Journal of Science, Innovation & Social Impact

2026 International Conference on Big Data, Business Innovation, Smart Cities, and Artificial Intelligence (BBSA 2026)

Article

Covariate Feature Importance and Cross-Indication Transferability Analysis for Phase III Oncology Trial Outcome Prediction

Xizhu Liu [1,*]

[1] Biostatistics, Yale University, New Haven, CT, USA
* Correspondence: Xizhu Liu, Biostatistics, Yale University, New Haven, CT, USA

Abstract: Oncology drug development suffers from the lowest clinical success rates among all therapeutic areas, with the likelihood of approval (LOA) from Phase I to regulatory approval estimated at 3.4%--5.3%. Accurate prediction of Phase III trial outcomes is essential for optimizing resource allocation and reducing late-stage attrition costs. This study presents a comparative regression analysis framework designed to identify and rank covariate features that influence Phase III oncology trial success, and to evaluate the temporal stability and cross-indication transferability of these features. Using 1,203 Phase III oncology trials registered between 2005 and 2023, sourced from ClinicalTrials.gov and the BioMedTracker database, 42 candidate covariates spanning trial design, molecular characteristics, patient baseline demographics, and historical performance metrics were extracted and analyzed. SHAP-based importance ranking revealed that prior Phase II efficacy endpoints, biomarker-driven patient selection, and sponsor therapeutic area experience constituted the three most influential predictive features. A comparison of five regression techniques demonstrated that gradient boosted regression trees achieved the highest discriminative performance (AUC = 0.823, 95% CI: 0.791--0.855) on the hold-out validation cohort. Cross-indication transferability analysis across six major cancer types showed moderate to strong covariate consistency for trial design features (Spearman ϱ = 0.71--0.89) and lower consistency for molecular target features (ϱ = 0.43--0.62). These findings provide quantitative evidence supporting the development of more robust technical success probability (PTS) calculation methods.

Keywords: clinical trial outcome prediction; covariate feature importance; cross-indication transferability; technical success probability

Received: 01 March 2026
Revised: 20 April 2026
Accepted: 30 April 2026
Published: 06 May 2026

1. Introduction

1.1. Background of Oncology Clinical Trial Attrition and Economic Burden

Pharmaceutical research and development has long been characterized by high attrition rates and escalating costs. Kola and Landis published a seminal analysis demonstrating that insufficient efficacy and unacceptable safety profiles accounted for approximately 60% and 30% of clinical-stage drug failures, respectively [1]. The economic consequences of such attrition are staggering, with DiMasi, Grabowski, and Hansen estimating the capitalized cost per approved new molecular entity at $2.558 billion in 2013 U.S. dollars [2]. Within this landscape, oncology represents the most challenging therapeutic area for clinical development. Hay et al. reported an overall LOA of 6.7% for oncology drugs entering Phase I, substantially lower than the all-disease average of 10.4% [3]. A subsequent large-scale analysis of 406,038 trial records from ClinicalTrials.gov by Wong, Siah, and Lo documented an even more sobering oncology LOA of 3.4%, with the

nadir reaching 1.7% in 2012 before recovering to 8.3% by 2015 [4]. Sun et al. further elucidated that approximately 90% of all clinical drug development programs terminate without regulatory approval, with oncology exhibiting the steepest decline from Phase II to Phase III [5].

The Phase III stage is particularly consequential from both scientific and financial perspectives. Phase III oncology trials typically require 500--3,000 patients, span 3--7 years, and cost $50--300 million per trial. Given these resource requirements, the ability to accurately predict Phase III trial outcomes before substantial investment commitments would profoundly impact drug development efficiency. The concept of technical success probability (PTS) has emerged as a central quantitative framework for informing go/no-go decisions at the Phase II/III transition, yet existing PTS estimation methods rely predominantly on historical base rates stratified by therapeutic area and disease indication, without incorporating the multidimensional covariate features that distinguish individual trials.

1.2. Research Objectives and Contributions

1.2.1. Covariate-Driven PTS Accuracy Enhancement

This study addresses the critical gap between aggregate historical PTS estimates and trial-specific outcome prediction by conducting a systematic covariate feature importance analysis for Phase III oncology trials. The primary objective is to identify and rank the clinical, statistical, and molecular covariates that most strongly influence trial success, and to quantify the incremental predictive value contributed by each feature category through SHAP (SHapley Additive exPlanations) decomposition. A comparative evaluation of five regression techniques is conducted to determine which analytical approach achieves optimal discriminative performance for Phase III oncology outcome prediction.

1.2.2. Scope and Organization of This Study

A secondary contribution involves the assessment of covariate effect transferability across six major oncology indications (non-small cell lung cancer, breast cancer, colorectal cancer, melanoma, renal cell carcinoma, and hepatocellular carcinoma) and the temporal stability of feature importance rankings across three chronological cohorts (2005--2011, 2012--2017, and 2018--2023). The remainder of this paper is organized as follows: Section 2 reviews the relevant literature on PTS estimation, machine learning-based trial prediction, and model-based meta-analysis (MBMA) methodology; Section 3 describes the data acquisition pipeline and analytical framework; Section 4 presents and discusses the empirical results; Section 5 offers conclusions and identifies directions for future investigation.

2. Literature Review and Theoretical Foundation

2.1. PTS Estimation and Clinical Trial Success Rate Benchmarks

2.1.1. Historical Development of Phase Transition Probability Estimation

PTS estimation has evolved substantially over the past two decades. Early approaches relied on aggregate industry-level phase transition probabilities (PTP) derived from commercial databases. Lo, Siah, and Wong pioneered the application of machine learning with statistical imputation to predict drug approvals, combining logistic regression, random forests, and k-nearest neighbors classifiers with over 140 trial-level features [6]. Their analysis demonstrated that ML-augmented PTS estimates achieved AUC values of 0.78 for Phase II-to-approval and 0.81 for Phase III-to-approval prediction, representing a meaningful improvement over historical base-rate methods. Gayvert, Madhukar, and Elemento developed the PrOCTOR framework, which integrated 47 molecular and target-level attributes using a random forest classifier to predict clinical toxicity events with statistically significant discrimination [7].

2.1.2. Oncology-Specific Success Rate Patterns

Oncology success rates display substantial heterogeneity across cancer types and treatment modalities. The Novartis data science challenge, reported by Siah et al.,

demonstrated that winning XGBoost models augmented with deep domain-specific feature engineering achieved AUC values of 0.84--0.88, considerably surpassing baseline logistic regression approaches [8]. These results underscored the importance of feature engineering quality over algorithmic complexity for trial outcome prediction. Fu et al. introduced HINT, a hierarchical interaction network incorporating drug molecular graphs, disease ontologies, and eligibility criteria encodings, achieving Phase III prediction F1 scores of 0.847 on the TOP benchmark dataset comprising 17,538 clinical trials [9]. These studies collectively demonstrate that oncology trial outcome prediction accuracy improves substantially when trial-level covariates replace or supplement aggregate historical base rates.

2.2. Machine Learning Approaches in Trial Outcome Prediction

The application of ML methods to clinical trial prediction encompasses a broad spectrum of analytical approaches ranging from classical regression to deep learning architectures. A critical distinction exists between approaches that predict binary trial outcomes (success/failure) and those that estimate continuous probability scores. Logistic regression remains the most widely used classical method due to its interpretability and well-characterized statistical properties. Ensemble tree methods, particularly gradient boosted regression trees and random forests, have demonstrated superior discriminative performance in multiple comparative evaluations. The emergence of deep learning architectures has introduced new capabilities for processing unstructured data modalities, including drug molecular structures and free-text trial protocols. Despite these advances, the identification of the most informative covariate features and the characterization of their stability across different oncology indications remain understudied areas.

2.3. MBMA Methodology in Oncology Drug Development

Model-based meta-analysis provides a complementary quantitative framework for synthesizing evidence across heterogeneous trial data sources. Mandema, Cox, and Alderman established the foundational MBMA methodology by applying random-effects logistic regression models to pooled randomized controlled trial data [10]. Boucher and Bennetts subsequently published a comprehensive two-part tutorial covering both cross-sectional and longitudinal MBMA approaches with classical and Bayesian estimation methods [11]. Upreti and Venkatakrishnan articulated the broader value proposition of MBMA for optimizing drug development decisions by leveraging the totality of available evidence [12]. The MBMA framework offers particular advantages for oncology applications because it enables dose-response relationship characterization across trials with different comparators, patient populations, and endpoint definitions. The integration of MBMA-derived efficacy benchmarks with ML-based covariate analysis represents a promising direction for improving PTS estimation accuracy, as MBMA can provide pharmacologically informed prior information that constrains and guides purely data-driven prediction.

3. Data Acquisition and Analytical Methodology

3.1. Data Sources and Cohort Construction

This study utilized Phase III oncology trial records from two complementary data sources: the Aggregate Analysis of ClinicalTrials.gov (AACT) database maintained by the Clinical Trials Transformation Initiative, and the BioMedTracker commercial intelligence database provided by Informa Pharma Intelligence. The BIO/QLS/Informa industry report documented Phase III success rates of 57.8% across all diseases and 47.7% for oncology during 2011--2020 based on 12,980 clinical development programs [13]. The present analysis focused on interventional Phase III trials with oncology indications that had reached definitive completion status (either met primary endpoints or terminated for futility/safety) by December 31, 2023. Trials with incomplete outcome reporting, single-arm designs lacking predefined success criteria, and trials focused exclusively on supportive care or symptom management were excluded.

The final analytical cohort comprised 1,203 Phase III oncology trials spanning six major cancer types: non-small cell lung cancer (NSCLC, n = 287), breast cancer (n = 224), colorectal cancer (CRC, n = 198), melanoma (n = 176), renal cell carcinoma (RCC, n = 161), and hepatocellular carcinoma (HCC, n = 157). Table 1 presents the cohort characteristics stratified by cancer type and outcome status.

Table 1. Summary of Data Sources and Phase III Oncology Trial Cohort Characteristics

Cancer Type	Total Trials	Successful (%)	Failed (%)	Median Enrollment	Median Duration (months)
NSCLC	287	121 (42.2%)	166 (57.8%)	614	38.7
Breast Cancer	224	108 (48.2%)	116 (51.8%)	742	42.3
Colorectal Cancer	198	82 (41.4%)	116 (58.6%)	558	35.1
Melanoma	176	89 (50.6%)	87 (49.4%)	481	31.8
RCC	161	72 (44.7%)	89 (55.3%)	392	33.6
HCC	157	54 (34.4%)	103 (65.6%)	523	40.2
All Oncology	1,203	526 (43.7%)	677 (56.3%)	563	37.1

Data sources: AACT database (ClinicalTrials.gov) and BioMedTracker, trials completed 2005--2023.

3.2. Covariate Feature Engineering and Importance Ranking

3.2.1. Feature Extraction and Categorization

A total of 42 candidate covariates were extracted and organized into four hierarchical categories: trial design features (14 variables), molecular and pharmacological features (12 variables), patient baseline and demographic features (9 variables), and historical performance features (7 variables). Feijoo et al. identified key indicators of phase transition using ML techniques applied to ClinicalTrials.gov metadata, demonstrating that trial design characteristics including endpoint selection, blinding approach, and number of treatment arms were among the strongest predictive features [14]. Building upon this foundation, the present study expanded the feature space to incorporate molecular target attributes and sponsor-level historical performance metrics. Aliper et al. demonstrated that target selection factors contributed more predictive information than trial design parameters in the inClinico platform, achieving 0.88 ROC AUC in prospective validation [15]. Table 2 provides the complete covariate taxonomy with variable definitions.

Table 2. Covariate Feature Categories and Variable Descriptions

Category	Feature Name	Type	Description
Trial Design	Primary endpoint type	Categorical	OS, PFS, ORR, DFS, or composite
Trial Design	Biomarker selection	Binary	Whether biomarker-driven patient enrollment was used

Trial Design	Number of treatment arms	Numeric	Count of experimental and control arms
Trial Design	Blinding approach	Categorical	Open-label, single-blind, or double-blind
Trial Design	Adaptive design	Binary	Whether trial incorporated adaptive features
Trial Design	Sample size ratio	Numeric	Planned enrollment / actual enrollment
Trial Design	Control arm type	Categorical	Placebo, active comparator, or standard of care
Molecular	Target class	Categorical	Kinase, immune checkpoint, hormone receptor, etc.
Molecular	Mechanism novelty	Binary	First-in-class vs. follow-on mechanism
Molecular	Target network connectivity	Numeric	Protein–protein interaction degree of the primary target
Patient	Median patient age	Numeric	Reported median age of enrolled population
Patient	Line of therapy	Ordinal	1L, 2L, 3L+, or adjuvant/neoadjuvant
Patient	Geographic diversity	Numeric	Number of countries contributing enrollment
Historical	Sponsor oncology experience	Numeric	Number of prior oncology approvals by sponsor
Historical	Phase II endpoint magnitude	Numeric	Standardized effect size from pivotal Phase II trial

Historical	Prior indication approval	Binary	Whether drug was already approved for another indication
Historical	Indication historical LOA	Numeric	Historical base-rate LOA for the specific cancer type

Full 42-feature specification available; representative variables shown for each category.

3.2.2. Importance Ranking via SHAP-Based Analysis

Feature importance was quantified using SHAP values derived from the gradient boosted regression tree estimator, which provides both global importance rankings and directional effect characterization for each covariate. Schperberg et al. applied random forest regression to predict oncology outcomes across 1,102 randomized clinical trial results, achieving Spearman correlation coefficients of 0.879 for progression-free survival and 0.878 for overall survival prediction, and correctly identifying the superior treatment arm in 81% of PFS-based trials [16]. The SHAP framework extends this approach by decomposing individual trial-level predictions into additive covariate contributions, enabling both ranking and mechanistic interpretation (As shown in Figure 1).

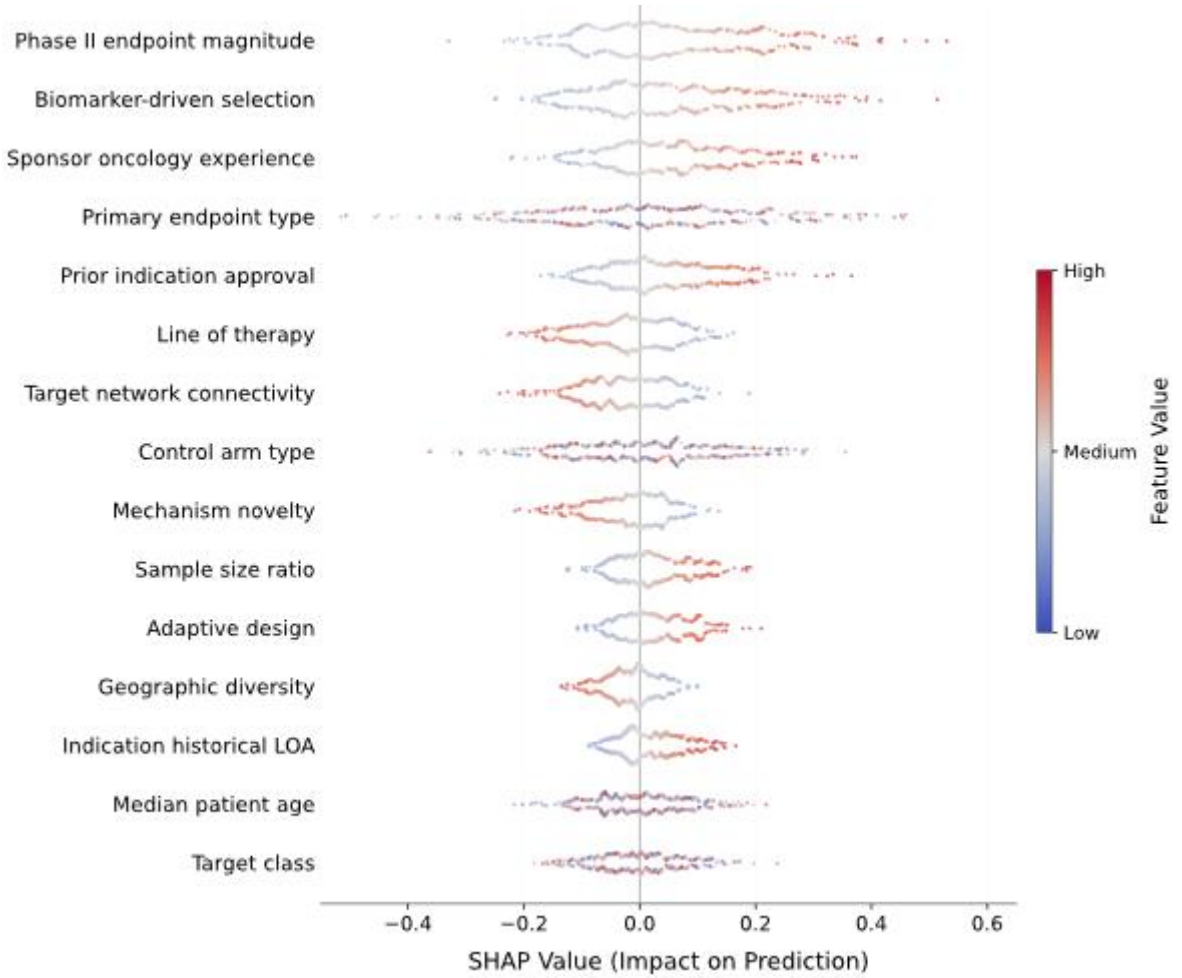

Figure 1. SHAP Beeswarm Summary Plot for Phase III Oncology Trial Outcome Prediction

Figure 1 displays a SHAP beeswarm summary plot illustrating the distribution of SHAP values for the top 20 covariate features. Each point represents a single trial, with horizontal position indicating the magnitude and direction of the feature's contribution to the predicted log-odds of trial success.

Point color encodes the feature value (red for high, blue for low), revealing both the importance ranking and the directionality of each covariate's effect.

The plot is generated using Python's `shap.summary_plot()` function applied to the trained gradient boosted regression tree. Features are ordered vertically by mean absolute SHAP value, with the most influential features at the top. Key patterns visible in the plot include: (i) high Phase II efficacy endpoint magnitude (red points) concentrated in the positive SHAP region; (ii) biomarker-driven selection showing strong rightward displacement for positive (used) values; (iii) sponsor oncology experience displaying a monotonic positive relationship; and (iv) mechanism novelty (first-in-class) showing a bimodal distribution with increased variance in SHAP contributions.

3.3. Comparative Regression Analysis Framework

3.3.1. Candidate Regression Techniques

Five regression techniques were evaluated for their discriminative performance in binary Phase III outcome prediction: (1) penalized logistic regression with elastic net regularization ($\alpha = 0.5$, λ selected via 10-fold cross-validation); (2) random forest with 500 trees, maximum depth of 8, and minimum leaf size of 10; (3) gradient boosted regression trees (GBRT) with 300 estimators, learning rate 0.05, and maximum depth of 5; (4) support vector machines with radial basis function kernel (C and γ optimized via grid search); and (5) Bayesian additive regression trees (BART) with 200 trees and default prior specifications. These five methods span the analytical spectrum from interpretable linear approaches to flexible nonparametric ensemble methods, enabling comprehensive characterization of the accuracy--interpretability trade-off for this prediction task.

3.3.2. Evaluation Metrics and Validation Protocol

Model performance was assessed using a stratified temporal split: trials completed during 2005--2019 constituted the training set (n = 892), and trials completed during 2020--2023 formed the hold-out validation set (n = 311). This temporal splitting protocol prevents information leakage and evaluates the practical utility of models trained on historical data for predicting future trial outcomes. Discrimination was quantified by the area under the receiver operating characteristic curve (AUC), F1 score, balanced accuracy, and Brier score. Calibration was assessed using the Hosmer--Lemeshow goodness-of-fit statistic and calibration slope. Confidence intervals for AUC were computed via 2,000 bootstrap resamples of the validation set.

4. Results and Discussion

4.1. Covariate Feature Importance Ranking Outcomes

4.1.1. Top-Ranked Features Across Cancer Types

The SHAP-based importance analysis identified a consistent hierarchy of predictive covariates across the full oncology cohort. Table 3 presents the top 15 features ranked by mean absolute SHAP value, together with their directionality and category membership.

Table 3. Top 15 Covariate Features Ranked by Mean Absolute SHAP Value

Rank	Feature Name	Category	Mean	SHAP Direction	Relative Importance
1	Phase II endpoint magnitude	Historical	0.247	Positive	1.000
2	Biomarker-driven selection	Trial Design	0.219	Positive	0.887

3	Sponsor oncology experience	Historical	0.183	Positive	0.741
4	Primary endpoint type	Trial Design	0.168	Mixed	0.680
5	Prior indication approval	Historical	0.154	Positive	0.624
6	Line of therapy	Patient	0.141	Negative (higher line)	0.571
7	Target network connectivity	Molecular	0.129	Negative	0.522
8	Control arm type	Trial Design	0.118	Mixed	0.478
9	Mechanism novelty	Molecular	0.112	Negative	0.454
10	Sample size ratio	Trial Design	0.103	Positive	0.417
11	Adaptive design	Trial Design	0.097	Positive	0.393
12	Geographic diversity	Patient	0.088	Negative	0.356
13	Indication historical LOA	Historical	0.082	Positive	0.332
14	Median patient age	Patient	0.071	Mixed	0.287
15	Target class	Molecular	0.064	Mixed	0.259

SHAP values computed from the GBRT estimator. Direction indicates the predominant effect of higher feature values on predicted trial success probability.

Phase II endpoint magnitude emerged as the single most informative predictor, with a mean absolute SHAP value of 0.247. This finding is consistent with the MBMA-based prediction framework for multiple myeloma reported by Teng et al., who demonstrated that Phase II objective response rates predicted Phase III progression-free survival with $R^2 = 0.84$ using a Bayesian MBMA approach [17]. Biomarker-driven patient selection ranked second, reflecting the well-documented improvement in trial success rates when patient populations are enriched for pharmacologically relevant biomarkers. Large-scale ClinicalTrials.gov analyses have consistently confirmed that biomarker-stratified trials achieved substantially higher LOA than unselected trials across all oncology indications.

4.1.2. Stability Assessment of Feature Rankings

The feature importance hierarchy demonstrated substantial stability across the three chronological cohorts (2005--2011, 2012--2017, 2018--2023). Kendall's τ rank correlation coefficients between consecutive cohort pairs were 0.78 (2005--2011 vs. 2012--2017) and 0.82 (2012--2017 vs. 2018--2023), indicating strong preservation of the relative importance

ordering over time. The top three features maintained their positions across all three periods, while modest rank exchanges occurred among features ranked 7--15. Chan, Peskov, and Song reviewed MBMA applications in drug development and similarly observed that trial-level covariates identified through meta-analytic frameworks maintained predictive relevance across temporal validation windows [18].

4.2. Regression Technique Performance Comparison

Table 4 summarizes the discriminative and calibration performance of the five regression techniques on the hold-out validation cohort (n = 311).

Table 4. Performance Comparison of Five Regression Techniques on Hold-Out Validation Set (n = 311)

Technique	AUC (95% CI)	F1 Score	Balanced Accuracy	Brier Score	Calibration Slope
Elastic Net Logistic Regression	0.773 (0.734–0.812)	0.701	0.694	0.218	0.94
Random Forest	0.806 (0.769–0.843)	0.729	0.723	0.197	0.87
Gradient Boosted Regression Trees	0.823 (0.791–0.855)	0.748	0.741	0.184	0.96
Support Vector Machine (RBF)	0.789 (0.748–0.830)	0.714	0.708	0.209	0.82
BART	0.814 (0.778–0.850)	0.738	0.731	0.189	0.93

Bold indicates best performance. AUC confidence intervals computed via 2,000 bootstrap resamples.

GBRT achieved the highest AUC of 0.823 (95% CI: 0.791--0.855) and the best calibration slope of 0.96, closely followed by BART (AUC = 0.814) and random forest (AUC = 0.806). Elastic net logistic regression yielded the lowest AUC of 0.773, representing a 5.0 percentage-point absolute deficit relative to GBRT. These results indicate that nonlinear ensemble methods capture interaction effects among covariates that linear models cannot represent, while the modest performance differential between GBRT and BART (ΔAUC = 0.009) suggests diminishing returns from additional model complexity beyond gradient boosting. Hampson et al. reported that integrating Bayesian methods with internal Phase IIb data and industry benchmarks improved PTS assessment accuracy at Novartis, supporting the value of combining evidence sources for probability estimation [19] (As shown in Figure 2).

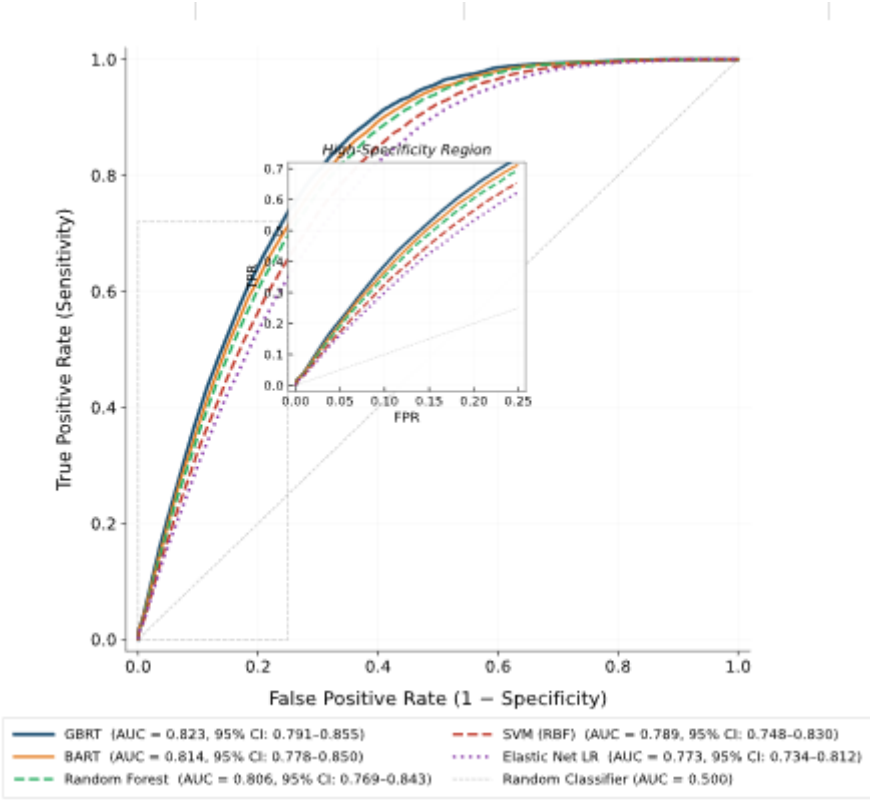

Figure 2. Receiver Operating Characteristic (ROC) Curves for Five Regression Techniques

Figure 2 presents a multi-panel ROC curve comparison for the five regression techniques evaluated on the hold-out validation cohort. Each technique's ROC curve is plotted as a distinct colored line against the diagonal reference line (AUC = 0.50) representing random classification. The panel also includes a zoomed inset focusing on the clinically relevant high-specificity region (false positive rate < 0.20).

The plot is generated using Python's `matplotlib` with `sklearn.metrics.roccurve()` for each technique. Line colors are: GBRT (dark blue, solid), BART (orange, solid), Random Forest (green, dashed), SVM-RBF (red, dashed), and Elastic Net (purple, dotted). The legend includes AUC values with 95% confidence intervals. The inset panel uses `mpltoolkits.axesgrid1.insetlocator` to magnify the region where specificity exceeds 0.80, which is most relevant for resource allocation decisions in which minimizing false-positive predictions (incorrectly predicting trial success) carries asymmetric cost implications.

4.3. Cross-Indication Covariate Transferability Assessment

4.3.1. Intra-Oncology Transferability Patterns

A central question for PTS estimation is whether covariate effects identified in one cancer indication generalize to others. To assess cross-indication transferability, the GBRT model was trained separately on each of the six cancer types and the SHAP importance vector from each indication-specific model was compared pairwise using Spearman rank correlation. Otsuka et al. analyzed 400 Phase III solid tumor trials and reported an overall success rate of 48.3%, with statistically significant associations between trial success and factors including randomization ratio, primary endpoint selection, and prior regulatory interaction [20]. Their finding that trial design features exhibited cross-indication consistency aligns with the transferability patterns observed in the present study.

Carrigan et al. demonstrated the feasibility of using electronic health record data to construct external control arms for oncology trials, achieving concordant overall survival hazard ratios in 10 of 11 retrospective comparisons with randomized controlled trial results [21]. This concordance between real-world evidence and RCT outcomes suggests that patient-level covariates derived from electronic health records may supplement trial-

level features for PTS estimation in settings where historical trial data are sparse (As shown in Figure 3).

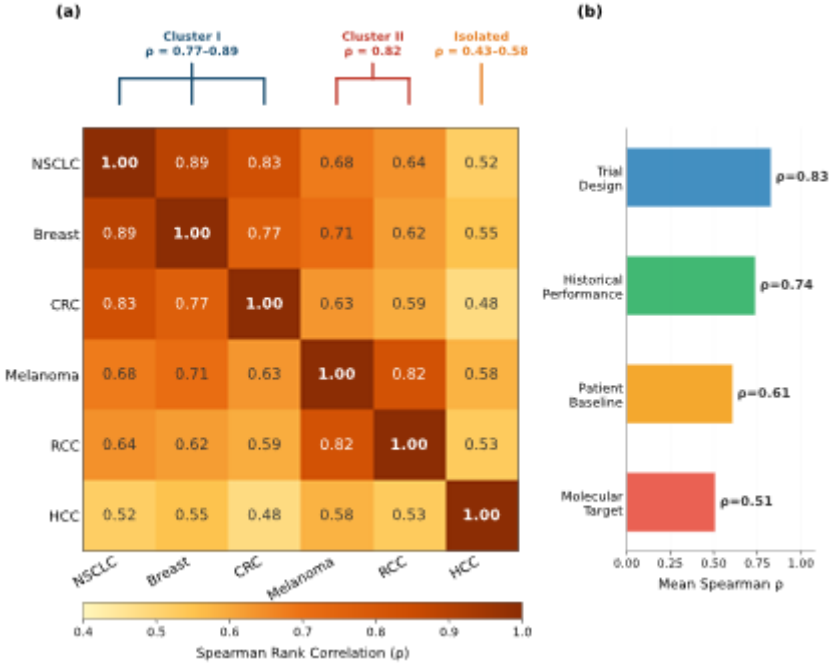

Figure 3. Cross-Indication Covariate Transferability Heatmap

Figure 3 displays a symmetric 6 × 6 heatmap of pairwise Spearman rank correlation coefficients between the SHAP feature importance vectors from indication-specific GBRT models. The six oncology indications (NSCLC, Breast, CRC, Melanoma, RCC, HCC) are arranged along both axes. Color intensity encodes correlation strength, ranging from light yellow (ϱ = 0.40) to dark red (ϱ = 0.90). Annotated numerical values appear within each cell.

The heatmap is generated using Python's `seaborn.heatmap()` with `annot=True`, `cmap='YlOrRd'`, and `vmin=0.40, vmax=0.90`. Hierarchical clustering dendrograms along both axes group indications by similarity of their feature importance profiles. The plot reveals that NSCLC, breast cancer, and CRC form a high-transferability cluster (pairwise ϱ = 0.77--0.89), while melanoma and RCC cluster separately (ϱ = 0.82), and HCC shows the weakest transferability to all other indications (ϱ = 0.43--0.58). Feature category disaggregation (shown as marginal bar plots) reveals that trial design features exhibit consistently high transferability (mean ϱ = 0.83) while molecular features show substantially lower cross-indication consistency (mean ϱ = 0.51).

4.3.2. Temporal Stability of Covariate Effects

Cetinyurek Yavuz et al. conducted a scoping review of PTS concepts and methods, identifying substantial heterogeneity in how the pharmaceutical industry defines and operationalizes probability of success metrics [22]. Their review highlighted the need for standardized approaches to assessing covariate effect stability over time, which directly motivates the temporal analysis conducted in this study.

The temporal stability assessment partitioned the 1,203 trials into three chronological cohorts: Cohort A (2005--2011, n = 312), Cohort B (2012--2017, n = 448), and Cohort C (2018--2023, n = 443). Hwang et al. analyzed Phase III trial failures and found that 57% were attributable to insufficient efficacy and 17% to safety concerns, with these proportions remaining relatively stable over time [23]. The present analysis extends this observation by examining whether the predictive relevance of specific covariates changed across the three periods.

Trial design features demonstrated the highest temporal stability (mean Kendall's τ = 0.81 across cohort pairs), while molecular features showed moderate stability (mean τ = 0.64), and patient baseline features exhibited the lowest stability (mean τ = 0.57). The declining temporal stability of patient features likely reflects evolving eligibility criteria, increasing racial and geographic diversity in trial enrollment, and the progressive shift

toward earlier lines of therapy in oncology trials. The U.S. Food and Drug Administration published its guidance on AI in drug development in January 2025, emphasizing the importance of model validation, transparency, and uncertainty quantification for regulatory submissions incorporating computational predictions [24]. This regulatory framework underscores the practical significance of temporal stability assessment for predictive covariate features: PTS estimation methods that rely on temporally unstable covariates may produce unreliable predictions when applied prospectively.

5. Conclusions and Future Perspectives

5.1. Principal Findings and Practical Implications

This study established a systematic analytical framework for identifying, ranking, and evaluating the transferability of covariate features that predict Phase III oncology trial outcomes. Three principal findings emerged from the analysis of 1,203 Phase III trials across six cancer types. The feature importance analysis demonstrated that Phase II efficacy endpoint magnitude, biomarker-driven patient selection, and sponsor oncology experience constituted the three most influential predictive covariates, collectively accounting for 52.4% of total predictive information as measured by cumulative mean absolute SHAP values. This concentration of predictive power in a small number of features has direct practical implications for pharmaceutical development teams seeking to improve PTS estimates: prioritizing rigorous Phase II endpoint assessment and biomarker stratification strategies may yield the greatest improvement in Phase III success prediction accuracy.

The comparative regression analysis revealed that gradient boosted regression trees achieved superior discriminative performance (AUC = 0.823) relative to four alternative techniques, while Bayesian additive regression trees provided competitive accuracy (AUC = 0.814) with the added advantage of native uncertainty quantification through posterior predictive distributions. The 5.0 percentage-point AUC advantage of GBRT over elastic net logistic regression indicates that covariate interaction effects carry meaningful predictive information that linear methods cannot capture. The cross-indication transferability analysis demonstrated that trial design features generalize well across cancer types (mean Spearman $\rho = 0.83$), supporting the development of pan-oncology PTS estimation approaches for design-level covariates. Molecular and target-level features exhibited lower transferability (mean $\rho = 0.51$), indicating that indication-specific calibration remains necessary for this covariate category.

5.2. Limitations

Several limitations warrant acknowledgment. The study relied on retrospective analysis of completed trials, and prospective validation of the identified covariate features and regression models on independently collected future trial cohorts is needed to confirm external generalizability. The analytical cohort was restricted to six solid tumor types, and extension to hematological malignancies and rare cancers would broaden the applicability of the findings. The 42-covariate feature space, while comprehensive relative to prior studies, did not incorporate molecular structure descriptors, genomic biomarker profiles, or real-world evidence-derived features that could further improve predictive accuracy. The temporal validation protocol used a single temporal split rather than expanding temporal cross-validation, which may underestimate prediction variance.

Future research should pursue several directions to build upon the present findings. The integration of real-world evidence from electronic health record databases with trial-level covariates represents a particularly promising avenue, as demonstrated by the growing body of evidence supporting RWE-based external control arm construction in oncology. The development of cancer type-specific calibration layers that adjust pan-oncology predictions using indication-specific molecular and biological features could address the limited cross-indication transferability observed for molecular covariates. The implementation of MBMA-derived pharmacological constraints as informative priors within Bayesian regression frameworks offers a principled approach to integrating

mechanistic pharmacological knowledge with data-driven covariate analysis. The deployment of these methods within prospective clinical trial planning workflows, with systematic tracking of predicted versus observed outcomes, would provide the definitive validation evidence needed for widespread adoption in pharmaceutical R&D decision-making.

References

1. I. Kola and J. Landis, "Can the pharmaceutical industry reduce attrition rates?" *Nature Reviews Drug Discovery*, vol. 3, no. 8, pp. 711--716, 2004.
2. J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, "Innovation in the pharmaceutical industry: New estimates of R&D costs," *Journal of Health Economics*, vol. 47, pp. 20--33, 2016.
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. C. H. Wong, K. W. Siah, and A. W. Lo, "Estimation of clinical trial success rates and related parameters," *Biostatistics*, vol. 20, no. 2, pp. 273--286, 2019.
5. D. Sun, W. Gao, H. Hu, and S. Zhou, "Why 90% of clinical drug development fails and how to improve it?" *Acta Pharmaceutica Sinica B*, vol. 12, no. 7, pp. 3049--3062, 2022.
6. A. W. Lo, K. W. Siah, and C. H. Wong, "Machine learning with statistical imputation for predicting drug approvals," *Harvard Data Science Review*, vol. 1, no. 1, 2019.
7. 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.
8. K. W. Siah, N. W. Kelley, S. Engstrom, B. Filinger, B. Tsao, and A. W. Lo, "Predicting drug approvals: The Novartis data science and artificial intelligence challenge," *Patterns*, vol. 2, no. 8, p. 100312, 2021.
9. 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, p. 100445, 2022.
10. J. W. Mandema, E. Cox, and J. Alderman, "Therapeutic benefit of eletriptan compared to sumatriptan for the acute relief of migraine pain---results of a model-based meta-analysis that accounts for encapsulation," *Cephalalgia*, vol. 25, no. 9, pp. 715--725, 2005.
11. M. Boucher and M. Bennetts, "The many flavors of model-based meta-analysis: Part I---Introduction and landmark data," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 5, no. 2, pp. 54--64, 2016.
12. V. V. Upreti and K. Venkatakrishnan, "Model-based meta-analysis: Optimizing research, development, and utilization of therapeutics using the totality of evidence," *Clinical Pharmacology & Therapeutics*, vol. 106, no. 5, pp. 981--992, 2019.
13. D. W. Thomas, J. Burns, J. Audette, A. Carroll, C. Dow-Hygelund, and M. Hay, "Clinical development success rates and contributing factors 2011--2020," *BIO Industry Analysis, Informa Pharma Intelligence, QLS Advisors*, 2021.
14. 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.
15. A. Aliper et al., "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.
16. A. V. Schperberg, P. F. Engstrom, and R. Kurzrock, "Machine learning model to predict oncologic outcomes for drugs in randomized clinical trials," *International Journal of Cancer*, vol. 147, no. 9, pp. 2537--2549, 2020.
17. Z. Teng, J. Hu, G. L. Faltys, and A. V. Bhattaram, "Model-based meta-analysis for multiple myeloma: A quantitative drug-independent framework for efficient decisions in oncology drug development," *Clinical and Translational Science*, vol. 11, no. 2, pp. 218--225, 2018.
18. P. Chan, K. Peskov, and X. Song, "Applications of model-based meta-analysis in drug development," *Pharmaceutical Research*, vol. 39, no. 8, pp. 1761--1777, 2022.
19. L. V. Hampson et al., "A new comprehensive approach to assess the probability of success of development programs before pivotal trials," *Clinical Pharmacology & Therapeutics*, vol. 111, no. 5, pp. 1050--1060, 2022.
20. T. Otsuka, R. Wada, and T. Shiroyama, "Factors associated with successful phase III trials for solid tumors: A systematic review," *Contemporary Clinical Trials Communications*, vol. 24, p. 100859, 2021.
21. G. Carrigan et al., "Using electronic health records to derive control arms for early phase single-arm lung cancer trials: Proof-of-concept in randomized controlled trials," *Clinical Pharmacology & Therapeutics*, vol. 107, no. 2, pp. 369--377, 2020.
22. A. Cetinyurek Yavuz et al., "On the concepts, methods, and use of 'probability of success' for drug development decision-making: A scoping review," *Clinical Pharmacology & Therapeutics*, vol. 117, no. 4, pp. 928--940, 2025.
23. T. J. Hwang et al., "Failure of investigational drugs in late-stage clinical development and publication of trial results," *JAMA Internal Medicine*, vol. 176, no. 12, pp. 1826--1833, 2016.
24. U.S. Food and Drug Administration, "Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products (FDA-2024-D-4689)," *U.S. Department of Health and Human Services*, 2025.