Predicting Clinical Trial Outcomes from US Patent Data Using a Large Language Model

Joonhyuk Cho, Matthew Gluckman, and Andrew W. Lo

MIT Laboratory for Financial Engineering (LFE), Computer Science and Artificial Intelligence Laboratory (CSAIL)

Abstract

This study uses large language models (LLMs) to predict clinical trial outcomes based on patent data. A rich source of early-stage drug information, this data was analyzed via an LLM to forecast the outcomes of clinical trials in phases 1 through 3. Preliminary results indicate that the patent data's predictive strength increases with each trial phase (AUC-ROC of 0.60 for phase 1, AUC-ROC of 0.72 for phase 3). This suggests that investors can assess financial risk based on findings from the preclinical stage, making therapeutic development a more attractive investment and bringing more capital into the sector. LLMs offer a new platform for risk mitigation and investment decision-making in the biopharma industry.

Introduction

Related Work

Machine Learning with Statistical Imputation for Predicting Drug Approvals *(Lo, Siah, and Wong 2019)*

- 140 features (**without patent data**)
- 15 disease groups
- 0.81 AUC

Drug Approval Prediction using Patents *(Kamijo et al. 2023)*

- Patent feature extraction with natural language processing (NLP) (**without LLMs**)
- Use first 512 words of Abstract, Claim, Description
- 0.8 ~ 0.9 F1 score

Novelty of the Work

- Employs an LLM for both patent summarization and feature extraction
- Predicts probability of transition between clinical trial phases

References

Lo, A. W., K. W. Siah, and C. H. Wong (2019), Machine Learning with Statistical Imputation for Predicting Drug Approvals, *Harvard Data Science Review*, https://doi.org/10.1162/99608f92.5c5f0525.

Kamijo, K., Y. Mitsumori, H. Kato, A. Kato (2023), Drug Approval Prediction Using Patents, *2023 Portland International Conference on Management of Engineering and Technology*, https://doi.org/10.23919/PICMET59654.2023.10216836.

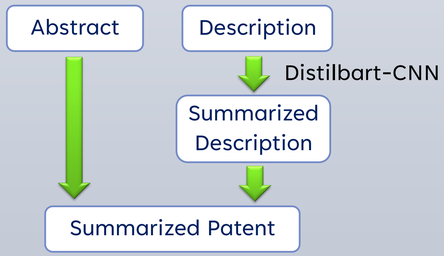
Methods

Data Statistics

- 7,527 Patents
  - 42% Launched
  - 17% Phase 2
  - 9% Phase 3
  - 6% Phase 1
- Biased towards higher phases (i.e., biased more towards success than failure)
  - Evaluation metric should be bias-aware
- Average length: 9,640 tokens (up to 38,000 tokens)
  - Need to be summarized in under 4,096 tokens to be an input of LongFormer model

Data Preprocessing

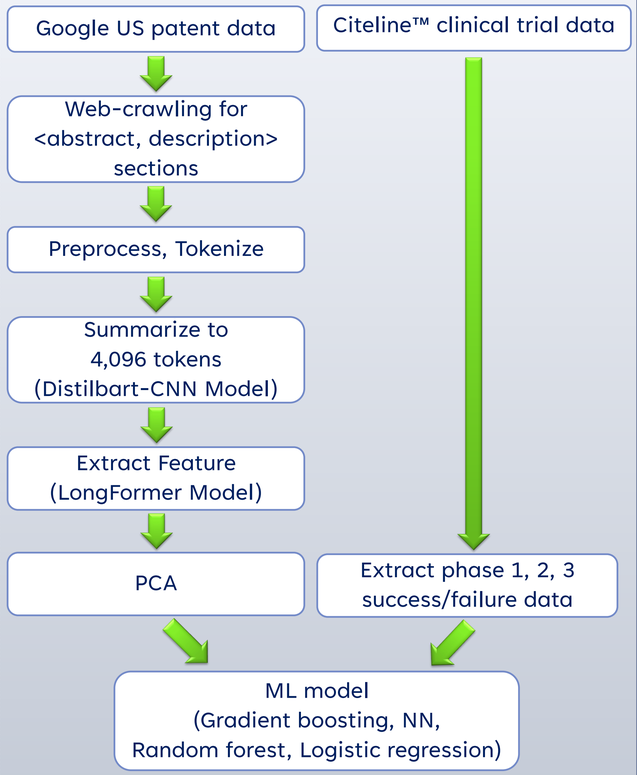
- Sections: Abstract and Description
- Abstract: Relatively short (~100 tokens), contains important information
- Description: Long (~thousands of tokens), summarization required for feature extraction



Feature Extraction

- Summarized patent is processed with pre-trained LongFormer model
- Feature extracted from the last layer of the model
- 1,024 features are converted into 32 features with principal component analysis (PCA)

Prediction Model



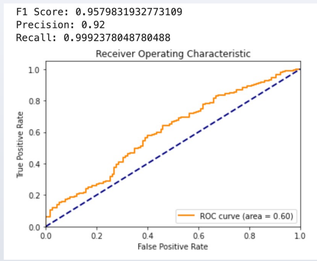
Performance Evaluation

F1 Score
Benchmark: 0.8~0.9
(Kamijo et al. 2023)

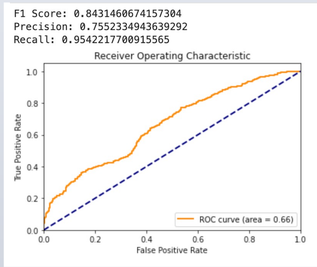
AUC-ROC
Benchmark: 0.81
(Lo et al. 2019)

Results

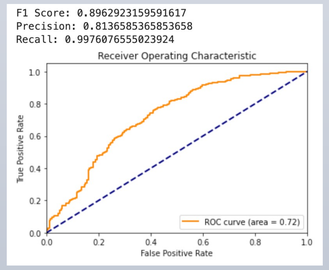
F1 Score/AUC-ROC



Phase 1
F1 Score: 0.96
AUC-ROC: 0.60



Phase 2
F1 Score: 0.84
AUC-ROC: 0.66



Phase 3
F1 Score: 0.90
AUC-ROC: 0.72

Conclusions

We have developed a prediction model that can predict clinical trial phase outcomes by analyzing patent data using an LLM.

Next Steps

Re-run the model with an extended dataset—beyond the ~7,500 patents and drugs utilized in this study—to generate more comprehensive results.