Use Artificial Intelligence to Predict Regulatory Approval Based on Phase III Oncology Clinical Trial Publications

Beverly Fuerte, Pharm.D., Da Sol Kim, Pharm.D., Kenneth Youens, M.D., MBA., Linda Chen, Pharm.D., MS., Tim Reynolds, Pharm.D., MS., Paul Godley, Pharm.D., FASHP, Harry Liu, Ph.D., MBBS
Baylor Scott & White

BACKGROUND

- Fifty percent of phase III clinical trials across all therapeutic areas fail to demonstrate adequate safety and efficacy for approval by the United States Food and Drug Administration (FDA)[1]
- The mean cost of developing a new drug ranges between $314 million to $2.8 billion[2]
- Improving the ability to predict regulatory outcomes using artificial intelligence (AI) could streamline research and development and could save the industry billions in wasted capital, while society can benefit from faster access to effective treatments

OBJECTIVES

1) Investigate the accuracy of AI to make predictions about regulatory outcomes based on phase III oncology clinical trial publications
2) Identify AI-assigned clinical trial features of importance with predicting regulatory outcomes
3) Compare the accuracy of cloud-hosted models versus that of local models

METHODS

- Literature Sample Characteristics (n= 208)
  - Phase III oncology clinical trials
  - Initiated between January 1990 and January 2021
  - Investigated overall survival endpoints
- Statistical analysis across different AI models:
  - Comparison of performance metrics such as F1 score, balanced accuracy, sensitivity, specificity, and C-statistics
  - Comparison of Receiver Operating Characteristic (ROC) curves and calibration curves

Prompt Summary

1. Binary FDA approval prediction (0 or 1)
2. Binary FDA approval prediction with retrieval-augmented generation (RAG)
3. Continuous probability of approval prediction (0-1)
4. Continuous probability of approval prediction with RAG
5. AI-identified decision weight assignment for each given clinical trial attribute

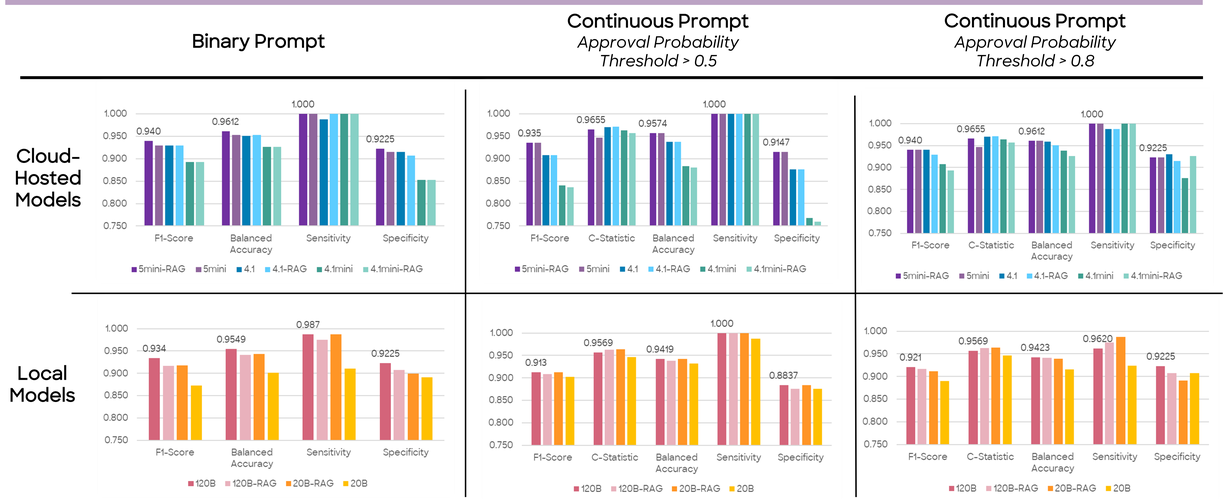
Large Language Models (LLMs)

Temperature = 0, no fine-tuning applied
Retrieval Corpus: FDA Guidance Documents
(1) Benefit-Risk Assessment for New Drug and Biological Products Guidance for Industry[3]
(2) Demonstrating Substantial Evidence of Effectiveness with One Adequate and Well-Controlled Clinical Investigation and Confirmatory Evidence Guidance for Industry[4]

Cloud-Hosted Models		
GPT-4.1 mini	GPT-4.1	GPT-5 mini
OpenAI, close sourced		
Local Models		
GPT-oss-20b	GPT-oss-120b	
OpenAI, open sourced, 21B parameters	OpenAI, open sourced, 117B parameters	

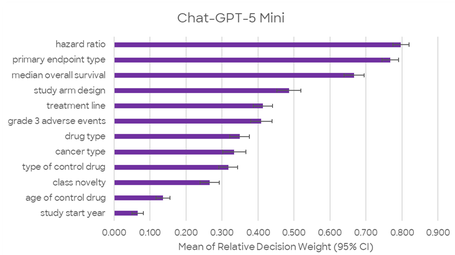
RESULTS

Performance Metrics

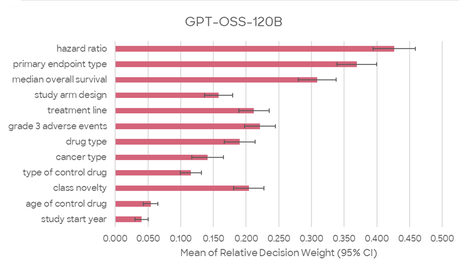


AI-Identified Clinical Attribute Importance

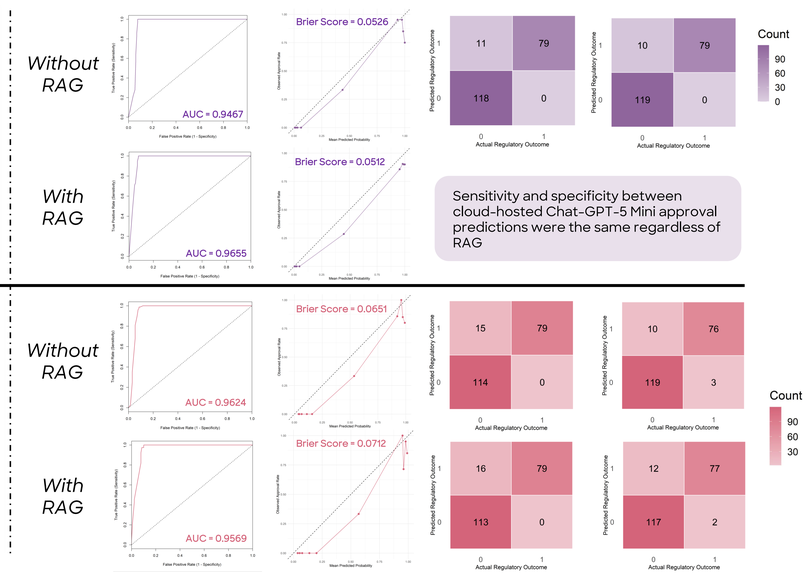
Cloud-Hosted Model: Chat-GPT-5 Mini



Local Model: GPT-OSS-120B



ROC, Calibration Curves, & Confusion Matrices



Sensitivity and specificity between cloud-hosted Chat-GPT-5 Mini approval predictions were the same regardless of RAG

CONCLUSION

- AI was able to predict regulatory outcomes with great sensitivity and specificity
  - Local and cloud-hosted models had comparable performance, but cloud-hosted model Chat-GPT-5 mini with RAG resulted in best approval prediction performance
- Between the cloud-hosted models, RAG marginally improved model performance
- Between the local models, RAG did not consistently improve model performance with respect to all metrics
- Increasing approval threshold from greater than 0.5 to 0.8 improved specificity across all models with greatest improvement in the cloud-hosted Chat-GPT-4.1 mini model
  - As a result of threshold increase, sensitivity remained stable in cloud-based models but decreased in local models
- Of the 12 clinical trial features assessed, performance of the primary endpoint, especially with regards to HR and OS, was identified as key factor in predicting outcomes

LIMITATIONS

- To limit model overload with too many tokens per request, clinical trials were converted from PDFs to pure text documents. Thus, LLMs did not have access to visual information such as figures or diagrams for decision-making
- Findings from this study may not be generalizable to publications outside of phase III clinical studies assessing overall survival within oncology drugs
- Open-source models were not specifically fine-tuned for regulatory decision making; hence, further tuning of the model may improve prediction accuracy

REFERENCES

1. Harrison RK. Phase II and phase III failures: 2013–2015. *Nature Reviews Drug Discovery*. 2016;15(12):817-818. doi:10.1038/nrd.2016.184
2. Wouters OJ, McKee M, Luyten J. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *JAMA*. 2020;323(9):844-853. doi:10.1001/jama.2020.1166
3. Research C for DE and. Benefit-risk assessment for new drug and biological products. August 9, 2024. Accessed April 2, 2026. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/benefit-risk-assessment-new-drug-and-biological-products
4. Research C for DE and. Demonstrating substantial evidence of effectiveness with one adequate and well-controlled clinical investigation and confirmatory evidence. November 30, 2023. Accessed April 2, 2026. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/demonstrating-substantial-evidence-effectiveness-one-adequate-and-well-controlled-clinical

DISCLOSURES

The authors of this study have the following disclosures:
Beverly Fuerte, Da Sol Kim - Paid Consultant (BeOne, Sanofi)
Linda Chen, Tim Reynolds, and Paul Godley - Paid consultant (BeOne, Sanofi, Pfizer)
Kenneth Youens, Harry Liu - Nothing to Disclose

Presented At ISPOR, Philadelphia 2026
Contact: beverly.fuerte@bswhealth.org