ChatCTP: Predicting Clinical Trial Phase Transitions Using Large Language Models

Author
Michael Reinisch
TU Graz

Supervisor
Bei Xiao
American University

Supervisor
Horst Bischof
TU Graz

Abstract

The typical regulatory process requires medical interventions to move through multiple phases of clinical trials. Passing a trial phase depends on several factors, such as safety, efficacy, and statistical significance, and can be influenced by the trial design. Since fruitless clinical trials are an unnecessary loss of money and time, we investigated Clinical Trial Outcome Prediction (CTOP). Previous CTOP-related works have two issues. First, trials are labeled based on their completion or termination status, which does not indicate if a drug progresses through the regulatory process. Second, the previous models used data that is prone to introduce look-ahead bias, as they rely on supplementary drug and trial information gathered in later stages of the process. To address these issues, we propose ChatCTP, the first attempt at using LLMs in CTOP. Our model, a fine-tuned version of GPT 3.5, can predict if a drug will transition from one clinical trial stage to the next, solely based on the original textual description of the trial design. Furthermore, we release the PhaseTransition Dataset with accurate labels to benefit future research. Our model shows an improvement of 4.20% on the F1-Score over baselines and demonstrates that fine-tuned GPT 3.5 can outperform specialized baseline models, while the original GPT 3.5 does not.

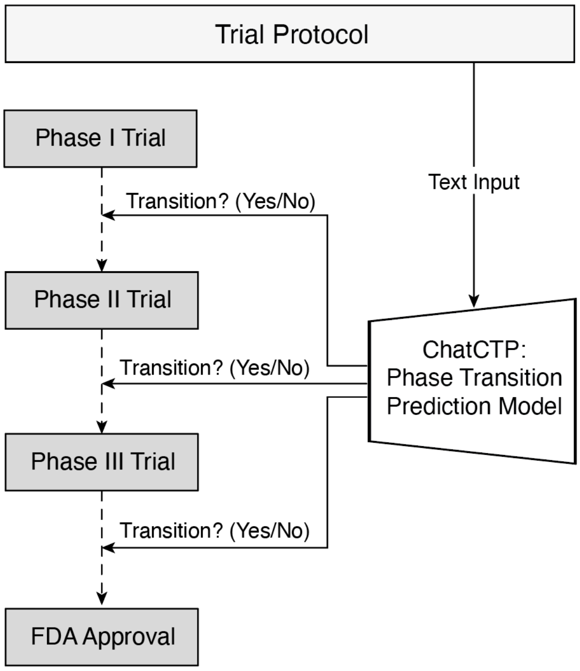

Figure 1: A new treatment is typically tested in three phases, starting with safety evaluation and dosage in Phase I, then assessing efficacy in Phase II, and finally confirming efficacy and safety in Phase III, before being evaluated by the FDA. However, the treatment can drop out in any phase for various reasons, wasting time and resources. We propose ChatCTP, an LLM-based model, to predict trial phase transitions, based solely on trial protocols, before starting a trial.

1 Introduction

A clinical trial is a systematic investigation conducted on human subjects to evaluate the safety and efficacy of medical interventions, typically categorized into Phase I, II, and III stages, aiming to obtain market approval. Phase transition prediction in clinical trials refers to forecasting whether a trial will progress from one phase to the next in the drug development process, typically from early phases (Phase I or Phase II) to later phases (Phase II or Phase III), based on the available trial data. Passing a trial phase depends on several factors, such as safety, efficacy, statistical significance, and the trial design (see Figure 1). Previous research indicates that trial protocol complexity, longer cycle times, and increased investigative site work burden also contribute to poor trial performance and failure (Gayvert et al., 2016; DiMasi et al., 2015). Therefore, predicting clinical trial phase transitions is essential for relevant stakeholders to anticipate the trajectory of drug development, allocate resources efficiently, and make informed decisions regarding further investment in promising treatments. Here, we develop a novel method based on large language models (LLMs) to predict trial outcomes and phase transition solely based on trial design. We also present a combined deep and shallow learning approach as a cost- and resource-efficient alternative that can be trained on most machines.

Two of the most comprehensive data sets of clinical trial designs and drug characteristics available (the Aggregate Analysis of ClinicalTrials.gov, AACT [34], based on https://clinicaltrials.gov/ and Biomedtracker [26] data sets) were used for this case study. ClinicalTrials.gov is a publicly available database of privately and publicly funded medical research studies conducted in the US and over 203 other countries. The current version contains information from over 481,000 studies and is maintained by the National Library of Medicine at the National Institutes of Health. Each record for a clinical trial includes information on over a hundred characteristics of the trial protocol. The protocol defines how the trial will be conducted and its design can be considered the first stage in the process of a clinical trial. Given that this documentation comprises textual descriptions of plans, objectives, and recruitment criteria, Natural Language Processing (NLP) can be leveraged to identify trial design flaws.

Previous work in clinical trial outcome prediction predominantly relied on data available at later stages of the process, such as participant enrollment details, or data limited to specific treatments, such as drug molecular composition. However, there remains a gap in approaches that effectively perform from the protocol design stage across all treatment types, including medical devices. Furthermore, since phase transition data is not easily available, previous approaches had to rely on alternative metrics to identify trial success or failure. As only phase transitions are a true indicator of success, previous efforts not incorporating this metric are inherently inaccurate predictors. Our objective is to address this shortcoming by introducing a method that accurately predicts clinical trial phase transitions at the earliest possible stage - the design of the clinical trial protocol.

Using the trial protocol as input poses several challenges. First, the complex nature of the text and its domain-specific vocabulary requires an NLP model that is pre-trained on medical data. Secondly, the length of the texts exceeds the capabilities of most transformer architectures, which are the backbone of the language models.

Previous approaches have circumvented these challenges by relying on hand-picked features from the trial protocols, such as the number of letters as a complexity measure or focusing on drug toxicity. However, this solution disregards data relevancy. Given the task's complexity, models must be able to autonomously extract information from the text. Furthermore, to solve the problem of domain-specific texts, previous approaches had to train their models on additional medical datasets. This is not only more resource-intensive but also poses other risks. The specific datasets used in previous approaches are sourced from clinical trial results and, therefore, pose the danger of introducing a look-ahead bias to the model. Previously, the challenge of labeling trials as successful or unsuccessful relied on their completion status (completed or terminated). However, this approach is unreliable, as a trial may terminate prematurely due to significant drug efficacy.

We propose the following two language models to address the challenges posed by clinical trial outcome prediction:

- **ChatCPT**. This model is an instruction fine-tuned Large Language Model (LLM) that harnesses the medical expertise embedded within GPT-3.5 to effectively analyze domain-specific texts encountered in clinical trials. With an attention window size of 4096 tokens, ChatCPT is equipped to comprehensively process lengthy clinical trial descriptions, ensuring a thorough understanding of the intricate details inherent in trial protocols.
- **BERT+RF**. Introducing a novel architecture, this model combines the capabilities of a medical transformer with the predictive power of a random forest classifier. By embedding chunks of the trial description using the medical transformer, BERT+RF enables the collective processing of information, allowing for a holistic analysis of trial data. This innovative approach is particularly adept at handling texts that surpass the processing capabilities of traditional LLMs, thereby expanding the scope and versatility of predictive modeling in clinical trial analysis.

These models represent a significant advancement in the field of clinical trial outcome prediction, offering distinct approaches to address the complexities inherent in analyzing trial data. By leveraging state-of-the-art language processing techniques and innovative architectural designs, ChatCPT and BERT+RF are poised to revolutionize the predictive modeling landscape, paving the way for more accurate and reliable predictions in clinical trial research.

Both models exhibit remarkable proficiency in predicting phase transitions, underscoring their efficacy in the realm of clinical trial outcome prediction. The LLM achieves an F1 score of 0.737, showcasing its ability in linking trial information across various phases, thereby facilitating a comprehensive understanding of the trial trajectory. On the other hand, BERT+RF not only demonstrates superior efficiency in training but also outperforms the LLM when trained on trials from a single phase. Notably, when exclusively trained on Phase III trials, BERT+RF achieves an exceptional F1 score of 0.847, underscoring its capacity to discern nuanced patterns and trends within this specific phase. BERT+RF presents a notable advantage due to its accessibility, as it can be readily utilized and trained on a wide array of computing platforms. This accessibility extends to various machines, enabling flexibility in deployment and usage across diverse computational environments. Consequently, users can leverage BERT+RF's capabilities without encountering significant barriers related to compatibility or resource constraints, thereby enhancing accessibility and facilitating widespread adoption in clinical trial outcome prediction tasks.

Our contributions are as follows:

- **Establish success of LLM in Clinical Trial Phase Transition.** We are the first to leverage the capabilities of LLMs for the task of clinical trial outcome prediction and introduce a benchmark for future research. Our instruction fine-tuned model demonstrates superior performance, outperforming comparable approaches.

- **Release of a new dataset for CTOP**. We introduce the PhaseTransition Dataset, a new resource specifically designed for the task of clinical trial outcome prediction. This dataset includes detailed information on trial phase transitions linked to trial and drug information, enabling researchers to evaluate and compare prediction models effectively.

- **Comprehensive experiments and a new benchmark.** As we present an improved task definition for CTOP, rigorous evaluation is necessary. Our extensive testing not only demonstrates the efficacy of our proposed method but also provides a solid foundation for further advancements in this field of study.

- **Novel labeling procedure for clinical trials**. We propose a novel method for labeling clinical trial outcomes, which involves tracking a medical intervention across multiple trials and considering a trial successful if the intervention reappears in a follow-up study. This labeling approach provides a more accurate reflection of trial success and failure, addressing the limitations of existing labeling metrics.

2 Related Work

2.1 Clinical Trial Outcome Prediction

Over the years, various strategies have emerged aiming to reduce the attrition rate in clinical trials. These approaches can generally be classified into two main categories: reducing the risk by identifying and preventing adverse events in the CT process or evaluating the risks of a particular CT without interfering with the process. Within the first category, several approaches focus on eligibility criteria classification to select more suitable participants (refer to Li et al. (2022); Tian et al. (2021); Zeng et al. (2021)), while others present methods to facilitate the design of successful CT protocols (refer to Wang et al. (2022); Wang and Sun (2022)). For the second category, the literature on CT outcome prediction is considerably more extensive. Various approaches exist, but all try to answer different questions. For example, (Artemov et al., 2016) and Gayvert et al. (2016)) link the outcome of the trial to drug toxicity and side effects, Follett et al. (2019) quantify the risk of trial termination through text mining, while Qi and Tang (2019) leverage deep learning to infer the outcome of Phase III trials by analyzing a drug's trough pharmacokinetic concentration and connecting them to patient characteristics. More and more machine learning approaches utilize publicly available datasets, such as *ClinicalTrials.gov*, and adjacent bodies of literature to discover patterns in the clinical trial process hinting at their outcome.

Fu et al. (2022) present a new graph-based neural network called Hierarchical Interaction Network (HINT). The model encodes multi-modal data, including information on drug molecules, target diseases, and trial eligibility criteria. These are then connected in a graph structure to capture interaction effects between the domains. Ferdowsi et al. (2023), focus on the historical evolution of the trial protocol. By tracking significant changes in trial protocols added during its run, they retrospectively derivate one of three risk-related labels. Even though they achieve good results, their method can only asses the risk of trials mid-execution.

However, the efficacy of current trial outcome prediction models is hindered by many constraints and limitations, which warrant thorough consideration and subsequent mitigation strategies. One such limitation pertains to the reliance on data that becomes accessible solely during or post-trial, as highlighted in studies such as Feijoo et al. (2020) and Ferdowsi et al. (2023). This temporal restriction impedes the model's ability to predict trial outcomes, hindering its utility in informing decision-making processes at earlier stages of drug development. Additionally, prevalent models often rely on hand-crafted features that demonstrate poor generalizability when applied to textual data, as evidenced by research conducted by Feijoo et al. (2020), Fu et al. (2022), and Kavalci and Hartshorn (2023). This limitation underscores the need for more robust feature engineering techniques to effectively capture the nuances inherent in clinical trial descriptions, thereby enhancing predictive accuracy and reliability. Furthermore, existing models often showcse a narrow scope, focusing exclusively on predictions for specific diseases and trial phases, as observed in studies such as Aliper et al. (2023) and Feijoo et al. (2020). This restricted applicability limits the model's versatility and hampers its potential to address broader challenges within the clinical trial domain. Lastly, certain models exclusively apply to molecular drugs with publicly available chemical structures, as highlighted by Fu et al. (2022). This constraint excludes a significant portion of drug candidates from consideration, thereby diminishing the model's overall applicability and relevance within the pharmaceutical landscape. Given these limitations, there is a pressing need to create trial outcome prediction models that are more comprehensive and adaptable, capable of overcoming these constraints and providing improved predictive capabilities across a wider range of drug development scenarios.

A common shortcoming of all aforementioned methods is that they do not predict the true phase transition of a clinical trial, which, to our knowledge, has so far only been attempted by Feijoo et al. (2020). Although the completion status of a CT is publicly available, this information does not indicate if an intervention for a specific disease will progress to the next trial phase. Pharmaceutical companies may decide not to advance a drug to the next trials for reasons unrelated to its efficacy or safety. Factors such as changes in market dynamics, competition, manufacturing challenges, reassessed development priorities, financial considerations, or the need for additional preclinical or clinical data may influence this decision (Friedman et al., 2015; Stallard et al., 2005). The actual phase transition can solely be learned by following the combination of drug and indication through several trials, which is only possible with access to proprietary data. Similar to our approach, they combine the data from *ClinicalTrials.gov* with the proprietary *Biomedtracker* database, allowing to monitor a drug over several clinical trial phases and discover trial terminations that cannot be caught through the official database. They furthermore propose a simple approach to rate the complexity of eligibility criteria, and their subsequent implementation of a random forest (RF) classifier achieves an average accuracy of 80% for specific diseases. Despite these results, their method is of limited use, as their classifier is dependent on data only available after trial completion (end date, trial duration, etc.) as well as on hand-crafted features, which is prone to introduce a human bias to the predicted outcome. Our approach offers a more direct and proactive solution by predicting phase transitions directly from the trial protocol itself, eliminating the need to wait for post-trial data availability and providing timely insights into the potential progression of clinical trials.

2.2 Large Language Models in the Clinical Trial Domain

The integration of Large Language Models (LLMs) within the clinical trial domain has been relatively limited, with only a select few methodologies thus far proposed. Among these, the CliniDigest model, introduced by White et al. (2023), represents a notable advancement, leveraging the GPT-3.5 architecture to condense extensive clinical trial descriptions, often spanning several thousand words, into succinct 200-word summaries. Similarly, Zheng et al. (2024) introduced an innovative approach to outcome prediction, using an LLM to translate multimodal clinical trial data into comprehensive natural language representations, subsequently classified by a shallow learning approach. Even though their approach and research goal aim to predict the outcome of clinical trials, they do not employ the LLM for this task. The LLM serves as pre-step in their model pipeline and only converts statistical trial information to textual data. The subsequent outcome prediction is performed by a Mixture-of-Experts model Zheng et al. (2024). Furthermore, Jin et al. (2023) proposed TrialGPT, a novel LLM architecture aimed at streamlining the matching process between free-text patient notes and clinical trial eligibility criteria, thereby enhancing the efficiency of participant selection procedures. Despite these notable strides, a comprehensive and robust approach leveraging LLMs for clinical trial outcome prediction remains an area ripe for exploration and development. As such, there exists significant potential for future research in investigating the capabilities and applications of LLMs within the realm of clinical trials, paving the way for enhanced predictive modeling and decision-making processes in the field.

3 Method

Our method aims to generate a model, represented by a function f, that predicts the target trial outcome $y_p \in \mathbb{R}^1$ based on the input x_D. Here, x_D denotes the trial textual description, a concatenation of several data elements. The approach consists of two stages: **dataset creation** and **model training**.

- Dataset creation refers to compiling an accurate CTOP dataset, providing x_D, that includes trial protocol data labeled with phase transition information inferred from drug performance data.

- In model alignment, both our models (BERT+RF and ChatCTP) are trained to predict the target phase transition according to a given input trial description, as described above.

It is important to highlight that the trial description is solely derived from textual data obtained from the publicly accessible *ClinicalTrials.gov* database. All used texts are created at the early stages of the trial design process, ensuring that our models can effectively predict the outcome of a trial before FDA approval is sought.

3.1 Model Overview

We present two models for the phase transition prediction task that differ in architecture, performance, and complexity. We have chosen both an LLM approach, as it represents the state-of-the-art in NLP, and a combined deep and shallow learning approach, chosen due to its accessibility and adaptability to a wide range of computing environments. The decision to incorporate the LLM approach stems from its capabilities to capture intricate linguistic nuances and contextual dependencies, making it an indispensable tool for tasks requiring a sophisticated understanding of complex natural language. In contrast, the combined deep and shallow learning approach offers practical advantages, such as ease of implementation and scalability, making it an attractive option for scenarios where computational resources are limited or when rapid prototyping and experimentation are paramount. By focusing on the strengths of both approaches, we aim to develop a robust predictive framework capable of delivering accurate and reliable phase transition predictions across diverse clinical trial datasets and computational settings. These models are:

BERT+RF BERT, or Bidirectional Encoder Representations from Transformers, represents a groundbreaking advancement in natural language processing (NLP) due to its distinctive bidirectional text processing capabilities. Unlike traditional transformer models, which typically process text in a unidirectional manner, BERT has the remarkable ability to simultaneously consider both the left and right context within a sentence during training. This bidirectional approach empowers BERT to develop a more comprehensive understanding of word meaning and context, enabling it to capture rich and nuanced semantic information embedded within textual data. By incorporating bidirectional processing, BERT transcends the limitations of previous NLP models, which often struggled to fully grasp the intricacies of language due to their unidirectional nature. Through its holistic examination of text, BERT effectively captures the interdependencies between words and phrases, discerning subtle nuances in meaning and context that might otherwise be overlooked. This enhanced contextual awareness enables BERT to generate more accurate representations of text, facilitating tasks such as language understanding, sentiment analysis, and information retrieval with unprece-

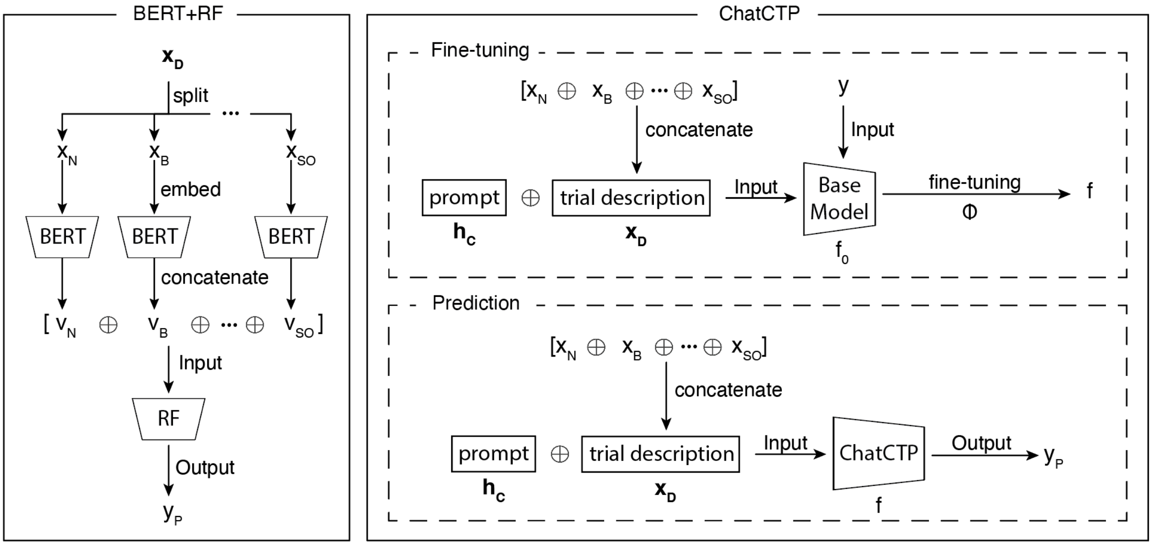

Figure 2: Overview of the two models. On the left is the BERT+RF approach, where the trial textual description x_D is divided into its entries, individually embedded by the clinical BERT, concatenated, and then inputted into the RF classifier. On the right are the two steps of the ChatCTP approach. First, the instruction fine-tuning of the base model f, using trial description x_D, the continuous prompt h_C, and the labels y as inputs to the fine-tuning function Φ, resulting in the fine-tuned model, ChatCTP (f). For an example of the continuous prompt, refer to Table 1. Below is the inference process, which only relies on the prompt and trial descriptions.

dented precision.(Devlin et al., 2018). One notable limitation for our specific use case is the restricted attention window size of BERT models, which typically permits the processing of text containing up to 512 tokens. This constraint poses a challenge when dealing with lengthy clinical trial descriptions that exceed the specified token limit, potentially resulting in the truncation or omission of crucial information during model training and inference. As a result, there is a risk of information loss or oversimplification, particularly in scenarios where detailed descriptions are essential for accurate prediction and analysis. This limitation underscores the importance of devising strategies to effectively manage and preprocess lengthy textual data to ensure compatibility with BERT-based models while preserving the integrity and informativeness of the original text. (Devlin et al., 2018; Dai et al., 2022). To overcome this limitation, we employ a hybrid approach by combining a clinical BERT embedding with Random Forest (RF) classification. Each trial's different entry categories are embedded separately using sentence transformers (Reimers and Gurevych, 2019), resulting in numerical representations. These representations are then concatenated into an 8,488-dimensional feature vector on which we train an RF classifier (see Section 3.3.1). This hybrid method enhances predictive accuracy by aggregating predictions from multiple decision trees. In our experiments, the BERT+RF model exhibited promising results, achieving an accuracy of 0.726 in predicting phase transitions across various clinical trials, a performance comparable to the LLM approach (see Table 2). Its prediction for specific phases outperforms even the LLM model, with an F1 score of 0.847 when trained to only predict Phase III transitions (see Table 4). The encoding of all 20.000 trial texts and training of the RF model takes approximately 20 minutes on a single RTX 4090 GPU, with instantaneous inference. This approach can be considered a cost- and time-efficient alternative to the LLM.

ChatCTP An issue of processing clinical trial descriptions is the specificity and inconsistency of the used vocabulary. Clinical trial documentation often encompasses a diverse array of terminology, reflecting the multifaceted nature of medical research and practice. This diversity presents a considerable challenge for NLP systems, which must contend with the vast spectrum of medical terminology, ranging from highly specialized technical terms to colloquial expressions. Moreover, inconsistencies in vocabulary usage across different trials further compound this challenge, as variations in terminology usage can lead to ambiguity and difficulty in interpreting textual data. Consequently, addressing this issue requires developing robust NLP techniques capable of effectively handling the nuanced vocabulary present in clinical trial descriptions, ensuring accurate and reliable analysis of trial data for predictive modeling and decision-making purposes. It is, therefore, preferred to fine-tune a model that is already familiar with medical literature. While multiple dedicated medical large language models are available to researchers, Zhou et al. (2024) demonstrated that GPT-3.5 Turbo exhibits robust performance on medical downstream tasks. The model was instruction fine-tuned for three epochs with the same continuous prompt (see Table 1) for all data points. We determined that a balanced set of 4000 samples (around 3 Mio tokens) is adequate to achieve a significant improvement in performance over the baselines while maintaining low costs. The fine-tuned model, ChatCTP, outperforms the baseline by 0.042 on the F1 score for all-phase outcome prediction (see Table 2), while further showcasing rudimentary reasoning abilities (see Section 4.2). The training took around 80 minutes through the dedicated API with associated costs of about $70.

3.2 Dataset creation

To ensure the creation of a robust dataset capable of effectively capturing the phase transitions of clinical trials, we must carefully address two crucial components: **obtaining comprehensive clinical trial protocols** and **establishing connections between medical interventions across multiple trials**. These aspects serve as the cornerstone of our dataset construction process, providing the necessary foundation for accurate prediction of trial outcomes. To achieve this, we gather data from two sources: **ClinicalTrials.gov** and **Biomedtracker**. This meticulous approach enables us to compile a comprehensive dataset that accurately reflects the dynamics of clinical trial progression, laying the groundwork for insightful analysis and prediction.

- **ClinicalTrials.gov** (http://clinicaltrials.gov/) is a publicly accessible repository maintained by the United States National Library of Medicine, offering comprehensive data on clinical studies worldwide. It is a vital resource for researchers, clinicians, and the broader medical community, offering a wealth of information on clinical studies conducted globally. With its user-friendly interface and extensive database, ClinicalTrials.gov facilitates access to valuable data on a wide range of medical interventions, including drugs, devices, procedures, and behavioral therapies. Researchers can explore detailed study records, including information on study design, participant eligibility criteria, intervention details, and outcome measures. Moreover, the platform provides transparency and accountability by requiring trial sponsors to register their studies and report key findings, contributing to the integrity and reliability of clinical research. Overall, ClinicalTrials.gov plays a pivotal role in advancing medical knowledge, fostering collaboration, and promoting evidence-based decision-making in healthcare. Presently, it houses 481,198 study records from 223 countries, making it the largest database of its kind globally.

- **Biomedtracker** (https://www.biomedtracker.com/), is a proprietary database compiled by Informa Business Intelligence Inc. It is a comprehensive resource that tracks and analyzes pharmaceutical and biotechnology industry developments, including clinical trials, regulatory milestones, drug approvals, and market trends. It is considered an indispensable tool for stakeholders across the pharmaceutical and biotechnology sectors, offering unparalleled insights into industry dynamics and trends. By tracking clinical trials, regulatory approvals, and market developments, Biomedtracker enables researchers, investors, and decision-makers to make informed decisions and effectively follow the development trajectory of novel medical interventions. Its user-friendly interface and comprehensive dataset provide users with access to critical information on drug development programs, enabling them to assess risk, identify opportunities, and navigate complex regulatory processes with confidence. With its extensive coverage and real-time updates, Biomedtracker serves as a trusted resource for industry professionals seeking to drive innovation and improve patient outcomes. In summary, it provides accurate insight into the progress of drug development programs and enables us to track a treatment's performance through multiple clinical studies. The version we used contains information on 20,016 unique drugs.

By merging information from Biomedtracker and ClinicalTrials.gov based on a common National Clinical Trial Identifier (NCT-ID) and excluding low-quality trials, we obtained an initial dataset comprising 25,000 entries. This merged dataset offers a comprehensive overview of clinical trials, combining detailed insights from Biomedtracker with the extensive study records available on ClinicalTrials.gov. By leveraging a common identifier and implementing stringent quality criteria, we ensured the accuracy and reliability of the dataset, laying a solid foundation for subsequent analysis and modeling efforts.

3.2.1 Labelling process

As previously noted, while all selected trials have known outcomes (completed or terminated), relying solely on this as an indicator of success or failure oversimplifies the intricate drug development process. Numerous external factors significantly influence a drug's trajectory, necessitating a more nuanced approach to evaluation. For instance, even if a Phase II trial is completed, it may not progress to Phase III due to various market considerations or evolving research needs. Conversely, terminating a Phase II trial does not necessarily denote failure; the collected data could potentially support concurrent Phase II trials testing the same intervention, ultimately propelling it towards becoming a market-ready product. Hence, it becomes evident that the completion status alone is insufficient for accurately labeling a trial. A more comprehensive assessment of trial outcomes requires consideration of the broader context, including external influences such as regulatory changes, competitive landscape shifts, and emerging clinical evidence. By adopting such an approach, stakeholders can gain deeper insights into the factors driving trial outcomes and make more informed decisions regarding drug development strategies.

A clinical trial is a rigorous examination of a medical treatment's efficacy in addressing a specific medical condition. This critical information is encapsulated within the Drug-Indication ID, providing a foundational basis for trial categorization and analysis. Leveraging the comprehensive database provided by Biomedtracker, we are equipped to establish meaningful connections between individual trials using this indicator. This strategic approach enables us to label each trial according to four distinct rules, facilitating a systematic and comprehensive assessment of their respective outcomes and contributions to medical research and innovation.

1. **Successful Phase Transition**: If, according to Biomedtracker, a drug advances to a certain ultimate phase, all trials in preceding phases featuring the same Drug-Indication ID are deemed successful.

2. **Incomplete Phase Information**: If, according to Biomedtracker, a drug advances to a certain ultimate phase, all trials of this phase featuring the same Drug-Indication ID are considered unsuccessful.

3. **Termination Status and Trial Success**: Trials labeled as terminated on ClinicalTrials.gov, even if considered successful, are considered unsuccessful.

4. **Completion Status and Trial Success**: Trials labeled as unsuccessful remain so, even if labeled as completed on ClinicalTrials.gov.

See Figure 1. While the third rule appears to contradict our earlier assertion regarding the potential for terminated trials to contribute to a market-ready product, we have opted not to categorize these trials as successful for two compelling reasons. Firstly, from a medical standpoint, the true impact of terminated trials on the progression of a specific drug remains uncertain and requires further investigation. Secondly, from a technical perspective, overlooking Rule 3 would lead to all trials being assigned the same label up to the final phase, potentially introducing bias into the model. Therefore, we have adopted a more cautious and restrictive approach in assigning the success label, prioritizing accuracy and integrity in our classification methodology.

3.2.2 Synthesis

Upon successfully assigning each trial an accurate phase transition label, identifying relevant trial information is the next step. Recognizing that excessively long texts can impede the efficacy of Natural Language Processing (NLP) models and considering the comprehensive nature of trial protocols that may encompass up to a hundred different entry categories (sponsors, locations, recruitment criteria, contact information, etc.), the necessity of crafting a concise yet informative trial description becomes increasingly apparent. Through comprehensive analysis, we have determined that certain entry

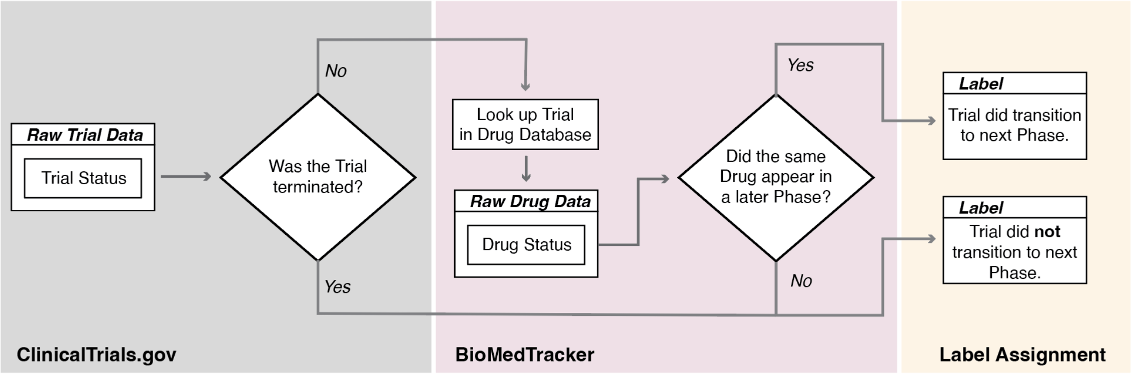
No
Raw Trial Data
Trial Status
Was the Trial terminated?
Yes
Look up Trial in Drug Database
Raw Drug Data
Drug Status
Yes
Did the same Drug appear in a later Phase?
No
Label
Trial did transition to next Phase.
Label
Trial did **not** transition to next Phase.
ClinicalTrials.gov
BioMedTracker
Label Assignment

Figure 3: Overview of the labeling method.

categories yield optimal performance while ensuring that their combined length remains below 4096 tokens. This strategic selection process ensures that the resulting trial descriptions strike a balance between informativeness and computational efficiency, thereby enhancing the overall effectiveness of our predictive models. Additionally, by focusing on key entry categories, we aim to provide our models with the most relevant and impactful information, enabling them to generate more accurate predictions regarding trial outcomes. Moreover, this approach facilitates a more streamlined and efficient workflow, minimizing computational resources while maximizing predictive performance.

Trial Name (x_N), Trial Brief (x_B), Drug Used (x_{DU}), Drug Class (x_{DC}), Indication (x_I), Target (x_T), Therapy (x_{Th}), Lead Sponsor (x_S), Criteria (x_C), Primary Outcome (x_{PO}), and Secondary Outcome (x_{SO}). By concatenating these entries, we create the final trial description x_D according to

$$x_D = (x_N \oplus x_B ... \oplus x_{SO}). \tag{1}$$

Let $\mathcal{D}$ be the resulting PhaseTransition dataset, where each row consists of the input x_D, with its corresponding phase transition label y. We can represent $\mathcal{D}$ as a set of ordered pairs as

$$\mathcal{D} = \{(x_{D1}, y_1), (x_{D2}, y_2), ..., (x_{Dn}, y_n)\}, \tag{2}$$

with each ordered pair $\mathcal{D}_i = (x_{Di}, y_i)$ representing a data point. An example of a data point can be seen in Table 5.

3.3 Model training

After introducing our two models in Section 3.1 and describing the creation of the PhaseTransition Dataset in Section 3.2, in this section, we now detail how we trained each model on the data.

3.3.1 BERT+RF

Let $x_N = (x_{N_1}, x_{N_2}, ..., x_{N_n})$ be the name of the clinical trial, where x_i represents the ith token in the text. We compute the embedding of the trial name by using a clinical BERT model from the sentence transformer library as

$$v_N = BERT(x_N). \tag{3}$$

With $v_N = (v_{N_1}, v_{N_2}, ..., v_{N_{768}})$, since the specific clinical BERT model we used produces embeddings of size 768. The embedding process is repeated for each data element in x_D, with the resulting embedding vectors being concatenated similarly to Equation 1 as

$$v_D = (v_N \oplus v_B \oplus ... \oplus v_{SO}). \tag{4}$$

Thus, $v_D = (v_{D_1}, v_{D_2}, ..., v_{D_{8,448}})$ be the embedded feature vector of x_D. By representing the associated label as a binary numerical value y_B according to

$$y_B = \begin{cases} 1, & \text{if label } y = \text{"Yes"} \\ 0, & \text{if label } y = \text{"No"} \end{cases}, \tag{5}$$

we can rewrite the dataset used for the RF classifier as

$$\mathcal{D}_{RF} = \{(v_{D1}, y_{B1}), ..., (v_{Dn}, y_{Bn})\}. \tag{6}$$

Following the RF algorithm, we randomly select m data points with replacement from the dataset $\mathcal{D}_{RF}$ to create $B = 100$ bootstrap samples $\mathcal{D}_b^*$, where $b = 1, 2, ..., B$. For each bootstrap sample $\mathcal{D}_b^*$ a decision tree T_b is grown from a random subset of features at each split until all leaves are pure, with Gini impurity being the splitting criterion. For a node t, if p_i represents the proportion of samples of class i in node t, then the Gini impurity $G(t)$ is defined as

$$G(t) = 1 - \sum_{i=1}^{C} p_i^2, \tag{7}$$

where $C = 2$ being the number of classes. Finally, the predictions of all decision trees are aggregated using the majority voting aggregation function Agg, and the predicted phase transition label y_p is calculated by

$$y_p = \text{Agg}(\{T_1, T_2, ..., T_B\}, v_D). \tag{8}$$

3.3.2 ChatCTP

In contrast to the BERT+RF model, which we train from scratch, ChatCTP is created by instruction fine-tuning GPT-3.5 Turbo on 4000 random samples from $\mathcal{D}$. Furthermore, we introduce the concept of a continuous prompt h_C (see Table 1), which serves as the model instructional component concatenated with x_D. The fine-tuning step is defined as

$$f(h_C, x_D) = (\Phi \circ f_0)(h_C, x_D, y), \tag{9}$$

whereby f_0 represents the base model, Φ denotes the fine-tuning operation, while f being the final phase transition prediction model. Thus, phase transition predictions are inferred by

$$y_p = f(h_C, x_D). \tag{10}$$

4 Experimental Results

To comprehensively evaluate the performance of our two models, we conducted cross-testing of the dataset on two related architectures: Longformer and LLama. This rigorous assessment allowed us to gain deeper insights into the comparative strengths and weaknesses of our models in relation to alternative approaches. By subjecting our models to rigorous scrutiny against these benchmarks, we aimed to provide a thorough and robust evaluation of their predictive capabilities. This holistic approach ensures a nuanced understanding of their performance across different architectural frameworks, ultimately enhancing the reliability and applicability of our findings in real-world settings.

4.0.1 Longformer

The Longformer model emerges as a transformative adaptation of the ubiquitous Transformer architecture, meticulously engineered to confront the formidable challenges inherent in processing extensive textual sequences. Departing from the constraints of conventional models like BERT, which are encumbered by rigid attention window sizes, the Longformer introduces a revolutionary global attention mechanism. This dynamic feature endows the Longformer with the unparalleled capability to discern intricate dependencies spanning the entirety of input sequences, regardless of their expansive length (Beltagy et al., 2020). By transcending traditional limitations, the Longformer becomes uniquely equipped to unravel intricate relationships between tokens dispersed across vast expanses of textual data (Beltagy et al., 2020), making it an ideal candidate for our specific task. Our exhaustive efforts entailed the meticulous training of a specialized clinical iteration of the Longformer (Li et al., 2022) over a span of seven epochs. This arduous endeavor consumed approximately 2.5 hours of computational resources on a singular RTX 4090 GPU. Despite the substantial investment of time and computational power, the model yielded a modest accuracy of 0.668, positioning it as the least performing among our cohort of models (refer to Table 2).

4.0.2 Llama 2

Llama 2 by Meta in 2023 marked a significant milestone, offering variants with 7B, 13B, and 70B parameters (Touvron et al., 2023). For our investigation, we deliberately chose to use the 7B model to mitigate memory overhead and expedite training duration. However, our initial attempts at fine-tuning this base model on the complete dataset for a duration of five epochs yielded suboptimal outcomes, as elucidated in Table 2. This outcome led us to an insightful realization: the inherent structure of the base model lacks the requisite medical domain understanding. Consequently, we ventured into an exploration of the 7B version of AlpaCare, a refined iteration of Llama 2 that underwent self-instruction on medical queries (Zhang et al., 2023).

System Prompt
You are a medical expert who specializes in analyzing clinical trials. Your role is to help the user predict whether a clinical trial will progress to the next phase. Answer only with 'Yes' if it progresses to the next phase or 'No' if it doesn't.

Table 1: LLM system prompt used for fine-tuning and inference.

Remarkably, this strategic adjustment heralded a noteworthy upswing in performance metrics, culminating in a prediction accuracy of 0.713 (as illustrated in Table 2). However, upon closer scrutiny of its phase-specific predictive prowess, it became evident that its capabilities lagged significantly behind those exhibited by our meticulously crafted models (refer to Table 3). The process of fine-tuning the model on the entire corpus of 20,000 samples entailed a considerable investment of computational resources, necessitating 18 hours of computation time on a single A1000 GPU. This laborious undertaking underscores the formidable computational demands associated with model refinement. Additionally, the prospect of fine-tuning either the 13B or 70B versions, while potentially promising superior results, remains largely unattainable due to logistical constraints and computational feasibility considerations.

Model	Accuracy	F1 Score
BERT + RF	0.726	0.682
Longformer	0.668	0.675
LLaMA 2 7B	0.567	0.535
LLaMA 2 7B (AlpaCare)	0.713	0.695
GPT 3.5 fine-tuned	**0.728**	**0.737**

Table 2: Comparison of model performances trained on trials from all phases.

4.1 Ablation Study

Here, we conduct an extensive ablation study to showcase the efficacy of our approach. The common path in CTOP is creating separate models for each phase, trained only on trials from the designated phase. This robs the models of any knowledge about previous trials and treatment performance. Our approach differs as we do not separate the training data into phases. However, to demonstrate the superiority of GPT-3.5 Turbo over all baselines, we have also trained each model on single-phase data, using only Phase III trials (see Table 4). Even though BERT+RF's F1 score is better by 0.019 points, it should be noted that GPT-3.5 Turbo was fine-tuned on only 35% of the data, while all other models were trained on the complete set of Phase III trials.

Compared to ChatCTP (see Table 3), which was trained on 4000 trials from all three phases, the dedicated Phase III version of GPT-3.5 Turbo only outperforms it by 0.025 points on the F1 score. Of the 4000 trials, 1120 belonged to Phase III. In comparison, the dedicated Phase III model was trained on 2000 trials. A 44% increase in training data. However, if we fine-tune GPT-3.5 Turbo on only the 1120 Phase III trials, the F1 score drops to 0.760. From this, we can conclude that the remaining 2880 Phase I and Phase II in ChatCTP's training data hold valuable information on predicting the transition from Phase III to Approval. Therefore, any outcome prediction model should be trained across phases and not be constrained to a single phase.

4.2 Reasoning Experiment

Using an LLM to predict a trial phase transition gives us the novel possibility of inquiring into the model's decision-making process and asking about the pivotal factors in the trial description. As there is no way to fact-check the model's answer, we trained a second version of ChatCTP on a modified dataset. This dataset included trials from ClinicalTrials.gov that provided explanations for trial termination, ranging from detailed descriptions to general reasons such as 'Strategic Decisions.' Despite the variability in information quality, we concatenated these explanations with the 'No' label. During the fine-tuning process, we instructed the model that whenever it predicts a trial will not transition to the next phase, it should provide an explanation of why. Since information on why a particular trial succeeded is typically unavailable, the labeling and prediction process for positive cases ('Yes' label) remains unchanged.

During evaluation, it became evident that the

	BERT + RF		LLaMA 2 7B (AlpaCare)		ChatCTP	
Phase	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**
Phase I	0.713	0.631	0.572	0.572	0.659	0.658
Phase II	0.675	0.646	0.588	0.586	0.618	0.619
Phase III	0.737	0.778	0.739	0.723	0.812	0.803

Table 3: Prediction accuracies per phase.

Model	Accuracy	F1 Score
BERT + RF	0.815	**0.847**
Longformer	0.788	0.780
Clinical Longformer	0.780	0.790
LLaMA 2 7B	0.567	0.535
LLaMA 2 7B (AlpaCare)	0.729	0.735
GPT-3.5 Turbo	**0.825**	0.828

Table 4: Comparison of model performances trained only on Phase III trials.

fine-tuned reasoning model was biased towards attributing trial termination to safety and efficacy concerns. However, there are also some cases where the model exhibits interesting reasoning abilities. Examples are given in Table 6. It is important to note that these findings lack validation and should be approached with caution. Nonetheless, they present an intriguing avenue for future research. Further investigation into the model's reasoning mechanisms and validation of its outputs could shed more light on its reliability and potential applications in clinical trial analysis. Additionally, exploring how the model's reasoning aligns with real-world clinical insights could enhance its practical utility and trustworthiness.

5 Conclusion

In this comprehensive study, we embarked on a critical endeavor to address the urgent demand for accurate prediction models in the realm of clinical trial outcome prediction (CTOP). Our primary objective was to combat the prevalent high failure rates and substantial resource wastage associated with unsuccessful trials, thereby aiming to revolutionize the landscape of clinical trial research. To achieve this goal, we introduced ChatCTP, a groundbreaking initiative marking the inaugural utilization of Large Language Models (LLMs) in CTOP. Concurrently, we developed a non-LLM-based approach, BERT+RF, to offer a diverse range of methodologies for comparison and evaluation.

Central to our methodology was the innovative labeling procedure for trial outcomes, which enabled us to track medical interventions across a multitude of trials. This approach afforded us a deeper and more nuanced understanding of the intricate dynamics surrounding trial success and failure. Leveraging the curated PhaseTransition Dataset, we conducted a series of comprehensive experiments to evaluate the efficacy and performance of our proposed methodologies thoroughly.

Despite the significant strides made in advancing our understanding and predictive capabilities, we recognize the need for further investigations into the impact of terminated trials on the efficacy of medical interventions in ongoing trials. Such endeavors are imperative to enhance the accuracy and reliability of outcome labels, ultimately facilitating more informed decision-making processes.

An inherent advantage of employing Language Model-based approaches lies in the potential for enhanced model interpretability. Leveraging LLMs enables us to delve deep into the inner workings of the model, allowing for the extraction of valuable insights into its decision-making processes. This, in turn, offers the tantalizing prospect of gaining a comprehensive understanding of the rationale behind the prediction of trial outcomes. While our initial forays into extracting reasoning from the model show promising results, we acknowledge the necessity for further refinement to achieve a level of interpretability that furnishes actionable insights for stakeholders involved in clinical trial decision-making.

Looking ahead, our study lays a solid foundation for future research endeavors in CTOP, offering invaluable insights into the untapped potential of LLMs and alternative methodologies in enhancing prediction accuracy. By addressing key challenges and introducing innovative methodologies and datasets, we aspire to catalyze the development of more robust and reliable prediction models in the dynamic and ever-evolving landscape of clinical trials.

Limitations

Predicting the outcome of a clinical trial is an important topic. Previous has mostly focused on using ClinicalTrials.gov, currently the largest database in the clinical trial domain, in combination with various supporting databases for specific topics. Even though ClinicalTrials.gov has more than 450,000 entries, we can only use a fraction of them. On the one hand, we are limited by the quality of the data. Sponsors and investigators are responsible for submitting their data to ClinicalTrials.gov. Ideally, the key information is registered at the start of the trial and updated regularly during its run, with the entry being closed as the study ends (Tse et al., 2018). In practice, the adherence to these requirements is inadequate, meaning that a large portion of entries lacks accurate and complete information (Tse et al., 2018). We are further limited by the second database, the BioMedTracker, as we can only use trials that are also featured in its much smaller collection. The nature of the trial report system introduces another limitation. Our approach focuses on predicting the outcome of a clinical trial from data available before the trial is started. This data is, in theory, available on ClinicalTrials.gov. However, the responsible parties can add, edit, or delete information at any time (Tse et al., 2018); therefore, not all trial descriptions we used might reflect the initial trial protocol created.

We further want to emphasize that the reasoning examples given in the experiment section are solely provided for exploratory purposes. While we tested the reasoning capabilities of the LLM we built, it is important to note that this aspect of the model was not the primary focus of our study. Further studies with this focus have to be conducted to create LLM's capable of stating reason for their predicted outcome.

Ethics Statement

Our study presents a predictive framework for clinical trial outcomes, offering insights into potential trajectories; however, it refrains from offering prescriptive advice on interpreting or acting upon these predictions. While our models showcase promising accuracy, it's crucial to acknowledge the inherent risks associated with their application. The possibility of disproportionately attributing significance to predictions looms large, given the inherent limitations of predictive modeling, thus warranting cautious utilization of our approach. At their current developmental stage, our models aren't suitable for definitively determining trial success or anticipating phase transitions. We acknowledge the potential for false positives or negatives, which could lead to misguided conclusions. Hence, we advocate for the use of our models as complementary tools alongside clinical judgment rather than sole determinants in decision-making processes.

Regarding data sourcing, our dataset relies on proprietary data from BioMedTracker, yet all information published in our study is meticulously sourced from publicly available data. While the labeling process entails the use of restricted information, stringent measures ensure that no confidential data is divulged in its raw form. This restricted data is exclusively employed for inference and labeling within our study, underscoring our commitment to upholding the confidentiality and integrity of proprietary information provided by BioMedTracker. Additionally, we adhere to ethical standards of transparency and accountability in our research methodology, thereby ensuring scientific integrity and responsible data usage in the development and dissemination of predictive models for clinical trial outcomes.

Moreover, it's essential to recognize that predictive modeling in the context of clinical trials is a rapidly evolving field. As such, our study represents just one step in a larger journey toward developing more accurate and reliable prediction models. Future research endeavors should focus on refining existing methodologies, exploring novel approaches, and incorporating additional data sources to further enhance prediction accuracy and robustness. Collaborative efforts between researchers, clinicians, and industry stakeholders will be instrumental in driving progress in this area and ultimately improving the efficiency and success rates of clinical trials.

Additionally, while our study leverages proprietary data from BioMedTracker, it's imperative to acknowledge the broader ethical considerations surrounding data usage in predictive modeling. Ensuring data privacy, transparency, and equitable access to information are paramount concerns that must be addressed to foster trust and accountability within the research community. As such, we advocate for greater transparency in data sourcing and sharing practices, along with the development of clear guidelines and ethical frameworks to govern the use of proprietary and publicly available data in predictive modeling research. By upholding these principles, we can ensure that predictive modeling continues to advance scientific understanding and benefit society while maintaining ethical integrity.

References

Alex Aliper, 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):972–980.

Artem V Artemov, Evgeny Putin, Quentin Vanhaelen, Alexander Aliper, Ivan V Ozerov, and Alex Zhavoronkov. 2016. Integrated deep learned transcriptomic and structure-based predictor of clinical trials outcomes. *BioRxiv*, page 095653.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting transformer-based models for long document classification. *arXiv preprint arXiv:2204.06683*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

JA DiMasi, JC Hermann, K Twyman, RK Kondru, S Stergiopoulos, KA Getz, and W Rackoff. 2015. A tool for predicting regulatory approval after phase ii testing of new oncology compounds. *Clinical Pharmacology & Therapeutics*, 98(5):506–513.

Felipe Feijoo, Michele Palopoli, Jen Bernstein, Sauleh Siddiqui, and Tenley E Albright. 2020. Key indicators of phase transition for clinical trials through machine learning. *Drug discovery today*, 25(2):414–421.

Sohrab Ferdowsi, Julien Knafou, Nikolay Borissov, David Vicente Alvarez, Rahul Mishra, Poorya Amini, and Douglas Teodoro. 2023. Deep learning-based risk prediction for interventional clinical trials based on protocol design: A retrospective study. *Patterns*, 4(3).

Lendie Follett, Simon Geletta, and Marcia Laugerman. 2019. Quantifying risk associated with clinical trial termination: A text mining approach. *Information Processing & Management*, 56(3):516–525.

Lawrence M Friedman, Curt D Furberg, David L DeMets, David M Reboussin, and Christopher B Granger. 2015. *Fundamentals of clinical trials*. Springer.

Tianfan Fu, Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. 2022. Hint: Hierarchical interaction network for clinical-trial-outcome predictions. *Patterns*, 3(4).

Kaitlyn M Gayvert, 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.

Qiao Jin, Zifeng Wang, Charalampos S Floudas, Jimeng Sun, and Zhiyong Lu. 2023. Matching patients to clinical trials with large language models. *ArXiv*.

Ece Kavalci and Anthony Hartshorn. 2023. Improving clinical trial design using interpretable machine learning based prediction of early trial termination. *Scientific Reports*, 13(1):121.

Jianfu Li, Qiang Wei, Omid Ghiasvand, Miao Chen, Victor Lobanov, Chunhua Weng, and Hua Xu. 2022. A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora. *BMC medical informatics and decision making*, 22(3):1–10.

Youran Qi and Qi Tang. 2019. Predicting phase 3 clinical trial results by modeling phase 2 clinical trial subject level data using deep learning. In *Machine Learning for Healthcare Conference*, pages 288–303. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Nigel Stallard, John Whitehead, and Simon Cleall. 2005. Decision-making in a phase ii clinical trial: a new approach combining bayesian and frequentist concepts. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, 4(2):119–128.

Shubo Tian, Arslan Erdengasileng, Xi Yang, Yi Guo, Yonghui Wu, Jinfeng Zhang, Jiang Bian, and Zhe He. 2021. Transformer-based named entity recognition for parsing clinical trial eligibility criteria. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–6.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tony Tse, Kevin M Fain, and Deborah A Zarin. 2018. How to avoid common problems when using clinicaltrials. gov in research: 10 issues to consider. *Bmj*, 361.

Zifeng Wang, Chufan Gao, Lucas M Glass, and Jimeng Sun. 2022. Artificial intelligence for in silico clinical trials: A review. *arXiv preprint arXiv:2209.09023*.

Zifeng Wang and Jimeng Sun. 2022. Trial2vec: Zero-shot clinical trial document similarity search using self-supervision. *arXiv preprint arXiv:2206.14719*.

Renee White, Tristan Peng, Pann Sripitak, Alexander Rosenberg Johansen, and Michael Snyder. 2023. Clinidigest: a case study in large language model based large-scale summarization of clinical trial descriptions. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, pages 396–402.

Kun Zeng, Yibin Xu, Ge Lin, Likeng Liang, and Tianyong Hao. 2021. Automated classification of clinical trial eligibility criteria text based on ensemble learning and metric learning. *BMC Medical Informatics and Decision Making*, 21(2):1–10.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2023. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*.

Wenhao Zheng, Dongsheng Peng, Hongxia Xu, Hongtu Zhu, Tianfan Fu, and Huaxiu Yao. 2024. Multimodal clinical trial outcome prediction with large language models. *arXiv preprint arXiv:2402.06512*.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Chenyu You, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A. Clifton. 2024. A survey of large language models in medicine: Progress, application, and challenge. *Preprint*, arXiv:2311.05112.

A Appendix

Even though we state that the regulatory process consists of three phases, we want to clarify that additional steps are involved in bringing a drug to market. Following the completion of Phase III trials, researchers analyze the data collected from the trials to give a final assessment of the drug (or device) safety and efficacy (Friedman et al., 2015). If the results are favorable and meet the predefined endpoints established by regulatory agencies, the sponsor submits a New Drug Application (NDA) or a Biologics License Application (BLA) to the regulatory authority for approval (Friedman et al., 2015).

Afterward, the responsible regulatory agency, commonly the U.S. Food and Drug Administration (FDA), carefully reviews the NDA or BLA to evaluate the clinical trial data, manufacturing processes, labeling, and proposed indications for use (Friedman et al., 2015). This review process ensures that the drug or device meets stringent safety and efficacy standards before being approved for market authorization. In some cases, the FDA convenes an advisory committee of independent experts to review the clinical trial data and provide recommendations regarding approving the drug or device (Friedman et al., 2015). The committee considers factors such as the risk-benefit profile, potential safety concerns, and unmet medical needs. If the data demonstrate that the benefits outweigh the risks and the product meets regulatory requirements, the agency may grant marketing approval (Friedman et al., 2015; Kavalci and Hartshorn, 2023).

Once approved, the drug or device can be launched into the market for widespread use by healthcare providers and patients. At this stage, the last phase of the clinical trial process, Phase IV, is launched, also known as post-marketing surveillance trials or post-market studies. Unlike earlier phases of clinical trials, which primarily focus on establishing safety and efficacy for regulatory approval, Phase 4 trials aim to monitor the drug or device's long-term safety profile and effectiveness in real-world settings (Friedman et al., 2015). Phase 4 trials typically involve larger and more diverse patient populations than earlier phases and may last for several years. They aim to identify rare or long-term adverse effects that may not have been detected during earlier clinical trials, as well as to gather additional data on the drug's efficacy in specific patient populations or clinical settings (Friedman et al., 2015). These trials often compare the new treatment with existing treatments or placebo to assess its relative benefits and risks in real-world conditions (Friedman et al., 2015).

The results of Phase 4 trials can lead to important updates to product labeling, changes in prescribing guidelines, or even the withdrawal of a drug or device from the market if serious safety concerns arise. Overall, Phase 4 trials are crucial in ensuring the ongoing safety and effectiveness of medical interventions once they are available to the general population.

In conclusion, the entire process, from preclinical development to the end of Phase IV, can span over ten years, with significant variability depending on the specific product and the associated regulatory requirements (Friedman et al., 2015).

Trial Description

TRIAL NAME: Phase II - X2202;
BRIEF: The purpose of this study was to determine if BVS857 is safe, tolerable and increases thigh muscle thickness in patients with spinal bulbar and muscular atrophy (SBMA).;
DRUG USED: BVS857;
DRUG CLASS: New Molecular Entity (NME);
INDICATION: Spinal Bulbar Muscular Atrophy (SBMA, Kennedy's Disease, X-linked spinal muscular atrophy type 1);
TARGET: IGF-1R (Insulin-like Growth Factor-1 Receptor);
THERAPY: Monotherapy;
LEAD SPONSOR: Novartis Pharmaceuticals;
CRITERIA: Key Inclusion Criteria: - Genetic diagnosis of SBMA with symptomatic muscle weakness - Able to complete 2-minute timed walk - Serum IGF-1 level less than or equal to 170 ng/mL Key Exclusion Criteria: - Medically treated diabetes mellitus or known history of hypoglycemia - History of Bell's palsy - Treatment with systemic steroids >10 mg/day (or equivalent dose); androgens or androgen reducing agents; systemic beta agonists; or other muscle anabolic drugs within the previous 3 months - History of cancer, other than non-melanomatous skin cancer - Retinopathy - Papilledema Other protocol-defined inclusion/exclusion criteria may apply;
PRIMARY OUTCOME: Number of Patients With Adverse Events (AEs), Serious Adverse Events (SAEs) and Deaths as a Measure of Safety and Tolerability;
SECONDARY OUTCOME: Mean Change From Baseline in Score on the Adult Myopathy Assessment Tool (AMAT) in Part B, Cohort 5

Phase Transition
No

Table 5: Example dataset entry. The field 'Trial Description' is the input the model receives. It holds the most important information on the clinical trial while not exceeding the 4096 token limit. The field 'Phase Transition' holds the associated label. In this example, the trial 'X2202' did not transition from Phase II to Phase III.

Reasoning Examples
Label: No. VTL-308 failed to meet its primary endpoint, all follow up studies were terminated. Predicted label: No. This clinical trial did not transition to the next phase. The reason is that the sponsor, Vital Therapies, Inc., discontinued the ELAD program, including the VTL-308 trial, due to lack of efficacy in previous studies.
Label: No. Strategic Predicted label: No. This trial will not transition to the next phase. The reason is that the lead sponsor, Gilead Sciences, has not shown any involvement in oncology-related clinical trials involving CYT997 or glioblastoma multiforme.

Table 6: Reasoning examples.