheng et al. (2024) utilize an ensemble of LLMs, their LIFTED framework, to manage multimodal data inputs for clinical trial predictions. By leveraging a Mixture-of-Experts (MoE) architecture, LIFTED adaptively selects the most suitable data source according to trial phase and particular prediction requirements. This enables the model to tailor its focal point, which in return increased predictive accuracy and robustness throughout the entirety of trial stages. This method extends *HINT* core by portraying the layers clinical trial data but also boosts versatility and predictive accuracy by processing diverse data through LLM ensembles.

As Aliper et al. (2023) and Zheng et al. (2014) employ general multi-modal datasets for phase predictions, Qi et al. (2019) aim at a thorough incorporation of target-based drug data with clinical features. With this, they are depicting patient responses and pharmacokinetics (PK) with the goal of predicting individual trial outcomes. This framework demonstrated improved prediction accuracy, which approach fills a gap in *HINT* comprehensive emphasis, by encapsulating detailed patient-level insights.

The usage of machine learning algorithms such as Random Forests and Gradient Boosting cannot only be found in relation to missing value imputation by Lo et al. Kavalci & Hartshorn (2023) also use machine learning algorithms to evaluate trial data and predict early termination based on recruitment rates, compliance, and trial characteristics dynamically. The study achieved high accuracy in selecting trials with high dropout probability, making room for preemptive refinement. Kavalci & Hartshorn (2023) not only complement the foundational factors of the *HINT* model by tackling with trial completion rates, but they are also able to detect risks prematurely, which consequently improves resource allocation and cost-efficiency in trials.

Building on Kavalci & Hartshorn (2023), leveraging machine learning algorithms to predict early trial dropout and manage resource allocation, Ghim & Ahn (2023) further boosts cost efficiency by implementing LLMs into clinical trial operations, as we have seen similarly in Aliper et al.'s (2023) study.

Ghim & Ahm (2023) have undertaken the task of automating clinical trial tasks concerning patient matching by eligibility criteria, protocol creation and explicit consent by taking advantage of AI driven chatbots simplifying complex trial information. The authors demonstrate a significant enhancement in trial efficiency, as patient matching times have reduced, while protocol generation and document processing reap the rewards of LLMs' ability to manage massive quantities of trial data. Additionally, they found that LLM-based chatbots enhanced subjects' comprehension by simplifying medical information, emphasizing that LLMs can contribute creating not only operational but also participant-centered improvements in clinical trials. These enhancements were outside of the scope of Fu et al.'s (2022) work but support their objective of enhancing trial success by reducing administrative delays.

In the same vein as Ghim & Ahm (2023), Reinisch et al. (2024) also addressed operational efficiency by predicting which clinical trials will advance from Phase III to final approval. This has been approached by finetuning CTP-LLM to clinical trial documents, reaching high accuracy and reliability. This indicates that CTP-LLM can aid in efficient resource allocation and improved decision-making throughout trial phases. Echoing Aliper et al.'s (2023) work, Reinisch et al. (2024) offer phase-specific insights and thus enriching *HINT* broad success estimation approach.

Efficient operations alone cannot compensate for the challenges of recruitment, which remains a dominant factor in trial timelines and completion rates. To address this, Bieganek et al. (2022) develop predictive models zeroing in on enrollment success forecasting, allowing for initiative-taking and operationally efficient recruitment planning. Through examining historical clinical trial data, the model predicted enrollment outcomes with enhanced accuracy, while also detecting influential trial characteristics leading to the recruitment’s success and enrollment changes at an early stage. This approach not only lessens delays related to under-enrollment, but it also aids to a more seamless trial completion, directly responding to a key factor of trial success highlighted by Fu et al. (2022).

Gayvert et al. (2017) shift the focus to an alternate key determinant in securing trial fulfillment: drug toxicity. This is executed by assessing toxicity risks influencing participant safety and retention. By including drug toxicity prediction leveraging target-based and structural features of compounds, the authors developed the PrOCTOR model, assessing drug properties to measure toxicity risk. The goal of this estimation is to forecast trial success by detecting compounds with manageable toxicity levels. The model displayed enhanced predictive power by differentiating FDA-approved drugs from those that failed due to toxicity with higher accuracy. Furthermore, PrOCTOR was also able to recognize network connectivity and target expression as crucial components in defining a compound's safety profile. Note, that *HINT* does include toxicity as part of the *ADMET* dataset, Gayvert et al. (2017) offer a directed approach to toxicity, delivering a detailed examination regarding patient safety.

As Gayvert et al. examines toxicity, Murali’s (2021) dataset includes general bioactivity metrics like drug-target interactions and pharmacodynamic properties. This study moves beyond the safety-focused perspective of toxicity to incorporate molecular-level relations impacting a drug's efficacy. The analysis highlighted that bioactivity-informed predictions significantly enhance the detection of drugs likely to succeed in clinical trials, specifically in early phases where biological characteristics are critical for predicting efficacy. This study complements *HINT* as it already uses the *ADMET* properties to broadly capture pharmacokinetics and safety. Nonetheless Murali (2021) can offer a deep dive into bioactivity metrics, such as drug-target interactions and pharmacodynamics, providing a more nuanced of a drug’s impact.

3. Exploratory Data Analysis

3.1 Data Collection and Description

In this chapter, we are diving into the exploratory data analysis which we conducted on the *ClinicalTrial.gov* dataset to uncover patterns, anomalies and structures, providing insights for our analysis. The dataset is highly relevant for clinical trial research, containing detailed attributes on study designs, phases, outcomes, and interventions.

Key columns like "primary_outcome_measures", "eligibility_criteria", and "intervention_model" provide substantial information to support diverse analyses.

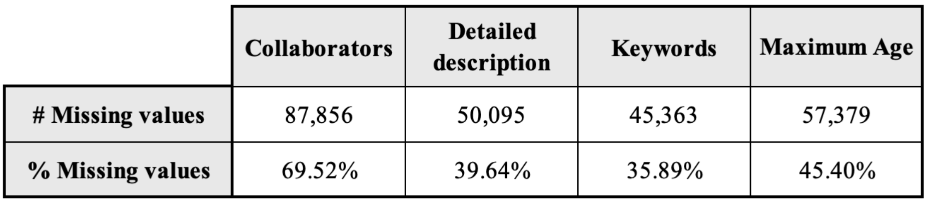
The data was sourced from ClinicalTrial.gov in the form of a JSON file, which was then transformed into a CSV file for easier analysis and manipulation. The dataset comprises 512,835 rows with 32 columns. However, given that our analysis is merely centered on interventional clinical trials dealing with small drugs and it includes only concluded clinical trials, we disregarded the rest, leaving us with a filtered dataset containing 132,129 rows and 32 columns. From this point onwards, we are referring to clinical trials, whereby we only understand this term to mean interventional studies. Also, the "study_status" only includes clinical trials with four different attributes that represent the trials that are concluded, such as "COMPLETED", "TERMINATED", "WITHDRAWN", and "SUSPENDED".

3.2 Data Cleaning and Preprocessing

In preparation for a detailed analysis, it is essential to execute preprocessing action to address issues such as missing values and duplicates ensuring robust data quality and integrity.

An initial look of the dataset was performed revealing that there were close to no missing values in any of the columns apart from the enrollment, having 2,304 missing values which will be maintained. Although these missing values would typically be removed in a standard approach to maintain data consistency, we decided to keep them since the column "enrollment" hold a central role in our dataset. A final decision will be made after progressing further, specifically when we are curating the final dataset including this data. Nonetheless, we also found the presence of “NA” values in the “phases” column, which could imply ambiguous classification for some trials and needs to be further investigated for accurate phase specific analysis. However, a closer inspection has uncovered a substantial number of blank entries in several columns, prompting more attention and thereby setting the stage for targeted data cleaning measures. These columns are for instance “collaborators” (69.52% blank), “detailed_description” (39.64% blank), “keywords” (35.89% blank), “maximum_age” (45.40% blank), and “secondary_outcome_measures” (21.51% blank). Given the share of blank spaces (20%) as well as their irrelevance for further analysis, they were discarded. Nevertheless, other columns such as “arms” (11.36% blank), “locations” (9.01% blank), and “primary_outcome_measures” (4.04% blank) also displayed blank spaces. Converting the format of “completion_date” and “start_date” has led to the same amount of missing data in both columns, which suggesting that the transformation has been successful. Upon addressing the missing values, we conducted further checks for data quality issues in regards of negative values and duplicated rows. Our analysis confirmed that the dataset is devoid of negative values and duplicated rows. The following table depicts the number of missing values in our *ClinicalTrial.gov* dataset.

Missing values overview



	Collaborators	Detailed description	Keywords	Maximum Age
# Missing values	87,856	50,095	45,363	57,379
% Missing values	69.52%	39.64%	35.89%	45.40%

Table 1 - The table presents the count of missing values and the corresponding percentage of missing values for each attribute.

Overall, we conclude that the dataset shows strong completeness in crucial areas, due to it being fully population in key columns such as “ntc_number”, “brief_title”, “study_status”, and “sponsor”, ensuring consistency in essential data attributes. It also demonstrates significant consistency across numerical and categorical data, which is proven by the absence of duplicates and negative values, confirming the integrity of the dataset. Nonetheless, the presence of "NA" values in "phases" indicates potential gaps in classification, which may impact the accuracy. In terms of accuracy we conclude that the dataset structure aligns with expected standards for clinical trials data. Despite this, some descriptive columns with a high number of missing records, such as "collaborators", are to be carefully analyzed to not bias the model outcome. Consistently populated fields for "start_date" and "completion_date", enabling chronological analyses of clinical trials suggest the great timeliness of the dataset, ensuring an effective study of trends over time, such as trial durations or historical shifts in trial designs. The dataset's high usability and relevance is suggested by its substantial information to support diverse analysis through its detailed and structured attributes. After this general analysis of the dataset, we are diving deeper into the contents of the dataset.

3.3 Insights and Patterns

Our effort of preprocessing and cleaning the data has set essential groundwork to explore the underlying patterns and insights, which are vital to understanding the complexities of our dataset.

3.3.1 Study Completion

By examining the historical progression of clinical trials over the years we discovered an increasing trend, starting in the year 2000, which was likely due to a rise in medical research (See Appendix, Graphic 1). This can be measured in multiple ways, although one of them in the rapid growth of scientific articles published in the Directory of Open Access Journal starting in the year 2000 (Laakso & Björk, 2012, 4-6). Additionally, we observed a notable rise in initiation and completion of clinical trials occurring around 2020, which could potentially have been impacted by the global COVID-19 pandemic, causing a massive increase in funding for vaccine research as seen by the investment of Operation Warp Speed with a total of $18 billion in COVID-19 related research and development (World Health Organization, 2023, 7).

Through this longitudinal analysis, we also uncovered that the amount of completed studies is slightly higher than the number of newly started trials, indicating that many recent trials are still in progress. Our investigation into temporal dynamics also gives us insights into completion dates extending into the 2030s, which illustrates how some studies collect data over elongated periods of time. Apart from completion dates being far in the future, we also found starting dates in the future, indicating the long-term project management roadmaps in clinical trials.

Moving beyond the historical insights gathered, we now broaden our discussion to the general landscape of study completion. While our analysis demonstrates a high completion rate of clinical trials across all phases due to most studies being completed, a small but noteworthy amount still failed to fall under this category as these studies have been either terminated or withdrawn. This is likely due to factors such as feasibility issues, safety concerns, or lack of efficacy. However, we can assert that a suspended study is a relatively rare occurrence in clinical trials. With general completion patterns established, it was also worth investigating how completion evolves throughout the year. It became apparent that completed trials showed a stable trend, while terminated and withdrawn trials were consistently depicted in smaller numbers and suspended trials are notably rare in any month. The month of December is the object of a significant surge in trial completion, which could be explained by year-end finalization. Next to yearly trends, we also narrowed our perspective to the structured phases of a trial. Phase II exhibits the highest amount of completed trials, followed by Phase I, III and IV, respectively. However, we still observed a considerable number of completed trials being unclassified under specific phases ("NA" with 10,235 trials). While phases I and II exhibit the highest amount of suspended and terminated trials, they occur the least in Phase IV. Although withdrawn trials are evenly spread throughout all phases, they occur the most in phases I and II.

3.3.2 Study Duration

After assessing the completion rates, we progress to explore the study duration, which is crucial to understand the efficiency of completion even more. Predominantly the clinical trials in our dataset have relatively short timelines with most of them being not longer than 8 years and the biggest among these trials falling into the duration of maximal 4 years (See Appendix, Graphic 2). Despite these observations, we discovered right-skewed distribution as well as a sizable number of outliers, indicating that a notable number of trials last longer than the average. Additionally, we discovered that suspended trials demonstrated a broader distribution and a higher median duration when compared to other statuses. This is a stark contrast to withdrawn trials, featuring a narrower range and a smaller number of outliers.

3.3.3 Study design

With an understanding of temporal aspects of clinical trials, we now proceed by shifting our focus to explore implications of study designs, to evaluate how different methodologies impact trial completion rates. Randomized studies (67,749) exhibit the highest completion count, followed by non-randomized studies (19,375). While suspensions were rather uncommon across study types but were slightly more prevalent in non-randomized studies (159), terminations also occur less often in randomized studies (9,227) than non-randomized (4,474). Despite this observation, withdrawals are relatively infrequent but still more notable in randomized (3,869) and non-randomized (1,760) studies. Therefore, our data underscores the effectiveness of randomized trials while indicating potential vulnerabilities to operational challenges of non-randomized studies.

Delving deeper into study designs has also shown us that the parallel model is the most used one while also exhibiting the highest number of completed studies. In this model, randomly assigned patients are given new therapy in the treatment group while control patients are given standard therapy (Cleophas & Vogel, 1998, 113). This is followed by the single group model, which is also widely used despite being impacted by higher termination and withdrawal rates. As already implied by the name, participants are here in only one group and receiving the same intervention. Since there is no comparison group, the outcomes are evaluated by comparing them to baseline measurements (Wang et al., 2024, 1-4). Sophisticated models such as crossover and factorial models are less commonly employed, although generally maintaining reasonable completion rates. While factorial models assess several interventions at the same time by assigning participants to multiple combinations of a treatment (Montgomery et al., 2003, 2), the crossover model assigns both new therapy and standard therapy to participants (Cleophas & Vogel, 1998, 113). Sequential models also see limited usage. The key difference between a sequential and a crossover design is the fact that participants receive one treatment only instead of all treatments in sequential order (Parson et al, 2024, 2-3). Ultimately, the evidence clearly points to a balance between study design complexity and stability with simpler models leading to more frequent utility while also occasionally coming short and consequently leading to early study termination or withdrawal. Given that the choice of the study model can inherently dictate the study's success or failure, we explored how these models are implemented in randomized and non-randomized studies. Essentially, parallel, crossover and factorial models heavily employ randomized study designs to maintain integrity in their comparative analysis. In contrast, single group models rely on non-randomized settings, due to their ideal utility for exploration or observational studies, characterized by the absence of a control group. Sequential models demonstrate a synthesis of comprehensive blend containing randomized and non-randomized allocations, enhancing their adaptability.

3.3.4 Primary purposes of clinical trials

While the study design lays the groundwork of the study design, it is also pertinent to delve into the primary purposes of clinical trials. Our analysis demonstrated that the dataset is dominated by trials dedicated to treatment (99,833 entries) as primary purpose, suggesting that a notable portion of attention in clinical trials is shifted towards evaluating the efficacy of therapeutic interventions to make progress in patient care. This is followed by prevention studies (7,808 entries), aiming to avoid diseases and conditions. Basic science (6,124 entries) is also prominent and focused on exploring biological processes. As we have defined the primary purpose of clinical studies, this led us to turn to the specific conditions they investigate. Delving into these conditions has shown that “Healthy” is the most frequently studied conditions, which is then followed by "Breast Cancer," "Healthy Volunteers," "HIV Infections, etc. Furthermore, our analysis suggested that clinical trials are heavily focused on grasping drug dynamics (pharmacokinetics), to address health challenges such as HIV or cancer. Most of clinical trials only observed one or two drugs, while the usage of ten or more drugs at once is a rare occurrence, suggesting the desire to focus on simpler regimens. Naturally, “Placebo” was used in most interventions in clinical trials, representing its fundamental role as a control substance. Due to our focus being set on predicting the success or failure of clinical trials regarding the testing of drugs, we discarded all entries titles “Placebo”. Cyclophosphamide, Carboplatin, and Cisplatin were among other frequently used drugs, which are mostly used in oncology trials, underscoring the significant role of cancer treatment.

3.3.5 Patient Characteristics

After detailing our findings regarding the trial’s purpose, it was imperative to shift our focus to the patient’s characteristics. While majority of the trials includes male and female participants, we found a small amount of gender-specific trials. Upon shedding light on this matter, our data also demonstrates that male-specific trials have the highest success rate, which is followed by female-specific trials and trials involving all genders. This could indicate that gender-specific trials may lead to successful outcomes, due to a more focused study design with clearer treatment effects. Nonetheless, gender-specific trials seem to inadvertently overlook older adults, as female-specific trials include participants between the ages of 21 to 56 years old. In contrast, trials including all genders exhibited a broader age range with ages spanning from 19 to 62 years. Aside from this, there are also several trials dedicated to pediatric conditions, suggested by the exceptionally low minimum ages which are stored in months instead of years. Generally, the minimum age of participants skews towards 18 years.

3.3.6 Enrollment

Patient characteristics directly inform patient enrollment which is why we now turn to exploring this attribute in our dataset. Following our analysis, we found 6,577 records of zero enrollments, which is why we consequently assume that trials with no participants are usually considered as a failure, due to no data being collected. Overall, we also concluded that enrollment values are generally low, suggesting that clinical trials must cope with challenges in recruiting large numbers of participants. After relating enrollment values to clinical trial outcomes, we demonstrated that successful trials (231.77) exhibit a higher average enrollment in comparison to failed trials (81.74). Consequently, this suggests a positive correlation between higher enrollment numbers and the success probability of trial success. Additionally, successful trials showcase a broader variability in enrollment, as suggested by the higher standard deviation. Our dataset also gives us insights into country-specific enrollment data, which showed that trials were often aimed at addressing widespread health issues such as infectious diseases, which is shown through the high enrollment in developing countries.

3.3.7 Contribution

The list of organizations with the highest number of studies was topped by Pfizer and Novartis, holding 2,442 and 2,436 studies respectively, which are followed by GlaxoSmithKline with 2,161 studies and the National Cancer Institute (NCI) with 1,968 studies. Top sponsors of clinical trials included a diverse mix of universities, pharmaceutical companies, and cancer research institutes, illustrating the broad range of organizations funding clinical trial research. Among these, universities and public health institutions hold a more prominent role in sponsoring. The presence of renowned medical institutions such as the Mayo Clinic and M.D. Anderson Cancer Center underlines the importance of clinical research in progressing patient care and medical innovation. Not only do sponsors hold a special role in advancing clinical trials, but collaborators are also of special consideration. Apart from that, we demonstrated that the National Cancer Institute (NCI) is the one with the highest number of collaborators in clinical trials, indicating their vital role in clinical trials while also suggesting the importance of clinical trials in cancer research. Pfizer and GlaxoSmithKline also frequently collaborate, underscoring their commitment to working with research institutions to advance drug development and clinical research.

4. Benchmark

Our starting point information is the clinical research studies data that we have retrieved from *ClinicalTrials.gov*, which is a publicly accessible website and online database that provides up-to-date information on clinical research studies conducted across all 50 states and in more than 200 countries worldwide.

We have defined and developed an up-to-date standardized benchmark which was essential to build a robust, interpretable and scalable model.

For each clinical trial, we have collected the following four data items: drug molecule information, disease information, trial eligibility criteria, number of enrollment and trial outcome information.

Our benchmark includes molecule information with SMILES of the drugs, the target disease information encoded into ICD-10 codes, the trial eligibility criteria and biomedical knowledge. The encoding into latent embedding vectors of all the data included in the final dataset differs according to their own specifications.

Standardizing clinical trial data by linking diseases with ICD codes and drug with SMILES (Weininger, 1988) medicine is a common practice. In fact, the approach ensures uniformity in data representation, facilitating the integration of new data from diverse sources and enhancing its utility for machine learning applications.

4.1 Drug molecule data

The drug molecule data is extracted from the collected clinical trials data on *ClinicalTrials.gov* and linked to the molecule structure, i.e. SMILES string. The SMILES collections with their own respective drug name and drug's description have been collected mainly from the GitHub dataset stored in the *HINT* repository, which retrieved the data from the Drug Bank Database (Fu et al. 2022).

The SMILES, i.e. Simplified Molecular Input Line Entry System, is a notation system that represents the structure of chemical compounds as concise ASCII strings. These SMILES strings can generate two-dimensional diagrams or three-dimensional models of the molecules (Weininger 2018).

The collected dataset on the Drug Bank Database results as a list of drugs with their respective SMILE molecules. We performed an algorithm which assigns ICD-10 codes to disease names using a heuristic matching approach. It first attempts an exact match, prioritizing precision by checking if the disease name directly corresponds to an ICD code. If no exact match is found, the algorithm performs a partial word match, focusing on significant words with seven or more characters, while limiting matches to simpler disease names to avoid ambiguity.

If these methods fail, the algorithm uses word-overlap matching, comparing the disease input name with candidate names by evaluating shared words and prioritizing those with the greatest overlap and combined word length. This step allows for flexibility in handling variations in terminology. When no suitable match is identified, the algorithm returns no result, ensuring it avoids incorrect associations. This combination of exact, partial, and overlap matching provides a structured approach to map disease names to ICD codes effectively.

4.2 Disease data

The disease data are extracted from the collected clinical trial data on *ClinicalTrials.gov* and linked to ICD-10 codes extracted from the Word Health Organization Database. The dataset contains disease information including ICD-10 codes and the disease description. ICD–10 is the “international standard for systematic recording, reporting, analysis, interpretation and comparison of mortality and morbidity” data published by the World Health Organization (WHO).

Recently, the ICD-10 has been replaced in January 2022 with the ICD-11 by the World Health Organization (WHO). However, the decision to use ICD-10 instead of the newer ICD-11 in our research was influenced by diverse factors.

Firstly, ICD-10 has a well-established ecosystem of tools, libraries, and datasets, making it the most pragmatic choice in the current time the study was elaborated. Secondly, the ICD-10 codes are still relevant and widely used in past and recent studies in the clinical trial landscape, therefore adapting to past studies guarantees to obtain relevant and interpretable results by the broader scientific and medical communities.

We have performed an algorithm, in the same ways as for the diseases, which maps drug names to their corresponding SMILES representations using a heuristic approach. It first checks for an exact match between the input drug name and a predefined drug-to-SMILES mapping. If an exact match is not found, it searches for partial matches by analyzing individual words within the drug name, focusing on words with seven or more characters that are more likely to yield valid matches. Likewise, the goal is to retrieve their ICD-10 codes to generalize and standardize the disease data information.

Linking diseases to ICD-10 codes and drugs to SMILES ensures standardization, enabling seamless integration of diverse datasets and supporting the addition of new data in the future. These representations allow machine learning models to incorporate novel information from various sources, enhancing their capacity to test and predict outcomes for emerging drugs or diseases.

4.3 Trial eligibility criteria

The eligibility criteria outline the parameters for selecting participants in a clinical trial, specifying both inclusion and exclusion requirements. They are presented in unstructured natural language, detailing participant characteristics such as age, gender, medical history, current health status, and the target disease or condition. Each clinical trial's eligibility criteria, along with its respective clinical case, are published on *ClinicalTrials.gov*. In this perspective, we separated the inclusion and exclusion criteria and applied language processing to the unstructured text data, preparing it for the encoding phase to generate embedding vector, as later explained in detail.

Eligibility criteria are crucial prerequisites for the success of clinical trials, as they significantly influence the outcomes. Poorly defined eligibility criteria can result in inadequate participant recruitment, a key factor contributing to the failure of many clinical trials (Su & Cheng 2023).

4.4 Trial outcome information

The TOP benchmark’s target label represents the success or failure of the trial, which indicates whether the trial meets the primary endpoints or not. The labels have been manually curated and are influenced by the feature p-value and why stop present in the XML file on ClinicalTrial.gov. Our research instead targets the completion trial prediction which focuses on determining whether the trial would be completed or not. Specifically, the labels are defined based on the clinical trial’s final status. To determine whether the trial would terminate, we excluded all cases where the status indicated an interim step prior to the study's conclusion. These cases could potentially result in either a positive outcome, such as completion, or a negative outcome, such as termination, withdrawal, or suspension.

4.5 ADMET

The pharmacokinetics (PK) knowledge is leveraged to pretrain embeddings, emphasizing how the body processes drugs after administration. This approach is driven by the critical role that PK factors and disease risks play in determining trial outcomes. Key properties such as Absorption, Distribution, Metabolism, Excretion, and Toxicity (*ADMET*) are central to this process. Integrating these *ADMET* factors into our pretraining ensures a deeper understanding of the interactions between drugs and the human body, enhancing the model's ability to predict the success or failure of clinical trials.

The different components of the *ADMET* could be described as follows: firstly, the absorption component refers to the process by which a drug is transported from the administration to the site where it will exert its pharmacological effect. Secondly, the distribution component demonstrates the movement of a drug to and from various tissues within the body, as well as the number of drugs present in these tissues.

Moreover, the metabolism component refers to how drug molecules are broken down by enzymatic systems, thereby influencing the duration and intensity of the drug's action. The excretion component describes the removal of drugs from the body via various excretion routes. Finally, the toxicity component refers to the extent to which a drug may cause harm to the body. An early assessment and prediction of these properties in the drug development workflow optimizes compounds with pharmacokinetics (*ADMET*) characteristics and minimizes toxicity (Vrbanac & Slauter 2017). As a result, it can assist in the prioritization of drug candidates, thus reducing the number of costly trial failures and optimizing resources for the successful development of new drugs.

We retrieved the *ADMET* data from the GitHub dataset stored in the *HINT* repository, which retrieved the data from the Drug Bank Database (Fu et al. 2022).

4.6 Enrollment

The enrolment feature is extracted directly from clinical trial data available on *ClinicalTrials.gov*, distinguishing itself from the TOP benchmark. Given the observed correlation between a high number of enrolments and an increased likelihood of clinical trial success, it has been included alongside other variables due to its potential influence on the model's outcomes. In fact, including a high number of participants poses challenges due to the cost and difficulty of recruiting individuals with the specific characteristics required for clinical studies, which can ultimately influence the likelihood of the clinical trial's success or failure.

Unlike the other variables, which are sequential in nature, the enrolment feature is a numerical, tabular variable. This distinction makes it unique in the dataset, as it does not require preprocessing before being embedded into the model.

4.7 Data preprocessing

As previously illustrated, we filtered the dataset to achieve a complete and clear dataset. In a first step, the number of trials was 512,835 in total. After filtering based on the study's status to determine the study's outcome label, the total number of clinical studies is 325,725.

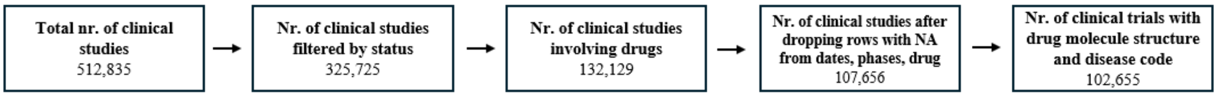

Total nr. of clinical studies
512,835
Nr. of clinical studies filtered by status
325,725
Nr. of clinical studies involving drugs
132,129
Nr. of clinical studies after dropping rows with NA from dates, phases, drug
107,656
Nr. of clinical trials with drug molecule structure and disease code
102,655

Figure 3 - Data preprocessing pipeline

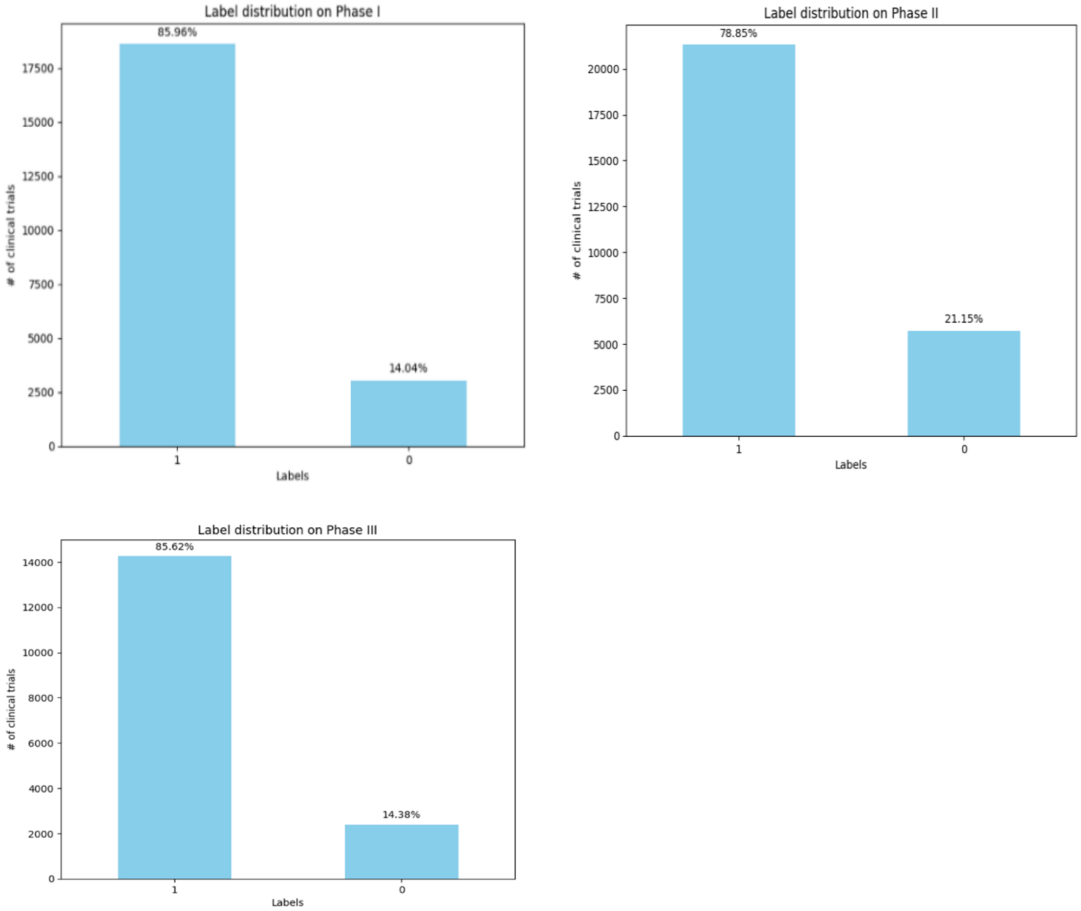
Since the model includes the molecule encoding to leverage the potential of the multi-model data source, it must handle only interventional trials involving small molecules, while other trial types such as medical devices and biologics trials are excluded. The total number of interventional studies including only those involving drugs is 132,129. Moreover, we have excluded those rows with missing values in the "phase", "start_date", "completion_date" and "drug" features, giving us a total of 107,656 clinical studies. Furthermore, we selected trials with known drug molecule structures and available disease codes, i.e. the trial we were able to match with the integrated datasets of the SMILES and ICD-codes, leaving us with a total of 102,655 clinical trials.

4.8 Data split

To evaluate the predictive model in an effective way, we split the data into three datasets: the training, validation, and test sets. This process has been done chronologically, meaning that we used the column "start_date" as an index so we can have more recent data in the test set. This splitting is following the rule of 70/15/15, consisting of having 70 % of the data in the training set, then 15% in the validation and the remaining 15% in the test set. This is a good approach since most of the data is utilized for the training set while the validation and the test have the same amount. The training set is used for the development of the model, the validation set is used for hyperparameter tuning and to prevent overfitting, while the test set is reserved for the final evaluation of the model. Since we are going to be focusing on looking at a specific phase, we also split the data according to their phases, leading to us having the data split in "PHASE1", "PHASE2" and "PHASE3". We did not include a dataset for "PHASE4", since its purpose is the monitoring of long-term effects of a treatment, which was already launched into the market according to the National Cancer Institute. Since our work is rather focused on supporting stakeholders to lead with an assertive strategy, to save resources like time and money, "PHASE4" data is out of scope of this work.

Consequently, our approach is to apply a filter for the column "Phases" with the result of only including entry with the value of a specific phase. This caused the data to be mapped to the data with specific phases, which was motivated by the fact that the column still contained a significant amount of data with more information apart from just one phase, such as "PHASE1|PHASE2" (6,155 entries), "PHASE2|PHASE3" (3,016 entries) and "EARLY_PHASE1" (1,834 entries).

The dataset demonstrates a pronounced imbalance in label distribution across all three phases, with a significantly higher proportion of clinical trials labeled as successfully completed compared to those labeled as not completed. Hence, the percentage of successful trials ranges from 78.85% to 85.96%, meaning that over one in five clinical trials is unsuccessful, as illustrated in Graphic 1. This imbalance is particularly evident in the training dataset and has a substantial impact on the model's performance. To mitigate this issue, appropriate evaluation metrics will be employed to ensure a fair and accurate assessment of the model’s predictive capabilities.



Graphic 1 - The table presents the label distribution of clinical trials categorized by their respective phases within the training dataset. The top-left image illustrates the distribution for Phase I, the top-right image corresponds to Phase II, and the bottom.

Additionally, we aimed to address the specific characteristics of each disease, therefore the data was further categorized by disease type. ICD codes had been used to map the clinical conditions with the aim of broadening the disease categories through the CCSR (Clinical Classification Software for Services and Procedures). The new information retrieved allowed us to refine the analysis by clustering similar conditions together, enabling us to create more focused and insightful evaluations according to the attributes of similar diseases and their trial outcomes. The disease areas selected are the nervous system (NVS), neoplasms (NEO), infectious and parasitic diseases (INF), respiratory system (RSP) and digestive system (DIG). These categories were chosen due to their significant impact on public health and because of their high frequency in the trial's dataset. This data has been split following the same process implemented to split the total dataset.

5. *HINTBasic* model

The *HINTBasic* model is a very sophisticated architecture design which was created to predict the clinical trial outcomes by integrating several data sources like diseases, drugs and trials protocols. This model is going to process and encode this information before integrating it into the final framework. Each feature is going to contribute in a unique way to capture the very complex relationship between the variables that define if a clinical trial will reach the end of their phases.

To include more medical information, we included a dataset of ICD codes. This means that we mapped these codes to their corresponding Clinical Classifications Software Refine (CCSR) categories. These codes are hierarchical, which means that they represent a relationship between diseases in a broader category which were then split into more specific categories: For example, the more general code "G90521" can also be split into variations such as "G9052", "G905" and "G90". This marked a crucial step in our modelling process, since it caused a reduction in the diversity of our data, which consequently enabled a division of clinical trials based on diseases and not the high quantity of conditions. For the trials with multiple ICD codes, we decided to create a one-hot vector to create a single composite vector that captured all related medical conditions. This approach allowed the model to take advantage of the hierarchical structure of medical conditions, enabling it to link trial outcomes with participants health conditions more efficiently. In this part we used a graph-based attention model, called GRAM, which is going to encode these hierarchical relationships. This did not only create embeddings for each ICD code, but it also captured the semantic similarities between related diseases. As a result, the model's performance is enhanced due to the added medical information. This is very important because it helped the model understand how certain diseases might influence the clinical trials outcomes, such as having a specific rare disease will decrease the probability of success, since it is very challenging to find eligible participants.

The protocol encoder addressed the complex text description that is in the eligibility criteria. This column includes information associated with inclusion and exclusion criteria, which gives insights into the criteria that patients need to have or not to be an eligible participant of the clinical trial. For example, trials testing for drugs to cure cancer only consider patients which not only have a specific cancer but are also in a specific phase of the disease. It is crucial to test the effectiveness of the drug in the treatment, since trial protocols are essential to determine the scope and the focus of the trial's proposal. To get all the information about the inclusion and exclusion, the data was split into two criteria, the inclusion and the exclusion criteria. After that, we employed Clinical Bert to generate the encoders at the sentence level (Guangyu. (nature medine, s.d.) 2023).

The protocol embedding used the neural embedding techniques to analyze the descriptions, applying layers of processing that contained highway networks, to extract and refine the semantic meaning of the text. By converting text data into high-dimensional embeddings, the protocol encoder ensures that the small details of trial design are accurately reflected in the model, providing valuable insights into how trial structure impacts its success or failure.

The molecular encoder is designed to handle data that is related to the chemical compounds, such as drugs that are being tested in the clinical trial. This encoder can capture the structural and chemical properties of the molecules and then transform that information into embeddings so it can be utilized in the model. This transformation is essential since the molecular properties of a drug significantly impact its safety, efficiency and the potential interactions within the trial. One illustrative example of the potential risks associated with clinical trials revolves around the drug “Trastuzumab Deruxtecan”. Despite the occurrence of several adverse effects, including pneumonitis, which has resulted in fatalities, the drug was ultimately approved following a comprehensive evaluation of its safety profile (Modi et al. 2022). This encoder uses the Message Passing Neural Network denominated MPNN, which is a type of graph-neural network designed to process the molecular structure. In the MPNN, molecules are represented as graphs, which depict atoms as nodes while the bonds are illustrated as edges. As a result, it can reflect the inherent molecular chemical properties and structures. This is crucial for capturing the relationships and the interactions between different atoms. The MPNN is divided into two phases, with the first one being the passing message and the second one being the readout. This means that during the propagates the information across the graph by exchanging messages between nodes (atoms) and updating their representation based on the properties of their neighbors and the edges that are connected to them. This repetitive process enables the model to capture the local features, that are the type and properties of a specific atom, and the global features such as the connectivity of the entire model. In the second phase the molecular encoder is aggregating the updated node representations into a single vector to represent the entire model. This aggregation was done to ensure that the embedding captures all the chemical and structural of the drug. Embedding is of high importance as it enables the model to contain a rich representation of molecular properties, influencing the drug ‘s behavior in the human body. This is due to its ability to provide information on how a drug might perform in a clinical trial, particularly regarding safety and efficacy.

The data loader plays an essential role in efficient preprocessing and batching of datasets for clinical trials analysis and *ADMET* prediction tasks. The dataset structure for clinical trials manages multi-modal clinical data. It contains features such as NCT ID, labels indicating the trial outcome and SMILES to represent the molecular compounds. Additionally, it also contains ICD-10 codes for disease identification, inclusion and exclusion criteria, and an additional feature, the enrollment number in its second version. The framework for *ADMET* focuses on predicting binary outcomes for *ADMET* properties, with SMILES strings as input features, and their corresponding label are binary (1 for positive and 0 for negative).

In addition, collation methods were used for batching and model training. In the context of *ADMET* prediction, the SMILES string and their corresponding labels are grouped together in batches, while labels themselves are converted into machine learning compatible tensors. This includes encoding ICD codes, protocols, and molecular structures. For instance, SMILES strings are transformed into structured lists and transforming text-based criteria into numerical feature vectors. All preprocessing steps are designed to ensure that the data is clean, organized, and suitable for our downstream tasks.

The data loading pipeline represents a crucial component of the machine learning workflow. It is responsible for preparing clinical trial data from CSV files and structuring it into training and evaluation batches. For *ADMET* prediction, it processes training and validation data for each property, creating task-specific datasets. This modular approach optimizes resource utilization of resources through efficient batching and merging of the data, while transforming data into structured, model-ready input, providing a robust foundation for machine learning models.

Combining *ADMET* with clinical trial data enables researchers to design more targeted and informed clinical trials, increasing the probability of successful drug development. Consequently, development is streamlined with predictive models that enhance clinical decision-making are supported. This also causes improved drug efficacy and safety and reduces costly clinical trials failures.

In the Graph Neural Network, the layers are designed to learn the interactions between different encoded features. Each node in the GNN represents an encoded feature like a disease, protocol information or a molecular property, while the edges represent the relationships between the features. Through iterative aggregation the GNN layers improve the node representation by adding more information from the neighbors that they are connected to. This process allows the model to build a comprehensive view of the relations between several trials' factors enabling the model to make better inform predictions.

Highway layers in the module are used as a mechanism to allow a smoother gradient flow during the training, which is crucial for preserving information across deep neural networks. These layers operate by applying an adaptative computation approach, where each transformation is balancing the nonlinear and the linear way. These specific layers calculate the weight combination of a non-linear transformation and a similar linear transformation. These weights are determined by a gating function that uses sigmoid activation to control the flow of information, which ensures that the essential features from earlier layers are retained while allowing the network to still learn complex patterns. Adding these layers is highly beneficial for the model since it will prevent the vanishing gradient. This is since holding that specific gradient creates more complex training in deep networks, while also enabling the model to refine representations effectively.

The model also uses the graph attention network that is denominated as GAT, to enhance the learning process by applying the attention mechanism to the graph. Unlike the other standard GNNs, GAT assigns the weights dynamically to edges based on their relevance, which allows the model to focus on the most important connections in the graph. Having this flexibility is very valuable since it can capture subtle interactions between clinical trials elements.

Joining the Highway layers and the GNN layers creates a cohesive framework that works inside the *HINTBasic* model. The highway refining the features during the early stage of processing can ensure efficient information flows and prevent the loss of critical data.

The model's primary focus is on its interactions between the clinical trials protocols, molecular data and the disease information using the advanced neural network components to produce meaningful predictions. After, the model initiated the encoding for each data point. This process raised for the protocol encoder regarding the text data of the eligibility criteria. The molecular encoder processed the chemical structure of the molecules, while the ICD code encoder operates regarding the hierarchical disease classifications. These encoders transformed the raw data into high dimensional embeddings that captured the essence of the information. Once the embeddings are generated, these are combined into an interaction representation, encapsulating the relationships between trial elements.

A critical step involves computing a "disease risk" embedding, which represents the likelihood of complications related to the target condition. This risk embedding is then integrated with the interaction representation in the "augmented interaction" node, producing a richer and more informative feature set.

Pharmacokinetics (PK) node leverages the *ADMET* framework, each property is modelled independently, and the resulting embeddings are combined into a unified PK representation, capturing the drug's pharmacological profile.

The enriched interaction embedding from the augmented interaction step is combined with the PK embedding to form a trial-level representation. This trial embedding is processed through a final node, which generates the predicted label, showcasing the probabilities of the clinical trials reaching the end of their phases, indicating their success or failure. Since the model employs a multi-task learning framework this allowed us to simultaneously optimize the model during the training, while also taking the used features into consideration.

6. *HINTPlus* Model

Enrollment is one of the mains factors leading to failure of clinical trials. To include this significantly impactful factor, we included the "enrollment" feature in our model, as this adds more crucial information about the success and failure of clinical trials (Kim et al. 2023)

It was necessary to process the “enrollment” feature in the data loader to create a complete clinical trial version with one additional variable compared to the original dataset. In the *HINTPlus* model the "enrollment" feature was treated as node on the GNN layers just like the other features in the previous *HINTBasic*. This allows the model to interact with all the features (ICD code, protocol and the molecular encoders) capturing the new relationships created by adding the enrollment.

In the *HINTPlus* model, the “enrollment” feature was added and shared across multiple tasks, enabling the model to learn how enrollment size influences a trial’s outcome. This integration ensures that the model incorporated an additional important feature, which contributes meaningfully to predictions and, in turn, enhanced the model’s ability to predict success or failure.

Furthermore, this model has been adapted to create specific versions according to the dataset featuring some selected diseases. The primary difference between the *HINTPlus* model and the specialized *HINTPlus* is to be found in the dataset specificity. This adjustment has been made to assess whether the specialized *HINTPlus* shows an improved performance than the generalized one. The following pipeline depicts the working processes of the developed models:

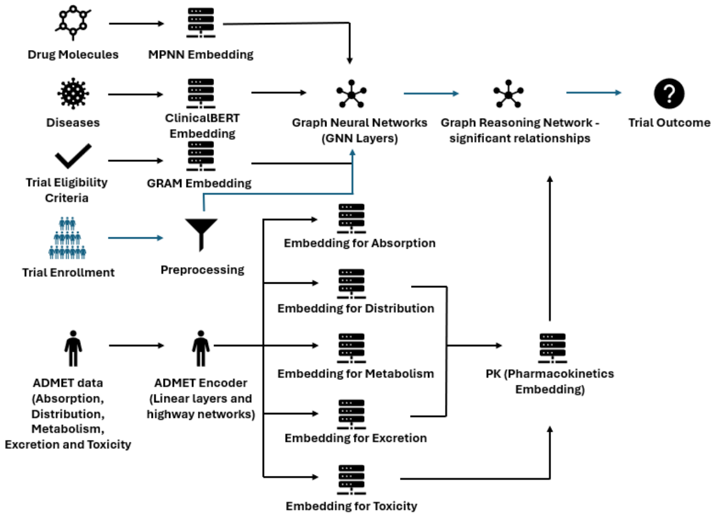

Drug Molecules
MPNN Embedding
Diseases
ClinicalBERT Embedding
Graph Neural Networks (GNN Layers)
Graph Reasoning Network - significant relationships
Trial Outcome
Trial Eligibility Criteria
GRAM Embedding
Trial Enrollment
Preprocessing
Embedding for Absorption
Embedding for Distribution
ADMET data (Absorption, Distribution, Metabolism, Excretion and Toxicity
ADMET Encoder (Linear layers and highway networks)
Embedding for Metabolism
PK (Pharmacokinetics Embedding)
Embedding for Excretion
Embedding for Toxicity

Figure 4: Pipeline of HINTBasic and HINTPlus

7. Results

Classical machine learning methods, such as Random Forest and LogisticRegression, and traditional deep learning models like Multi-Layer Perceptron's, often rely on limited features and fail to capture complex interactions among trial components. In contrast, the *HINTBasic* model leverages multi-modal data sources, integrates deep interaction mechanisms through weighted nodes, and incorporates pretraining with the *ADMET* dataset. This approach allows *HINTBasic* to dynamically model complex relationships, prioritize critical features, and utilize domain-specific drug knowledge, resulting in more accurate and reliable clinical trial outcome predictions.

We tested three different models: we focused on traditional predictive models, such as the *Logistic Regression*, the *XGBoost* (Chen & Guestrin 2016) and modern neural network models, such as the *Multi-Layer Perceptron*. Additionally, we have included the *HINT* model, as it was trained initially with the old dataset. The models have been tested using the PyTrial, which is a "Python package providing benchmarks and open-source implementation of a series of ML algorithms for clinical trial designs and operation (pytrial.readthedocs.io) including trial outcome predictions. It includes an evaluation of clinical trial outcome predictions using the TOP benchmark (Fu et al. 2022) involving different machine learning prediction models and the *HINT* itself (Wang et al. 2023).

The evaluation metrics to compare the diverse models are ROC - AUC, F1 and PR-AUC for evaluating the binary classification, i.e. the trial outcome labels. Those three metrics are widely used to evaluate the performance of predictive models, for classification tasks as in our case. Additionally, given the unbalanced nature of our dataset, F1 and PR-AUC are particularly well-suited as they are designed to handle imbalanced data by focusing on precision, recall, and the performance of the minority class.

The metrics we used are detailed, as follows: firstly, the Receiver Operating Characteristic –

Area Under the Curve (ROC - AUC) measures a model's ability to distinguish between positive and negative classes. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various thresholds. The AUC (Area Under the Curve) summarizes this curve into a single value, where 1 indicates perfect discrimination and 0.5 represents no better than random guessing. ROC-AUC is popular for its threshold independence, making it suitable for evaluating models when the relative costs of false positives and false negatives are unclear (Fawcett 2006; Huang & Ling 2005). Secondly, it balances precision and recall, providing a single measure of a model's accuracy that accounts for both false positives and false negatives. Then, the F1 score is particularly valuable in situations with imbalanced datasets (Van Rijsbergen 1979; Sokolova & Lapalme 2009). Lastly, the Precision-Recall Area Under the Curve (PR-AUC) evaluates the relationship between precision and recall across thresholds, focusing specifically on the performance of the positive class. Unlike ROC-AUC, PR-AUC ignores the true negative rate and is particularly suited to imbalanced datasets, where the positive class is underrepresented. PR-AUC provides a clearer picture of a model’s ability to capture positive cases, especially when false positives and false negatives carry significant costs (Davis & Goadrich 2006; Saito & Rehmsmeier 2015)

7.1 Evaluation of Our Models

To assess the efficacy of the *HINTBasic* models in diverse scenarios, we conducted experiments on both single phases (Phase I, Phase II, and Phase III) and on the combined dataset, containing all three phases.

The two models adhere to the same structured methodology, involving data preparation, model initialization, training, and evaluation. About the single-task *HINTBasic* model, each phase was processed independently, with datasets partitioned into training, validation, and test sets. Prior to training, the model was pre-trained on relevant data, thus enhancing its capacity to capture characteristics specific to each feature.

7.2 Comparison between *HINTBasic* and *HINTPlus*

Results comparison between *HINTBasic* and *HINTPlus*

	Phase I		Phase II		Phase III	
	HINTBasic	HINTPlus	HINTBasic	HINTPlus	HINTBasic	HINTPlus
PR-AUC	0.717 ± 0.007	0.878 ± 0.005	0.645 ± 0.007	0.783 ± 0.007	0.745 ± 0.007	0.822± 0.005
F1	0.835 ± 0.005	0.931 ± 0.003	0.784 ± 0.005	0.865 ± 0.005	0.854 ± 0.005	0.899 ± 0.003
ROC-AUC	0.724 ± 0.008	0.878 ± 0.007	0.547 ± 0.006	0.846 ± 0.007	0.535 ± 0.010	0.807 ± 0.009

Table 2 - Results of HINTBasic and HINTPlus

The inclusion of the enrollment feature consistently improved the performance of the *HINT* model across all three phases. An examination of the PR-AUC and ROC-AUC metrics, which assess the model's capability to differentiate between successful and unsuccessful trials, revealed enhancements across over each phase when comparing the two models. Furthermore, the F1 score, which balances precision and recall, exhibited gains across phases, indicating that the integration of the enrollment feature resulted in enhanced prediction quality.

The model demonstrated optimal performance in Phase I and Phase III, with or without the enrollment feature. Although, it outperformed across all three phases, achieving robust results with the inclusion of the enrollment feature. Precisely, it achieved F1 scores of 0.931, 0.865, and 0.899 in Phases I, II, and III, respectively. The inclusion of the enrollment feature consistently enhanced performance across all metrics, demonstrating its critical role in improving predictive accuracy. Overall, the enrollment feature proved to be a valuable addition, providing meaningful information that enhanced the model's ability to generate more accurate predictions.

7.3 Evaluation of the *HINTBasic* with baseline models

We trained the models using our up-to-date benchmark dataset, employing the pre-split dataset created earlier for consistent evaluation.

The tables provided represent the performance of five different models (*LogisticRegression*, *XGBoost*, *MLP*, and *HINTBasic*) across three clinical trial phases (Phase I, Phase II, and Phase

III), evaluated using metrics such as ROC-AUC, F1, and PR-AUC.

7.3.1 Phase I

The table represents the results of the mentioned metrics according to the five different models ingesting the clinical trials in Phase I. *HINTBasic* outperforms all other models across all metrics, achieving the highest values for ROC-AUC (0.717), F1 (0.835), and PR-AUC (0.724).

Phase I Baseline Results

	Models			
	LogReg	**XGBoost**	**MLP**	**HINTBasic**
ROC - AUC	0.608 ± 0.000	0.615 ± 0.000	0.599 ± 0.000	**0.717 ± 0.007**
F1	0.764 ± 0.291	0.763 ± 0.269	0.765 ± 0.260	**0.835 ± 0.005**
PR - AUC	0.7118 ± 0.008	0.7121 ± 0.008	0.7105 ± 0.007	**0.724 ± 0.008**

Table 3 - Table of results of the clinical trials prediction on Phase

7.3.2 Phase II

The table represents the outcome results of the mentioned metrics according to the five different models ingesting the clinical trials in Phase II. Likewise *Phase I*, *HINTBasic* remains the best-performing model, showing significant improvements in all metrics. It achieves 0.645 in ROC-AUC, 0.784 in F1, and 0.547 in PR-AUC. The performance gap between *HINTBasic* and the other models is especially pronounced for ROC - AUC, further emphasizing its ability to handle unbalanced datasets effectively.

Phase II Baseline Results

	Models			
	LogReg	**XGBoost**	**MLP**	**HINTBasic**
ROC - AUC	0.523 ± 0.000	0.523 ± 0.001	0.523 ± 0.002	**0.645 ± 0.007**
F1	0.748 ± 0.246	0.748 ± 0.231	0.748 ± 0.273	**0.784 ± 0.005**
PR - AUC	0.589 ± 0.000	0.596 ± 0.004	0.599 ± 0.021	**0.547 ± 0.006**

Table 4 - Table of results of the clinical trials prediction on Phase II

7.3.3 Phase III

The table represents the results of the mentioned metrics according to the five different models ingesting the clinical trials in *Phase III*. In the final phase, *HINTBasic* continues to slightly outperform the other models with ROC-AUC (0.745), F1 (0.854), and PR-AUC (0.535). The consistent improvement in these metrics across phases indicates that *HINTBasic* maintains its reliability as the dataset grows larger and more complex.

Phase III Baseline Results

	Models			
	LogReg	**XGBoost**	**MLP**	**HINTBasic**
ROC - AUC	0.523 ± 0.000	0.519 ± 0.000	0.528 ± 0.000	**0.745 ± 0.007**
F1	0.883 ± 0.274	0.832 ± 0.254	0.834 ± 0.251	**0.854 ± 0.005**
PR - AUC	0.416 ± 0.003	0.414 ± 0.003	0.515 ± 0.003	**0.535 ± 0.010**

Table 5 - Table of results of the clinical trials prediction on Phase III

7.4 Disease groups evaluation

Similarly to past studies (Fu et all, 2022), we have included an evaluation of the model's performance on different datasets separated per groups of diseases. Given that the best performing model is *HINTPlus*, we present the results in Table 6.

Disease groups evaluation with HINTPlus

	HINTPlus		
Diseases Groups	ROC - AUC	F1	PR - AUC
DIG - Phase_I	0.8571 ± 0.0196	0.8405 ± 0.0133	0.7251 ± 0.0195
DIG - Phase_II	0.7777 ± 0.0293	0.8546 ± 0.0116	0.7464 ± 0.0176
DIG - Phase_III	0.7747 ± 0.0321	0.8382 ± 0.0174	0.7219 ± 0.0261
NVS - Phase_I	0.7760 ± 0.0374	0.8297 ± 0.0262	0.7388 ± 0.0374
NVS - Phase_II	0.8309 ± 0.0235	0.8668 ± 0.0182	0.7811 ± 0.0273
NVS - Phase_III	0.8469 ± 0.0362	0.8511 ± 0.0181	0.7413 ± 0.0275
NEO - Phase_I	0.7754 ± 0.0360	0.8294 ± 0.0213	0.7382 ± 0.0314
NEO - Phase_II	0.8349 ± 0.0221	0.8687 ± 0.0122	0.7838 ± 0.0162
NEO - Phase_II	0.8338 ± 0.0430	0.8371 ± 0.0232	0.7205 ± 0.0344
INF - Phase_I	0.7486 ± 0.0338	0.8927 ± 0.0087	0.8190 ± 0.0150
INF - Phase_II	0.7783 ± 0.0181	0.8854 ± 0.0074	0.8092 ± 0.0118
INF - Phase_III	0.7730 ± 0.0231	0.9109 ± 0.0068	0.8425 ± 0.0126
RSP - Phase_I	0.8653 ± 0.0739	0.8504 ± 0.0483	0.7428 ± 0.0712
RSP - Phase_II	0.8131 ± 0.0525	0.7137 ± 0.0419	0.5565 ± 0.0500
RSP - Phase_III	0.8326 ± 0.0854	0.8408 ± 0.0374	0.7271 ± 0.0549

Table 6 - Table of results of the clinical trials prediction on HINTPlus diseases

If the metrics achieved by the *HINTPlus* across specific diseases are compared with the ones accomplished by the general model, it can be denoted the superior performance of the disease specific one. For instance, if the infectious diseases (INF) are considered, the model shows an improved of F1 score for Phase III and an enhanced PR-AUC across all phases, suggesting that model may return less false positives results (a false positive occur when a trial is mistakenly predictive as a success, while it is a failure).

Similarly, in the cases of nervous system diseases (NVS in Phase II and III), neoplasms (NEO in Phase II and III), respiratory diseases (RSP in Phase I and III) and digestive system diseases (DIG in Phase I and III), a better ROC-AUC is observed.

These results demonstrate that the diseases specific model maintains robust accuracy in the predictions of trials outcomes across the diverse phases despite the degree of complexity of the relationship established.

This performance early confirms the reliability of *HINTPlus* as an analytical toll, making it suitable even for highly specific, complex scenarios like clinical trials focusing on individual diseases.

7.5 Results overview

Our models have been compared to the *HINT* model from the original paper, as well as to other baseline models, to better assess the quality of our approach.

HINTBasic consistently outperforms all other baseline models across all phases and metrics, demonstrating its superiority in clinical trial outcome prediction. Its ability to handle unbalanced datasets is evident in the significantly higher F1 and PR-AUC scores. Traditional models like LogisticRegression and XGBoost lag significantly in performance, as they do not include pretraining with the *ADMET* dataset, which limits their ability to incorporate domain-specific drug-related properties into their predictions.

Furthermore, the *HINTBasic* achieves better results compared to the original *HINT*, primarily due to the upgrades operated on the encoders. Specifically, the protocol encoder has been enhanced by utilizing the *ClinicalBERT* model (Wang et al. 2023), which has enabled more accurate and effective encoding of the clinical trial protocols. Indeed, it demonstrates the significant advantage of improving encoders to boost performance in prediction models that rely on textual data.

The *HINTPlus* model, with the addition of the enrollment feature, demonstrates relevant improvements over the *HINTBasic* results. It is to be mentioned that the trend observed in *HINTBasic*, where Phase III and Phase I performed better than Phase II, is also evident in the models we developed. This consistency highlights potential intrinsic differences in trial characteristics across phases that influence the predictive framework performance. Notably, the *HINTBasic* model can overcome most of these challenges, turning it into the most effective approach for predicting outcomes across each phase.

Furthermore, it is noteworthy that the *HINT* paper used the *p_value* and *why_stop*, as well as IQVIA, to categorize outcomes as either successful or unsuccessful, whereas our model relies on a different labeling method developed by us. This distinction in criteria may impact results and partially explain performance variability between our models and those reported in the *HINT* model.

To conclude, these findings illustrate that *HINTPlus* model is our best model, as it clearly showcases the value of implementing tailored approaches and feature enhancements in the field of clinical trial prediction. This is further emphasized by the model behavior on preselected diseases, as argued above, as the *HINTPlus* demonstrated meaningful efficacy by adapting to the distinctive characteristics of each disease, resulting in enhanced predictive accuracy across diverse trial phases and complexities.

8. Discussion

Our work's foundation has been data. Its quality and comprehensiveness have highly influenced our outcomes. However, our data has also limited our work. To provide a balanced interpretation of our findings, we are going to discuss the limitations associated with the data and models used. Additionally, we are going to provide a cost-benefit analysis to measure the model's effect. After that we will dive deeper into the implications of the model's implementations for involved stakeholders.

8.1 Data-related limitations

In terms of data-related limitations, dependency on high-quality and complete data is a central limitation. Due to our model having a multi-modal data architecture, which is leveraging ICD codes, clinical protocols and molecular data, it is highly sensitive to missing values. As a result, this could lead to errors and thus to a significant decrease in predictive accuracy. The fact that our work is based on a neural network model makes it challenging to interpret and validate the reasons for the results. This might also have implications when it comes to changing management, as stakeholders could potentially not have enough trust in the model's accuracy, due to its complex nature and their limited knowledge (Chopra et al. 2024)

Additionally, our work is focused on interventional clinical trials testing drugs. Since we are not limiting prediction on a specific disease, it might fall short due to not considering specific factors influenced by a disease. As a result, our model might lack granularity to predict the clinical trial's success.

This limitation is closely connected to another one we are encountering, which is data scarcity. Although we are working with multi-modal data architecture, it is still possible that our model still does not meet the necessary amount of data to make an accurate prediction. This is even more prominent when considering that our dataset is imbalanced, having more success cases (83,360) than failure cases (19,295). This could happen when researchers want to predict a clinical trial outcome which concerns a rare disease. Due to insufficient data about rare diseases, our model might have to face difficulties since it is not able to learn from robust patterns (Nestor et al. 2018).

8.2 Cost/Benefit analysis

After addressing data-related limitations, it is now equally important to evaluate the economic feasibility of our proposed model. This is done by assessing both costs and benefits to see if stakeholders would benefit from implementation. This calculation’s purpose is to provide critical insights into the model’s value.

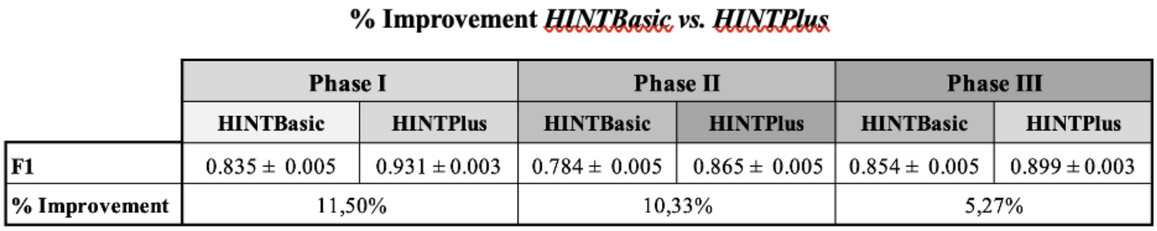
8.2.1 Benefits

The predicted results demonstrated by our *HINTPlus* model can be leveraged to improve the overall efficiency of the trial process. Indeed, one of the main benefits consists in shortening the trial timeline.

Given the integration of this model into a trial streamline, the reduction of the number of failed trials can be assumed to range from 10 % to 15 %. This estimation is achieved through a calculation, broken down as follows.

Given the results in terms of F1 score, observed in Phase III for both the *HINTBasic* and *HINTPlus*, it can be stated that the result achieved by the latter shows better differentiation between successful and failed trials.

Thus, the improvement pursued can be quantified through the following:



% Improvement *HINTBasic vs. HINTPlus*

	Phase I		Phase II		Phase III	
	HINTBasic	HINTPlus	HINTBasic	HINTPlus	HINTBasic	HINTPlus
F1	0.835 ± 0.005	0.931 ± 0.003	0.784 ± 0.005	0.865 ± 0.005	0.854 ± 0.005	0.899 ± 0.003
% Improvement	11,50%		10,33%		5,27%	

Table 7 - Improvements in percentage of HINTBasic compared with HINTPlus

HINTPlus demonstrates an ability of distinguishing successful from unsuccessful higher than 11,50%, 10.33% and 5.23%, respectively to their phases, in comparison to the *HINTBasic*. The comparison has been based on the F1-score metric, as it is the most suitable method to assess the general performance in the case of highly imbalanced datasets, like the one here employed. Since real-world factors such as operational complexity, uncertainty and regulatory implications need to be considered and given that multiple and reliable sources estimate AI improvements on clinical trials to be in the range from 10 % to 30 % (McKinsey 2023), we will assume a net 10 % to 15 % reduction of failed trials. Taking this into consideration, the overall duration may be reduced by 15 % to 30 % (McKinsey 2023).

Specifically, patient recruitment and enrollment stages would be more targeted, and the process would be consequentially shortened. Predictive accuracy may also improve the decision-making process, leading to faster Go/No-Go decisions, which will additionally decrease timelines. If looking at Phase-specific statistics, earlier Phases trials (Phase I and Phase II) could feature time-savings spanning between 6 and 12 months; similarly, Phase III trials could benefit from a reduction approximately equal to the 20-25 % of their average duration (2-4 years).

Hence, if a total development timeline of 15 years is assumed, the model integration could lead to a time saving estimation spanning from 1.5 to 4.5 years.

This will eventually ensure improved efficiency, consequently reducing the operational burden on the workflow. Furthermore, it might be helpful to tailor specific strategies to avoid early trial termination and to enhance patient recruitment.

8.2.2 Lower costs implications

In terms of cost, given the benefits analyzed above, potential savings may be computed by making some assumptions. Assuming that 10 %-15 % of failed trials could be early detected and avoided across all phases, which could potentially save $ 260 to $ 390 million (10 % to 15 %) per each drug development process, if the average $ 2.6 billion expenditure is taken into account. To breakdown this estimate, we can make the following calculations:

Patient recruitment accounts from 30 % to 40 % of the overall trial budget expenditure (Getz et al. 2016), therefore if the model is applied, the cost can be reduced by 15 % at least (e.g. if the cost is equal to $ 15 million, $ 2.25 million could be saved). Phase-specific expenditure span from $ 4 to $ 52 million (DiMasi et al. 2016), depending on the Phase stage; if an average cost of $20 million is considered, savings could be estimated in $3 million (15 %) per phase. Other operational expenses, such as data handling, logistical and outsourcing support, and staff salaries account for almost $1 million per month (Sertkaya et al. 2016), equal to 12 million per year; thus, a timeline reduction of 2 years can reduce the operational charges for up to $24 million.

Despite the positive metrics achieved, our *HINTPlus* model can make errors and provide unreliable answers. Specifically, the error rate is estimated to be around 0.101 (1-F1 score = 1-0.899).

8.3 Business implications based on stakeholders

Building on this calculation, it is also crucial to consider all the potentially involved stakeholders in the usage of our model. Our goal is to examine implications, while also providing actionable insights to support informed decision-making in a real-world setting. The following stakeholders are the focal point of this discussion: Researchers, Pharmaceutical Companies, Regulators, Healthcare Providers and Patients.

Before delving into the potential impact of our work regarding stakeholders, the implication represented by adoption costs need to be addressed; for instance, embracing a machine learning tool into a clinical trial workflow brings numerous limitations, ranging from technology costs to legal allegations. Therefore, adopters might face substantial expenses in terms of resources such as technology acquisition, infrastructure set up, trained personnel recruitment, operational recurring costs and possible outsourcing costs (McKinsey 2023).

Hence, the initial investment to adopt a ML-based model is undeniably resource-intensive, even though it may be argued that the payback period is short, as benefits can be observed in the immediate period following the integration.

Secondly, another implication that has to be highlighted lies in the fact that collaboration and communication across all stakeholders are of significant importance when it comes to successfully implementing machine learning methods into clinical research. An open culture for benefits and drawbacks has to be embraced, while best practice methods need to be shared in order to successfully apply these methods (Weissler et al. 2021, 11-12).

The integration of our model could impact researchers and health care providers work significantly, as it could potentially enhance the success and efficiency of clinical research. The resource optimization that comes with the implementation of our model would also positively impact their work, as the prediction of success or failure would maybe prevent them from starting a clinical trial which is not going to be successful according to our prediction.

However, researchers should also be alerted when it comes to the data used in training the model. To prevent overfitting from happening, which in return causes a poor performance of the model, it is of high importance that researchers validate their models in different settings. This practice ensures that the model generalizes well and is not merely capturing noise in the training data, while not fully considering underlying patterns of the data (Kappen et al. 2018, 4).

The accelerated timelines of drug development do not only benefit researchers but also pharmaceutical companies, as they are able to launch new drugs to the market faster, improving their position in the market as well as their competitiveness and thus potentially increasing their ROI. This is especially vital, since it has been found that AI can significantly reduce drug development. The optimized resource allocation, which comes with the use of AI, can lead to efficient trial design, which impacts savings made in the process of drug development (Weissler et al. 2024, 4-5). The economic impact of drugs being launched into the market on pharmaceutical companies can be measured in a company’s stock prices. The company “BridgeBio Pharma” has seen a significant surge in stock after the FDA has approved their drug based on their clinical trial results, treating a rare heart condition, which might influence market dominance. The company is expected to reach global sales of $ 2.5 billion by 2035 (Investor's Business Daily 2024).

This stands in stark contrast to the company “Cassave Science”, which experienced a failed clinical trial, focused on testing an Alzheimer’s drug. The announcement was followed by a stock drop of 84 %, indicating that the company’s value was highly influenced by the outcome of this clinical trial. Although pharmaceutical companies can highly benefit from the integration of AI in clinical trials, this also comes with limitations. These include data privacy concerns, regulatory and ethical constraints (GlobeNewswire 2024).

This leads us to examine implications and limitations of regulators. The usage of machine learning in clinical trials is highly relying on evaluations or FDA representatives. These suggest that ML practices in clinical trials are seen as high-risk use cases. The reason for this classification is the potential of errors and biases which are influencing the algorithm. While pharmaceutical companies and researchers are limited by regulators such as the FDA, the FDA is limited by the lack of an existing guideline. However, they are open to collaborating with sponsors and stakeholders on a case basis to support implementations (Weissler et al. 2021, 10-11). Implications for patients would firstly be a minimized treatment burden, as researchers already know beforehand that a trial is not going to succeed, which then prevents unnecessary treatments or therapies from happening. This has the potential to save patients from being impacted by the trial physically and/or emotionally. Additionally, prediction of clinical trial outcomes is linked to early identification of risk factors, leading to adjustments of clinical trials which in return could improve overall health outcomes. Since ML implementations also have a positive effect biomedical evidence, they could potentially save humans and reduce their suffering (Kappen et al. 2024).

Future research can focus on addressing these challenges by developing solutions using other frameworks, which could be tailored to regulatory and ethical requirements, while also having higher interpretabilities for all stakeholders. While generalizability and scalability can be improved by training the models on more diverse datasets, which include an even wider range of data, it should also be considered that disease-specific models could be able to make an enormous contribution to the prediction of clinical trial outcomes.

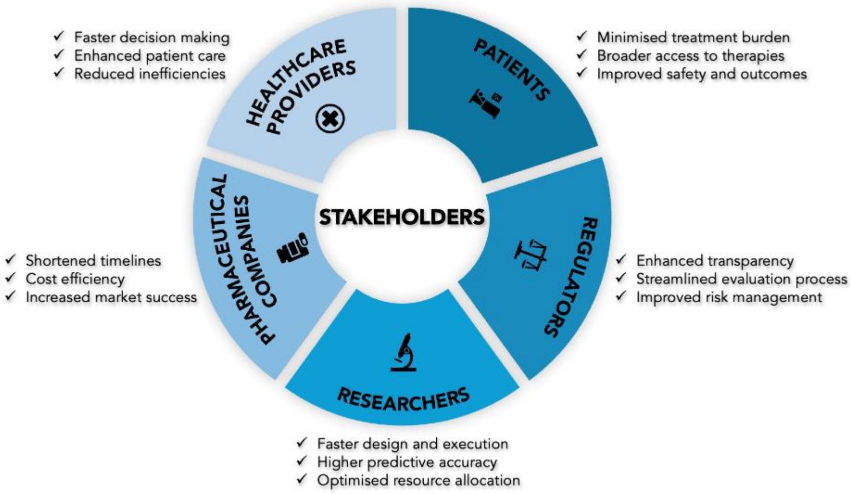

STAKEHOLDERS
HEALTHCARE PROVIDERS
✓ Faster decision making
✓ Enhanced patient care
✓ Reduced inefficiencies
PATIENTS
✓ Minimised treatment burden
✓ Broader access to therapies
✓ Improved safety and outcomes
PHARMACEUTICAL COMPANIES
✓ Shortened timelines
✓ Cost efficiency
✓ Increased market success
REGULATORS
✓ Enhanced transparency
✓ Streamlined evaluation process
✓ Improved risk management
RESEARCHERS
✓ Faster design and execution
✓ Higher predictive accuracy
✓ Optimised resource allocation

Figure 5- Identification of the stakeholders

9. Conclusion

The thesis aims to provide a novel approach at predicting clinical trial outcomes by integrating Large Language Models techniques and leveraging the Hierarchical Interaction Network. The goal of this research is to enhance the process of drug development, which is achieved by proposing the integration of LLMs predictive framework into the trial streamline.

Our *HINTBasic* model's ability to handle a wide range of diverse datasets including integrate ICD codes, molecular data, and trial protocols resulted in significant enhancements in predictive performance. These advancements were exhibited by significant improvements in PR-AUC and ROC-AUC scores in comparison to baseline models, conveying robust differentiation between successful and failed trials. Moreover, by comparing *HINTBasic* with *HINTPlus,* which includes enrollment, we saw improvements underscoring the importance of this new feature in achieving better predictions values, causing it to be our best model. By accelerating trial processes, the model is forecasted to decrease the duration of drug development by 15-30 %, resulting in savings of 6-12 months in earlier phases (Phase I and II) and a 20-25 % duration reduction in Phase III. Given its predictive power, our *HINTBasic* model can handle inefficiencies and significantly lower the high costs associated with failed trials. Indeed, the integration of the framework into a clinical workflow is estimated to reduce trials failures by 10 % to 15 %, potentially resulting in cost-savings equal to $ 260 to $ 390 million for each drug development process, if considered a total expenditure of $ 2.6 billion. For instance, patient recruitment, which is a major cost center accounting for 30-40 % of the trial budget, could feature a cost reduction equal to 15 % due to a more targeted approach.

Although the model exhibits impactful benefits, it still must face challenges such as the model's reliance on high-quality. Data scarcity, specifically for rare diseases, is a significant risk to the prediction's performance.

Several stakeholders face significant implications, while also having to evaluate ethical considerations. While Researcher's and Health Care Provider's work is influenced by better resource allocation, while also reducing the initiation of unlikely successful trials, Pharmaceutical Companies benefit from accelerated drug launches, enhancing market competitiveness and ROI. Patients' health outcome might also be improved through early risk identification, which can have an impact on their whole lifespan. All these stakeholders are affected by regulators, which can enhance the implementation of AI in clinical trials by collaboratively developing regulatory frameworks.

In terms of future outlooks, we believe that actions that might be performed on our current work, to accomplish further improvement, include dataset expansion, as diversity is required to achieve generalizability and scalability.

An example of future work to be implemented is the development of personalized models to address clinical research for specific diseases and in techniques refinement to ensure granularity and accuracy of predictions. Additionally, ethical bias must be addressed, which can be executed through the integration of equity-focused frameworks and guidelines. Conclusively, the thesis depicts benefits and challenges associated with the adoption of machine learning techniques into the field of clinical trials.

Specifically, our work showcases the potential predictive ability of the *HINT* model to revolutionize the trial process, opening the frontier to a more operationally efficient and cost-effective drug development. Its implementation could lead the way for a major adoption of AI-driven models in the healthcare industry, ultimately enhancing the patients' journey by reducing the leap time between treatments development and go-to market, while ensuring optimized efficiency.

10. Incorporating Missing Value Imputation

10.1 Introduction

Handling missing data is a critical challenge in clinical trial predictions, where reliable and accurate models are essential to minimize costly failures. The variable of interest, Enrollment, represents the number of participants in a trial and helps as a key predictor of clinical trial outcomes. Inadequate patient recruitment is one of the main factors contributing to trial failures, especially in the later phases where operational and financial consequences are higher (Fogel 2018; DiMasi et al. 2016). The transition from the *HINTBasic* to *HINTPlus* showed the importance of integrating Enrollment into predictive models. However, the dataset contains 1,183 missing Enrollment values across the training, validation, and test datasets, representing only 1.27% rows with missing values of over 93,386 rows. The *HINTPlus* model addressed these gaps by defaulting the missing values with 0, a simple approach that can introduce bias, distorts feature distributions, and reduces predictive performance, but also compromising the validity of study conclusions (Mercaldo 2018).

This research is based on the *HINTPlus* model by applying advanced imputation techniques, such as K-Nearest Neighbors (KNN), Multiple Imputation (MI), and Decision Tree-based imputations, to replace missing Enrollment values to effectively provide more plausible estimates. By enhancing data quality, this study aims to improve the predictive reliability of the *HINTPlus* model. Accurate predictions are particularly critical for stakeholders because they help optimize resource allocation, reduce trial failures, and support evidence-based decisions in clinical trial design and execution.

10.2 Literature Review

Missing data in clinical trials can result from patient-related factors (e.g., withdrawal, refusal), investigator errors, or unreported responses. Mechanical problems and physician decisions, such as not ordering tests, also contribute to incomplete data (Austin 2021).

Rubin (1976) provided a foundational framework for addressing missing values by separating them into three types. Missing Completely at Random (MCAR) occurs when the probability of missing data is completely independent of both observed and unobserved variables. Missing at Random (MAR) applies when missing values depends only on observed variables. Missing Not at Random (MNAR) happens when missingness depends on unobserved variables or the variable itself. But distinguishing between MAR and MNAR can be challenging, as it often requires understanding relationships between observed and unobserved factors (Van Buuren 2018). The impact of missing data on outcomes can be profound, leading to biased estimates, reduced statistical power, and unreliable conclusions if not appropriately addressed (Mercaldo 2018).

Lo et al. research (2018) on drug approval predictions investigated statistical imputation techniques for addressing missing data while integrating them with multiple models. Initially, basic techniques such as listwise deletion and single imputation methods, including unconditional mean replacement, were employed. While easy approaches, these approaches often lead to significant data loss and reduced variance, introducing bias and underestimates uncertainty in the data (Burren 2018; Austin 2021). However, these simpler methods served as benchmarks for evaluating more advanced techniques.

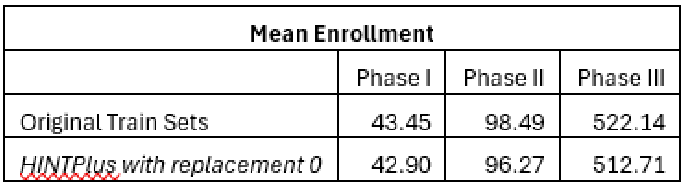
Multiple imputation (MI) is effective under the missing at random (MAR) assumption. By generating multiple plausible datasets for missing values, MI accounts for imputation uncertainty and enhances the robustness of subsequent analyses (Van Buuren & Groothuis-Oudshoorn 2011). The k-nearest neighbors (KNN) method, specifically with five nearest neighbors using Gower distance combined with a random forest, demonstrated high predictive accuracy. However, while effective, KNN can introduce false associations or struggle with precision in high-dimensional data (Emmanuel et al. 2021). Additionally, decision tree-based methods were applied to address nonlinear relationships by partitioning data into homogeneous regions. Despite their utility, these algorithms risk overfitting or bias when used with smaller datasets (Emmanuel et al. 2021).

10.3 Methodology

10.3.1 Data

The data used for this study includes nine files that were previously divided for the *HINT* models, which include three distinct datasets: a training set, a validation set, and a testing set. These datasets have been organized across the three clinical trial phases, which include Phase I, Phase II, and Phase III. A total of 1,184 missing values are present in the Enrollment column across approximately 93,000 rows accounting for all the 9 files. Most of these missing values are present in the training datasets, with 274, 609, and 301 missing values in Phase I, Phase II, and Phase III, respectively (See Appendix, Table 1).

It is noticeable that replacing these missing values with zero, as done in the original *HINTPlus* model, decreased slightly the mean of enrollment across all phases when compared to the mean of the enrollment (non-missing rows) in the original train sets, which may distort the dataset's overall distribution and influence or results in prediction.



Mean Enrollment			
	Phase I	Phase II	Phase III
Original Train Sets	43.45	98.49	522.14
HINTPlus with replacement 0	42.90	96.27	512.71

Table 8 - Enrollment means of train sets before and after replacement

In order to address more effectively the issue of missing enrollment values, several contextual variables have been employed. Eligibility Criteria provides essential context for participant selection, can directly influence on the enrollment numbers, trials with strict eligibility requirements often experience lower enrollment (Fogel 2018). The integration of the International Classification of Diseases (ICD) codes, which describe the diseases under study, offers insights into the enrollment difficulties, especially rare diseases where patient recruitment is challenging (Augustine et al. 2013). The ICD and Eligibility Criteria to derive meaningful insights as predictors, previous embeddings used for the model were applied to these 2 variables. The Status Study applies as another potential predictor, as trials with low or no enrollment is often associated with early termination of the trials, in our case failure in completing the trials (Fogel 2018). This variable was encoded using hot encoding to be treated as a categorical variable in all imputation techniques. The Start Date was included as a feature, extracting the year, acknowledging that trials conducted in earlier years tended to have lower enrollment numbers compared to more recent years (Bieganek et al. 2022). Also, the Phase of the clinical trial affects enrollment expectations. Phase I trials typically involve fewer participants due to their experimental nature, while Phases II and III generally have higher enrollment targets (FDA 2021). Although phase information was not employed as a feature in the imputation models, the datasets were pre-split by phase, thereby ensuring its contextual influence was preserved.

10.3.2 Imputation Techniques

To address the issue of missing values in the Enrollment column, which had previously been substituted with a default value of zero, more advanced imputation methods were employed to more effectively capture the underlying patterns and relationships in the data. In this study, missing data in the "Enrollment" feature is assumed to follow the MAR mechanism, as its absence is related with observed variables such Study Status and Start Date (See Appendix, Table 1 and Graphic 3) (Buuren 2018).

Inspired by Lo et al. (2018), this study employed multiple imputation strategies, such as KNN, Multiple Imputation, and Decision Tree-based methods, to address missing enrollment values in clinical trials. For all techniques, the missing enrollment values were estimated using imputation on the training sets. The imputed training sets were employed as a reference for the imputation of missing values in the test and validation sets, to avoid data leakage. To ensure consistency of our data, all imputed enrollment values were guaranteed to start from 0, to be positive values. Following this, the imputed files for each phase, generated by these techniques were employed to train the *HINTPlus* model, thus facilitating a comparative assessment of their influence on model performance. This enabled us to discover whether more advanced imputation methodologies could enhance the predictive precision of the model in predicting the success or failure of clinical trials.

The KNN imputation approach used Gower's distance metric (Gower 1971) instead of the traditional Euclidean distance to effectively deal with mixed type variables, both numerical and categorical data (Lo et al. 2018), ensuring accurate neighbors' selection in the KNN imputation process. This metric effectively captured the characteristics of both numerical and categorical variables, including normalized numeric embeddings (such as eligibility criteria and ICD codes) and one-hot encoded categorical variables like Study Status. For each row with missing values, the Gower distances were calculated to identify the most similar rows with observed values (k=5 and k=10). Afterwards, the mean enrollment value of the neighbors was calculated and employed to impute the missing value, ensuring that the imputed values reflected patterns within the dataset and aligned with the overall data distribution.

The Multiple Imputation method was implemented using the Iterative Imputer from the Scikit-learn library, based on the Multivariate Imputation by Chained Equations (MICE) framework (Van Buuren & Groothuis-Oudshoorn, 2011). This method performs an iterative process modelling missing values in one variable as a function of the other, which enables the generation of multiple plausible datasets while addressing uncertainty in predictions. To guarantee the reliability of the estimates, the imputer was configured to perform a maximum of ten iterations per imputation and three imputed datasets were generated as the missing values are not in big proportion (Austin 2021). The subsequent phase of the study entailed the analysis of the imputed datasets. This was conducted by training the *HINT*Plus model on each of the three imputed training data sets and evaluating its performance independently on the validation and test sets. Ultimately, combine estimates from multiple analyses using Rubin's rules (Lo et al. 2018). Having $\bar{Q}$ being the combined estimate, m the number of imputed datasets, and $\tilde{Q}_l$ the estimate from the lth analysis with the pooled estimate.

$$\bar{Q} = \frac{1}{m}\sum_{l=1}^{m}\widetilde{Q_l}$$

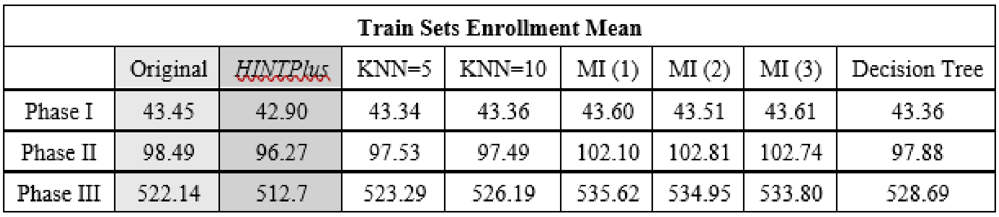
I: Average of estimates for m imputed datasets according to Rubin's Rule

The Decision Tree based imputation was chosen for its ability to capture intricate feature relationships without extensive preprocessing (Emmanuel et al. 2018), and its ability to address nonlinear relationships. The tree divides the data into regions based on the values of the features, with the aim of reducing the variance within each region to make an accurate prediction regarding enrollment. To ensure reproducibility, a fixed random seed (random state=42) was employed.

Evaluation Metrics

Given the imbalance in the dataset's success and failure cases, the F1 score was chosen as the main evaluation metric in order to be comparable with our *HINTPlus*. The F1 score balances precision and recall, offering a unified measure of a model's performance that considers both false positives and false negatives (Van Rijsbergen 1979; Sokolova & Lapalme 2009). Since the new data imputed is really a small amount, the not rounded statistical values are used to compare in order to capture the minor changes that will happen when compared to our baseline *HINTPlus*.

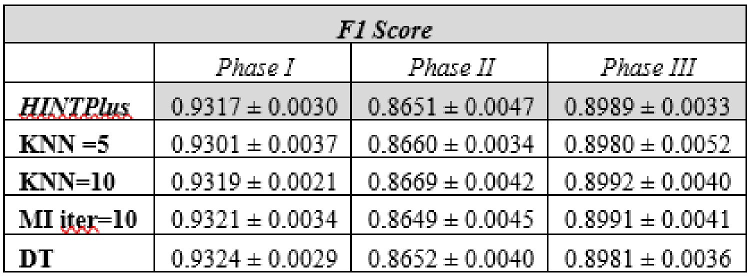
10.4 Results



Train Sets Enrollment Mean								
	Original	*HINTPlus*	KNN=5	KNN=10	MI (1)	MI (2)	MI (3)	Decision Tree
Phase I	43.45	42.90	43.34	43.36	43.60	43.51	43.61	43.36
Phase II	98.49	96.27	97.53	97.49	102.10	102.81	102.74	97.88
Phase III	522.14	512.7	523.29	526.19	535.62	534.95	533.80	528.69

Table 9 - Train set Enrollment means after imputation

After the datasets are being imputed mainly on the train sets where majority of the missing are located, a shift in the mean of the Enrollment column is observed. Compared to the dataset used inside of our model *HINTPlus* where missing values were substituted per 0 directly, the imputed datasets show closer means to the original train sets values, observed in Table 10.



F1 Score			
	Phase I	*Phase II*	*Phase III*
HINTPlus	0.9317 ± 0.0030	0.8651 ± 0.0047	0.8989 ± 0.0033
KNN =5	0.9301 ± 0.0037	0.8660 ± 0.0034	0.8980 ± 0.0052
KNN=10	0.9319 ± 0.0021	0.8669 ± 0.0042	0.8992 ± 0.0040
MI iter=10	0.9321 ± 0.0034	0.8649 ± 0.0045	0.8991 ± 0.0041
DT	0.9324 ± 0.0029	0.8652 ± 0.0040	0.8981 ± 0.0036

Table 10 - F1 Score comparison between HINTPlus after imputation of missing values in Enrollment

The table presents a comparison of F1 scores across Phase I, Phase II, and Phase III for a variety of imputation techniques. The techniques under consideration are KNN with k=5 and k=10, multiple imputation (MI) which the combine estimate was calculated (see Appendix, Table 3), and decision tree-based imputation (DT). Although only 1,183 cells show missingness, the results indicate that robust imputation strategies are essential even for small proportions of missing data, as they have shown slight improvements in predicted performance across some phases.

In Phase I, the KNN method with k=5 presented a lower performance relative to the baseline, with an F1 score of 0.9301. While with k=10 demonstrated a slight improvement, with an F1 score of 0.9319. The MI method yielded comparable results with 0.9321, while the DT method demonstrated the highest performance, with a score of 0.9324 compared with *HINT*Plus.

In Phase II, all imputations displayed marginal differences compared to the baseline of 0.8651.

The KNN method with a neighborhood k=5 exhibited a slight improvement in performance, achieving a score of 0.8660. While the k=10 demonstrated the highest score of 0.8669. In contrast, the values for MI and DT were 0.8649 and 0.8652, respectively, which remained close to the baseline.

In Phase III, *HINT*Plus achieved an F1 score of 0.8989. The KNN (k=5) method demonstrated the lowest performance, with an F1 score of 0.8980. While the KNN (k=10) and MI methods demonstrated small improvements, with F1 scores of 0.8992 and 0.8991, respectively. The DT, with an F1 score of 0.8981, performed similarly to the KNN(k=5) both with lower values compared to baseline.

10.4.1 Overview

The results across all phases demonstrate slight, yet statistically significant, in most cases improvements over the baseline *HINTPlus* where missing values are replaced directly when training with 0, with performance variations contingent on the imputation technique and phase. It is to be mentioned that KNN (k=10) consistently demonstrated better performance across all the 3 phases, achieving the highest F1 score in Phase II 0.8669 and Phase III 0.8992, which are phases where failure costs are particularly critical. Multiple imputation (MI) also performed well in Phase I and Phase III, while KNN (k=5) and decision tree-based imputation (DT) showed less robust performance, especially in Phase III.

10.5 Discussion

The results illustrate the efficacy of advanced imputation techniques in enhancing model performance, even in the presence of minimal missing data. Among all the methods tested, KNN (k=10) consistently yielded the most reliable results, achieving the highest F1 scores in Phase II (0.8669) and Phase III (0.8992), where predictive stability and financial stakes are of particular importance due to the high costs associated with trial failures in later phases (DiMasi et al. 2016). Its outstanding performance can be attributed to its capacity to capture a broader neighborhood of similar data points, which serves to reduce bias and produce imputations that are both smoother and more accurate.

Multiple Imputation (MI) demonstrated also an increased performance, particularly in Phase I and Phase III, through the generation and combination of multiple estimates for missing values. This approach effectively addresses uncertainty, ensuring stable and reliable predictions. In comparison, KNN (k=5) performed less effectively due to its smaller neighborhood size, limiting the amount of contextual data available for imputations. Similarly, the decision tree-based imputation (DT) method encountered difficulties with data variability, particularly in Phase III, where predictive stability is critical due to the consequences in terms of costs associated with trial failures (DiMasi et al. 2016). While effective in earlier phases, these methods exhibited reduced adaptability in later phases with higher demands for better performance.

For our stakeholders, the use of better imputation methods can increase the credibility of predictive outputs as well as data quality, which can help on supporting the decision-making process and the clinical trials plan and execution (Afkanpour et al. 2024; Wang et al. 2022). In addition, it helps to allocate better resources, minimize trial failures and reduce costly timelines. In addition, proper handling of missing data ensures that external tools, such as automated reporting systems that use our data sets, produce reliable results. This is critical because incorrect data handling, such as replacing missing values with zeros, can lead to erroneous conclusions (Mercaldo 2018).

Despite the improvements observed, this study shows some limitations, including the relatively small proportion of missing data, which may have constrained the extent of the improvements. Furthermore, the complex imputation methods used, while demonstrated slight improvements, are inherently time-consuming, particularly with the *HINTPlus* model estimated run time of 8 hours per execution. This has constrained the opportunity to compare with a greater number of techniques or even alternatives parameters within the techniques used.

Future research could consider listwise deletion or single imputations, especially given the very small volume of missing data, to determine if these could also yield accurate results as compared with more complex imputations that are more time computational costly (Heymans et al. 2022), although under the assumption the MAR is not the best approach (Buuron 2018). Or even be used as benFurthermore, implementing more efficient imputation strategies earlier in the data processing pipeline, such as immediately after data retrieval from sources like *ClinicalTrials.gov*, might ensure cleaner, more complete datasets for analysis. Lastly, treating missing values with some advanced imputation methods not only improved tried to improve the model performance but also increased the quality and the reliability of the data, allowing for more informed and robust decision making in clinical trial predictions.

10.6 Conclusion

The objective of this study was to address the issue of missing values, which *HINTPlus* had initially substituted with 0, by applying more plausible estimates using advanced imputation techniques, including KNN, multiple imputation, and decision tree-based methods.

The KNN (k=10) technique was identified as the optimal approach, demonstrating consistently superior prediction values across all three phases when compared to other methods. Also, it outperformed the default approach of replacing missing values with 0. Although the small proportion of missing data, the findings underscore the critical role of robust imputation techniques in enhancing both model performance and data quality. Proper handling of missing values allows for more reliable and trustworthy outputs, which can later assist the stakeholders in making informed decisions about clinical trial design, resource allocation, and execution.

However, the study faced some constraints particularly due to computational intensity, which restricted further exploration of alternative techniques or parameters. Further research should be conducted to ascertain whether more straightforward techniques that are based on the missing-at-random (MAR) assumption can achieve comparable results to more complex methods when the proportion of missing values is relatively low. Additionally, integrating imputation into the data collection process in clinical trials could provide more data to improve model performance.

11. Factors Driving Predictive Outcomes

11.1 Introduction

In recent years, the application of deep neural networks to various fields has grown, along with growing attention to their interpretability (Wang 2023). Although the concept of interpretability is neither unified nor clearly identified, it could commonly be explained as “the degree to which an observed can understand the cause of a decision” (Miller 2019). In recent years, the application of deep neural networks to various fields has grown, along with growing attention to their interpretability (Wang 2023). Despite the concept of interpretability is neither unified nor clearly identified, it could commonly be explained as “the degree to which an observed can understand the cause of a decision” (Miller 2019). Interpretable models benefit the relationship between the humans and the machines in gaining trust and assisting in the decision-making process (Wu 2022), thanks to their ability to clearly define causal relationships.

This present work aims to investigate the interpretability of the *HINTPlus* model in predicting clinical trial outcomes, with a particular emphasis on assessing its capacity to provide meaningful insights and identifying the key factors that influence its predictive performance.

11.2 Literature review

Deep learning models, particularly Graph Neural Networks (GNNs), are often criticized as "black boxes" due to their intricate architectures, characterized by extensive connections between neurons. While this complexity drives their high performance across various tasks, it also makes their causal interactions challenging to interpret.

Consequently, researchers have developed a range of methods to improve their interpretability. One common approach is the black-box method, which simplifies the model’s interpretability by replacing it with a more interpretable surrogate, such as a linear regression or decision tree. For instance, PGM-Explainer (Minh N. Vu 2020) uses probabilistic graphical models to approximate GNN decisions, providing clearer insights into their behavior. Another widely used technique is approximation-based methods, which rely on gradients and model parameters to determine the importance of features. An example of this is Integrated Gradients, proposed by Sundararajan et al. (2017), which calculates the contribution of individual features to the output, addressing challenges like gradient saturation to offer robust explanations.

Some models also incorporate interpretability directly into their structure, in this respect, researchers have leveraged this property, as demonstrated in the case of Heterogeneous Attention Networks (HAN) (Xiao Wang 2019), where attention mechanisms were used to assign weights to nodes and edges, enabling researchers to understand their importance to the prediction task.

Furthermore, evaluating the trustworthiness of the prediction results is essential in the context of decision-making processes. In this regard, researchers have developed *fidelity metric*, as outlined by Pope et al. (2019), which measures the impact of removing important features on the model's performance, ensuring that explanations effectively highlight key components.

11.3 Objective

In this paper, we aim to explore the interpretability of the *HINTPlus* model in predicting clinical trial outcomes, focusing on its insights and key predictive factors.

As previously highlighted, interpreting the model is key to making informed decisions, particularly in the context of predicting the outcomes of clinical trials. More specifically, it allows for the identification of critical factors that most significantly influence the success or failure of a trial. This, in turn, enables the potential optimization of the trial, leading to a higher probability of success and a consequent saving of resources.

HINTPlus is designed as a Graph Convolutional Network (GCN) model, a specialized type of Graph Neural Network that incorporates an attention weight mechanism (Fu 2022). The model combines Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs) to process multi-modal data within a graph structure, dynamically refining relationships through an attention matrix. The inclusion of the multi-head attention mechanism further enhances interpretability by assigning weighted importance to edges between nodes. Furthermore, the attention matrix offers a clear understanding of how node relationships influence predictions, significantly contributing to the model's relevance.

Despite its intricate and complex architecture, the model’s layered design enables a step-by-step interpretability process, making its analysis more manageable to study.

11.4 Methodology

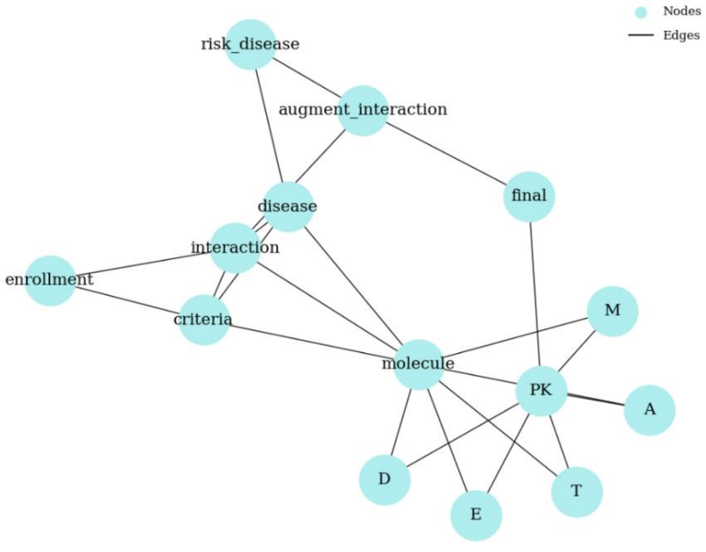


Figure 6 - The image illustrates the graph structure of the HINTPlus model, showcasing its nodes and edges.

Building on the prior discussion, the analysis of the *HINTPlus* follows a step-by-step approach aligned with the model's layered structure. Specifically, we began by evaluating the overall attention matrix, which captures the relationships across the entire model. Subsequently, we segmented the model into its two primary components—the GAT and the GCN—and conducted a separate analysis, focusing on their respective contributions to the model's predictions and on the importance of the individual features. Before diving into the step-by-step analysis, a comprehensive overview of the model's structure is presented to provide context. The model's graph integrates hierarchical processing levels with defined nodes and edges in its structure. The nodes represent key components of clinical trial data, such as molecules, diseases, enrolment, and criteria, which are the primary features. Additional nodes, such as interaction node, risk disease, and augment interaction, represent intermediate processing layers where features are combined or transformed. The *ADMET* nodes, labelled as

A, D, M, E and T (absorption, distribution, metabolism, excretion, and toxicity) capture specific drug properties, while PK node aggregates these into a *pharmacokinetic* representation. The final node integrates all processed information to predict the clinical trial outcome. The edges in the graph structure represent the relationships between nodes, linking the features and intermediate transformations in a way that reflects their dependencies and interactions. These edges guide the flow of information throughout the network, as illustrated in Figure 1.

11.4.1 Global – layer analysis: Attention Matrix Heatmap on GNN

Attention matrices are a fundamental tool in Graph Neural Networks (GNNs) for quantifying the importance of connections between nodes by integrating their intrinsic features with the structure of their relationships. They serve not only as an interpretive mechanism but also as a decision-making tool, enabling models to focus on the most relevant relationships (Xiao Wang 2019). In GNNs, attention matrices dynamically assign weights to edges based on their relevance in the context of the data, allowing the model to adapt to complex and noisy environments.

To provide a global perspective, the attention matrices for each instance in the analyzed model are aggregated by averaging, which computes the element-wise mean of attention scores across batches. This technique ensures stability and robustness by reducing the influence of noise and outliers, preserving the general structure of node relationships. A similar technique has been applied in Transformer models for text classification, where averaging attention scores supports consistent and interpretable outcomes (Mylonas 2023). Following aggregation, the attention matrix is visualized as a heatmap to highlight the most relevant features for all the three different phases of the clinical trials. These visualizations make the model's feature importance interpretability straightforward by clearly indicating which relationships are most influential.

11.4.1 Layer - level analysis: GCN and GAT

Building on these architectures, the *HINTPlus* model combines GCNs and GATs to process multi-modal data within a graph structure, dynamically refining relationships through attention mechanisms. Graph Convolutional Networks (GCN) employ localized graph convolutions to combine graph structure with node features, enabling effective information propagation among neighboring nodes using a fixed adjacency matrix. This layer-by-layer approach enriches node representations with contextual information, ensuring both scalability and computational efficiency (Kipf 2016). Whereas Graph Attention Networks (GAT) address the limitations of spectral-based graph neural networks by incorporating multi-head self-attention mechanisms to dynamically weight the importance of neighboring nodes during the aggregation process, enhancing both stability and learning effectiveness (Petar Veličković 2018).

The analysis of feature importance on the two models across the three phases was conducted using Captum's interpretability framework, which provides state-of-the-art algorithms that enable researchers and developers with an easy solution to understand how the model learns. This framework emphasizes scalability, evaluation, and the adaptation of feature importance metrics across modalities, making it especially suitable for analyzing GNNs (Reblitz-Richardson 2020). Moreover, the two models were implemented using PyTorch and PyTorch Geometric (Matthias Fey 2019), leveraging the Captum library for attribution methods. All computations were executed on an NVIDIA GeForce GPU with CUDA 12.0 support.

The experiment evaluated the contributions of features to the model’s predictions across the dataset’s three phases using an approximation-based method, such as Integrated Gradients. This method measures how much each feature of the input contributes to the model’s prediction. By moving from a neutral baseline to the actual input and summing up the gradients along the way, it provides a clear and fair explanation of the model’s behavior (Sundararajan 2017).

11.5 Results

In this section, we applied the previously described methodologies to analyze *HINTPlus* and explained the results obtained at each step of the process. This approach follows the structured step-by-step framework outlined earlier, ensuring consistency in methodology.

11.5.1 Global – layer analysis: Attention Matrix Heatmap Analysis

The aggregated attention heatmaps for Phases I, II, and III provide a clear visualization of the attention weights assigned by the model across datasets (See Appendix**,** Graphic 4 – 5 – 6**)**.

A key pattern is the consistent prominence of the augment interaction to absorption node (A) connection, with high weights of 0.99 in Phase I and 0.96 in later phases, highlighting its universal importance in the model's decision-making process. Interestingly, an asymmetry emerges in the reverse connection (absorption to augment interaction), where the weight drops decrease slightly from 0.96 in Phase I to 0.50 in Phase II. This suggests directional dependency, with A relying more heavily on augment interaction than vice versa.

In Phase I, other significant relationships include the risk disease to absorption node(A) connection with high weights of 0.95 and molecule to interaction node with 0.86, emphasizing their influence. Connections like interaction node to disease play moderate roles, while criteria to unrelated nodes show minimal impact. *ADMET* nodes also contribute moderately to the final node and enrolment, with weights ranging from 0.50 to 0.55. Phase II shifts attention to molecule node to interaction node and interaction node to disease, showing a phase-specific emphasis on interaction node. The final node to enrolment connection rises to 0.64, indicating an evolving priority in this phase. While in Phase III, the final node to enrolment connection becomes a focal point, rising sharply to 0.90. Meanwhile the risk disease to A and molecule to interaction node connections remain relevant, the importance of interaction node to disease drops to 0.50, indicating reduced relevance in this phase.

Overall, the model demonstrates a balance between universal patterns, such as the augment interaction to absorption node (A) connection, and phase-specific adaptations, like the increased importance of interaction node in Phase II and enrolment prioritization in Phase III.

11.5.2 Layer - level analysis: Graph Convolutional and Graph Attention Networks

The analysis of the *GCN* and *GAT* components using *Integrated Gradients* was conducted across the three phases of the model. The results of the integrated gradients, presented in Table 1, are normalized in the code, scaling the attribution values to range between 0 and 1 for consistent interpretation and visualization. The attribution scores quantify each feature's direct impact on the model's predictions: positive scores indicate features that increase the model's confidence, while negative scores identify features that decrease it**.**

Integrated Gradients results among all nodes in GCN and GAT models

	GNC			**GAT**		
	Phase I	*Phase II*	*Phase III*	*Phase I*	*Phase II*	*Phase III*
molecule	-3.890	-6.809	-4.381	-1.551	-2.242	-1.517
disease	0.084	0.019	0.000	2.511	-0.011	0.000
enrollment	-0.001	-0.007	0.000	0.002	0.000	0.034
criteria	1.020	1.018	1.219	1.238	0.842	-0.538
Interaction	0.019	-0.029	-0.010	0.010	-0.012	0.000
risk_disease	-0.004	0.000	-0.033	-0.014	-0.018	0.000
augment_interaction	0.050	0.000	0.000	-1.155	2.451	0.000
A	-0.727	-2.131	-1.601	0.088	-0.078	0.026
D	-0.920	-3.239	-2.119	0.250	-0.305	0.039
M	-0.924	-1.097	-1.983	0.084	-0.052	0.054
E	-0.610	-1.298	-0.455	0.091	-0.081	0.392
T	-0.571	-1.641	-0.698	0.421	-0.095	0.590
PK	-0.450	-3.791	-0.588	3.712	-0.253	1.891
final	-1.270	-0.040	-3.559	-0.350	-0.060	-0.083

Table 11 - The table presents the integrated gradients computed for each node in the GCN and GAT models within the HINTPlus framework.

In the *GCN, criteria node* consistently emerges as the most significant feature across all phases, especially in Phase I and II with normalized attribution value of 1.0, and 0.679 in Phase III, underscoring its central role in predictions. Another relevant contribution is given by the augmented interaction node in Phase II and Phase III with 0.604 and 1.0 normalized attribution values, respectively. Conversely, molecules consistently exhibit strong negative contributions, particularly in Phase I, suggesting it detracts from prediction quality. Other features, such as augment interaction, and nodes like interaction and risk disease show minimal contributions, reinforcing the GCN’s dependence. While the *ADMET* and pharmacokinetics nodes show consistently a slight negative contribution in all three phases, particularly in Phase III.

GAT demonstrates a dynamic adaptation of feature importance across phases. In Phase I, the criteria node is dominant with a score of 1.0, supported by pharmacokinetics (0.712) and disease (0.511). Additionally, the toxicity node contributes positively with a score of 0.421. In Phase II, augment interaction becomes critical, reflecting its importance in interaction-heavy tasks. Conversely, the trial criteria node detracts significantly from prediction quality with a score of -0.842, alongside pharmacokinetics and distribution nodes, which contribute -0.253 and -0.305, respectively. By Phase III, pharmacokinetics takes precedence with a score of 1.0, supported by toxicity at 0.590, while the criteria node detracts with a score of -0.538. By Phase III" is repeated twice. By Phase III, pharmacokinetics (PK) takes precedence with a score of 1.0, supported by toxicity (0.590), while the criteria node contributes negatively (-0.538).

Overall, the analysis highlights key differences between the GCN and GAT models in their feature attribution and adaptation strategies across phases. The GCN demonstrates a reliance on stable features, such as the criteria node, which consistently drives predictions, while molecules and *ADMET* nodes exhibit persistent negative contributions.

In contrast, the GAT dynamically adapts to phase-specific demands, emphasizing criteria, pharmacokinetics, and disease in Phase I to drive performance. However, in Phase II, criteria and pharmacokinetics shift to detract from the model, while by Phase III, pharmacokinetics dominates alongside toxicity, with criteria contributing negatively. This adaptability contrasts sharply with GCN’s reliance on stable features.

These insights underscore the models' differing strengths in leveraging feature importance within the *HINTPlus* framework.

11.6 Discussion

HINTPlus represents a significant advancement in leveraging machine learning for clinical trial optimization. By incorporating diverse data modalities—such as molecular properties, disease characteristics, number of enrolment, and trial protocols—it provides a comprehensive framework for analyzing and predicting trial outcomes.

Interpretability is a crucial feature of the *HINTPlus* model, setting it apart from many other machine learning frameworks often criticized as 'black boxes.' By employing tools such as attention heatmaps and integrated gradients method, the model highlights key relationships and validates feature importance, aligning its predictions with clinical domain knowledge. These insights not only improve confidence in the model's outputs but also could offer improvements in the guidance for clinical trial design. While the *HINTPlus* model demonstrates itself to be efficient, the analysis conducted has highlighted areas where improvements are needed. These insights provide valuable direction for enhancing the model's performance and addressing its improvements.

One significant challenge is the consistent negative contribution of the molecule feature across all trial phases. This negative gradient indicates that the molecular data, as currently represented in the model, detracts from prediction accuracy rather than enhancing it. This could stem from limitations in how molecular features are encoded, such as insufficient granularity or a failure to capture complex biochemical interactions. Additionally, this negative molecular score might indirectly influence *ADMET* nodes in the GCN, causing them to exhibit negative contributions as well. Since *ADMET* properties depend on molecular characteristics, inaccuracies or insufficient representations in the molecular data may propagate through the graph, reducing the reliability of these nodes.

In addition to the molecule challenge, the feature importance analysis shows the complementary strengths and weaknesses of the GCN and GAT components in the *HINTPlus* model. The GCN consistently focuses on stable features like trial criteria but struggles to adapt to the changing demands of different trial phases. In contrast, the GAT adjusts feature importance dynamically based on phase-specific needs, effectively highlighting augment interaction in Phase II and pharmacokinetics information in Phase I and III.

To better align the strengths of these components with the specific needs of each phase, a phase-specific weighting strategy could be implemented. This strategy would leverage the stability of the GCN and the adaptability of the GAT to maximize predictive accuracy across all trial phases**.** In Phase I and II, the GCN should be prioritized over the GAT, as its reliance on the criteria node ensures a stable foundation aligned with early-phase trial priorities. By contrast, in Phase II, the GAT should take precedence to emphasize augment interaction, a critical feature for tasks involving complex relationships and dynamic trial demands. In Phase III, the GAT's ability to capture intricate phase-specific dynamics becomes increasingly relevant, allowing it to focus on pharmacokinetics—a dominant factor in later-phase predictions—and toxicity. Simultaneously, enrolment should be given greater importance in Phase III to reflect its growing impact on trial outcomes at this stage.

Implementing this adjusted collaboration between GCN and GAT can be achieved through mechanisms such as "a priori weighting", which would assign fixed importance levels to the GCN and GAT based on phase-specific priorities. For example, GCN could dominate in Phases I and II by emphasizing stable foundational features like criteria, while GAT's flexibility could allow it to excel in Phases II and III by highlighting dynamic features such as augment interaction and pharmacokinetics.

By deploying this phase-specific collaboration between GCN and GAT, the *HINTPlus* model can effectively balance stable features with dynamic adjustments, enhancing its ability to address the shifting priorities of clinical trials across different phases.

11.6.1 Limit of the study

A significant limitation of the study is the absence of a second validation layer to confirm the correctness of feature importance analysis. The current study demonstrates significant advancements in enhancing the interpretability of the *HINTPlus* model, but it lacks a comprehensive ablation study to systematically validate the findings. By isolating individual features, removing them one at a time, and assessing their impact on model performance, this analysis would provide deeper insights into the unique roles each feature plays.

Moreover, the attention matrix framework, while effective in stabilizing feature importance assessments, presents certain trade-offs. Although averaging attention scores reduces noise and highlights general patterns, it may dilute the significance of rare but impactful relationships. Exploring alternative approaches to balance stability with the preservation of these critical connections will be an important focus in future work. Together, these steps will deepen the understanding of *HINTPlus* predictive processes and strengthen its applicability to clinical trial optimization.

11.6.2 Future directions

To address these challenges and further enhance the *HINTPlus* model, several solutions for improvement can be pursued. First, as discussed earlier, the representation of molecular data should be revisited. By incorporating more granular features—such as drug-target interactions, biochemical pathways, or 3D molecular structures—could significantly improve the model's overall performance.

Second, alternative aggregation methods for attention matrices could be explored. Besides the current averaging method, different techniques should be experimented and applied to provide a better representation of relationships, preserving critical insights.

Another promising direction involves refining the model's architecture to align with the specific demands of different trial phases. Building on the complementary strengths of GCN and GAT, structural adjustments can further improve adaptability and prediction precision across phases, ensuring a balance between stable foundational features and dynamic adjustments.

Building on these observations, future research should aim to address these limitations and in addition extend the scope of interpretability analysis. Indeed, the scope of evaluation metrics should be broadened: for instance, adopting fidelity as a metric, as proposed by Pope et al. (2019), would help validate whether the explanations truly capture the essential features influencing the model's performance. Furthermore, exploring advanced interpretability techniques like the PGM-Explainer (Vu and Thai, 2020) could complement existing methods and provide deeper insights into the model's decision-making processes. Such methods could enhance the capacity to explain complex relationships and further unravel the "black box" nature of the model.

11.6.3 Future implications

The practical implications of these improvements are significant. In fact, by enhancing its ability to predict trial outcomes, the *HINTPlus* model could significantly reduce the time and cost associated with clinical trials. Early identification of trials with a high likelihood of failure would enable researchers to allocate resources more efficiently, focusing on studies with greater potential for success. This could lead to substantial financial savings while accelerating the development of effective treatments. Moreover, the model's interpretability features, and actionable insights would empower stakeholders across the clinical trial ecosystem, from researchers and clinicians to regulatory bodies, to make more informed decisions.

11.7 Conclusion

This study presents the *HINTPlus* model, a graph-based neural network framework designed to predict clinical trial outcomes and provide interpretability to its decision-making process.

The model's interpretability development follows a structured methodology, analyzing the model's components to address its complexity. Through global analysis, attention matrices are aggregated and visualized to capture overarching feature relationships across all trial phases. Layer-level analysis allows for the independent evaluation of Graph Convolutional Networks

(GCN) and Graph Attention Networks (GAT), demonstrating their complementary roles in balancing stability and adaptability.

The primary contribution of this paper is demonstrating *HINTPlus* ability to dynamically assess relationships within clinical trial data by leveraging the complementary strengths of Graph Convolutional Networks for stability and Graph Attention Networks for adaptability. This enables the model to effectively balance foundational features with phase-specific dynamics. Secondly, the use of attention matrices and integrated gradients provides actionable insights into feature prioritization, enhancing the transparency and interpretability of the model's predictions. Lastly, the establishment of a robust methodological framework for systematically analysing feature importance ensures a comprehensive evaluation of the factors driving the model's decision-making process.

Overall, the *HINTPlus* model represents a powerful tool for transforming clinical trial design and outcome prediction. Its strengths in multi-modal integration and interpretability make it an asset in an increasingly data-driven field. However, addressing the identified limitations will be critical for realizing its full potential. By pursuing the proposed improvements, the *HINTPlus* model could not only enhance the efficiency and effectiveness of clinical trials but also contribute to the broader goal of advancing medical innovation while reducing costs and improving patient outcomes.

12. Enhancing Trial Predictions through Uncertainty Quantification

12.1 Introduction

This paper focuses on improving the predictive performance of the *HINTPlus* model by addressing uncertainty quantification through the integration of the selective classification technique.

Assessing uncertainty guarantees reliability of the model results, which are of extreme importance in the field of clinical trials, where predictions have a direct influence on the study process and quality of outcomes; indeed, poor results accuracy has been identified as the most frequent root cause for failures in later stage clinical studies, with trials not reaching their endpoints due to inconsistent results (Arrowsmith 2011).

Furthermore, uncertainty can impact the process of patient recruitment and retention, since models may fail to properly classify participants, leading to inappropriate inclusion or exclusion criteria and resulting in skewed trials outcomes (McHugh et al. 2019).

In economic terms, the costs associated with uncertainty are substantially high, since include the expenses for extended trials, further analysis and the opportunity costs related to the potential delays in the launch of the new treatment (DiMasi et al. 2016).

In cardiovascular drug trials, for example, reducing uncertainty and streamlining trial design have been identified as key strategies to lower costs while maintaining trial integrity (Califf et al. 2005).

Therefore, uncertainty implies multiple consequences for clinical trials, which span from the reliability of study outcomes to the economic and operational impacts on the trial pipeline.

To mitigate these effects, uncertainty quantification mechanisms can be employed with the aim of improving prediction accuracy and ensuring the compliance with the integrity standards. Such techniques are fundamental to shorten the overall drug development process and guarantee the efficiency needed to deliver effective treatments to patients. (Chen et al. 2024)

Selective classification is employed with the aim of allowing the model to withhold predictions when faced with input samples characterized by ambiguity or low confidence.

Conversely, this method improves the overall accuracy of predictions by ensuring that the model only makes decisions when it is confident enough about the outcome. (Chen et al. 2024)

12.2 Literature Review

Uncertainty quantification has emerged as a vital aspect of machine learning, particularly in domains where high-stakes decisions are made, such as healthcare and clinical trials.

Numerous methods to address the issues associated with uncertainty in the clinical research industry have been developed in the last decades and provided multiple efficient solutions.

One approach is indeed represented by the conformal prediction methods, which are model-agnostic and establish the threshold based on the calibration sets to guarantee reliable and rigorous predictions. Hence, conformal prediction methods generate prediction sets that include the true label with a user-specified probability.

These frameworks have been formalized by Vovk et al. (2005) and their application mostly involved binary classification problems, including clinical research and diagnostic, mainly due to their ability to provide explicit, non-asymptotic guarantees.

Another technique to quantify uncertainty lies in the Bayesian methodologies, in which model parameters are not treated as fixed values, but rather as probabilistic distributions (Nemani et al. 2023).

The Bayesian method appear to be particularly suitable in the case of quantifying epistemic uncertainty, which occurs from limited knowledge in the process of generating data. In the field of clinical trials, Bayesian models have been employed with the aim of estimating the likelihood of success for the studies, by combining prior knowledge and observational data. It has been argued by Nemani et al. (2023) how Bayesian methodologies may lead to significantly improve robustness if integrated in the model streamline, despite their computational demands might frequently require the aid of approximations such as Monte Carlo sampling or variational inference.

Ultimately, Lakshminarayanan et al. (2017) argued the efficacy of deep ensembles application in the context of uncertainty quantification; this method is implemented by training numerous neural networks using the same architecture but employing different initializations. This approach allows ensembles to capture both epistemic and aleatoric uncertainties, the latter deriving from inherent noise in the data. Deep ensembles have been applied in the domain of clinical trials to predict outcomes by aggregating predictions across multiple models. This methodology was particularly emphasized in Hong et al. (2020) for drug toxicity predictions.

The methodology here employs the selective classification technique explored by Chen et al. (2024) to manage uncertainty in the *HINTPlus* model.

This approach operates based on the assumption that predictions should all be treated equally, since certain inputs may intrinsically carry ambiguity or even be outliers, leading the model to return outputs with low-confidence scores (confidence scores are indicators of the model's certainty regarding a given prediction) (Chen et al. 2024, 1).

Indeed, selective classification enables the model to stop from making predictions when confidence level is not as high.

To accomplish this, confidence scores are compared against a pre-selected threshold; the threshold is established based on the application, with higher thresholds guaranteeing the acceptance of only the predictions associated with high confidence scores. However, it needs to be taken into account that the higher the threshold, the lower the coverage, as a great number of cases for which the model provides a decision is to be excluded (Chen et al. 2024, 2).

By withholding predictions for uncertain cases, the method ensures that the overall accuracy of the model is improved, therefore also the reliability of the outputs is enhanced.

For instance, the application of selective classification is valuable in contexts associated with a substantial cost of incorrect predictions. However, increasing the confidence threshold may result in a higher abstention rate, therefore reducing the overall coverage. This methodology not only enhances trust in automated systems but also mitigates the risks associated with incorrect predictions, making it an essential tool for applications where the cost of errors is high. However, its effectiveness depends on appropriate threshold selection, the availability of fallback mechanisms, and a balanced understanding of the trade-off between coverage and accuracy (Chen et al. 2024, 2).

12.3 Methodology

The methodology begins with the *HINTPlus*, an end-to-end framework designed for clinical trial outcome prediction. *HINTPlus* integrates multi-modal data sources, including drug molecules, disease information, and trial protocols. The data is processed through an input embedding module and is pretrained with external pharmacokinetic knowledge, including absorption, distribution, metabolism, excretion, and toxicity (*ADMET*).

Selective classification is integrated in the model to enhance its predictive ability, since it allows the framework to withhold predictions when the confidence score is below a pre-established threshold. The threshold operates as a decision boundary and is derived through several key steps which include data preprocessing, model training, uncertainty estimation, and selective decision-making.

The mechanism is applied to Phase I, Phase II and Phase III studies to evaluate its impact across the overall trial process.

The threshold mechanism is implemented by first deriving prediction confidence scores using softmax probabilities. These confidence scores are then calibrated to align with realistic probabilities, achieved through techniques such as temperature scaling and binomial statistics. The calibrated scores allow for the application of a threshold to determine whether the model should return a prediction or abstain. Then, fine-tuning is iteratively applied during the training and validation phases. The training data is split into calibration and validation subsets (with the calibration set consisting of n=200 samples), with the aim of facilitating the computation of selective risk. Selective risk allows to quantify the error rate of the model for predictions meeting a specific confidence threshold and is defined as 1−α, where α represents the risk tolerance (in this specific case, α is set to 0.1, to assure a maximum error rate of 10%, given that the context in which the model operates is characterized by high-stakes decisions) (Mindee 2024). After computing selective risk, the optimal threshold lambda (λ^) is identified by evaluating a range of candidate thresholds lambda (λ) between 0.3 and 0.9; this range (0.3 - 0.9) is selected by taking into consideration empirical observations, and by assuming that it is aligned with the confidence scores distributions expected in this context. In the field of clinical trials, this spectrum ensures both inclusivity and reliability. (Johnson et al. 2020)

For each threshold, the selective accuracy (equal to 1−selective risk) and the fraction of points retained are computed. In correspondence of this optimal threshold, the optimal balance between accuracy and coverage is accomplished. Higher thresholds lambda (λ) account for more reliable predictions by ensuring the return of only reliable predictions, though at the cost of reduced coverage. Conversely, lower thresholds allow for greater coverage but increase the risk of erroneous predictions.

These computations visualize the trade-off between accuracy and coverage, enabling the identification of the highest lambda (λ), which is the optimal lambda (λ^), for which the selective risk is steadily below 1−α, thereby guaranteeing that the model operates within the predefined risk tolerance while seeking accuracy optimization.

Ultimately, the model's performance is assessed through traditional metrics such as PR-AUC, F1 score, and ROC-AUC, as well as selective accuracy, which measures the accuracy of predictions retained after applying the threshold. This ensures that the model's reliability improves by selectively abstaining from low-confidence predictions.

In conclusion, the integration of uncertainty quantification through selective classification significantly enhances the predictive reliability of the *HINTPlus* framework. By abstaining from low-confidence predictions, the model ensures higher accuracy and reliability, particularly in critical contexts like clinical trials, where erroneous predictions can have profound consequences. This methodology provides a systematic approach to reach the trade-off between accuracy and coverage, ensuring that predictions meet rigorous reliability standards while remaining practically applicable.

12.4 Results

The results achieved in each phase are here discussed as follows.

The graphs provided showcases the dynamics between accuracy and coverage. Accuracy is represented by the blue curve, while the orange curve represents the coverage (the fraction of points kept). The optimal threshold lambda (λ^) is illustrated by the vertical line, while the horizontal line represents the risk tolerance (1−α). Given the high stakes context of the clinical research fields, the error rate needs to be minimized, thereby the selective risk is set to the fixed value of 0.1, (corresponding to a confidence level of 90%) to ensure high consistency and robustness. Since α (selective risk) is here fixed at 0.1, the algorithm naturally selects as the optimal lambda (λ^), the threshold that guarantees to reach the highest accuracy within the allowed risk tolerance.

This intersection between the curves serves as an indicator of the point in which accuracy begins to be favored against inclusivity. Beyond this threshold, precision is prioritized, therefore a higher number of predictions are withheld, assuring higher reliability. While the intersection itself is not necessarily the optimal threshold lambda (λ^), it provides valuable insight into the model's performance.

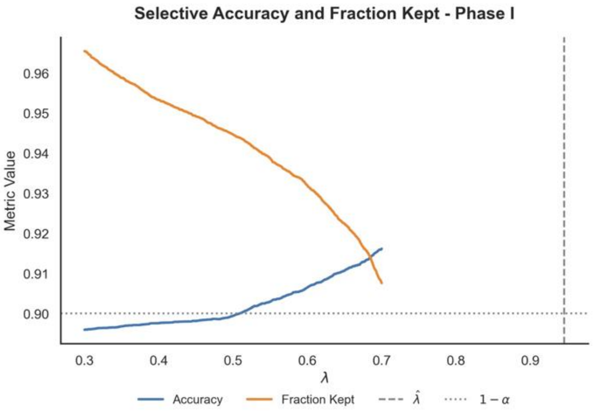
Generally, a pattern can be observed throughout all the phases, as lower thresholds involve extended coverage but reduced accuracy, whereas higher thresholds enhance accuracy at the expense of reduced coverage.

12.4.1 Phase I

In terms of precision and accuracy metrics, the *HINTPlus* model showed a significant performance. Indeed, the PR-AUC, Precision-Recall Area Under Curve, shows a major improvement, with a mean of 0.9705, signaling that the model successfully identifies the factors influencing trials outcomes (such as specific criteria). Likewise, the F1 score returned a higher value, equal to 0.977, indicating that both true positives and true negatives are correctly predicted. Another relevant KPI demonstrating the reliability of the predictive framework, is the selective accuracy, which is equal to 0.9546.

Contrarily, the ROC-AUC observed is equal to 0.6143, notably lower than the value obtained without selective classification (0.8780), potentially implying a better model behavior in general class discrimination rather than on the entire dataset, mostly due to its ability to process more points, including those that may be ambiguous.

Ultimately, it is to be mentioned the limitation of the selective classification approach, which is the trade-off between accuracy and coverage; indeed, the fraction of points kept exhibits a mean of 43.62%.



Graphic 2 - Selective Accuracy and Fraction Kept – Phase I

This can be observed in the graph, which shows the balanced between accuracy and the fraction of data points retained as a function of the confidence threshold lambda (λ). It can be noticed that accuracy improves as lambda (λ) increases, therefore the model correctly withholds predictions with low confidence scores. On the contrary, as lambda (λ) increases, the fraction of points kept decreases, meaning that the model priorities precision. In this case, the intersection falls on the left of the optimal lambda (λ^), meaning that a meaningful portion of coverage is still preserved, despite prioritizing precision (the fraction of points kept in correspondence of the selective accuracy indicates that the 43.62% of the points are retained).

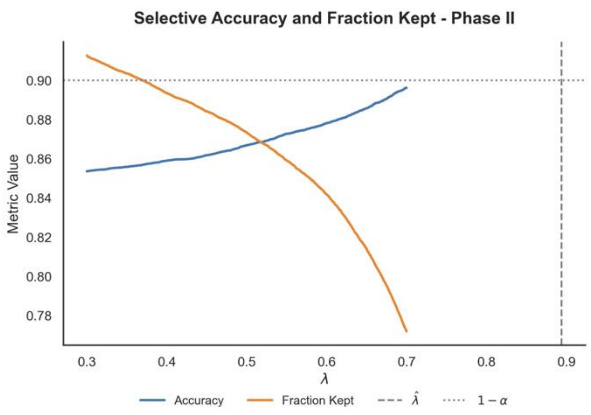
Without selective classification, the model returns a less precise performance; the PR-AUC drops to 0.8778, and F1 score is also lower, at 0.9317, reflecting a greater balance between precision and recall (See Appendix, Graphic 7).

12.4.2 Phase II

Selective classification has then been applied to Phase II, and the following results have been obtained.

Still, the model showcased a strong performance in precision-related metrics, accomplishing major improvements if compared to the metrics returned without selective classification; for instance, the PR-AUC shows a mean 0.9445, while the F1 score achieved a mean 0.9632.

Likewise, selective accuracy is equal to 0.9290, enhancing the reliability of predictions. However, in this phase, it is to be underlined the low fraction of points kept for which it is observed a mean of 0.3025, meaning that an extremely conservative approach has been employed when selecting predictions (leading to high precision and accuracy score).



Graphic 3 - Selective Accuracy and Fraction Kept – Phase II

Without selective classification, broader coverage is achieved at the cost of reduced precision. The PR-AUC and F1 score both show lower values, equal to 0.7835 and 0.8651 respectively (See Appendix, Graphic 8). Similarly to Phase I, also in Phase II the ROC-AUC observed is remarkably low at 0.5624, yet highlighting the challenges of operating with the entire dataset.

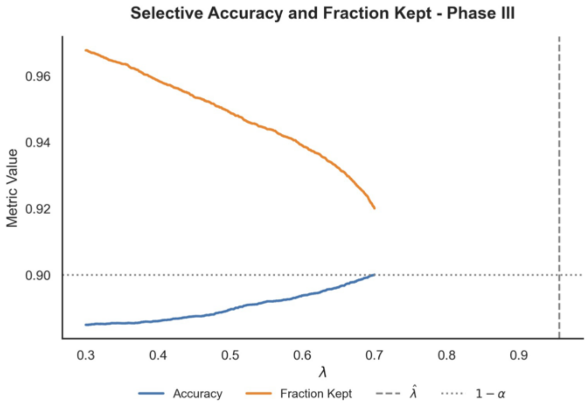
The graph provided shows the trade-off between accuracy and coverage, and it can be noticed that increasing lambda (λ) leads accuracy to improve, while the fraction of data points kept sharply decreases, thereby sacrificing coverage. In this case, the optimal lambda (λ^) is located to the right of the intersection, furtherly confirming that the chosen threshold prioritizes high confidence in predictions over inclusivity.

12.4.3 Phase III

Ultimately, selective classification has been applied to Phase III, and the following outcomes have been observed.

Likewise to the other phases, the model demonstrates substantial improvements in metrics related to precision; for instance, it has been accomplished a PR-AUC of 0.9452, and an F1 score of 0.9667, significantly higher than the ones obtained without selective classification (equal to 0.823 and 0.899 respectively). These results signal the balance in precision and recall, which may to lead to the minimization of false positives, extremely crucial in late phases trials. Indeed, selective accuracy scored a value equal 0.9357, furtherly underlying predictions reliability. However, similarly to what has been observed in the other phases, the ROC-AUC returned a lower value, equal to 0.5647, than the ROC-AUC achieved by the model before applying selective classification (See Appendix, Graphic 9).

With regards to coverage, the fraction of points kept averages 0.4118, yet reflecting a conservative approach, which is actually aligned with the Phase III trials scope of complying with regulatory priorities and avoiding incorrect predictions.

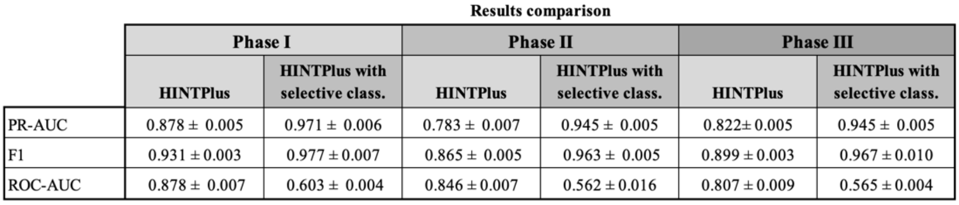


Graphic 4 - Selective Accuracy and Fraction Kept – Phase III

As shown by the graph, accuracy increases as the threshold lambda (λ) increases; this demonstrates the model increasingly conservative approach, which prioritizes predictive accuracy at the expense of coverage, since the fraction of data points kept significantly decreases. The absence of an intersection of the curves indicates that the fraction of data points retained decreases faster than accuracy improves. This reflects the model's steady focus on achieving high confidence in predictions at higher thresholds, resulting in a prioritization of reliability over inclusivity.

In contrast, when selective classification is not applied, the model prioritizes broader coverage but at the expense of reduced precision and reliability.

By implementing this approach, it is ensured the compliance of the model outcomes with the rigorous standards required by regulators (such as EMA and FDA) for late stages trials, even if coverage is reduced, in order to minimize false positives and avoid severe consequences on the study streamline.



Results comparison

	Phase I		Phase II		Phase III	
	HINTPlus	HINTPlus with selective class.	HINTPlus	HINTPlus with selective class.	HINTPlus	HINTPlus with selective class.
PR-AUC	0.878 ± 0.005	0.971 ± 0.006	0.783 ± 0.007	0.945 ± 0.005	0.822± 0.005	0.945 ± 0.005
F1	0.931 ± 0.003	0.977 ± 0.007	0.865 ± 0.005	0.963 ± 0.005	0.899 ± 0.003	0.967 ± 0.010
ROC-AUC	0.878 ± 0.007	0.603 ± 0.004	0.846 ± 0.007	0.562 ± 0.016	0.807 ± 0.009	0.565 ± 0.004

Table 12 - Results comparison between HINTPlus and HINTPlus with selective classification

12.5 Discussion – Cost/Benefit analysis

The integration of an uncertainty quantification technique, such as selective classification, into clinical trials streamline, offers an innovative approach which may lead to enhanced efficiency and effectiveness in the study processes. Indeed, selective classification implements a mechanism which allows the model to return exclusively high-confidence prediction, thereby guaranteeing increased reliability and potentially lowering the costs associated with incorrect outcomes. Thus, selective classification lowers the risks associated to both false negatives and false positives outcomes, which is of significantly relevant especially in later phase trials. According to Geifman & El-Yaniv, 2017, the application of machine learning techniques featuring selective classification into the field of clinical research, has demonstrated the ability to decrease error rates by up to 20% while assuring selective accuracy values of over 95%;

consequently, the avoidance of a single false positive in a Phase III trial could produce savings for $100 million, resulting in a more efficient resource allocation.

Furthermore, trials adopting uncertainty quantification techniques may experience a shortened timeline, as the duration of Phase II and Phase III is estimated to be reduced by 10%-15% (Thorlund et al. 2018). Furthermore, this approach may foster trust among stakeholders, as decisions are backed by rigorous, interpretable, and confidence-weighted predictions.

In spite of the several benefits, the implementation of selective classification into a trial workflow implies barriers and limitations. Firstly, monetary costs associated to integration of a machine learning model may range from $500,000 to $2 million per each trial, and include infrastructure, algorithm implementation, system maintenance and trained personnel with the required expertise (Muehlematter et al. 2021).

Secondly, as previously argued, selective classification operates a trade-off between accuracy and coverage, therefore the consequences of data points exclusion need to be carefully evaluated; for instance, in Phase I trials, if a low fraction of data points kept is assumed, critical insights may be overlooked, leading to biased results.

12.6 Conclusion

Uncertainty quantification is a critical factor affecting predictive frameworks, and it is of major relevance in high-sensitive contexts, such as the one of clinical research. The existing methodologies, reviewed above, offer diverse approaches to assess uncertainty and provide viable solutions to increase model outcome’s reliability. This may lead not only to improved performances, but also to a more efficient resource allocation and shortened timelines, enhancing the patient journey and relieving the burden on the actors involved in the drug development process.

This study specifically explored the integration of the selective classification method on the *HINTPlus* model, arguing the results gathered and the potential costs and benefits of implementing such technique into the trial streamline.

Selective classification has been applied to Phase I, II and III and has proved to boost precision and accuracy across all phases, as it can be denoted from the precision-related metrics (PR-AUC and F1 score), despite significantly reducing coverage (as indicated by the fraction of points kept). Indeed, by setting a confidence threshold this methodology ensures the return of only reliable results, since the model withholds predictions with lower confidence scores. Therefore, the risk of both false positives and false negatives is minimized, suggesting the adoption of this approach in settings in which precision is non-negotiable, such as later phases clinical trials.

Conversely, selective classification requires the trade-off between accuracy and coverage, thereby achieving higher reliability implies the exclusion of potentially relevant data.

Thus, the implementation of this practice depends on the objective of the specific clinical trial phase; earlier-stage studies, such as Phase I trials, may indeed require broader coverage in order to guarantee data inclusivity for the development process. Contrarily, in later stage studies the error rate must be minimized to avoid impactful consequences such delay in the approval process, even at the expense of less coverage.

Ultimately, the integration of uncertainty quantification techniques into clinical trials implies the assessment of potential costs and benefits. Despite the initial investment may seem substantial, the long-term advantages (enhanced accuracy, optimized resource allocation, improved operational efficiency and reduced timelines) outweigh the costs, particularly in later-phase trials. Numerous analyses underlined the economic benefit of this approach, estimating potential savings in the hundreds of millions, offering significant competitive advantages in the field of clinical research.

13. A What-if Scenario Approach on the Enrollment

13.1 Introduction

Clinical trials are essential to advance medical knowledge and to develop new treatments. However, trial success depends on the enrollment of participants in order to gather enough data from patients and make sustained conclusions. Enrollment is a challenge to trials, since approximately 80% of the clinical studies do not meet their initial objectives leading to substantial delays and financial losses.

This paper delves into the complexities of the enrollment feature, examining its role in assessing the success and efficiency of clinical trials.

There are several barriers to the enrollment process, including patient hesitancy due to the uncertainty of randomized treatment protocols and also the inclusion and exclusion criteria. These challenges show the necessity for innovated strategies to enhance enrollment rates, which may include improvements in the patient engagement practice, the implementation of synthetic "patients" using generator of data or even the use of potential financial incentives.

This chapter aims to examine the incremental increases in enrollment across different phases of clinical trials, by using the *HINTPlus* model to make predictions to analyze if changing the enrollment affects the trial outcomes. Through a scenario simulation and predictive modeling, the paper seeks to offer a what-if analysis to understand what the impacts on trial outcomes are, if enrollment values are increased by 10%, 20% and 30%. By exploring the dynamic of trial enrollment and their repercussions on the trial success, this research will contribute to understand the role of this feature in the trials, suggesting a way of dealing with the results achieved, ultimately, a cost benefit analysis is provided to compare if the use of synthetic data is beneficial to this area.

13.2 Literature Review

Kim et al. (2023) classifies a clinical trial success as a clinical trial "that passes on to the next stage after the regulatory body has approved the use of the drug or the clinical trial has been completed" (Kim et al. 2023, 3). This paper recognizes key success factors in the patients registered in the trials. The enrollment goal is crucial to accomplish success, however approximately 80% of the clinical trials do not reach the sufficient number of participants to achieve their primary objectives. This inevitably leads to delays, which are registered to cause losses for the equivalent of $8 million in revenues per day.

Therefore, it is of major importance to include the enrollment feature in the analysis, as it is a significant root cause for the majority of terminated studies.

Logan et al. (2016) deeply explores the barriers related to the enrollment, specifically focusing on the area of oncology on the prospective of randomized control trials. This paper aims to look into the problem of the enrollment increment in this specific area, more precisely on the factors that are affecting the enrollment during the whole process (from the patient recruitment to the patient enrollment in the study). The paper identifies, as one of the most challenging obstacles, the patient preference, which can lead to a hesitation towards randomized treatments. Ultimately, in this paper it is found that there is a high enrollment rate when the patients are approach several times during the course of the treatment.

Campbell et al. (2007) reports that one third of the publicly funded trials did require a time-extension due to the studies not meeting their initial criteria for the enrollment goal. Looking to specific areas the same pattern occurs, for instance in cancer related trial 25% of trials fail due to the lack of enrollment (Fogel 2018).

Fogel (2018) reported this as a long-standing problem and argues about the measures that were used to improve enrollment rates. Some studies started to give an allowance to the participants, to cover the patients time and expenses that may occur, despite being several studies showing that there is no relation between monetary compensation and patient participation and retention. However, Edwards et al. (2023) reports that monetary compensation increases the participation and engagement in questionnaires that were developed to improve the retention of the participants in the trials. Despite this, it is also stated that high compensation is associated with higher risk in the patient’s minds, leading to a reluctance in enrolling.

13.3 Methodology - What-if analysis

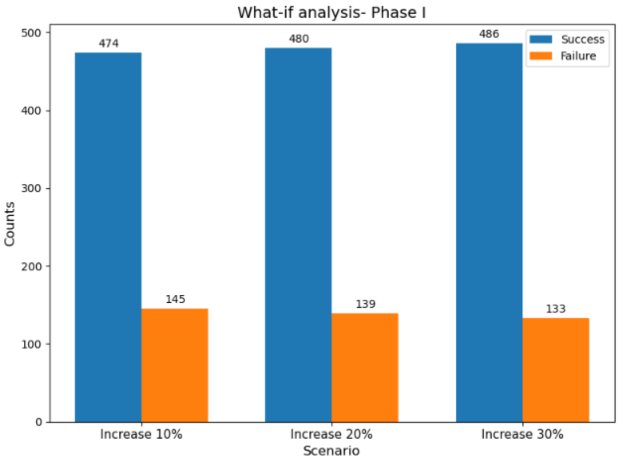
This analysis delves into how predictive scenarios to evaluate incremental increases in enrollment (10%, 20% and 30%) could potentially affect the trial outcomes, utilizing the *HINTPlus* model to predict the impacts of these adjustments on trials that initially failed.

13.3.1 Data Preparation

The sample was created from the test set which includes the most recent trials (between 2019 and 2024) and a sample was created for each of the phases. These particular trials were selected due to their failure to reach the end of their respective phases. Trials where the number of patients enrolled was recorded as zero were excluded from this sample to focus more on trials that had some participants involved. This decision was made in order to examine trials that had operational engagement but still have failed to achieve a trial outcome of success. The use of this sample was critical to ensure accurate representation of the most recent trials and the robustness of the data, thereby ensuring reliable predictions.

13.3.2 Predictive modeling and what-if analysis

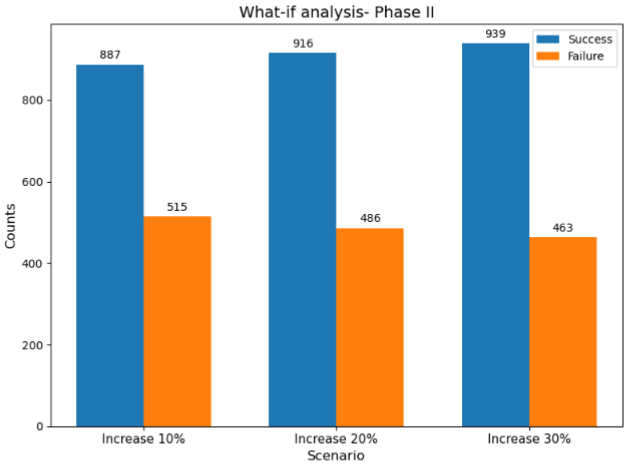
Using the sample data for each phase, the predictions were made using the *HINTPlus* model. This phase of the analysis was essential to predict how changes in enrollment could affect trial outcomes. The model robust predictive capability allows to simulate several scenarios where the enrollment was increased from 10% to 30%. These scenarios were quantified by the number of fail trails that change outcomes due to the enrollment increase and evaluate the success rate.



Graphic 5 - What-if analysis for Phase I

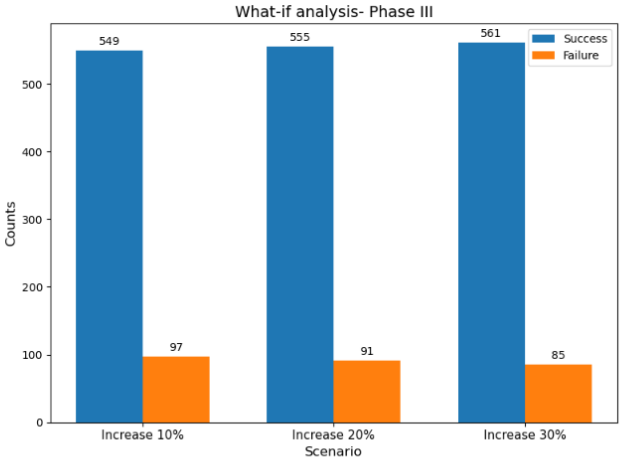
In Graphic 1 representing Phase I (sample with 619 trial), there is a remarkable decline in the number of failures when the percentage of enrollment increases. This trend suggests that even modest increases in enrollment can significantly reduce the likelihood of trial failure. In each scenario the success rate increases from 76.58% to 77.54% and then to 78.51%.

Similarly in Phase II (Graphic 2 have an sample with 1.402 trials) in each increase in enrollment corresponds to a decrease in failures, and it is also observed a success rate increase over the incremental values (63.27%, 65.34% and 66.98% respectively). The most significant decrease in failures was between 10% and 20% enrollment increases, having an approximately increase of 2.07% in the success rate of those cases.



Graphic 6 - What-if analysis for Phase II

In Graphic 3 (sample with 646 trials) that represent Phase III, there is also a consistent decrease in the number of failures with each incremental enrollment. In this case the success rate follows the same increasing pattern (84.98%, 85.91% and 86.84%, respectively) over the different increases on the enrollment. This pattern highlights the continued effectiveness of increasing the participant



Graphic 7 - What-if analysis for Phase III numbers, specifically in advanced phases, where archiving conclusive results are critical.

The increase in enrollment across each phase showed that most of the trials change their outcome, however some remain failures. This information provides insights into the threshold of enrollment increase, which is necessary to potentially convert failed trials into successful ones.

This analytical approach exhibits the potential of targeted enrolment increase to change the trial outcome but also highlights the need for a deeper understanding of this feature. By integrating this what-if analysis into predictive modeling, it is demonstrated that employing strategic enrollment practices can enhance the efficacy and efficiency of clinical trials. This analysis can affect the future structure of the trials in order to become more efficient and effective and even reduce the incidence of trial failures, by just allowing the trials to test different levels of enrollment for upcoming trials.

13.4 Synthetic data

As demonstrated in the cases previously argued and analyzed, if the trials registered as failure were to increase the enrollment, the probability of reaching the end phase would increase. However, the process of increasing the enrollment size is significantly challenging, as it relies on the patient’s availability and on the protocol information, especially on the inclusion and exclusion criteria defined for a specific study. A workaround to this problem can be found in the application of synthetic data in clinical trials at the patients’ level, resulting in the creation of artificial “patients”.

The concept of Generative Adversarial Network (GANs) was introduced by Goodfellow et al. (2014). This revolutionized the generation of synthetic data by leveraging adversarial training. Even though this methodology has been initially used to generate imaginary datasets, it has later been adapted to serve several scopes and industries, including the healthcare and life-science field. Indeed, GAN-based techniques generate synthetic data at a patient’s level, addressing the challenges of data privacy and small sample sizes.

Xu and Veeramachaneni (2018) proposed a Tabular GAN (TGAN), designed to generate synthetic tabular data by using an LDTM-based generator (Large Diameter Trunk Main) and a multi-layer perceptron discriminator. TGAN focuses on mixed data types, including continuous and categorical variables and introduces a more specific normalization technique to handle multimodal distributions. By leveraging an attention mechanism and reversible data transformation, TGAN produces synthetic tables preserving correlations between the variables. This model is evaluated on multiple datasets and demonstrates excellent performance in capturing data distributions if compared to the traditional methods, such as Bayesian network and Gaussian copulas.

Lei Xu et al. (2019) developed the Conditional Tabular GAN (CTGAN), specifically designed to generate synthetic tabular data, consequently addressing the challenge of mixed data types, imbalanced categorical variables and multimodal continuous distributions. This model introduces a mode specific normalization to efficiently handle non-Gaussian and multimodal continuous variables and ensures an accurate representation and generation. The CTGAN utilizes a conditional generator alongside a training by sampling approach to handle imbalance categorical data in an effective way. This ensures that the model gives adequate attention to underrepresented classes. The innovated design of the CTGAN surpasses Bayesian networks and others GAN-based approaches, providing a versatile and efficient method for creating high quality synthetic tabular data.

Krenmayr et al. (2022) created a new model architecture to generate synthetic data, denominated as GANerAid. This approach is based on the Generative Adversarial Network (GAN) that generates realistic patient data specifically tailored for clinical trials. This new model enables to overcome the obstacle of preserving variable correlations and accommodation for small sizes. GANerAid uses a Long Short-Term Memory (LSTM) network to return data that mimics very well the statistics properties of the real-world data from patients, even for the imbalanced variables. In the mimic part, two models referred as Generator and the Discriminator are employed. The Generator creates new data that is as similar as possible to the original data and the Discriminator identifies if the data was created artificially by the generator. The process initiates with the Generator, proceeds with the Discriminator and reiterates until the discriminator is no longer able to identify the artificial data. One of the issues encountered, when using generative data, is regard to the data privacy of the patient, as this data is considered as sensible and personal information and needs to strictly adhere to the General Data Protection Regulatory (GDPR). The study shows that the utilization of the GANerAid to augment datasets can substantially improve the statistical validity of the trials, while simultaneously deal with the problem of patients' privacy.

To conclude, the use of synthetic data is a powerful tool to address the challenges related to the clinical trials designs, particularly the limitations on the patient enrollment and data availability. Models such as TGAN, CTGAN and GANerAid demonstrate how advancements in generative AI may produce realistic, high-quality data while still preserving the statistical properties and correlations of the real-world datasets. These methods enhance the statistical validity of clinical trials, ensure at the same time compliance with data privacy regulations. By leveraging synthetic data, clinical trials can be optimized to improve their efficiency, reduce the recruitment barriers and support more robust and ethical research methodologies.

13.5 Discussion – Cost/Benefit analysis

As argued in the previous paragraphs, the integration of synthetic data in clinical trials workflow is undeniably a transformative opportunity for the industry of medical research, due to its multiple benefits, which range from cost reduction to shortened timelines. To achieve a more complete understanding of synthetic data impact on this field, an analysis of the potential implications, from an economical and operational perspective is provided here. Synthetic data have significant power to lower the costs related to patient recruitment, which is one of the major costs in the drug development process. For instance, the replacement of placebo groups with synthetic control arms allows pharmaceutical companies (or other study conductors) to minimize the number of participants required (this has been done by Roche, which employed a synthetic control arm during the successful approval process for Alecensa) (Accenture & Phesi 2021). Consequently, resource allocation can be optimized, leading savings to be used for multiple purposes, including trial expansion or R&D processes for new medications. Nevertheless, another benefit of major importance lies in the shortening of the overall study timeline, resulting in enhanced ROI (Return On Investment), this is achieved by the possibility of pre-defining patient cohorts using synthetic datasets, thereby reducing the time required by traditional data collection methods.

With regards to operational efficiency, synthetic data may allow an enhanced usage of existing data, since reliable synthetic control arms can be generated by leveraging data lakes and repositories. This might result in reduced logistical complexities related to the phase of patients enrolment. The synthetic data can be used to enhance predictive modeling, which leads to refined trial protocols, achieved by testing several inclusion and exclusion criteria. This technique may enable the selection of an optimal patient population, ultimately increasing the likelihood of trial success. Additionally, regulatory authorities such as the FDA and EMA have increasingly recognized the validity of synthetic data, further creating an efficient process of approval for treatments that rely on this methodology. The use of synthetic data in the development of Amgen's Blincyto, which utilized historical data from 2,000 patients, underscores its growing acceptance in the regulatory landscape (Accenture & Phesi 2021).

The use of synthetic data addresses several long-standing ethical issues in clinical research. Traditional trial-bases often require patients to receive placebo treatments, raising ethical dilemmas about withholding potentially life-saving therapies. The use of synthetic control arms eliminates this concern, allowing all participants to receive active treatments while historical data serves as the comparator. This innovation reduces the burden on patients, particularly in diseases with high unmet needs or limited patient populations, such as rare diseases and oncology. Furthermore, the reduced reliance on experimental treatments minimizes patient risk, making synthetic trials a more patient-centric approach to clinical research.

To the challenges presented, the implementation of synthetic data in clinical trials may face several barriers. One lies in the lack of a unified global dataset to favor health data sharing, for instance, pharmaceutical companies and the other stakeholders involved frequently have to work with siloed datasets, which do not allow to exploit the full potential of synthetic data. Another significant obstacle is to be found in the long adoption timeline, since the pharmaceutical industry has shown a slow integration of AI practices and machine learning methods into its workflows.

Nevertheless, the integration and development of synthetic data into clinical trials is associated with substantial costs. For instance, an article by Eularis highlights that AstraZeneca's use of over 300 million synthetic patient records led to a 30% increase in clinical trial success rates for a new cancer drug target, reducing average development costs by an estimated $100 million per drug. The cost estimated considers the investments needed to implement the data infrastructure, analytics tool and dedicated personnel with the related expertise (Accenture and Phesi 2021). Ultimately, skepticism surrounding synthetic data is still a relevant limitation. In fact, stakeholders (which comprehend regulators, clinicians, and patients), need significant evidence and transparency in order to place trust in these innovative practices. Related to this is the ethical concern surrounding data privacy and regulatory compliance. In spite of the major costs and limitations that might be encountered, the integration of synthetic data may lead to a substantial cost reduction (between 30% and 50%) in the overall drug development process and is able to facilitate patient enrollment and can lead to shortened timelines (Eularis 2024;

Oualikene-Gonin et al. 2024).

13.6 Conclusion

The study explores extensively the crucial role of enrollment in determining trial outcomes in clinical trials. Through a what-if analysis and predictive modeling using *HINTPlus* model, it was demonstrated that this feature is associated with trial success. The findings indicate that the enrollment size is crucial to determinate trial outcomes for each phase, showing that the incremental increases done can change most of the trials outcomes. The what-if analysis highlights how changes in the enrollment criteria can have an impact on trial outcome, particularly for trials that initially failed. This insight provides a concrete solution to enhance the trial outcomes. Furthermore, the integration of synthetic data via Generative Adversarial Network (GANs) represents an important step to overcome enrollment challenges. By generating high quality, realistic synthetic data from the original patients using models like TGAN, CTGAN and GANerAid have the potential to revolutionize clinical trial designs. This offers solutions to patients enrollment barriers while ensuring compliance with data privacy regulations. The use of synthetic data improves trial results and brings financial benefits, according to the cost-benefit analysis. The synthetic data can decrease the operating and recruitment costs by lowering the requirements for patient enrollment. These adjustments make it possible to re-allocate funds to other areas of research and development. Additionally, a shorter trial period may accelerate the process and reduce logistical issues, resulting in a faster market launch. In addition to saving money, this helps patients by facilitating rapid access to the new therapies. The financial and operational benefits of using synthetic data outweighs the initial expense, making it worthwhile to invest in order to increase the effectiveness and cost-effectiveness of clinical studies.

In conclusion, this thesis contributes to the academic understanding of clinical trials outcomes and also offers practical solutions to improve trial efficiency and success rates. The models and techniques created provide a useful perspective that may impact on how future trials are planned and executed. These advances promise to improve trial outcomes and increase the efficacy of medical research by optimizing the trial process and taking ethical considerations into account.

Jasmin Hamrouni

14. Incorporating RAG Techniques for Predictive Modeling of Clinical Trial Outcomes

14.1 Introduction

Chopra et al. (2024) give reasonable arguments for why predictive models, and especially neural network models face thLewis et al. (2020) were early to introduce the Retrieval-Augmentation-Generation (RAG) architecture, which has been defined as a hybrid model architecture, combining pre-trained language models and retrieval-based external memory components for natural language generation.

The architecture's key components are the following: The retrieval mechanism reflects the process of fetching information or documents from an external knowledge base. The foundation is a dense vector representation of the input query. The augmentation represents the retrieved information from the input query. The generator processes the augmented input and is often built using a transformer-based architecture in order to create responses based on the retrieved knowledge (Gupta et al. 2024).

Historically, a surge in development of traditional Natural Language Processing models has been observed. Nonetheless, these models greatly rely on training data, causing limitation when they are tasked to provide accurate answers regarding data the models were not trained on. As a result, traditional models could generate plausible yet incorrect information, reflecting the phenomenon of hallucinating large language models (Gupta et al. 2024).

In comparison to generative models, RAG has the ability to leverage external knowledge dynamically without requiring retraining, making it suitable for domains in which information is constantly evolving such as the medical field. However, this also requires it to handle growing datasets. While the retrieval mechanism is a powerful step, it can also lead to flawed results, which can be traced back to the vector representation of data, causing the retrieval of irrelevant documents. Furthermore, the generative module has the potential to be unable to incorporate retrieved information into its responses, which may create discrepancies between retrieval and generation. (Binns 2018). The objective is to incorporate RAG techniques in order to contextualize predictions leveraged by the *HINTPlus* and *HINTPlus per Disease*. Two RAG-Systems were developed in order to achieve this. The performance of the RAG-Systems was evaluated in comparison to each other. Moreover, it was tested how both systems react to general and more detailed prompts. Lastly, a Chatbot was employed to enable users to ask follow-up questions. The Generation system of the RAG-Systems alone as well as the Chatbot-enhanced RAG-Systems were evaluated in order to see if the system hallucinates answers.

14.2 Literature Review

Lu et al. (2024) introduced ClinicalRAG, a multi-agent pipeline in which every agent was in charge of a defined process stage, aiming to increase diagnosis accuracy. Not only was ClinicalRAG able to extract medical data from the user's input, it also had the capability to dynamically incorporate related medical knowledge in the text generation process, leading to the collection of high-quality information. Lu et al. (2024) observed that ClinicalRAG was able to significantly outperform traditional methods such as prompt learning, leading to enhanced reliability in clinical decision-making. However, limitations such as interpretability were stated, as they require attentions when it comes to an implementation.

Similar to Lu et al. (2024), Ong et al.'s (2024) employed a more general look on decision support. Ong et al.'s (2024) study aimed at evaluating the effectiveness of RAG-based clinical decision support system regarding the identification of medication-related errors, by developing a simple and an advanced RAG-LLM. These two systems differed in numerous ways such as the chosen vector database, the retrieval scope or the embedding model. Their performances was evaluated on their own and as an assistive tool of a pharmacist. The study classified the simple RAG-LLM as the optimal choice, as it balanced accuracy, recall and precision. The usage as an assistive tool yielded superior results in comparison to the standalone variation, highlighting the potential of hybrid human-AI systems in clinical trial decision-making. However, the study concluded that data quality and scalability issues should be taken into account. While Ong et al. (2024) focused on patient safety in clinical care, Unlu et al. (2024) assessed the ability of a RAG-System to automate patient’s eligibility screening by using electronic health records. The goal was the enhancement of accuracy, efficiency and cost-effectiveness regarding participant screening in comparison to traditional methods. They used GPT-4 for the generation aspect, while also comparing its performance to GPT-3.5 and manual screening done by staff. The results showed that their GPT-4 enhanced RAG-System demonstrated the highest accuracy, leading to relatively reduced screening costs. Although the GPT-4 enhanced RAG-System provided benefits such as accuracy and cost-effectiveness, it also came with challenges such as generalization and regulatory compliance.

Wang et al. (2023) took the implementation of GPT-based RAG systems one step further by exploring its integration with clinical-data into AI chatbots and comparing it to a baseline Chatbot, aiming to enhance patient guidance. AI chatbots such as ChatGPT demonstrated a deficiency of specificity in medical advice, which was conveyed by the frequent generation of generic responses. In contrast, the RAG-enhanced alternative showed enhanced accuracy and specificity by integrating relevant medical knowledge into its generated answers, entailing more detailed diagnostics and treatment recommendations. Yet, potential accelerated treatments might also involve technical, ethical and scalability limitations.

14.3 Methodology

14.3.1 Database

With the goal of building a knowledge base, data curation steps were taken which were carried out for both RAG-Systems in the same manner. Contrary to conventional RAG-Systems, the knowledge bases are static. This was attributed to the fact that prediction as well as training data of both leveraged models were already up-to-date.

The derived results from both models, the *HINTPlus* and the *HINTPlus per Disease,* included NCT numbers and their respective predictions for all three phases separately. Since the results were saved in pkl.files, they were then converted to CSV files in order to make data handling easier. In addition, the developed benchmark dataset was utilized to retrieve the training data. Consequently, all columns which were not used for training the model were dropped, resulting in a filtered dataset containing 103,847 rows and 6 columns titled “NCT Number”, “Enrollment”, “Eligibility Criteria”, “Success/ Failure”, “SMILES” and “icd_code”. The CSV files containing NCT numbers and predictions were joined into one dataset. Since there are instances in which NCT numbers appeared in multiple phases, duplicated NCT numbers were dropped, ensuring the existence of unique NCT numbers and their respective predictions. Upon the described preparation steps, a left merge was performed. A significant amount of rows with no corresponding NCT numbers had missing value in the “Prediction” column, which were then eliminated. This resulted in the curation of a dataset with 12,817 rows for the *HINTPlus* RAG-System and 615 rows for the *HINTPlus per Disease* RAG-System while containing 7 columns respectively, serving as knowledge bases for both systems.

14.3.2 RAG Architectural Structure

The proposed RAG Architecture was built on insights from current literature such as Ong et al. (2024). As their simple RAG achieved good performance metrics, it will be the foundation for the following proposed RAG Architecture. The curation of the dataset was followed by a data preprocessing step, in which all the columns of each row were batched into one single text string, containing all the information of each row. This simplified the system’s performance of semantic similarity searches based on the users query, as every consolidated text serves as a representation of each trial. The clear data structure results in an efficient document retrieval (Burdick et al. 2021).

Once the preprocessing was completed, the sentence transformer “all-MiniM-L6-v2” was employed, a well text encoder, ideal for detecting textual similarity. The model converted the combined text into numerical embeddings. Consequently, semantic relationships and contextual meanings were captured, enabling efficient document comparisons (Wang et al. 2020).

The numerical embeddings were then stored as vectors in a FAISS index, serving as an efficient library for similarity search. Its ability to retrieve similar documents based on the embeddings by performing fast nearest neighbor searches was highly important, especially because the L2 distance metrics was also employed to handle the large data set. The same sentence transformer generated an embedding for a made query which prompted a search in the FAISS index, retrieving the top k, in this case 5, most relevant documents (Douze et al. 2024). These documents were then transferred to OpenAI’s language model GPT-4o, with the ability to process complex information while generating clear and human like answers (Shahriar et al. 2024). The model was provided with the user’s query as well as the retrieved relevant documents in order to ensure a grounded answer. This was also reinforced by instructing the model to specifically use details from the retrieved documents. In order to enhance the user experience of the RAG-systems, a chatbot was added for both systems, allowing users to ask follow-up questions whilst maintaining context. Its foundation is a loop, handling the user input. It also reflected the RAG framework’s ability to retrieve the most relevant documents, to then generate a context-based response with GPT-4-o and outputs the answer (Wang et al. 2023). The same architectural structure was utilized for both RAG-Systems, the *HINTPlus* RAG-System and the *HINTPlus per Disease*. The following pipeline describes the proposed architecture:

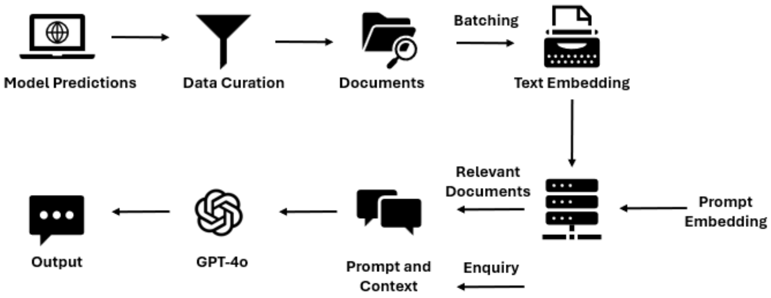


Figure 7 - RAG-System Architecture

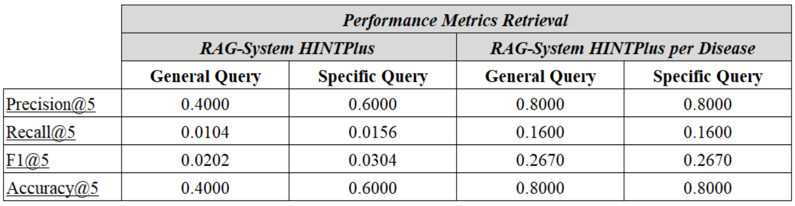
14.3.3 Experimental Setup and Hypotheses

The goal was to compare the ability of both developed RAG-Systems. First, we wanted to evaluate, if a disease-specific RAG-System depicts higher performance. Moreover, we wanted to test how the RAG-Systems reacted to Query-Engineering. The following hypotheses were tested to assess the retrieval component of both RAG-Systems: (1) The *HINTPlus per Disease* RAG-System will perform better than the *HINTPlus* RAG-System. (2) Queries detailing the user's role and goal of the query will yield a better performing RAG-System than a general query. In order to test these hypotheses, two queries were formulated. The first query entailed that the user wanted to know how high the enrollment number had to be regarding a Crohn's disease clinical trial in order to achieve a prediction of 0.8 or higher. The second query further included that the user is a medical researcher and that the answer would assist him in assessing success in future clinical trials. Apart from the added information about the user's role and the goal, the query stayed the same. Additionally, the Chatbot's potential to hallucinate answers was tested, by asking questions, which it could not be able to answer, based on the provided knowledge bases. In case of hallucination, we were refining the Chatbot by applying a feedback loop and thus giving users the ability to rate the Chatbots answers or refining the prompt and system message (Petridis et al. 2023).

14.4 Results

14.4.1 Retrieval

In order to test the hypotheses, we are following Ong et al.’s (2024) example by utilizing the same metrics. The first metrics is Precision@k, which quantifies the proportion of relevant documents within the top k retrieved documents. This metrics is especially useful to get insights into whether the top results presented are actually relevant. The second measure is Recall@k, measuring the proportion of relevant documents retrieved out of all relevant documents in the dataset. The third used metrics is the F1@k score, which is calculated as the weighted average between Precision@k and Recall@k. Lastly, the Accuracy@k metric measures the proportion of correctly classified documents within the top k results. The following results were found by running the experiment of both RAG-Systems:



	Performance Metrics Retrieval			
	RAG-System HINTPlus		*RAG-System HINTPlus per Disease*	
	General Query	**Specific Query**	**General Query**	**Specific Query**
Precision@5	0.4000	0.6000	0.8000	0.8000
Recall@5	0.0104	0.0156	0.1600	0.1600
F1@5	0.0202	0.0304	0.2670	0.2670
Accuracy@5	0.4000	0.6000	0.8000	0.8000

Table 13 - Metrics measuring Retrieval Performance

While both RAG-Systems demonstrate relatively good performance in terms of Precision@5 and Accuracy@5, the same cannot be said about the Recall@5. This can be explained by the large number relevant data in the dataset, lying at 195. However, due to the set k=5, not all of them can be considered in the retrieval. A higher k could potentially lead to an improvement of this metric. Consequently, the low value in Recall@5 has also influenced the F1@5 score negatively. Overall, the resulting metrics suggested that the *HINTPlus per Disease* RAG-System demonstrated better performance overall, validating the first hypothesis. Additionally, queries detailing the user’s role and goal led to higher performance achieved by the *HINTPlus*

RAG-System. However, the same cannot be stated about the *HINTPlus per Disease* RAG-System, which demonstrated the same results for both the general and the detailed query. Consequently, the hypothesis can only be partially supported by the results of the *HINTPlus* RAG-System. The difference in performance might explainable by the complexity of the datasets used for both RAG-Systems. While the *HINTPlus* RAG-System contained a rage of clinical trials regarding different diseases, the *HINTPlus per Disease* RAG-System dataset is merely focused on diseases about the human's digestive system.

14.4.2 Generation

We applied human evaluation in order to see if both RAG-Systems are generating correct outputs, while also answering the formulated query. Although, the RAG-Systems were at times prone to retrieving irrelevant documents, they consistently generated the correct outputs, responding to the query. In cases, in which irrelevant documents were retrieved, it was still able to filter them out in the generation step, leading to a correct output, which consistently answered the questions. This is especially important when considering challenges such as hallucination. Inputting queries, which were impossible to answer with only the documents inside the RAG-Systems, led to text generation which highlighted the fact that such questions cannot be answered based on the documents the RAG-Systems has been provided with. Overall, the generation demonstrated good performance consistently, while also highlighting its limitations thus avoiding hallucinations. However, the same cannot be said about the RAG-enabled Chatbot, as it has shown signs of hallucinated answers. Although, the same queries were prompted to the RAG-enhanced Chatbot, it was hallucinating answers which were not part of the provided datasets. Nonetheless, they were correct after verifying the information with ClinicalTrial.gov data. After refining the model's explicit system message and the prompt by highlighting the fact that the Chatbot is only supposed to use the retrieved documents while also applying a feedback loop which rewards the Chatbots highly ranked answers, no hallucination was detected after asking the same question again. Similar, to the RAG-System did the Chatbot inform the user about its incapability of answering the question due to insufficient data providing this information.

14.5 Discussion

Our results demonstrated that the retrieval mechanism of the *HINTPlus per Disease* RAG-System performed better than the *HINTPlus* RAG-System. Moreover, detailed queries including the user's role and goal only yielded a better performing retrieving for the *HINTPlus* RAG-System. While the generating mechanism of both RAG-Systems showed great performance and no hallucination, this does not account for the RAG-enhanced-Chatbots. Both Chatbots showed signs of hallucination in their generated answers. However, refinements such led to no hallucinated answers after asking the same question again. It is important that these results have to be interpreted in the context of the following limitations.

14.5.1 Architectural Limitations

A higher setting of k could potentially lead to a higher Recall@k score, which would in return influence the F1@k score and thus the overall evaluation of both RAG-Systems. Additionally, the retrieval component could also benefit from a domain-specific embedding model, such as ClinicalBERT, as they demonstrate a higher understanding of the used vocabulary and the relationships within the data due to them being trained on it (Huang et al. 2019). Consequently, this could potentially impact the generational aspect positively as more relevant information is retrieved. The usage of FAISS for efficient similarity search could have also potentially influenced the findings. An alternative such as the HNSW (Hierarchical Navigable Small Word) algorithm, could be beneficial due to its ability to handle large datasets fast and efficiently (Malkov & Yashunin 2018). Another point to note is the lack of chunking in the proposed architecture. Chunking would provide the system with the capability of breaking down text elements or documents into smaller pieces before preprocessing them in the pipeline.

This would provide the RAG-System to put more emphasis on specific information in large text and document, which could potentially increase the chance that the system ranks relevant information higher (Yepes et al. 2024) However, literature showed that a straightforward approach yielded better performance metrics which is why chunking was not applied in this case (Ong et al. 2024).

The proposed RAG-Architecture lacks an important component of conventional RAG-Systems, which is the ability to dynamically update its database. Since this work is an extension of set prediction derived from machine learning models, dynamic changes were considered.

14.5.2 Stakeholder Implementation

Although these constraints may impact the RAG-Systems performance and interpretability, there are also important implications. Machine learning predictive models are often seen as a "black box", as they lack interpretability (Chopra et al. 2024). The proposed RAG-Systems could potentially enhance the confidence in the predictions made by the *HINTPlus* and *HINTPlus per Disease* models, due to its ability to answer a variety of queries, while also providing a contextualized answer, giving insights into the reasons of the prediction. The lack of hallucinated answers also supports this argument, especially since the RAG-Systems communicated its limitations when answering queries, going beyond the scope of the retrieved documents (Venkit et al. 2024). Enhanced usability by adding a Chatbot should also be considered, since this could potentially contribute to a higher understanding of the predictions, as users can ask follow up questions while still being able to see the whole conversation. The added feedback loop in order to reward the Chatbot contributes to the efforts of continuously improving the system. All of these points could help Healthcare Providers and Researchers to make informed decisions in their clinical trials, in terms of setting hypotheses, planning enrollment numbers comparing eligibility criteria which could also decrease errors per study. As a result, patients could benefit from receiving appropriate and safe care more efficiently.

Companies would in return also benefit from the employment of the RAG-Systems as they could potentially influence drug discovery and thus the company's ROI. The lack of hallucinated information ensures that the system is generating trustworthy information, which could potentially make Regulators gain trust in the system.

Building on these limitations and implications, future research can explore several areas more in depth. Considering that this architecture does not link a database, which changes dynamically, future research could design a RAG-architecture which hold its data up-to-date. Additionally, different embedding models, language models or vector databases could be employed with the goal to improve the system's performance.

14.6 Conclusion

The aim of this work was to build a RAG-System in order to leverage predictions made by the *HINTPlus* and *HINTPlus per Disease* model. Additionally, we created two Chatbots which were connected to their respective RAG-Systems. The two RAG-System's retrieval performance was compared overall and in dependance of the created query. Findings showed that the *HINTPlus per Disease* RAG-System performed better in terms of the retrieval component. Detailed queries only led to improved retrieval performance in the *HINTPlus per Disease* RAG-System. Generated answers were free of hallucinations. However, this changed after employing a RAG-enhanced Chatbot. Refinements by including a feedback mechanism and reformulating explicit service message and prompt caused the disappearance of hallucinated answers. Although the proposed RAG architecture could be improved in terms of its retrieving capability, it also provides beneficial implications for involved stakeholders in healthcare, such as the increased interpretability of predictions due to contextualized answers. Future research could refine this proposed architecture by implementing a key advantage of RAG-System, which is the ability to keep datasets up to date.

Bibliography

“BridgeBio Catapults After Snagging Approval for Its Rival to Pfizer’s Heart Drug - Bing.” n.d. Bing. https://www.bing.com/search?pglt=299&q=BridgeBio+Catapults+After+Snagging+Approval+For+Its+Rival+To+Pfizer%27s+Heart+Drug&cvid=7099336a5ec8472389f9c78d44c09d64&gs_lcrp=EgRlZGdlKgYIABBFGDkyBggAEEUYOdIBBzQ0MmowajGoAgCwAgA&FORM=ANNTA1&ucpdpc=UCPD&PC=ASTS.

“Cost of Clinical Trials: A Breakdown (Infographic) – CRIO.” 2018. July 1, 2018. Accessed November 30th 2024 https://clinicalresearch.io/blog/running-a-study/cost-of-clinical-trials-breakdown/.

“Flexible Imputation of Missing Data, Second Edition | Stef Van Buuren.” 2018. Taylor & Francis. July 17, 2018. https://www.taylorfrancis.com/books/edit/10.1201/9780429492259/flexible-imputation-missing-data-second-edition-stef-van-buuren.

“Understanding The Benefits Of Clinical Trials For Cancer.” 2021. Johns Hopkins Medicine. October 16, 2021. https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/understanding-the-benefits-of-clinical-trials-for-cancer.

Accenture and Phesi. 2021. "Faster and Cheaper Clinical Trials: The Benefit of Synthetic Data." Accenture. Accessed November 20, 2024. https://www.accenture.com/content/dam/accenture/final/a-com-migration/r3-3/pdf/pdf-148/accenture-insilico-faster-and-cheaper.pdf.

Afkanpour, Marziyeh, Elham Hosseinzadeh, and Hamed Tabesh. 2024. “Identify the Most Appropriate Imputation Method for Handling Missing Values in Clinical Structured Datasets: A Systematic Review.” *BMC Medical Research Methodology 24 (1)*. https://doi.org/10.1186/s12874-024-02310-6.

Aliper, Alex, Roman Kudrin, Daniil Polykovskiy, Petrina Kamya, Elena Tutubalina, Shan Chen, Feng Ren, and Alex Zhavoronkov. 2023. “Prediction Of Clinical Trials Outcomes Based On Target Choice And Clinical Trial Design With Multi-Modal Artificial Intelligence.” *Clinical Pharmacology & Therapeutics* 114 (5): 972–80. https://doi.org/10.1002/cpt.3008.

Anagnostopoulos, Chris, David Champagne, Alex Devereson, Thomas Devenyns, and Heikki Tarkkila. 2023. “How Artificial Intelligence Can Power Clinical Development.” McKinsey & Company. November 22, 2023. Accessed November 30th 2024 https://www.mckinsey.com/industries/life-sciences/our-insights/how-artificial-intelligence-can-power-clinical-development

Arrowsmith, John. 2011. "Phase III and submission failures: 2007–2010” *Nature Reviews Drug Discovery*, 10(2): 87–87. https://www.nature.com/articles/nrd3375

Augustine, Erika F., Heather R. Adams, and Jonathan W. Mink. 2013. “Clinical Trials in Rare Disease.” *Journal of Child Neurology* 28 (9): 1142–50. https://doi.org/10.1177/0883073813495959.

Austin, Peter C., Ian R. White, Douglas S. Lee, and Stef Van Buuren. 2020. “Missing Data in Clinical Research: A Tutorial on Multiple Imputation.” *Canadian Journal of Cardiology* 37 (9): 1322–31. https://doi.org/10.1016/j.cjca.2020.11.010.

Bieganek, Cameron, Constantin Aliferis, and Sisi Ma. 2022. “Prediction Of Clinical Trial Enrollment Rates.” *PLoS ONE* 17 (2): e0263193. https://doi.org/10.1371/journal.pone.0263193.

Binns, Rdp. 2018. “Fairness in Machine Learning: Lessons From Political Philosophy.” *Journal of Machine Learning Research*, January. https://ora.ox.ac.uk/objects/uuid:2ff2785b-b0d4-447a-8326-a1fcc4c80840.

Burdick, Laura, Jonathan K. Kummerfeld, and Rada Mihalcea. 2021. “To Batch or Not to Batch? Comparing Batching and Curriculum Learning Strategies Across Tasks and Datasets.” Mathematics 9 (18): 2234. https://doi.org/10.3390/math9182234.

Campbell, M, C Snowdon, D Francis, D Elbourne, A McDonald, R Knight, V Entwistle, et al. 2007. “Recruitment to Randomised Trials: Strategies for Trial Enrolment and Participation Study. The STEPS Study.” *Health Technology Assessment* 11 (48). https://doi.org/10.3310/hta11480.

Cassava Sciences, Inc. 2024. “Cassava Sciences Topline Phase 3 Data Did Not Meet Co-Primary Endpoints.” *GlobeNewswire News Room*, November 25, 2024. https://www.globenewswire.com/news-release/2024/11/25/2986578/8339/en/Cassava-Sciences-Topline-Phase-3-Data-Did-Not-Meet-Co-Primary-Endpoints.html.

Chen, Tianqi & Guestrin, Carlos (2016). “XGBoost : A Scalable Tree Boosting System”. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Tianyi, Nan Hao, Yingzhou Lu, and Van Rechem Capucine. 2024. “Uncertainty Quantification On Clinical Trial Outcome Prediction.” *arXiv (Cornell University),* January. https://doi.org/10.48550/arxiv.2401.03482.

Chopra, Hitesh, None Annu, Dong K. Shin, Kavita Munjal, None Priyanka, Kuldeep Dhama, and Talha B. Emran. 2023. “Revolutionizing Clinical Trials: The Role Of AI in Accelerating Medical Breakthroughs.” *International Journal of Surgery* 109 (12): 4211–20. https://doi.org/10.1097/js9.0000000000000705.

Cleophas, Ton J.M., and Ed M. De Vogel. 1998. “Crossover Studies Are A Better Format For Comparing Equivalent Treatments Than Parallel-group Studies.” *Pharmacy World & Science* 20 (3): 113–17. https://doi.org/10.1023/a:1008626002664.

Clinicaltrials.gov. (n.d.). https://clinicaltrials.gov/about-site/about-ctg

Council, Science. 2023. “Potential Benefits And Limitations Of mRNA Technology For Vaccine Research And Development For Infectious Diseases And Virus-induced Cancers.” December 13, 2023.https://www.who.int/publications/i/item/9789240084551.

Davis, Jesse, and Goadrich, Mark 2006. "The Relationship Between Precision-Recall and ROC Curves." *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 233–240. DOI:10.1145/1143844.1143874

DiMasi, Joseph A., Henry G. Grabowski, and Ronald W. Hansen. 2016. “Innovation in the Pharmaceutical Industry: New Estimates of R&Amp;D Costs.” *Journal of Health Economics* 47 (February): 20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012.

Douze, Matthijs, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. "The Faiss Library." *arXiv (Cornell University)*, January. https://doi.org/10.48550/arxiv.2401.08281.

Edwards, Philip James, Ian Roberts, Mike J Clarke, Carolyn DiGuiseppi, Benjamin Woolf, and Chloe Perkins. 2023. "Methods to Increase Response to Postal and Electronic Questionnaires." *Cochrane Library* 2023 (11). https://doi.org/10.1002/14651858.mr000008.pub5.

Eisenstein, Eric L., Lemons, Philip W., Tardiff, Barbara E., Schulman, Kevin A., Jolly, M. King, Califf., Robert M. 2005 "Reducing the costs of phase III cardiovascular clinical trials." *American Heart Journal*, 149(3): 482–488. doi: 10.1016/j.ahj.2004.04.049.

Emmanuel, Tlamelo, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago, and Oteng Tabona. 2021. "A Survey on Missing Data in Machine Learning." *Journal of Big Data* 8 (1). https://doi.org/10.1186/s40537-021-00516-9.

Eularis. 2024. "How synthetic data and digital twins are transforming research and clinical trials." Eularis. Accessed December 8th, 2024. https://eularis.com/how-synthetic-data-and-digital-twins-are-transforming-research-and-clinical-trials/

Faizullabhoy, Mariam, and Gauri Wani. 2024. "Clinical Trials Market - by Phase (I, II, III, IV), Study Design (Interventional, Observational, Expanded Access), Service Type (Outsourcing, In-house), Therapeutic Area (Oncology, Dermatology, Neurology, Cardiology) – Global Forecast (2024 – 2032)." *Global Market Insights Inc.* Accessed November 30th 2024 https://www.gminsights.com/industry-analysis/clinical-trials-market.

Fawcett, T. (2005). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Financial Times. 2024. "Europe Falls Behind China in Playing Host to Clinical Drug Trials." Accessed November 30th 2024 https://www.ft.com/content/3d861acb-8e7d-4157-b845-81124254da8a.

Fogel, David B. 2018. "Factors Associated With Clinical Trials That Fail and Opportunities for Improving the Likelihood of Success: A Review." *Contemporary Clinical Trials Communications* 11 (August): 156–64. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review - PMC.

Fu, Tianfan, Kexin Huang, Cao Xiao, Lucas M. Glass, and Jimeng Sun. 2022. "HINT: Hierarchical Interaction Network for Clinical-trial-outcome Predictions." *Patterns* 3 (4): 100445. https://doi.org/10.1016/j.patter.2022.100445.

Gayvert, Kaitlyn M., Neel S. Madhukar, and Olivier Elemento. 2016. "A Data-Driven Approach To Predicting Successes And Failures Of Clinical Trials." *Cell Chemical Biology* 23 (10): 1294–1301. https://doi.org/10.1016/j.chembiol.2016.07.023.

Geifman, Yonatan., El-Yaniv, Ran. 2017. "Selective Classification for Deep Neural Networks." *arXiv (Cornell University)*, arXiv:1705.08500.

Getz, Kenneth A., Stella Stergiopoulos, Mary Short, Leon Surgeon, Randy Krauss, Sybrand Pretorius, Julian Desmond, and Derek Dunn. 2016. "The Impact of Protocol Amendments on Clinical Trial Performance and Cost." *Therapeutic Innovation & Regulatory Science* 50 (4): 436–41. https://doi.org/10.1177/2168479016632271.

Ghim, Jong-Lyul, and Sangzin Ahn. 2023. “Transforming Clinical Trials: The Emerging Roles Of Large Language Models.” *Translational and Clinical Pharmacology* 31 (3): 131. https://doi.org/10.12793/tcp.2023.31.e16.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. “Generative Adversarial Nets.” 2014. https://papers.nips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

Gower, J. C. 1971. "A General Coefficient of Similarity and Some of Its Properties." *Biometrics 27* (December): 857-74. https://members.cbio.mines-paristech.fr/~jvert/svn/bibli/local/Gower1971general.pdf

Grandperrin, Jonathan. 2024. “How to Use Confidence Scores in Machine Learning Models.” Mindee. December 3, 2024. (Accessed December 17th, 2024) https://www.mindee.com/blog/how-use-confidence-scores-ml-models.

Guo, Wenjing, Liu, Jie, Dong, Fang, Song, Meng, Li, Zoe, Khan, Md Kamrul Hasan, Patterson Tucker A., Hong Huixiao 2023. "Review of machine learning and deep learning models for toxicity prediction. " *Experimental biology and medicine (Maywood, N.J.),* 248(21), 1952–1973. https://doi.org/10.1177/15353702231209421

Gupta, Shailja, Rajesh Ranjan, and Surya Narayan Singh. 2024. “A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions.” *arXiv (Cornell University),* October. https://doi.org/10.48550/arxiv.2410.12837.

Heymans, Martijn W., and Jos W.R. Twisk. 2022. “Handling Missing Data in Clinical Research.” *Journal of Clinical Epidemiology* 151 (September): 185–88. https://doi.org/10.1016/j.jclinepi.2022.08.016.

Huang, Jin, and Ling, C. X. 2005. "Using AUC and Accuracy in Evaluating Learning Algorithms." *IEEE Transactions on Knowledge and Data Engineering*, 17(3): 299–310. doi: 10.1109/TKDE.2005.50

Huang, Kexin, Jaan Altosaar, and Rajesh Ranganath. 2019. “ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission.” arXiv (Cornell University), January. https://doi.org/10.48550/arxiv.1904.05342.

John Hopkins Medicine. n.d. "Understanding the Benefits of Clinical Trials for Cancer." Accessed November 30th 2024 https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/understanding-the-benefits-of-clinical-trials-for-cancer.

Johnson, Justin M., and Taghi M. Khoshgoftaar. 2020. “Thresholding Strategies for Deep Learning With Highly Imbalanced Big Data.” In *Advances in Intelligent Systems and Computing*, 199–227. https://doi.org/10.1007/978-981-15-6759-9_9.

Kappen, Teus H., Wilton A. Van Klei, Leo Van Wolfswinkel, Cor J. Kalkman, Yvonne Vergouwe, and Karel G. M. Moons. 2018. “Evaluating The Impact Of Prediction Models: Lessons Learned, Challenges, And Recommendations.” *Diagnostic and Prognostic Research* 2 (1). https://doi.org/10.1186/s41512-018-0033-6.

Kavalci, Ece, and Anthony Hartshorn. 2023. “Improving Clinical Trial Design Using Interpretable Machine Learning Based Prediction Of Early Trial Termination.” *Scientific Reports* 13 (1). https://doi.org/10.1038/s41598-023-27416-7.

Kim, Eungdo, Jaehoon Yang, Sungjin Park, and Kwangsoo Shin. 2023. “Factors Affecting Success of New Drug Clinical Trials.” *Therapeutic Innovation & Regulatory Science* 57 (4): 737–50. https://doi.org/10.1007/s43441-023-00509-1.

Kipf, T. N. (2016). Semi-supervised classification with graph Convolutional Networks. ICLR 2017. arXiv. doi:https://doi.org/10.48550/arXiv.1609.02907

Krenmayr, Lucas, Roland Frank, Christina Drobig, Michael Braungart, Jan Seidel, Daniel Schaudt, Reinhold Von Schwerin, and Kathrin Stucke-Straub. 2022. "GANerAid: Realistic Synthetic Patient Data for Clinical Trials." January 1. Accessed December 1, 2024. https://doaj.org/article/6e6d854321a146ae8db29c8420ebc642.

Laakso, Mikael, and Bo-Christer Björk. 2012. “Anatomy Of Open Access Publishing: A Study Of Longitudinal Development And Internal Structure.” *BMC Medicine* 10 (1). https://doi.org/10.1186/1741-7015-10-124.

Lakshminarayanan, Balaji, Pritzel, Alexander, and Blundell, Charles. 2017. "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles." *Advances in Neural Information Processing Systems*, 30: 1–12. https://papers.nips.cc/paper_files/paper/2017

Lewis, Patrick S. H., Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. 2020. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” *Neural Information Processing Systems* 33 (May): 9459–74. https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Lo, Andrew W., Kien Wei Siah, and Chi Heem Wong. 2018. “Machine-Learning Models for Predicting Drug Approvals and Clinical-Phase Transitions.” *SSRN Electronic Journal*, January. https://doi.org/10.2139/ssrn.2973611.

Lo, Andrew W., Kien Wei Siah, and Chi Heem Wong. 2019. “Machine Learning With Statistical Imputation For Predicting Drug Approvals: Supplementary Materials.” *Harvard Data Science Review* 1 (1). https://doi.org/10.1162/99608f92.5c5f0525/2a3160be.

Logan, Jennifer K., Chad Tang, Zhongxing Liao, J. Jack Lee, John V. Heymach, Stephen G. Swisher, James W. Welsh, Jianjun Zhang, Steven H. Lin, and Daniel R. Gomez. 2016. “Analysis of Factors Affecting Successful Clinical Trial Enrollment in the Context of Three Prospective, Randomized, Controlled Trials.” *International Journal of Radiation Oncology*Biology*Physics* 97 (4): 770–77. Analysis of Factors Affecting Successful Clinical Trial Enrollment in the Context of Three Prospective, Randomized, Controlled Trials - PubMed.

Malkov, Yu A., and D. A. Yashunin. 2018. “Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42 (4): 824–36. https://doi.org/10.1109/tpami.2018.2889473.

Matthias Fey, J. E. (2019). Fast Graph Representation Learning with Pytorch Geometric. Department of Computer Graphics.

McHugh, Leo C., Kevin Snyder, and Thomas D. Yager. 2019. “The Effect of Uncertainty in Patient Classification on Diagnostic Performance Estimations.” *PLoS ONE* 14 (5): e0217146. https://doi.org/10.1371/journal.pone.0217146.

Mercaldo, Sarah Fletcher, and Jeffrey D Blume. 2018. “Missing Data and Prediction: The Pattern Submodel.” *Biostatistics* 21 (2): 236–52. https://doi.org/10.1093/biostatistics/kxy040.

Miller, J. S. (2019). “Law’s semantic self-portrait: Discerning doctrine with co-citation networks and keywords”. *University of Pittsburgh Law Review*, 81(1). doi:https://doi.org/10.5195/lawreview.2019.660

Minh N. Vu, M. T. (2020). “PGM-Explainer: Probabilistic Graphical Model Explanations for Graph Neural Networks”. *arXiv*. doi:https://doi.org/10.48550/arXiv.2010.05788

Modi, Shanu, William Jacot, Toshinari Yamashita, Joohyuk Sohn, Maria Vidal, Eriko Tokunaga, Junji Tsurutani, et al. 2022. “Trastuzumab Deruxtecan in Previously Treated HER2-Low Advanced Breast Cancer.” *New England Journal of Medicine* 387 (1): 9–20. https://doi.org/10.1056/nejmoa2203690.

Montgomery, Alan A, Tim J Peters, and Paul Little. 2003. “Design, Analysis And Presentation Of Factorial Randomised Controlled Trials.” *BMC Medical Research Methodology* 3 (1). https://doi.org/10.1186/1471-2288-3-26.

Muehlematter, Urs J. Daniore, Paola Vokinger, Krestin N. 2021 “Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): a comparative analysis”. *Lancet Digit Health*., 3(3) doi: 10.1016/S2589-7500(20)30292-2. Epub 2021 Jan 18. PMID: 33478929.

Mukund Sundararajan, A. T. (2017). “Axiomatic Attribution for Deep Networks”. *ArXiv*, 3. doi:https://doi.org/10.48550/arXiv.1703.01365

Murali, V. 2021. “Data - Predicting Clinical Trial Outcomes Using Drug Bioactivities.” *IEEE DataPort*. December 27, 2021. https://ieee-dataport.org/documents/data-predicting-clinical-trial-outcomes-using-drug-bioactivities.

Mylonas, N. M. (2023). An attention matrix for every decision: Faithfulness-based arbitration among multiple attention-based interpretations of transformers in text classification. Data Mining and Knowledge Discovery, 38(1), 128–153. doi:https://doi.org/10.1007/s10618-023-00962-4

Nemani, K. V., et al. 2023. "Bayesian Models in Clinical Trials: Estimating Success Probabilities Using Priors." *Statistics in Medicine*, 42(5): 1231–1247.

Nestor, Bret A., Matthew B. A. McDermott, Geeticka Chauhan, Tristan Naumann, Michael C. Hughes, A. Goldenberg, and M. Ghassemi. 2018. “Rethinking Clinical Prediction: Why Machine Learning Must Consider Year of Care and Feature Aggregation.” 2018. https://www.semanticscholar.org/paper/Rethinking-clinical-prediction%3A-Why-machine-must-of-Nestor-McDermott/393b41ea791f5a70ee87b361dec1970745d1908c#:~:text=This%20work%20augments%20MIMIC%20with%20the%20year%20in,aggregates%20of%20raw%20features%20significantly%20mitigate%20future%20deterioration.

Office of the Commissioner. 2018. “Step 3: Clinical Research.” U.S. Food And Drug Administration. January 4, 2018. https://www.fda.gov/patients/drug-development-process/step-3-clinical-research

Ong, Jasmine Chiat Ling, Liyuan Jin, Kabilan Elangovan, Gilbert Yong San Lim, Daniel Yan Zheng Lim, Gerald Gui Ren Sng, Yuhe Ke, et al. 2024. “Development and Testing of a Novel Large Language Model-Based Clinical Decision Support Systems for Medication Safety in

12 Clinical Specialties." *arXiv (Cornell University)*, January. https://doi.org/10.48550/arxiv.2402.01741.

Oualikene-Gonin, Wahiba, Marie-Christine Jaulent, Jean-Pierre Thierry, Sofia Oliveira-Martins, Laetitia Belgodère, Patrick Maison, and Joël Ankri. 2024. "Artificial Intelligence Integration in the Drug Lifecycle and in Regulatory Science: Policy Implications, Challenges and Opportunities." *Frontiers in Pharmacology* 15 (August). https://doi.org/10.3389/fphar.2024.1437167.

Parsons, Nick R., Joydeep Basu, and Nigel Stallard. 2024. "Group Sequential Designs For Pragmatic Clinical Trials With Early Outcomes: Methods And Guidance For Planning And Implementation." *BMC Medical Research Methodology* 24 (1). https://doi.org/10.1186/s12874-024-02174-w.

Petar Veličković, Guillem Cucurull. 2018. "Graph Attention Networks." *International Conference on Learning Representations*. https://openreview.net/forum?id=rJXMpikCZ

Petridis, Savvas, Ben Wedin, James Wexler, Aaron Donsbach, Mahima Pushkarna, Nitesh Goyal, Carrie J. Cai, and Michael Terry. 2023. "ConstitutionMaker: Interactively Critiquing Large Language Models by Converting Feedback Into Principles." arXiv (Cornell University), January. https://doi.org/10.48550/arxiv.2310.15428.

Pope, Phillip. E. Kolouri, Soheil Rostam, Mohammad Martin, Charles E. Hoffman, Heiko 2019. Explainability methods for graph convolutional neural networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 10764-10773. doi: 10.1109/CVPR.2019.01103.

Pytrial. “Welcome to PyTrial documentation!”. Accessed October 11, 2024. https://pytrial.readthedocs.io/en/latests

Reblitz-Richardson, N. K. 2020. “Captum: A unified and generic model interpretability library for PyTorch.” *arXiv:2009.07896*. doi:https://doi.org/10.48550/arXiv.2009.07896

Reinisch, Michael, Jianfeng He, Chenxi Liao, Sauleh Ahmad Siddiqui, and Bei Xiao. 2024. “CTP-LLM: Clinical Trial Phase Transition Prediction Using Large Language Models.” *arXiv.Org.* August 20, 2024. https://arxiv.org/abs/2408.10995.

Rubin, Donald B. 1976. “Inference and Missing Data.” *Biometrika* 63 (3): 581–92. https://doi.org/10.2307/2335739https://www.jstor.org/stable/2335739.

Saito, Takaya, and Rehmsmeier, Marc 2015. "The Precision-Recall Plot Is More Informative Than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets." *PLoS One*, 10(3): e0118432. DOI: 10.1371/journal.pone.0118432

Shahriar, Sakib, Brady D. Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. 2024. “Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency.” *Applied Sciences* 14 (17): 7782. https://doi.org/10.3390/app14177782.

Shanu, M., et al. 2023. "Trastuzumab Deruxtecan in Previously Treated HER2-Low Advanced Breast Cancer." *New England Journal of Medicine*, 387(1): 9–20.

Simonyan, K. A. 2014. “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”. *arXiv:1312.6034.* doi:https://doi.org/10.48550/arXiv.1312.6034

Singh, Nathan, Noelle V. Frey, Stephan A. Grupp, and Shannon L Maude. 2016. “CAR T Cell Therapy in Acute Lymphoblastic Leukemia and Potential for Chronic Lymphocytic Leukemia.” *Current Treatment Options in Oncology* 17 (6). https://doi.org/10.1007/s11864-016-0406-4.

Sokolova, Marina, and Lapalme, Guy 2009. "A Systematic Analysis of Performance Measures for Classification Tasks." *Information Processing & Management*, 45(4): 427–437. DOI:10.1016/j.ipm.2009.03.002

Su, Qianmin, Gaoyi Cheng, and Jihan Huang. 2023. “A Review of Research on Eligibility Criteria for Clinical Trials.” *Clinical and Experimental Medicine* 23 (6): 1867–79. https://doi.org/10.1007/s10238-022-00975-1.

Sundararajan, M. A. 2024. “Axiomatic Attribution for Deep Networks: Proceedings of the 34th International Conference on Machine Learning - Volume 70.” doi:https://dl.acm.org/doi/10.5555/3305890.3306024

Thorlund, Kristian, Haggstrom, Jonas, Park, Jay J.H., Mills, Edward J. 2018. "Key design considerations for adaptive clinical trials: a primer for clinicians."*BMJ*, 360:k698. https://doi.org/10.1136/bmj.k698

Unlu, Ozan, Jiyeon Shin, Charlotte J Mailly, Michael F Oates, Michela R Tucci, Matthew Varugheese, Kavishwar Wagholikar, et al. 2024. “Retrieval Augmented Generation Enabled Generative Pre-Trained Transformer 4 (GPT-4) Performance for Clinical Trial Screening.” *medRxiv (Cold Spring Harbor Laboratory)*, February. https://doi.org/10.1101/2024.02.08.24302376.

Van Rijsbergen, C. J. 1979. *Information Retrieval*. London: Butterworth-Heinemann.

Venkit, Pranav Narayanan, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. 2024. "An Audit on the Perspectives and Challenges of Hallucinations in NLP." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, January, 6528–48. https://doi.org/10.18653/v1/2024.emnlp-main.375.

Vickery, B.C. 1979. "Reviews : Van Rijsbergen, C. J. Information Retrieval. 2nd Edn. London, Butterworths, I978. 208pp." *Journal of Librarianship* 11 (3): 237. https://doi.org/10.1177/096100067901100306.

Vovk, Vladimir, Gammerman, Alexander, and Shafer, Glenn 2005. *Algorithmic Learning in a Random World*. New York: Springer.

Vrbanac, J., & Slauter, R. (2016). ADME in drug discovery. In *Elsevier eBooks* (pp. 39–67). https://doi.org/10.1016/b978-0-12-803620-4.00003-7

Wang, Calvin, Joshua Ong, Chara Wang, Hannah Ong, Rebekah Cheng, and Dennis Ong. 2023. "Potential for GPT Technology to Optimize Future Clinical Decision-Making Using Retrieval-Augmented Generation." *Annals of Biomedical Engineering* 52 (5): 1115–18. https://doi.org/10.1007/s10439-023-03327-6.

Wang, Guangyu, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, et al. 2023. "Optimized Glycemic Control of Type 2 Diabetes With Reinforcement Learning: A Proof-of-concept Trial." *Nature Medicine* 29 (10): 2633–42. https://doi.org/10.1038/s41591-023-02552-9.

Wang, Huimin, Jianxiang Tang, Mengyao Wu, Xiaoyu Wang, and Tao Zhang. 2022. "Application of Machine Learning Missing Data Imputation Techniques in Clinical Decision

Making: Taking the Discharge Assessment of Patients With Spontaneous Supratentorial Intracerebral Hemorrhage as an Example.” *BMC Medical Informatics and Decision Making* 22 (1). https://doi.org/10.1186/s12911-022-01752-6.

Wang, Minyan, Huan Ma, Yun Shi, Haojie Ni, Chu Qin, and Conghua Ji. 2024. “Single-arm Clinical Trials: Design, Ethics, Principles.” *BMJ Supportive & Palliative Care*, June, spcare-004984. https://doi.org/10.1136/spcare-2024-004984.

Wang, Tao, Xiangwei Zheng, Lifeng Zhang, Zhen Cui, and Chunyan Xu. 2023. “A Graph-based Interpretability Method for Deep Neural Networks.” *Neurocomputing* 555 (August): 126651. https://doi.org/10.1016/j.neucom.2023.126651.

Wang, Wenhui, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. “MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers.” *Neural Information Processing Systems* 33 (February): 5776–88. https://proceedings.neurips.cc/paper/2020/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, Xiao, Houye Ji, Chuan Shi, Bai Wang, Peng Cui, P. Yu, and Yanfang Ye. 2019. “Heterogeneous Graph Attention Network.” *arXiv (Cornell University)*, January. https://doi.org/10.48550/arxiv.1903.07293.

Wang, Zifeg Theodorou, Brandon Fu, Tianfan Xiao, Cao & Sun Jimeng. (2023). “PYTRIAL: Machine Learning Software and Benchmark for Clinical Trial Applications.” *arXiv:2306.04018.* https://doi.org/10.48550/arXiv.2306.04018

Weininger, David. 1988. “SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules.” *Journal of Chemical Information and Computer Sciences* 28 (1): 31–36. https://doi.org/10.1021/ci00057a005.

Weissler, E. Hope, Tristan Naumann, Tomas Andersson, Rajesh Ranganath, Olivier Elemento, Yuan Luo, Daniel F. Freitag, et al. 2021. “Correction To: The Role Of Machine Learning in Clinical Research: Transforming The Future Of Evidence Generation.” *Trials* 22 (1). https://doi.org/10.1186/s13063-021-05571-4.

Wong, Chi Heem, Kien Wei Siah, and Andrew W Lo. 2017. “Estimation Of Clinical Trial Success Rates And Related Parameters.” *Biostatistics* 20 (2): 273–86. https://doi.org/10.1093/biostatistics/kxx069.

World Health Organization. 2024. "Global Cancer Burden Growing, Amidst Mounting Need for Services." Accessed November 30th 2024 https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services.

Wu, Lingfei, Peng Cui, Jian Pei, and Liang Zhao. 2022. *Graph Neural Networks: Foundations, Frontiers, and Applications*. Springer Nature.

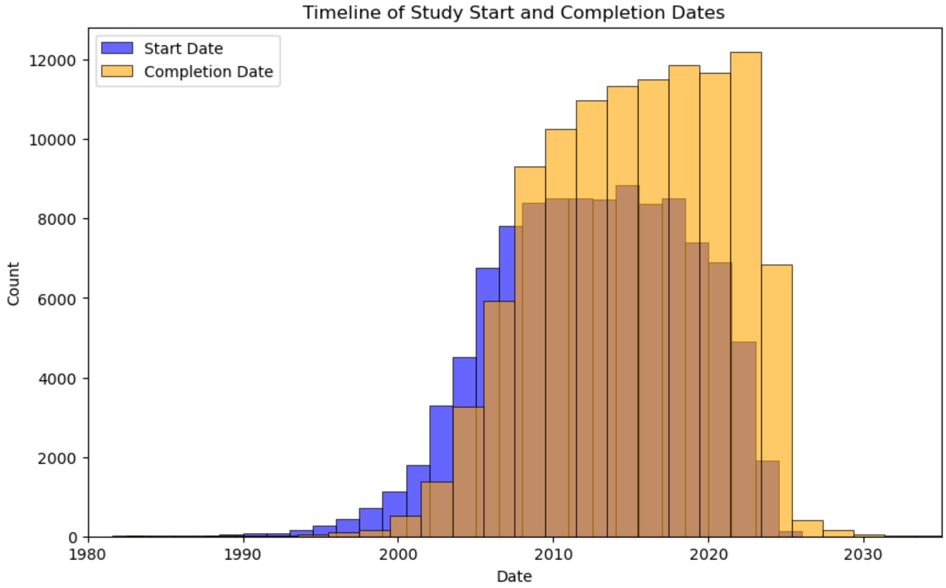
Xu, Lei, and Veeramachaneni, Kalyan 2018. “Synthesizing Tabular Data Using Generative Adversarial Networks.”. https://www.semanticscholar.org/paper/Synthesizing-Tabular-Data-using-Generative-Networks-Xu-Veeramachaneni/d0c63d02bd63ecc0fb9debfe2b93b2e6be983ed8.

Xu, Lei, Skoularidou, Maria, Cuesta-Infante, Alfredo and Veeramachaneni, Kalyan. 2019. “Modeling Tabular Data Using Conditional GAN.”. https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html.

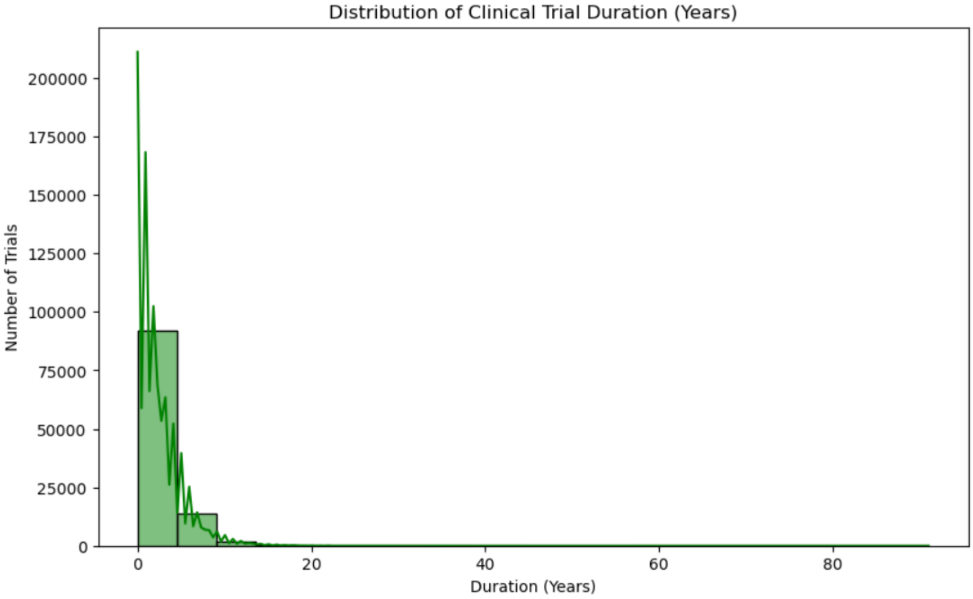
Yepes, Antonio Jimeno, Yao You, Jan Milczek, Sebastian Laverde, and Leah Li. 2024. “Financial Report Chunking for Effective Retrieval Augmented Generation.” *arXiv (Cornell University)*, February. https://doi.org/10.48550/arxiv.2402.05131.

Zheng, Wenhao, Dongsheng Peng, Hongxia Xu, Hongtu Zhu, Tianfan Fu, and Huaxiu Yao. 2024. “Multimodal Clinical Trial Outcome Prediction With Large Language Models.” *arXiv (Cornell University)*, February. https://doi.org/10.48550/arxiv.2402.06512.

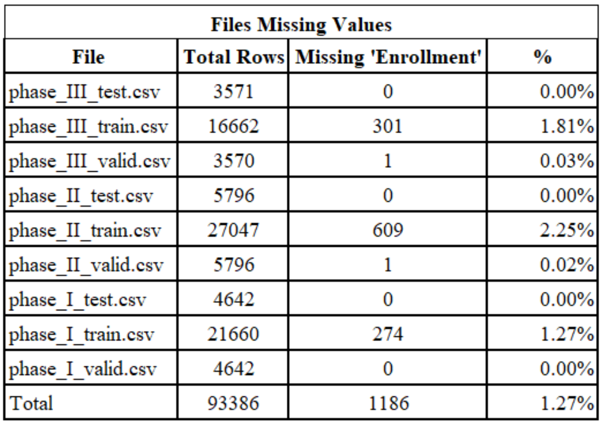
Appendix



Graphic 1 - The graph shows an increasing trend in clinical studies over the years, with relative completion date according to ClinicalTrials.gov.

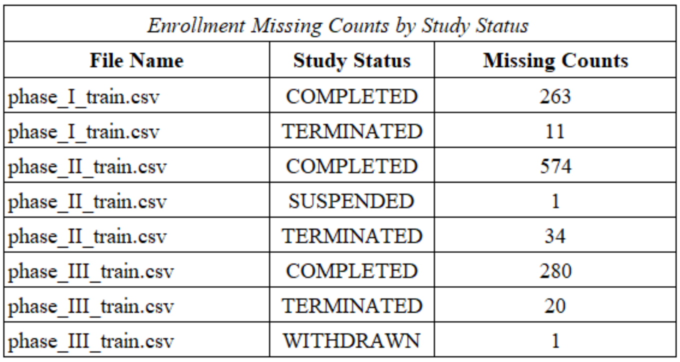


Graphic 2 - These graphs show the right-skewed distribution of the clinical trial duration in terms of year.



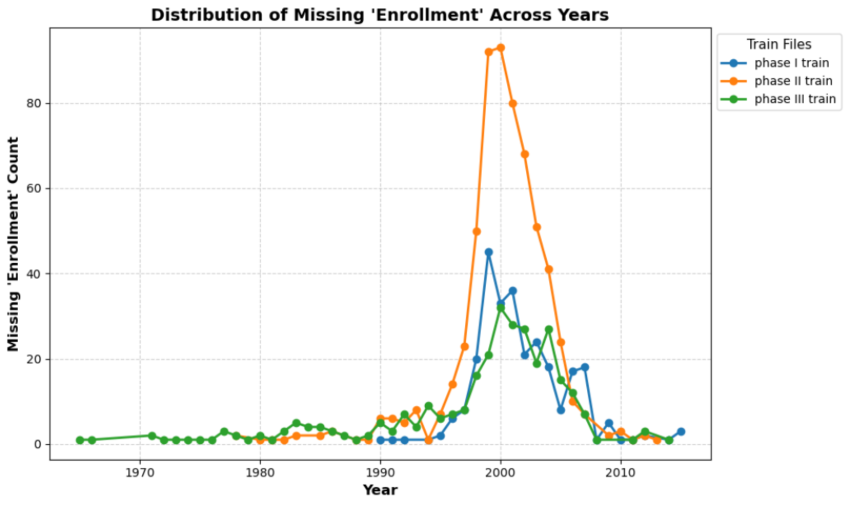
Files Missing Values			
File	**Total Rows**	**Missing 'Enrollment'**	**%**
phase_III_test.csv	3571	0	0.00%
phase_III_train.csv	16662	301	1.81%
phase_III_valid.csv	3570	1	0.03%
phase_II_test.csv	5796	0	0.00%
phase_II_train.csv	27047	609	2.25%
phase_II_valid.csv	5796	1	0.02%
phase_I_test.csv	4642	0	0.00%
phase_I_train.csv	21660	274	1.27%
phase_I_valid.csv	4642	0	0.00%
Total	93386	1186	1.27%

Table 1 - Number of rows with missing values

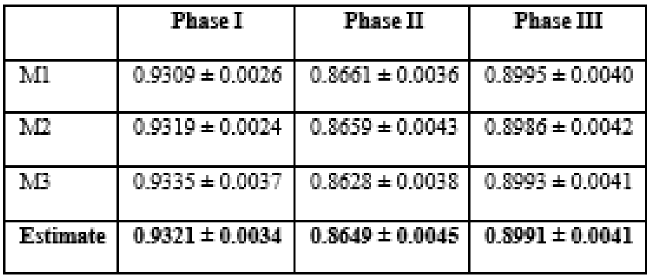


Enrollment Missing Counts by Study Status		
File Name	**Study Status**	**Missing Counts**
phase_I_train.csv	COMPLETED	263
phase_I_train.csv	TERMINATED	11
phase_II_train.csv	COMPLETED	574
phase_II_train.csv	SUSPENDED	1
phase_II_train.csv	TERMINATED	34
phase_III_train.csv	COMPLETED	280
phase_III_train.csv	TERMINATED	20
phase_III_train.csv	WITHDRAWN	1

Table 2 - Enrollment missing counts over Study Status

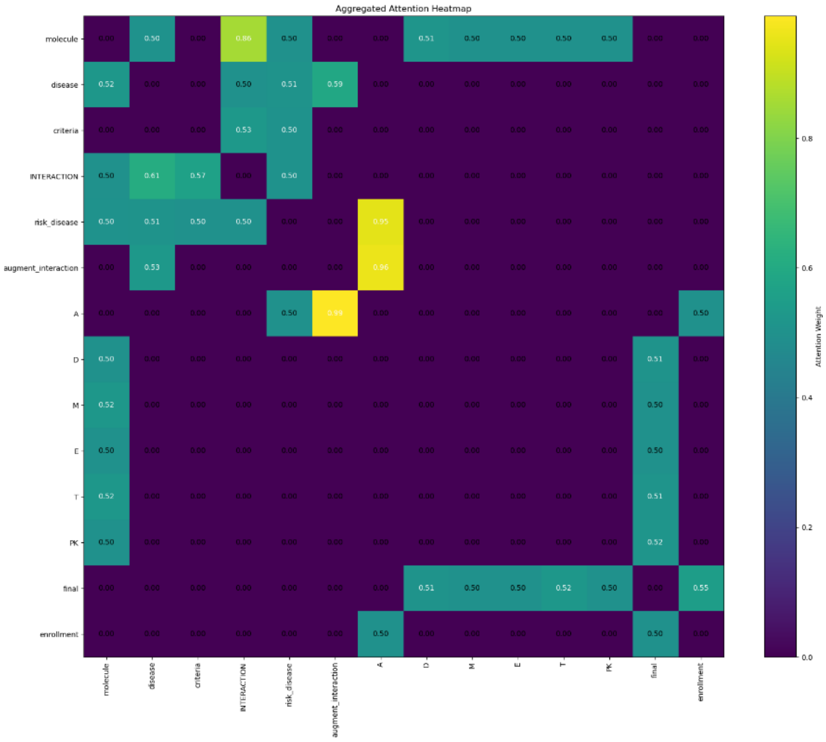


Graphic 3 - Enrollment missing values over years.



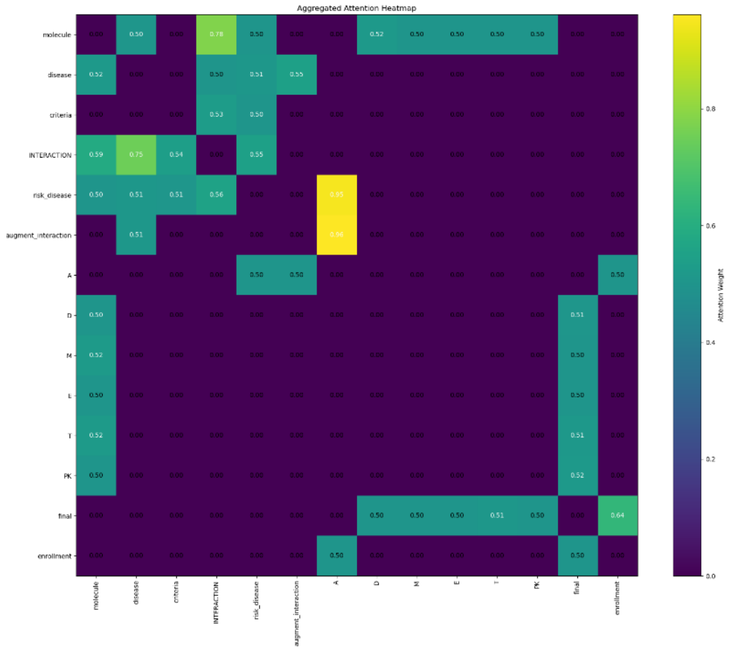
	Phase I	Phase II	Phase III
M1	0.9309 ± 0.0026	0.8661 ± 0.0036	0.8995 ± 0.0040
M2	0.9319 ± 0.0024	0.8659 ± 0.0043	0.8986 ± 0.0042
M3	0.9335 ± 0.0037	0.8628 ± 0.0038	0.8993 ± 0.0041
Estimate	**0.9321 ± 0.0034**	**0.8649 ± 0.0045**	**0.8991 ± 0.0041**

Table 3 - Estimate of the multiple imputations of the three imputed datasets using Rubin's Rule

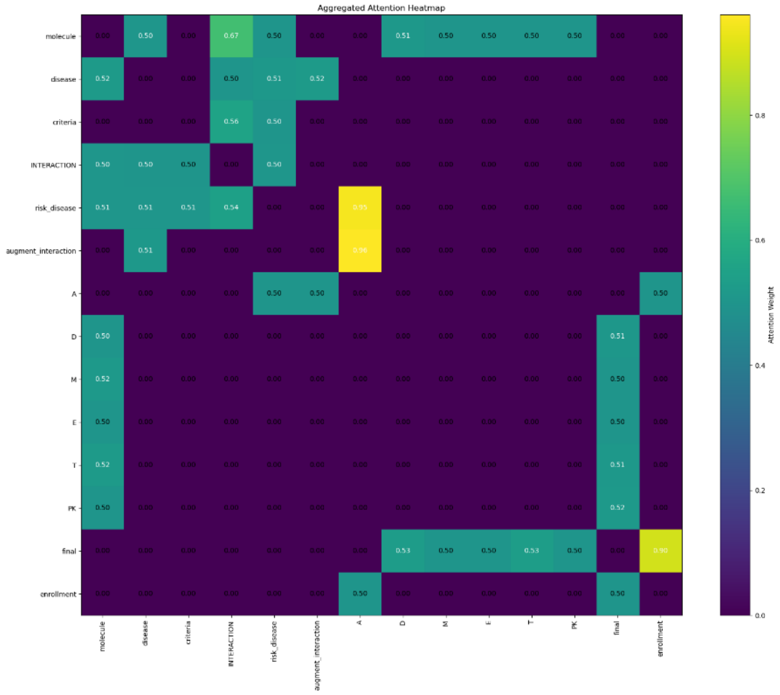


	molecule	disease	criteria	INTERACTION	risk_disease	augment_interaction	A	D	M	E	T	PK	final	enrollment
molecule	0.00	0.50	0.00	0.86	0.50	0.00	0.00	0.51	0.50	0.50	0.50	0.50	0.00	0.00
disease	0.52	0.00	0.00	0.50	0.51	0.59	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
criteria	0.00	0.00	0.00	0.53	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
INTERACTION	0.50	0.61	0.57	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
risk_disease	0.50	0.51	0.50	0.50	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00
augment_interaction	0.00	0.53	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A	0.00	0.00	0.00	0.00	0.50	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50
D	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.51	0.00
M	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00
E	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00
T	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.51	0.00
PK	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.52	0.00
final	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.51	0.50	0.50	0.52	0.50	0.00	0.55
enrollment	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.50	0.00

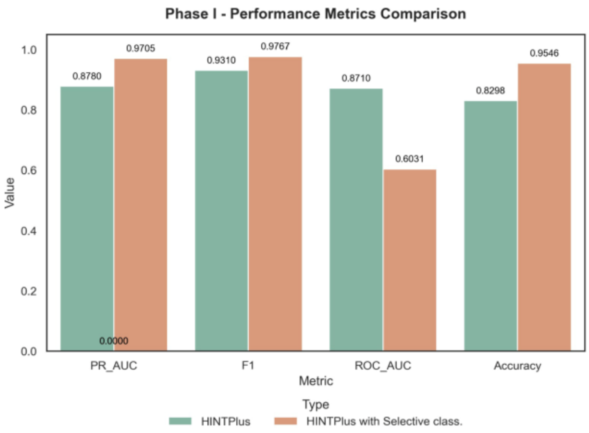
14Graphic 4 - Attention matrix heatmap in Phase I.



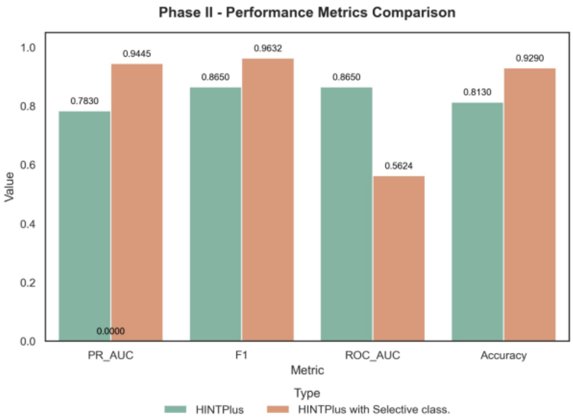
Graphic 5 - Attention matrix heatmap of Phase II



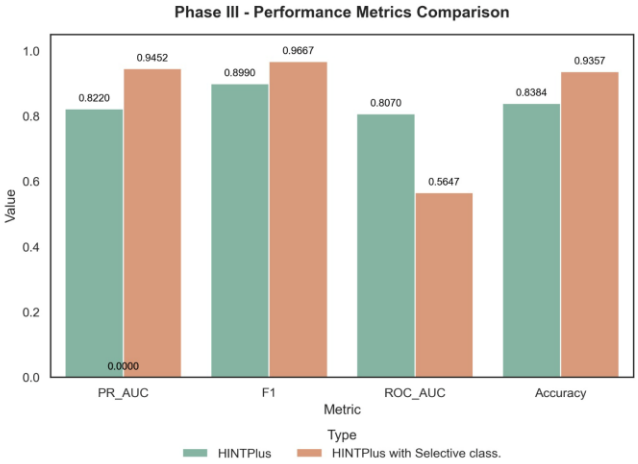
Graphic 6 - Attention matrix heatmap in Phase III.



Graphic 7 - Phase I – Performance Metrics Comparison



Graphic 8 - Phase II – Performance Metrics Comparison



Graphic 9 - Phase III – Performance Metrics Comparison