Journal of Population Therapeutics & Clinical Pharmacology

RESEARCH ARTICLE
DOI: 10.53555/79vkpq57

# ROLE OF ARTIFICIAL INTELLIGENCE IN PREDICTING OUTCOMES OF CLINICAL TRIALS IN ONCOLOGY

**Dr Mayank Pancholi[1*], Dr. Himanshu Patidar[2], Dr Kavin Rawal[3], Dr Vinod Dhakad[4], Dr Dharmesh Chouhan[5], Dr Jay Patel[6]**

[1*]Designation professor and HOD, Department Surgical Oncology, College/University Name Sri Aurobindo Institute of Medical Sciences, Indore, City- Indore,
Email : dr.mayank.pancholi@gmail.com
[2]Designation Associate Professor, Department Surgical Oncology, College/University Name Sri Aurobindo Institute of Medical Sciences, Indore, City Indore, Email dr_himanshu1@yahoo.com
[3]Designation Senior Resident, Department Surgical Oncology, College/University Name Sri Aurobindo Institute of Medical Sciences, Indore, City Indore, Email : drkavinrawal@gmail.com
[4]Designation Professor, Department Surgical Oncology, College/University Name Sri Aurobindo Institute of Medical Sciences, Indore, City Indore, Email  drvnod7@gmail.com
[5]Designation Senior Resident, Department Surgical Oncology, College/University Name Sri Aurobindo Institute of Medical Sciences, Indore, City Indore,
Email: dharmesh_chauhan777@yahoo.com
[6]Designation Senior Resident, Department Surgical Oncology, College/University Name Sri Aurobindo Institute of Medical Sciences, Indore, City Indore, Email:drjaypatel1609@gmail.com

## Abstract

**Background:** The field of oncology continues to face high attrition rates in clinical trials, with nearly 90% of investigational agents failing before market approval. This places enormous financial and ethical burdens on research institutions and patients alike. The advent of Artificial Intelligence (AI), particularly large language models (LLMs) and specialized neural networks, presents a promising strategy to enhance predictive accuracy for trial outcomes.

**Objective:** This study evaluates and compares the predictive performance of multiple AI architectures—including GPT-3.5, GPT-4, GPT-4mini, GPT-4o, LLaMA3, and the HINT model—in forecasting success or failure of oncology clinical trials using real-world datasets.

**Methods:** Oncology trial data were extracted from ClinicalTrials.gov. Structured trial inputs—phase, sponsor type, duration, endpoints, intervention class—were standardized across models. GPT-family and LLaMA3 models were accessed via API and prompted to classify trial outcomes. HINT, a structured AI model, was used with molecular and ICD-10 inputs. Performance was evaluated using Balanced Accuracy, Matthews Correlation Coefficient (MCC), Recall, and Specificity.

**Results:** GPT-4o demonstrated the highest balanced accuracy (0.573) and robust recall (0.931), while HINT achieved the best specificity (0.541) and stable MCC (0.111). All models underperformed in predicting outcomes for complex oncology subtypes and longerduration trials.

**Conclusion:** AI models hold significant promise in improving predictive capabilities for oncology trial outcomes. GPT-4o and HINT exhibit complementary strengths, and ensemble approaches may optimize trial design, patient selection, and resource allocation.

## Introduction

Cancer remains a leading cause of morbidity and mortality worldwide, with its burden disproportionately affecting low- and middle-income countries like India. With approximately 1.4 million new cancer cases and nearly 900,000 deaths annually, India faces a significant public health challenge [1]. This burden is compounded by infrastructural limitations— overcrowded hospitals, prolonged diagnostic delays, insufficient trained oncologists— especially in rural and underserved regions. Furthermore, the limited penetration of screening programs and low health literacy hinder early detection, contributing to high mortality rates [2].

In this context, **Artificial Intelligence (AI)**, driven by advances in machine learning, deep learning, and natural language processing, holds promise in transforming oncology care and research. AI applications in cancer care extend across the continuum: from risk stratification and early diagnosis to treatment planning, prognostication, and monitoring. Particularly in resource-constrained countries, AI could mitigate systemic inefficiencies and optimize clinical workflows [3]. Moreover, India's diverse and large population provides a rich data reservoir, offering immense potential for training AI models that are generalizable and clinically robust.

[4]

A critical yet underutilized domain for AI integration is oncology clinical trials. These trials represent a high-risk and high-investment endeavor, with over **85% of cancer drug candidates failing to progress beyond early-phase trials**, despite promising preclinical data [5]. Reasons for these failures include poor trial design, inadequate endpoint selection, and suboptimal patient stratification. Consequently, the **average cost of developing a successful oncology drug exceeds $2.5 billion**, with timelines stretching over a decade [6].

To address these challenges, AI models—particularly **Large Language Models (LLMs)** like GPT-4o and specialized architectures such as HINT—are increasingly being employed to simulate, model, and predict clinical trial outcomes. These models integrate structured data (e.g., trial phase, endpoint type) and unstructured data (e.g., study summaries) to assess trial success probabilities, thereby supporting early decision-making and resource allocation [7,8]. Despite the promise, few comparative studies have assessed the performance of these models in oncology-specific trials, where complexity and heterogeneity are inherently higher.

This study addresses this gap by evaluating and comparing the predictive performance of six AI models on real-world oncology clinical trial data. The findings aim to inform AI-assisted trial design strategies and reduce failure rates in cancer drug devel

## Materials and Methods

### *Study Design and Objectives*

This was a retrospective, comparative modeling study evaluating the predictive performance of six Artificial Intelligence (AI) models in forecasting outcomes of oncology clinical trials. The primary objective was to assess their accuracy, reliability, and model-specific strengths across different phases, endpoints, and durations of trials.

### *Data Source and Curation*

Data were extracted from the publicly accessible **ClinicalTrials.gov** database. The inclusion criteria were:

- Interventional oncology trials (Phases I–III)
- Completed status with results posted between January 2015 and December 2023
- Clearly defined primary outcomes (e.g., OS, PFS, ORR, AE/SAE)
- Human subjects only (all genders, all age groups)

Trials with statuses such as "Terminated", "Suspended", or lacking outcome data were excluded to ensure labeling consistency.

Each trial record was curated for:

- Title and brief summary
- ICD-10 coded disease classification (neoplasms only)
- Intervention description and modality
- Sponsor type (industry vs. academic)
- Primary outcome type
- Start and end dates (to compute duration)

A total of **2,163 trials** were included, of which **712 were oncology-specific trials**.

***AI Models Evaluated***

The following models were assessed:

| **Model** | **Architecture Type** | **Version** | **Training Cut-Off** |
|---|---|---|---|
| GPT-3.5 | LLM (Closed-source) | gpt-3.5-turbo-0125 | Sep 2021 |
| GPT-4 | LLM (Closed-source) | gpt-4-turbo-2024-04-09 | Dec 2023 |
| GPT-4mini | LLM (Lightweight) | gpt-4o-mini-2024-07-18 | Oct 2023 |
| GPT-4o | LLM (Multimodal) | chatgpt-4o-latest | Oct 2023 |
| LLaMA3 | LLM (Open-source) | LLaMA-3-8B-Instruct | Mar 2023 |
| HINT | Domain-specific AI | Hybrid Inference Neural Transformer | Apr 2022 |

***Trial Outcome Labeling***

The outcome of each trial was annotated as either **"Success"** or **"Failure"** based on:

- Whether the primary endpoint was met as per reported results
- Consistency across study conclusion and outcome measures
- Independent verification through NLP-assisted manual review by two clinical experts

***Model Input and Prediction Process***

**For LLMs (GPT and LLaMA3):**

Each model was prompted with structured trial data using a standardized instruction:

"Use the following clinical trial data to predict whether the outcome will be a success or failure. Reply with only one word: success or failure."

Input fields included: trial title, summary, disease category, phase, intervention, endpoint type, sponsor type, and trial duration.

**For HINT Model:**

Data were preprocessed using SMILES representations for drug molecules and ICD-10 codes for diseases, as per Fu et al. [1]. Prediction outputs were binary (success/failure) based on a probability threshold of 0.5.

***Evaluation Metrics***

Performance of each model was measured using:

- **Balanced Accuracy** = (Sensitivity + Specificity)/2
- **Matthews Correlation Coefficient (MCC)**: A robust metric for imbalanced classification tasks
- **Recall (Sensitivity)**: Ability to identify successful trials

- **Specificity**: Ability to correctly detect failed trials

## RESULTS

This section presents the comparative evaluation of six AI models—GPT-3.5, GPT-4, GPT4mini, GPT-4o, LLaMA3, and HINT—on their ability to predict the outcomes of oncology clinical trials. Model performance was assessed using **Balanced Accuracy**, **Matthews Correlation Coefficient (MCC)**, **Recall (Sensitivity)**, and **Specificity**.

### *Overall Model Performance*

The overall predictive accuracy across all clinical trial data is summarized in **Table 1**. GPT-4o emerged as the top-performing model with the highest **balanced accuracy** (0.573) and a strong **recall** (0.931), suggesting a high capability in correctly identifying trials with successful outcomes. However, its relatively low **specificity** (0.214) indicates a tendency to over-classify trials as successful.

In contrast, the HINT model exhibited **the highest specificity** (0.541), showing superior performance in correctly identifying failed trials, which is critical for de-risking trial portfolios. Although its recall (0.586) was lower than that of the GPT-based models, HINT's **balanced accuracy** (0.563) and **MCC** (0.111) indicate a more stable performance across both outcome classes. GPT-4 also performed consistently with a balanced accuracy of 0.542 and MCC of 0.234. LLaMA3 and GPT-3.5 demonstrated lower predictive power, with LLaMA3 yielding an MCC of 0.058 and GPT-3.5 showing a significant bias toward predicting positive outcomes, as indicated by its near-zero specificity (0.011).

**Table 1: Overall AI Model Performance**

| Model | Balanced Accuracy | MCC | Recall | Specificity |
|---|---|---|---|---|
| GPT-3.5 | 0.504 | 0.049 | 0.997 | 0.011 |
| GPT-4 | 0.542 | 0.234 | 1.000 | 0.083 |
| GPT-4mini | 0.500 | 0.000 | 1.000 | 0.000 |
| GPT-4o | 0.573 | 0.212 | 0.931 | 0.214 |
| LLaMA3 | 0.517 | 0.058 | 0.949 | 0.085 |
| HINT | 0.563 | 0.111 | 0.586 | 0.541 |

### *Phase-wise Model Performance*

To further assess the adaptability of AI models across trial development stages, we analyzed performance by clinical trial phase. As shown in **Table 2**, GPT-4o maintained consistent performance across all phases, with its best results in Phase III trials (**balanced accuracy: 0.667**, **MCC: 0.509**). HINT outperformed all other models in **Phase III**, with the highest **balanced accuracy (0.699)** and **MCC (0.312)**, suggesting its strength in handling more complex, multi-center trials where data heterogeneity is higher.

Interestingly, both models showed moderate predictive accuracy in early phases, which may reflect limited and more homogeneous datasets.

**Table 2: Phase-wise Model Performance**

| Phase | GPT-4o (Balanced Acc) | HINT (Balanced Acc) | GPT-4o (MCC) | HINT (MCC) |
|---|---|---|---|---|
| Phase I | 0.557 | 0.562 | 0.120 | 0.126 |
| Phase II | 0.556 | 0.516 | 0.277 | 0.030 |
| Phase III | 0.667 | 0.699 | 0.509 | 0.312 |

### *Endpoint-specific Model Performance*

Trial outcomes were also stratified by primary endpoints—namely Overall Survival (OS), Objective Response Rate (ORR), Progression-Free Survival (PFS), and Adverse Events (AE/SAE). The performance of GPT-4o and HINT across these subgroups is presented in **Table 3**.

GPT-4o showed **the best results in trials with OS as the primary endpoint**, with a **balanced accuracy of 0.614** and **MCC of 0.278**, reflecting the model's strong ability to predict longterm survival benefits. HINT demonstrated **more stable performance across AE/SAE endpoints**, indicating its potential utility in safety prediction. All models underperformed in ORR and PFS endpoints, where the trial data tend to be more variable and surrogate-based.

**Table 3: Endpoint-specific Model Performance**

| **Endpoint** | **GPT-4o (Balanced Acc)** | **HINT (Balanced Acc)** | **GPT-4o (MCC)** | **HINT (MCC)** |
|---|---|---|---|---|
| Overall Survival (OS) | 0.614 | 0.548 | 0.278 | 0.085 |
| Objective Response Rate | 0.438 | 0.488 | -0.167 | -0.022 |
| Progression-Free Survival | 0.500 | 0.446 | 0.000 | -0.120 |
| Adverse Events (AE/SAE) | 0.583 | 0.516 | 0.169 | 0.034 |

### *Trial Duration-based Performance*

The performance of AI models varied with trial duration, as shown in **Table 4**. GPT-4o performed best in **short-term trials (<1000 days)** with a **balanced accuracy of 0.875**, while HINT's best performance was also in the short-term category (**balanced accuracy: 0.675**, **MCC: 0.285**). However, both models experienced a performance decline in **long-term trials (>2000 days)**, with HINT's MCC dropping to **-0.342**, indicating difficulty in modeling extended trial timelines with multiple confounding factors.

This trend suggests that AI models may require additional training or longitudinal data integration to handle complex, long-term oncology studies effectively.

**Table 4: Trial Duration-based Model Performance**

| **Duration** | **GPT-4o (Balanced Acc)** | **HINT (Balanced Acc)** | **GPT-4o (MCC)** | **HINT (MCC)** |
|---|---|---|---|---|
| Short-term (<1000 days) | 0.875 | 0.675 | 0.000 | 0.285 |
| Medium (1001–2000 days) | 0.625 | 0.519 | 0.316 | 0.034 |
| Long-term (>2000 days) | 0.562 | 0.333 | 0.265 | -0.342 |

### *Summary of Findings*

- **GPT-4o** is most effective for generalizable outcome prediction, particularly for short- and medium-term trials with OS endpoints.
- **HINT** excels in **specificity and Phase III trial prediction**, making it suitable for identifying trials at high risk of failure.
- All models underperform in surrogate endpoint prediction (e.g., ORR, PFS) and longer-duration trials, emphasizing the need for domain-specific fine-tuning and incorporation of longitudinal data.

These findings provide a strong foundation for recommending hybrid or ensemble modeling approaches for real-world trial prediction in oncology.

## REFERENCES

1. Miller KD, Nogueira L, Mariotto AB, et al. Cancer treatment and survivorship statistics, 2019. *CA Cancer J Clin*. 2019;69(5):363–385.
2. Mallath MK, Taylor DG, Badwe RA, et al. The growing burden of cancer in India: epidemiology and social context. *Lancet Oncol*. 2014;15(6):e205–12.
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44–56.
4. Goel I, Bhaskar Y, Kumar N, Singh S, Amanullah M, Dhar R, Karmakar S. Role of AI in empowering and redefining the oncology care landscape: perspective from a developing nation. Front Digit Health. 2025 Mar 4;7:1550407.
5. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters in oncology. *Biostatistics*. 2019;20(2):273–286.
6. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J Health Econ*. 2016;47:20–33.
7. Siu LL, et al. Challenges and opportunities in adapting clinical trial design for immunotherapies. *Clin Cancer Res*. 2017;23(17):4990–5000.
8. Fu J, et al. Predicting clinical trial outcomes using deep learning. *Nat Mach Intell*. 2022;4:834–845.