2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10822746

CTP-LLM: Clinical Trial Phase Transition Prediction Using Large Language Models

Michael Reinisch
Department of
Computer Science
American University
Washington DC, USA
reinisch@american.edu

Jianfeng He
Department of
Computer Science
Virginia Tech,
Falls Church, USA
jianfenghe@vt.edu

Chenxi Liao
Department of
Neuroscience
American University
Washington DC, USA
cl6070a@american.edu

Sauleh Siddiqui
Department of
Environmental Science
American University
Washington DC, USA
sauleh@american.edu

Bei Xiao
Department of
Computer Science
American University
Washington DC, USA
bxiao@american.edu

***Abstract*—New medical treatment development requires multiple phases of clinical trials. Recent literature indicates that the design of the trial protocols significantly contributes to trial performance. We investigated Clinical Trial Outcome Prediction (CTOP) using trial design documents to predict trial phase transitions automatically. We propose CTP-LLM, the first Large Language Model (LLM) based model for CTOP. We also introduce the PhaseTransition (PT) Dataset; which labels trials based on their progression through the regulatory process and serves as a benchmark for CTOP evaluation. Our fine-tuned GPT-3.5-based model (CTP-LLM) predicts clinical trial phase transition by analyzing the trial's original protocol texts without requiring human-selected features. CTP-LLM achieves a 67% accuracy rate in predicting trial phase transitions across all phases and a 75% accuracy rate specifically in predicting the transition from Phase III to final approval. Our experimental performance highlights the potential of LLM-powered applications in forecasting clinical trial outcomes and assessing trial design.**

***Index Terms*—clinical trial outcome prediction; deep learning; text mining; large language models.**

I. Introduction

A clinical trial systematically evaluates the safety and efficacy of medical interventions on human subjects, categorized into Phases I, II, and III, and the final FDA approval. High drug attrition in the regulatory process is well-documented in the literature [1]. This paper focuses on predicting the Clinical Trial Phase Transition from Trial Protocols, an early forecast for clinical trial success, which assists trial designers in making more informed decisions and allocating resources efficiently.

All clinical trial phases follow a textual protocol outlining the study objectives, treatment plan, participant recruitment criteria, and procedures to be followed during the study. Although the primary objective of a clinical trial is to assess a drug's efficacy, effectiveness, and safety, qualitative evidence from stakeholders has shown that protocol design is critical for a drug's successful progression through the regulatory phases of the trials [2]. Research suggests that trial protocol complexity, the type of company performing the trial, and barriers erected by regulatory agencies contribute to trial failure [3]. Our paper proposes two language models to predict clinical trials from one phase to the next based on protocol-time known variables (see Figure 1), an under-explored area [4].

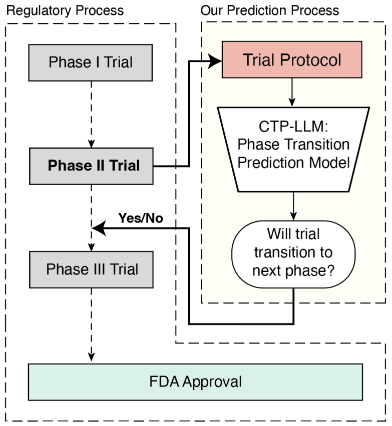

Fig. 1. Overview of our framework, CTP-LLM, for clinical trial phase transition prediction. A trial protocol is a comprehensive document that outlines the plan for conducting the trial. A treatment is typically tested in three phases, starting with safety evaluation and dosage in Phase I with a small group of people, then assessing efficacy in Phase II with a larger group, and finally confirming efficacy and safety in Phase III with a large population before FDA approval. However, the treatment can drop out in any phase. Our model takes a protocol of a given phase as input and predicts whether it can successfully transition to the next phase before the trial starts.

Significant challenges exist in predicting the clinical trial phase transition automatically from protocol design documents as inputs. The first challenge is the need for precise phase transition labeling in existing datasets. Previous studies have relied on trial recruitment status (e.g., "Completed" or "Terminated") as a proxy for success [5]–[8]. It only indicates if the trial was completed as scheduled but does not predict if it successfully transitioned to the next phase. A trial can have been completed as planned while proving that the test treatment has no efficacy. Similarly, a trial could have been terminated because the tested treatment showed promising effects earlier than expected. Second, clinical trial success is influenced by complex variables [4], [5], making it difficult for experts to identify the most impactful variables across diverse medical fields. Therefore, a data-driven approach au-

This work was done before Jianfeng He joined Amazon.

979-8-3503-8622-6/24/$31.00 ©2024 IEEE

tomatically identifying relevant features from trial documents is advantageous for creating a successful prediction and can be better evaluated later by experts.

To tackle these challenges, we propose two models for predicting trial phase transition success or failure by directly using protocol documents as input, eliminating the need for human-selected features and inference (see Figure 3 and section III-C for details). Our first model, CTP-LLM, is a specialized version of GPT-3.5 Turbo, trained on our PhaseTransition (PT) Dataset. Our second model, BERT+RF, combines a clinical Bidirectional Encoder Representations from Transformers (BERT) with a Random Forest (RF). BERT+RF offers the advantage of low computational cost, while CTP-LLM achieves higher performances overall. Our proposed models show robust accuracy in predicting phase transitions. The CTP-LLM model achieves an F1 score of 0.67 when integrating trial information across all phases. When specifically trained to predict transitions to Phase III - the most costly phase in the regulatory process - the model demonstrates even higher accuracy with an F1 score of 0.75.

Our contributions are as follows:

- **Establish the framework of using language models in Automatic CTOP.** We are the first to leverage the capabilities of LLMs for CTOP by predicting clinical trial phase transitions as an indication of trial outcome.
- **Build a new open-source dataset for CTOP**. We introduce the PT Dataset, a new resource specifically designed for clinical trial phase transition prediction. The dataset links trial protocols with information on whether the trial advanced to the next phase.
- **A new benchmark for CTOP evaluation.** We present an improved task for CTOP and experimentally demonstrate that our proposed methods accurately predict clinical trial phase transitions.

II. Related Work

The utilization of LLMs in the clinical trial domain is an emerging topic, with only a few published papers focusing on different aspects of the domain [9]–[11]. Notably, [12] links the outcome of the trial to drug toxicity and side effects, [13]–[15]) try to improve trial design through simulation, [6] quantify the risk of trial termination through text mining, while [16] leverage deep learning for CTOP by analyzing pharmacokinetic concentrations and connecting them to patient characteristics. However, the effectiveness of these existing CTOP methods is affected by similar limitations, such as training on data that only becomes available during or after the trial [4], [5], relying on human-selected features that do not generalize well to textual data [4], [7], [17], focusing only on predictions for specific diseases and phases [4], [18], or being only applicable to molecular drugs with a publicly available chemical structure [7], [16]. Our approach applies to trials of all phases and treatment types.

Moreover, the majority of the approaches mentioned above share a commonality in labeling trials based on their completion or termination status, which, in reality, does not serve as a reliable indicator of trial success. Ferdowsi et al. proposed a risk assignment methodology based on historical clinical trial statistics as well as termination status to label the trial protocols [5]. In contrast, we label the trials based on trial phase transitions with an additional dataset that tracks the phase progression of the trials. So far, only [4] have attempted to use phase transitions to indicate success for CTOP. However, their model relies on human-selected features. Our approach predicts phase transitions directly from the trial protocol documents.

III. Method

We introduce two independent language models, BERT+RF and the CTP-LLM, for the clinical trial phase transition prediction task that differ in architecture, performance, and complexity (Figure 2).

A. Problem Setup

Our approach aims to develop a model, represented by a function $f \in \mathbb{Z}_2$, that predicts the target trial outcome $y_p \in \mathbb{Z}_2$, as a binary output, based on the input x_D. Here, x_D is the trial 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. Every trial that enters each phase has a unique protocol.
- In model training, both our models (BERT+RF and CTP-LLM) are trained to predict the target phase transition according to a given input trial description.

The trial description is solely derived from textual data obtained from the publicly accessible *ClinicalTrials.gov* database. All used texts are created in the trial design process before the trial starts, ensuring that our models can effectively predict the outcome of a trial before approval from the Food and Drug Administration (FDA).

B. Overview of Models

- **BERT+RF** The BERT is limited by its restricted attention window size, typically allowing text up to 512 tokens. To overcome this limitation, we employ a hybrid approach by combining clinical BERT embedding with Random Forest (RF) classification. Each trial protocol's different attributes (e.g., name, description, recruitment criteria) are embedded separately using a clinical BERT model [19], resulting in numerical representations. These representations are then concatenated into a high-dimensional feature vector on which we train an RF classifier (see Section III-D1).
- **CTP-LLM** We fine-tune GPT-3.5 Turbo by concatenating individual attributes extracted from the trial protocol, accompanied by a prompt. To make GPT-3.5 Turbo adapt to our task, we prompt the model to return a binary response, ”yes” or ”no”, indicating whether a clinical trial progresses to the next phase.

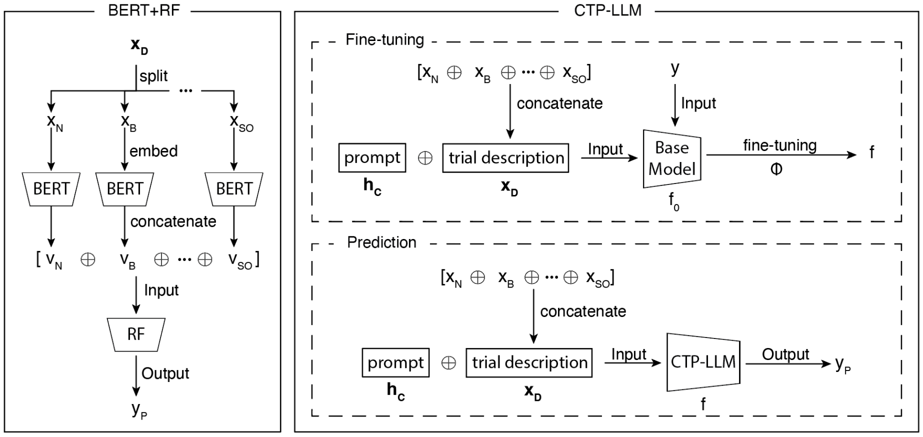

Fig. 2. Overview of models. **Left:** BERT+RF approach, where the trial textual description x_D is composed of individual attributes (e.g., $x_N, ...x_SO$, see VI), individually embedded by the clinical BERT, concatenated, and then inputted into the RF classifier. **Right:** are the two steps of the CTP-LLM approach. First, the instruction fine-tuning of the base model f, using trial description x_D, the prompt h_C, and the labels y as inputs to the fine-tuning function Φ, resulting in the fine-tuned model, CTP-LLM (f). CTP-LLM only requires the prompt and a trial description to generate a prediction.

C. PhaseTransition Dataset Construction

Two aspects are essential for constructing an accurate phase transition dataset: **gathering clinical trial protocols** and **establishing connections between medical treatments across trials**. We gather the relevant information from two distinct resources:

- **ClinicalTrials.gov** (https://clinicaltrials.gov/http://clinicaltrials.gov/) is an English-language clinical repository in the public domain maintained by the United States National Library of Medicine, offering comprehensive data on clinical studies worldwide. Presently, it houses 481,198 study records from 223 countries.
- **Biomedtracker** (https://www.biomedtracker.com/https://www.biomedtracker.com/) is an English-language proprietary database compiled by Informa Business Intelligence Inc. It is a comprehensive resource that tracks and analyzes pharmaceutical and biotechnology industry developments. Biomedtracker allows us to track a treatment's performance through multiple clinical studies. The version we used contains information on 20,016 unique drugs.

By merging trial information from Biomedtracker and ClinicalTrials.gov based on the common National Clinical Trial Identifier (NCT-ID) and excluding low-quality trials, we obtained an initial dataset comprising 20,000 entries.

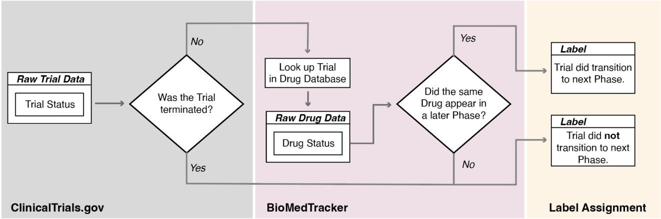
Raw Trial Data
Trial Status
Was the Trial terminated?
No
Yes
Look up Trial in Drug Database
Raw Drug Data
Drug Status
Did the same Drug appear in a later Phase?
Yes
No
Label
Trial did transition to next Phase.
Label
Trial did not transition to next Phase.
ClinicalTrials.gov
BioMedTracker
Label Assignment

Fig. 3. Overview of the labelling process as described in Section III-C.

Labeling process As discussed in I, the previous CTOP work labeled phase transition as the recruitment status, which is not accurate [4]. In this work, we employ the Drug-Indication ID from Biomedtracker to link individual trials that utilize the same medical intervention and target the same disease. This approach allows us to label phase transitions for these trials according to the following rules:

1) **Successful Phase Transition**: If a drug advances to a certain ultimate phase, all trials in preceding phases featuring the same Drug-Indication ID have performed a successful phase transition.
2) **Incomplete Phase Information**: If a drug advances to a certain ultimate phase, all trials of this phase featuring the same Drug-Indication ID have not performed a successful phase transition.
3) **Unsuccessful Phase Transition**: Trials labeled as terminated on ClinicalTrials.gov have not performed a successful phase transition, even if indicated otherwise by the Biomedtracker.

See Figure 3. While the third rule contradicts our earlier assertion that terminated trials can still result in a successful phase transition, we have opted against labeling these trials as such. Since the model is not given information on which trials are connected, Rule 3 further ensures that trial labels are not sequential, preventing model bias when training on all phases at once.

Data Synthesis After assigning each trial an accurate phase transition label, the next step is preparing the trial protocols for processing. We filtered out data that was only available after the trial had started, such as the number of participants, to prevent any potential look-ahead bias. ClinicalTrial.gov entries are of varying quality [20]. Therefore, we choose attributes of the protocols that are of high quality (e.g., present, complete, informative) throughout most trial entries, such as name, brief description, and recruitment criteria. For a list of all selected attributes and their explanations, see VIA. We create the trial description x_D by concatenating these attributes.

Let $\mathcal{D}$ be the resulting PT dataset, where each row consists of the input x_D, with its corresponding phase transition label y, either *Yes*, for a successful phase transition, or *No* for not having performed a phase transition. We can represent $\mathcal{D}$ as a set of ordered pairs as

$$\mathcal{D} = \{(x_{D1}, y_1), (x_{D2}, y_2), ..., (x_{Dn}, y_n)\}, \tag{1}$$

with each pair $\mathcal{D}_i = (x_{Di}, y_i)$ representing a data point, and n is the total number of training samples.

D. Model Training

In this section, we now detail how we trained each model.

1) BERT+RF: We compute the embedding of the trial protocols with a clinical BERT model. For each trial in the dataset, we generate the feature embeddings, $v_N, ...v_{SO}$ of the corresponding attributes, $x_N, ...x_{SO}$. For example, the feature embedding of the attribute "trial name" x_N is $v_N = BERT(x_N)$, where $v_N \in \mathbb{R}^h$ and $h = 768$, reflecting the BERTs defined embedding size. Hence, we obtain the overall embedding v_D by concatenating the embeddings from all attributes (i.e., $v_N, ...v_{SO}$). Finally, we represent the associated binary label y_B as "1" if the phase transition label is "Yes" and 0 if it is "No". Note that B could be any trial transition from I → II,II → III.

We can rewrite the dataset used for the RF classifier as

$$\mathcal{D}_{RF} = \{(v_{D1}, y_{B1}), ..., (v_{Dn}, y_{Bn})\}. \tag{2}$$

Following the RF algorithm, we randomly select m data points with replacement from the dataset $\mathcal{D}_{RF}$ to create $K = 100$ bootstrap samples $\mathcal{D}_k^*$, where $k = 1, 2, ..., K$. For each bootstrap sample $\mathcal{D}_k^*$, a decision tree T_k is grown from a random subset of features at each split until all leaves are pure, with Gini impurity being the splitting criterion. 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

$$y_p = \text{Agg}(\{T_1, T_2, ..., T_K\}, v_D). \tag{3}$$

2) CTP-LLM: In contrast to the BERT+RF model, which we train from scratch, CTP-LLM is created by instruction fine-tuning GPT-3.5 Turbo on ∼6000 random samples from $\mathcal{D}$. Furthermore, we introduce the prompt h_C, which serves as the model instructional component concatenated with x_D. The fine-tuning step is defined as

$$f(h_C, x_D) = (\Phi \circ f_0)(h_C, x_D, y), \tag{4}$$

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

$$y_p = f(h_C, x_D). \tag{5}$$

IV. Experimental Results

We first evaluate the performance of our two models on the PT dataset and compare it to the performance of two off-the-shelf transformer-based models: Longformer and Llama 2.

Model	Accuracy	F1 Score
Longformer	0.515	0.508
Clinical Longformer	0.644	0.599
LLaMA 2 7B (AlpaCare)	0.518	0.515
CTP-LLM	**0.667**	**0.665**
BERT + RF	0.626	0.617

TABLE I
Model performances trained on trials from all phases.

We balanced the PT dataset to include an equal amount of successful and failed trials. We split the PT dataset into 65% training, 15% validation, and 20% test data. To fine-tune GPT-3.5 Turbo (see Section IV-B), we have used a balanced subset of 6250 entries with the same split ratio. To avoid biasing the models, we sorted all trial entries chronologically by their last modification date. The training and validation sets include trials from 2005 to July 2022, while the test sets consist of trials dated from August 2022 onwards. The following experiments (A-E) are conducted and we report the results from a single run.

A. BERT+RF

As mentioned in Section III-C, we concatenated 11 high-quality attributes of the trial protocols, resulting in an 8,488-dimensional trial description feature vector on which we train an RF classifier (see Section III-D1). 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 F1 score of 0.617 in predicting phase transitions across various clinical trials (see Table I). Even though its performance is similar to that of the Clinical Longformer, our approach can be trained nearly seven times faster. It further excels in individual phase outcome prediction, slightly outperforming even CTP-LLM by 0.005 points in F1 score (see Table II). The encoding of all 20,000 trial texts and training of the RF model takes approximately 20 minutes on a single RTX 4090 GPU.

B. CTP-LLM

We determined that a balanced set of ∼ 6000 samples (around 18M tokens) is sufficient to achieve a significant improvement in performance over the baselines while maintaining low costs. An even larger model might perform better, but it is beyond the resources allowed for the study. The fine-tuned model, CTP-LLM, outperforms BERT+RF by 0.048 on the F1 score if trained on trials from all phases simultaneously (see Table I) and by 0.061 when fine-tuned for Phase III trials (see Table III). The training took around 150 minutes through the dedicated API with associated costs of about $100.

	BERT + RF		Clinical Longformer		CTP-LLM	
Phase Transition	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**
Phase I to II	0.642	0.580	0.586	0.386	**0.732**	**0.730**
Phase II to III	0.580	0.533	**0.610**	0.501	0.604	**0.600**
Phase III to Approval	0.665	**0.700**	0.638	0.681	**0.707**	0.695

TABLE II
We compare how well the models predict individual phase transitions when trained on data from all three phases simultaneously. CTP-LLM outperforms the other models in predicting Phase III to Approval because it effectively connects information from earlier phases. See Section IV-E.

Model	Accuracy	F1 Score
Longformer	0.686	0.691
Clinical Longformer	0.682	0.689
LLaMA 2 7B (AlpaCare)	0.624	0.625
BERT + RF	0.677	0.689
CTP-LLM	**0.751**	**0.750**

TABLE III
BERT+RF and CTP-LLM versus the baseline models when trained only on Phase III trials.

C. *Longformer*

We trained both the classical Longformer [21] ($\sim$149M parameters) and a clinical version [22] ($\sim$207M parameters) for seven epochs, with a training duration of 150-180 minutes on a single RTX 4090 GPU. While the Clinical Longformer performs comparably to BERT+RF when trained on a single phase (see Table III), it is significantly outperformed by our models when trained across multiple phases (see Table II).

D. *Llama 2 (AlpaCare)*

We conducted our experiments using AlpaCare 7B [23], a pre-fine-tuned version of LLaMA 2 that was self-instructed on medical queries. Despite its considerably larger size than the Longformer architectures, its performance is inferior to all other evaluated models (see Tables I and III). Even though better results can be expected when using the 13B and 70B versions, we decided only to utilize the 7B model due to resource limitations.

E. *Ablation Study*

Previous approaches in CTOP [4], [7], [16], [24] developed three separate models to predict the outcomes of Phase I, II, and III trials. In contrast, we train our models on trials from all three phases simultaneously. In reality, the same drug can be tested in multiple trials across all three phases at the same time, with the results of these trials influencing each other [4]. Therefore, we hypothesize that CTOP models should be trained across phases. To avoid look-ahead bias, we split the training and testing data according to protocol modification dates (see Section IV). Table II shows how well the models predict the three individual phase transitions when trained on data from all three phases.

a) Single phase training versus all phases: To test our theory, we trained all models on $\sim$6000 Phase III trial samples to investigate how well they predict the Phase III to Approval transitions (see Table III). To reduce cost, CTP-LLM was trained on only 2300 Phase III samples. Our BERT+RF model shows an improved performance of 11.67% on the F1 score compared to the version trained on all three phases.

b) Influence of previous phases: The performance of CTP-LLM in predicting Phase III to Approval is improved by 12.78% in F1 score when only trained on Phase III trial data (see Table III), compared to CTP-LLM trained on data from all three phases (see Table II). This suggests a detrimental effect of cross-phase training on performance. However, these results are not directly comparable, as the cross-phase training set contains only $\sim$1120 Phase III trials, 49% less than the dedicated Phase III training set. If we train CTP-LLM on only these exact 1120 Phase III trials, the F1 score for predicting Phase III to Approval drops by 5.33%. We conclude that the remaining **Phase I and Phase II trial data in CTP-LLM's training set holds valuable information on predicting the transition from Phase III to Approval.** The reverse scenario is also valid: trials from later regulatory phases (chronologically earlier) can influence outcome predictions for earlier phases (chronologically later). However, this bias is favorable. A drug can re-enter the regulatory process multiple times over several years or even decades. Consequently, understanding the performance of similar trials from later phases in the past is advantageous. These results indicate that an outcome prediction model based on LLMs benefits from training across multiple phases rather than being constrained to information from a single phase, whereas classification approaches based on shallow learning classifiers show enhanced performance when tailored to individual phases.

V. Conclusion

We propose a novel approach for automatically predicting clinical trial outcomes based on trial protocol descriptions. We developed a labeled PhaseTransition dataset comprising protocol details for 20,000 trials, with labels indicating successful transitions between phases. Using this dataset, we find that CTP-LLM, a fine-tuned GPT-3.5 Turbo model, achieves optimal performance when trained on data from all three phases concurrently, effectively reflecting the real-world regulatory process by integrating information across phases.

Acknowledgment

This work is supported by U.S.National Science Foundation (NO. 2050727).

References

[1] F. Pammolli, L. Magazzini, and M. Riccaboni, "The productivity crisis in pharmaceutical r&d," *Nature reviews Drug discovery*, vol. 10, no. 6, pp. 428–438, 2011.

[2] N. Vischer, C. Pfeiffer, J. Kealy, and C. Burri, "Increasing protocol suitability for clinical trials in sub-saharan africa: a mixed methods study," *Global health research and policy*, vol. 2, pp. 1–15, 2017.

[3] M. Hay, D. W. Thomas, J. L. Craighead, C. Economides, and J. Rosenthal, "Clinical development success rates for investigational drugs," *Nature biotechnology*, vol. 32, no. 1, pp. 40–51, 2014.

[4] F. Feijoo, M. Palopoli, J. Bernstein, S. Siddiqui, and T. E. Albright, "Key indicators of phase transition for clinical trials through machine learning," *Drug discovery today*, vol. 25, no. 2, pp. 414–421, 2020.

[5] S. Ferdowsi, J. Knafou, N. Borissov, D. V. Alvarez, R. Mishra, P. Amini, and D. Teodoro, "Deep learning-based risk prediction for interventional clinical trials based on protocol design: A retrospective study," *Patterns*, vol. 4, no. 3, 2023.

[6] L. Follett, S. Geletta, and M. Laugerman, "Quantifying risk associated with clinical trial termination: A text mining approach," *Information Processing & Management*, vol. 56, no. 3, pp. 516–525, 2019.

[7] T. Fu, K. Huang, C. Xiao, L. M. Glass, and J. Sun, "Hint: Hierarchical interaction network for clinical-trial-outcome predictions," *Patterns*, vol. 3, no. 4, 2022.

[8] J. Luo, Z. Qiao, L. Glass, C. Xiao, and F. Ma, "Clinicalrisk: A new therapy-related clinical trial dataset for predicting trial status and failure reasons," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 5356–5360.

[9] K. Zeng, Y. Xu, G. Lin, L. Liang, and T. Hao, "Automated classification of clinical trial eligibility criteria text based on ensemble learning and metric learning," *BMC Medical Informatics and Decision Making*, vol. 21, no. 2, pp. 1–10, 2021.

[10] R. White, T. Peng, P. Sripitak, A. Rosenberg Johansen, and M. Snyder, "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*, 2023, pp. 396–402.

[11] W. Zheng, D. Peng, H. Xu, H. Zhu, T. Fu, and H. Yao, "Multimodal clinical trial outcome prediction with large language models," *arXiv preprint arXiv:2402.06512*, 2024.

[12] K. M. Gayvert, N. S. Madhukar, and O. Elemento, "A data-driven approach to predicting successes and failures of clinical trials," *Cell chemical biology*, vol. 23, no. 10, pp. 1294–1301, 2016.

[13] Q. Jin, C. Tan, M. Chen, X. Liu, and S. Huang, "Predicting clinical trial results by implicit evidence integration," *arXiv preprint arXiv:2010.05639*, 2020.

[14] Z. Wang, C. Gao, L. M. Glass, and J. Sun, "Artificial intelligence for in silico clinical trials: A review," *arXiv preprint arXiv:2209.09023*, 2022.

[15] Z. Wang and J. Sun, "Trial2Vec: Zero-shot clinical trial document similarity search using self-supervision," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 6377–6390. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.476

[16] Y. Qi and Q. Tang, "Predicting phase 3 clinical trial results by modeling phase 2 clinical trial subject level data using deep learning," in *Machine Learning for Healthcare Conference*. PMLR, 2019, pp. 288–303.

[17] E. Kavalci and A. Hartshorn, "Improving clinical trial design using interpretable machine learning based prediction of early trial termination," *Scientific Reports*, vol. 13, no. 1, p. 121, 2023.

[18] A. Aliper, R. Kudrin, D. Polykovskiy, P. Kamya, E. Tutubalina, S. Chen, F. Ren, and A. Zhavoronkov, "Prediction of clinical trials outcomes based on target choice and clinical trial design with multi-modal artificial intelligence," *Clinical Pharmacology & Therapeutics*, vol. 114, no. 5, pp. 972–980, 2023.

[19] O. Rohanian, M. Nouriborji, H. Jauncey, S. Kouchaki, F. Nooralahzadeh, L. Clifton, L. Merson, D. A. Clifton, I. C. C. Group *et al.*, "Lightweight transformers for clinical natural language processing," *Natural Language Engineering*, pp. 1–28, 2023.

[20] T. Tse, K. M. Fain, and D. A. Zarin, "How to avoid common problems when using clinicaltrials. gov in research: 10 issues to consider," *Bmj*, vol. 361, 2018.

[21] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[22] J. Li, Q. Wei, O. Ghiasvand, M. Chen, V. Lobanov, C. Weng, and H. Xu, "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*, vol. 22, no. 3, pp. 1–10, 2022.

[23] X. Zhang, C. Tian, X. Yang, L. Chen, Z. Li, and L. R. Petzold, "Alpacare: Instruction-tuned large language models for medical application," *arXiv preprint arXiv:2310.14558*, 2023.

[24] N. Stallard, J. Whitehead, and S. Cleall, "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*, vol. 4, no. 2, pp. 119–128, 2005.

VI. Appendix

A. Definitions of attributes of the trial protocols used to create the trial description x_D.

Below are definitions of the key features used in our model.

- **Trial_Name** (x_N)**.** The official title of the clinical trial as listed on ClinicalTrials.gov. This often includes key information about the study purpose, population, phase, and intervention.
- **Brief** (x_B)**.** Refers to a shortened clinical trial description provided on ClinicalTrials.gov. It gives an overview of the study's objectives, design, and key characteristics.
- **Indication** (x_I)**.** Describes the broad medical field or condition targeted by the trial, such as Solid Tumors, Breast Cancer, etc.
- **Target** (x_T)**.** Specifies the medical condition or disease that the clinical trial is targeting.
- **Drug_Used** (x_{DU})**.** Refers to the specific drug or drugs being tested in the clinical trial.
- **Drug_Class** (x_{DC}) Within our dataset, we find clinical trials focusing on drugs categorized into five groups: Biologic, Biosimilar, New Molecular Entity, Non-New Molecular Entity, and Vaccine. Additionally, some trials are categorized as "Unknown."
- **Therapy** (x_{Th})**.** Describes the type of therapeutic approach being employed in the trial. Examples are Monotherapy, Targeted Therapy, etc.
- **Lead_Sponsor** (x_S)**.** The organization or individual responsible for conducting the clinical trial and ensuring it adheres to regulatory requirements.
- **Criteria** (x_C)**.** The inclusion and exclusion criteria determine which participants are eligible to join the clinical trial.
- **Primary_Outcome** (x_{PO})**.** The main result that the trial is designed to measure is usually specified as an endpoint in the study protocol. Primary outcomes are critical for assessing the efficacy and safety of the intervention.
- **Secondary_Outcome** (x_{SO})**.** Additional results measured in the trial to provide more information on the intervention's effects. Secondary outcomes can offer insights into other benefits or risks associated with the treatment and help support the primary outcome findings.