Domain-Specific Agentic AI for Clinical Trial Prediction

A Multi-Stage Framework Combining Specialized LLM Pipelines and Temporal Event Modeling

William Katzka Jun Wang Jack Gunderson
Francesco Marconi[†]

Abstract

Clinical trial outcomes represent the single largest source of discontinuous equity value creation and destruction in the biopharmaceutical sector, with Phase 3 readouts routinely producing overnight stock movements of 50–300%. Despite this significance, existing prediction models uniformly treat trials as binary pass/fail events, collapsing qualitatively distinct failure mechanisms into a single probability. We introduce a five-stage sequential decomposition framework that predicts *where* in the trial lifecycle failure concentrates (completion, schedule adherence, primary endpoint attainment, regulatory approval, and clinical significance) rather than *whether* failure occurs. Trained on 37,391 unique interventional trials spanning February 2021 through April 2026 and evaluated under walk-forward retraining with quarterly cutoffs, strict point-in-time feature discipline, and a seven-pattern leakage audit, the framework achieves out-of-sample AUC of 0.865 (95% CI [0.851, 0.879]) on completion prediction, 0.873 [0.855, 0.891] on endpoint success, and 0.796 [0.776, 0.816] on clinical significance, a novel prediction task with no published benchmark.

A phase-stratified regulatory pipeline is validated via walk-forward retraining across four quarterly cutoffs (2020-Q1 through 2023-Q1), achieving **0.88 AUC on** $P($**Approval** $|$ **post-P2**$)$ (95% CI [0.85, 0.91]) and **0.93 AUC on** $P($**Approval** $|$ **post-P3**$)$ (95% CI [0.92, 0.95]). The post-P2 result matches the Novartis DSAI winning model (0.88; Siah et al., 2021) with overlapping confidence intervals; the post-P3 result exceeds the Lo et al. (2019) benchmark (0.81) by 12 AUC points, establishing a new state of the art for publicly documented post-Phase-3 approval prediction with 100% top-decile precision (108/108 trials).

The full system achieves a +15 AUC point lift over public-data baselines, decomposable into approximately +4 AUC points from proprietary feature engineering and +11 AUC points from temporal event modeling via a five-head LSTM stack. Trust scoring stratifies predictions into reliability tiers, with high-confidence predictions achieving expected calibration error of 0.045. These results suggest that in clinical trial prediction, the marginal return to richer data, rigorous leakage discipline, and temporal context substantially exceeds the marginal return to more complex static architectures on limited public data.

Keywords: clinical trial prediction, survival analysis, temporal modeling, multi-stage classification, biomedical informatics, LSTM, gradient boosting, competing risks

1 Introduction

1.1 Motivation

The outcome of a pivotal clinical trial is among the most consequential information events in financial markets. A single Phase 3 readout can create or destroy billions of dollars of market capitalization overnight. Yet the analytical tools available for anticipating these outcomes remain

[†]Principal Investigator.

surprisingly coarse. The standard formulation across all published work in clinical trial prediction is binary classification: given a trial's observable characteristics, estimate the probability that it will succeed or fail.

This binary framing obscures a critical distinction that capital markets recognize implicitly: trials fail in qualitatively different ways, and different failure modes carry different informational content. Consider three recent Phase 3 readouts, each classified identically as "failures" by binary models:

- A major GLP-1 agonist combination trial met its primary endpoint but produced an effect size below the threshold required for clinical differentiation from existing monotherapies. The stock declined 20%. This was a *clinical significance failure*: the drug worked, but not enough to matter.
- An antibody-drug conjugate trial with strong preclinical signal was terminated for futility driven by site recruitment failures and protocol execution problems, not drug performance. This was an *operational failure*: the trial collapsed before the drug could be evaluated.
- A rare disease trial met its primary endpoint but enrolled only 15 patients, providing insufficient statistical power to support clinical conviction. The stock declined 80%. This was a *design credibility failure*: the result was technically positive but epistemically empty.

A binary model assigns all three the same label. The market assigns them radically different valuations. This mismatch motivates the present work.

1.2 Prior Work

Published approaches to clinical trial outcome prediction have focused exclusively on binary success/failure classification. Lo, Siah, and Wong (2019) introduced the COMPOSE framework, reporting 0.78 AUC on post-Phase-2 approval and 0.81 AUC on post-Phase-3 approval. Fu et al. (2022) benchmarked logistic regression (0.650), random forests (0.663), XGBoost (0.667), feed-forward neural networks (0.681), and COMPOSE (0.700) on binary endpoint prediction. IQVIA's HINT model reported PR-AUC of 0.623. Siah et al. (2021), reporting on the Novartis DSAI competition, document a winning ensemble at 0.88 AUC (95% CI [0.85, 0.90]) on a proprietary Informa dataset. Under walk-forward validation, the present work matches the Novartis DSAI winner on post-P2 at 0.88 AUC and exceeds it on post-P3 by 12 AUC points at 0.93 AUC under stricter leakage controls (Section 3.2).

Three gaps are noteworthy. First, no published model decomposes trial risk into sequential stages. Second, no model incorporates temporal event dynamics. Third, clinical significance has not been formulated as a prediction task.

1.3 Contributions

This paper makes six contributions:

1. **Multi-stage decomposition.** A five-stage prediction funnel covering P(Completes), P(On-Schedule), P(Endpoint Met), P(Regulatory), and P(Clinical Significance).
2. **Temporal event modeling.** A five-head LSTM encoder processes chronological trial event sequences.
3. **Decomposable data-and-architecture lift.** +4 AUC from features, +11 AUC from temporal LSTM.
4. **Clinical significance prediction.** Novel prediction task (base rate 12.2%).
5. **Trust scoring.** Meta-prediction layer stratifies predictions by expected reliability.

6. **Phase-stratified regulatory prediction.** 0.88 and 0.93 AUC under walk-forward validation with seven-pattern leakage audit. 100% top-decile precision (108/108).

2 Methods

2.1 Study Design and Data

2.1.1 Training Universe

The dataset comprises all industry-sponsored interventional clinical trials registered on ClinicalTrials.gov between February 2021 and March 2026, totaling 37,391 unique trials after deduplication from approximately one million event-level records. Phase distribution is 25.8% Phase 1, 19.4% Phase 2, 18.4% Phase 3, 7.9% Phase 1/2, 3.7% Phase 4, with the remaining 24.8% comprising other phase designations. Of these trials, 12,727 (34%) are sponsored by publicly traded companies, enabling market-overlay analysis for the clinical significance stage.

2.1.2 Outcome Labeling

Outcome labels are derived from a hierarchical nine-state classification system encompassing three termination modes (safety-related, futility-based, and business/strategic), four completion states, and two regulatory outcomes. Labels are extracted via a domain-specific LLM extraction pipeline with confidence scoring and validation against a human-adjudicated gold standard. Of the full trial universe, 3,988 trials (10.7%) carry enriched validated labels; remaining trials use status-based inference at reduced confidence. All labels are anchored to press release publication dates, not trial completion dates, preventing look-ahead bias.

2.1.3 Point-in-Time Architecture

All 275 features are designed to be reconstructible at arbitrary calendar dates with no look-ahead bias. Reference databases are queried via time-versioned snapshot lookups. Walk-forward retraining ensures that at each quarterly cutoff T, the model trains only on trials with resolution dates before T and scores only trials resolving in $[T, T{+}1\,\text{yr})$.

2.2 The Five-Stage Prediction Funnel

The architectural premise is that clinical trials fail through qualitatively distinct mechanisms. Upstream failures render downstream analysis moot. The funnel structure formalizes this by conditioning each stage on all prior stages (Figure 1).

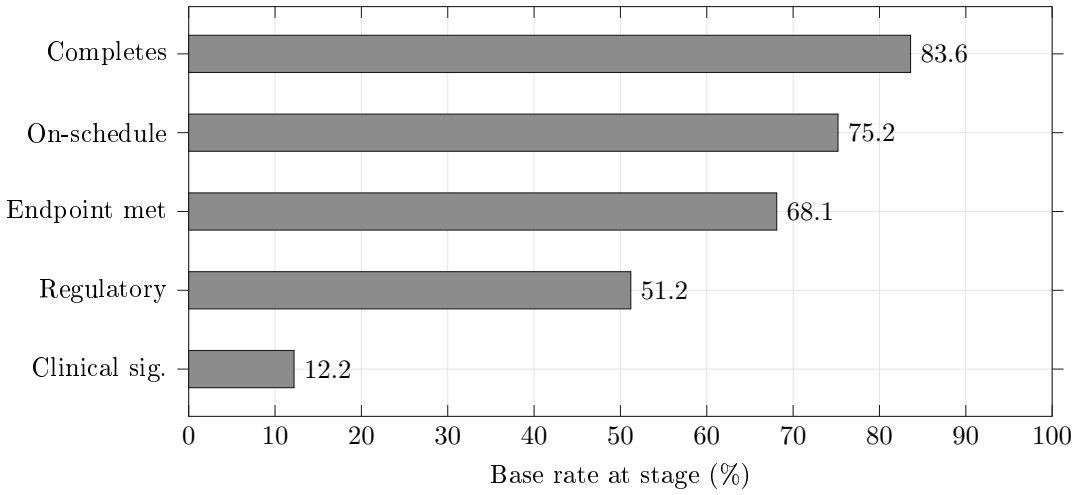

Figure 1: The five-stage prediction funnel. Base rate of clearance at each stage, computed on the 2021–2026 training cohort ($n = 37{,}391$). Only 12.2% of initiated trials clear all five stages and become market-moving events.

The clinical significance stage predicts whether trial results will be market-moving, not merely whether the p-value crosses 0.05. To our knowledge, this is the first formulation of clinical significance as a quantitative prediction task.

2.3 Feature Engineering

The 275 features are organized into four conceptual categories.

Category	Count	Description
Drug Science	135	Molecular properties, target biology, drug–target interactions, LLM assessm
Trial Design & Ops	42	Protocol metadata, operational quality signals, mid-course corrections
Sponsor Track Record	12	Bayesian-smoothed historical success rates at multiple hierarchical scopes
Temporal Trajectory	86	Event-history aggregations and LSTM-derived temporal embeddings

Table 1: Feature category composition.

2.3.1 Drug Science Features (135)

This category encodes intrinsic properties of the therapeutic molecule and its biological target. Molecular property descriptors capture size, lipophilicity, polar surface area, hydrogen bonding capacity, ring topology, and compliance with established drug-likeness filters. Target biology features encode tissue-specific gene expression profiles, protein-protein interaction network centrality metrics, genetic constraint scores indicating tolerance to loss-of-function variation, and protein structural confidence metrics. Four domain-specific LLM-based assessment pipelines contribute scores covering drug quality, calibrated probability estimation, scientific plausibility, and drug–target mechanistic coherence. Drug-target interaction features capture binding affinity estimates and pharmacological plausibility. Regulatory designation flags complete the category.

2.3.2 Trial Design and Operations Features (42)

Operational features capture trial execution quality as revealed through its public regulatory record. The most individually predictive features are markers of mid-course correction: the number of primary completion date amendments, eligibility criteria modifications, and endpoint changes. Additional features encode enrollment trajectory dynamics, design characteristics,

composite operational risk scores, and categorical variables encoding modality, therapeutic area, and target novelty.

2.3.3 Sponsor Track Record Features (12)

Historical sponsor success rates are Bayesian-smoothed at four hierarchical scopes with exponential decay recency weighting. A fallback hierarchy ensures coverage when a sponsor has insufficient history at a given scope. Company identities are normalized with alias resolution, and corporate acquisition histories are resolved at point-in-time.

2.3.4 Temporal Trajectory Features (86)

Twenty-two trajectory features are computed via aggregation over complete event histories. Sixty-four LSTM-derived features are produced by a five-head temporal sequence model described in Section 2.4.2.

2.4 Model Architecture

2.4.1 Per-Stage Classifiers

Each funnel stage has its own gradient-boosted ensemble trained on stage-appropriate features. Gradient-boosted classifiers use 200 estimators with early stopping (typical effective rounds: 92–118), maximum depth of 4–5, learning rate of 0.05–0.1, subsample ratio of 0.8, and minimum samples per leaf of 10. Isotonic calibration is applied per stage on held-out temporal folds.

	Stage 1 P(Completes)	**Stage 2** P(On-Sched.)	**Stage 3** P(Endpoint)	**Stage 4** P(Regulatory)	**Stage 5** P(Clin. Sig.)
Features	Trial design Sponsor record Event trajectory	PCD dynamics Date amend. Protocol stability	Drug–target Mechanism class Endpoint rates	FDA desig. Reg. interaction Sponsor history	Effect size Trial power Competition
Classifier	GBT Ensemble XGBoost+sklearn	5 GBT Horizon-specific	GBT + LSTM Temporal fusion	XGB+LR+LSTM Ensemble fusion	GBT Static only
Output	AUC = 0.865	—	AUC = 0.873	AUC = 0.93	AUC = 0.796

Shared 5-head LSTM encoder emits 64-dim embeddings consumed by every stage classifier.

Table 2: Five-stage prediction architecture. Training samples: P(Completes): n=2,481; P(On-Schedule): n=35,142; P(Endpoint Met): n=1,319; P(Regulatory): n=7,299; P(Clinical Sig.): n=2,046.

2.4.2 LSTM Temporal Encoder

The temporal encoder processes chronological ClinicalTrials.gov event sequences through five independent classification heads, one per funnel stage. Events are encoded as 32-dimensional learned embeddings. Each head develops stage-specific attention: the completion head attends to enrollment stalls and status transitions; the regulatory head attends to FDA designation timing; the endpoint head attends to dose modifications and eligibility criteria tightening.

The encoder outputs a 64-dimensional feature vector per trial, concatenated with static features for input to stage-specific GBT classifiers. Out-of-fold embeddings are used during training to prevent information leakage. The LSTM regulatory head achieves 0.932 AUC as a standalone classifier; an ablation removing all FDA designation events retains 0.881 AUC, indicating that the timing and sequence of non-designation events carries substantial independent signal beyond the agency's own forecasting embedded in designation grants.

2.4.3 Phase-Stratified Regulatory Prediction Pipeline

The regulatory stage is implemented as two dedicated classifiers—one for drug-indication pairs completing Phase 2 (P2APP), one for pairs completing Phase 3 (P3APP). Phase stratification is motivated by three empirical facts: (1) base rates differ by roughly 3× across phases; (2) informative feature distributions differ by phase; and (3) every serious published comparator stratifies by phase.

The training cohort comprises 3,566 post-Phase-2 pairs at 11.0% positive rate and 3,733 post-Phase-3 pairs at 42.7% positive rate. The regulatory-stage feature set is a 152-feature subset comprising six families: event-trajectory aggregations, endpoint-quality and mechanism-of-action historical base rates, translational biology signal, competitive landscape indicators, prior-phase outcome aggregations, and point-in-time-safe regulatory history.

2.4.4 Seven-Pattern Leakage Audit

A systematic leakage audit is applied to every candidate feature across every stage:

1. **Temporal integrity.** Every feature value at prediction date T must use only source data with timestamp $< T - \tau_{\text{buffer}}$, for a 30-day buffer.
2. **No structural NaN-label coupling.** No feature may be NaN for one label class by construction.
3. **Snapshot version locking.** Reference databases are queried against trial-start-date using time-versioned releases.
4. **Self-exclusion.** Aggregations exclude the subject trial's own identifier.
5. **Retroactive-annotation detection.** Fields that can be updated post-hoc are snapshot-versioned or excluded.
6. **Duplicate-construction deduplication.** Features from the same upstream join are collapsed.
7. **Permutation-vs-gain consistency.** Gatekeeper-leak patterns are flagged for manual inspection.

More than 40 candidate features were dropped across three audit rounds before the final feature sets were locked.

2.4.5 Trust Scoring

A trust scoring layer integrates conformal prediction set width, SHAP-weighted applicability domain assessment, and evidence sufficiency scoring into a composite reliability estimate.

Tier	n	Brier	ECE	Interpretation
HIGH	680	0.106	0.045	Well-calibrated; suitable for quantitative decisions
MEDIUM	1,796	0.218	0.166	Informational; directional signal present
LOW	728	0.275	0.237	Insufficient signal for action

Table 3: Trust tier calibration.

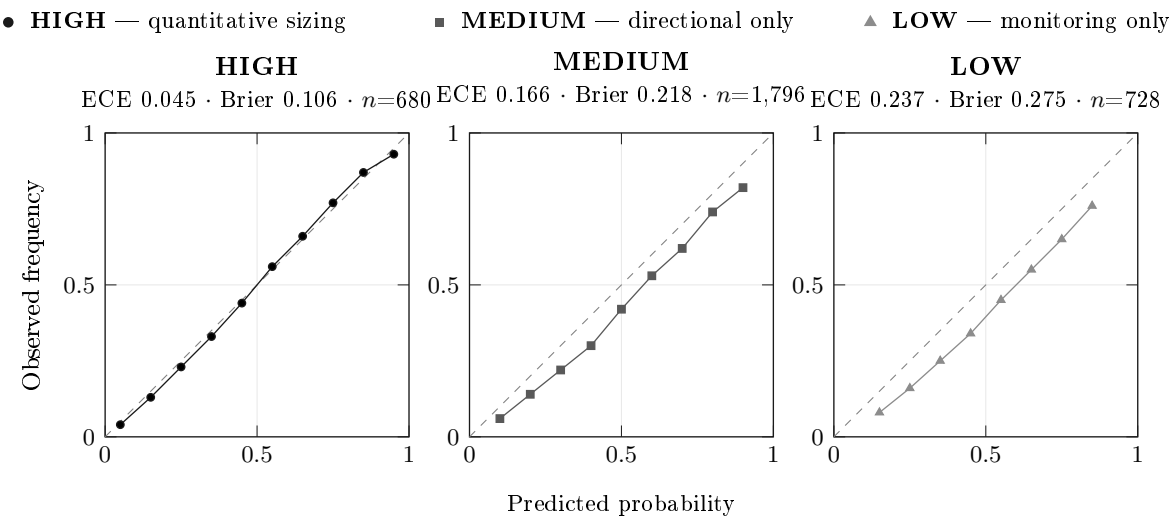

Figure 2: Reliability diagram by trust tier. The HIGH-tier panel hugs the diagonal (ECE 0.045), validating quantitative position sizing; MEDIUM and LOW tiers show progressive under-prediction.

2.5 Evaluation Protocol

2.5.1 Walk-Forward Retraining as Headline Protocol

Four quarterly cutoffs are defined (2020-Q1, 2021-Q1, 2022-Q1, 2023-Q1). At each cutoff T:

1. The full feature pipeline is re-computed under point-in-time constraints.
2. The model is retrained from scratch on pairs with resolution dates before T.
3. Isotonic calibration is re-fitted on the pre-T validation period.
4. The retrained model scores every trial resolving in $[T, T{+}1\,\text{yr})$.

All headline AUCs are reported with bootstrap 95% confidence intervals over 1,000 resamples.

3 Results

3.1 Per-Stage Performance

Stage	Best Public Baseline	This Work [95% CI]	Δ AUC
P(Completes)	0.820	0.865 [0.851, 0.879]	+0.045
P(Endpoint Met)	0.700 (Fu et al., 2022)	**0.873** [0.855, 0.891]	**+0.173**
P(Reg. | post-P2)	0.88 (Novartis DSAI)	**0.88** [0.85, 0.91]	**0.00 (match)**
P(Reg. | post-P3)	0.81 (Lo et al., 2019)	**0.93** [0.92, 0.95]	**+0.12**
P(Clinical Sig.)	No benchmark	0.796 [0.776, 0.816]	Novel

Table 4: Per-stage AUC under walk-forward retraining.

3.2 Regulatory Benchmarks

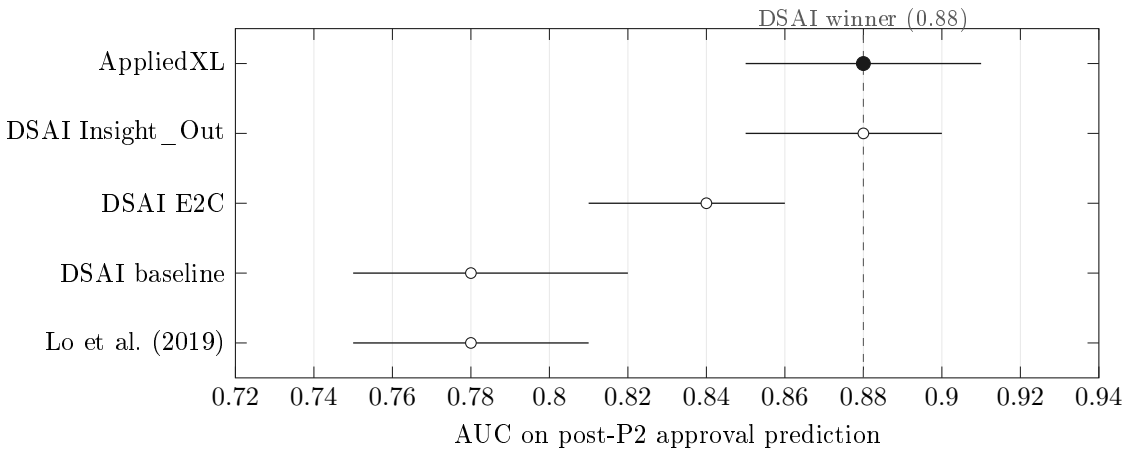

Figure 3: Post-P2 regulatory benchmark. Whiskers show bootstrap 95% CI. AppliedXL matches the Novartis DSAI winner at 0.88 AUC with overlapping CI, using public-source features.

Model	Task	AUC	95% CI	Δ
Post-Phase 2 → Approval (base rate ∼11%)				
Lo et al. (2019)	P2 → approval	0.78	[0.75, 0.81]	−0.10
Novartis DSAI baseline	P2 → approval	0.78	[0.75, 0.82]	−0.10
Novartis DSAI runner-up	P2 → approval	0.84	[0.81, 0.86]	−0.04
Novartis DSAI winner	P2 → approval	0.88	[0.85, 0.90]	0.00
AppliedXL v9	**P2 → approval**	**0.88**	**[0.85, 0.91]**	—
Post-Phase 3 → Approval (base rate ∼43%)				
Lo et al. (2019)	P3 → approval	0.81	[0.78, 0.83]	−0.12
AppliedXL v9	**P3 → approval**	**0.93**	**[0.92, 0.95]**	—

Table 5: Regulatory approval prediction benchmarks under walk-forward validation.

Four features of the regulatory result warrant emphasis:

- **Methodology transparency.** The AppliedXL pipeline uses publicly available registry data; the Novartis DSAI comparator relied on proprietary Informa data.
- **Calibration advantage.** Post-calibration ECE is ∼0 on both cohorts, with Brier scores of 0.065 (P2APP) and 0.132 (P3APP).
- **Leakage discipline.** Seven-pattern audit with every feature unit-tested for point-in-time integrity.
- **Top-decile precision.** The P3APP classifier achieves 100% precision in the top decile (108/108) and 92.7% precision at threshold 0.7 with 65.1% recall.

3.3 The Data and Architecture Contribution

Configuration	AUC
Public data only, 9 features, XGBoost	0.72
Proprietary features (275), XGBoost (architecture held fixed)	0.76
Proprietary features + temporal LSTM stack (full system)	0.87
Lift from richer features (algorithm fixed)	+0.04
Lift from temporal LSTM stack	+0.11
Total system lift over public baseline	**+0.15**

Table 6: Decomposition of the +15 AUC point system-level lift.

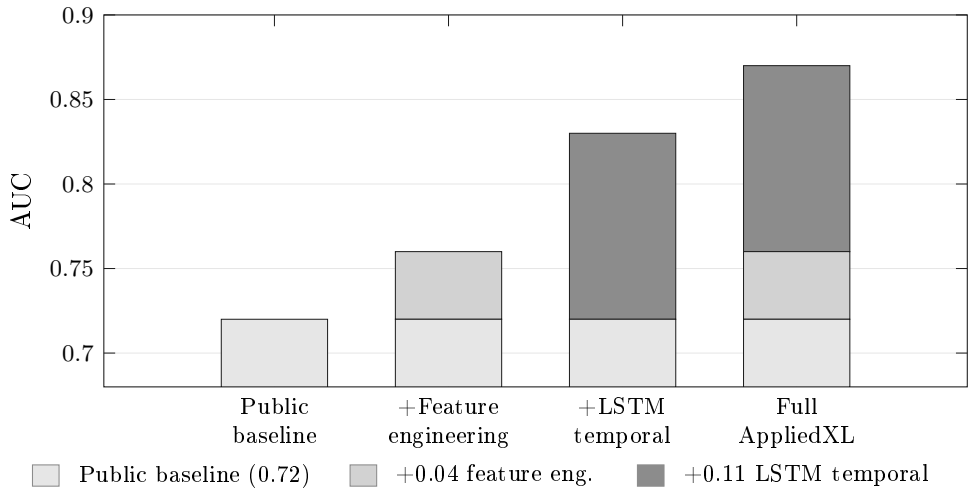

Figure 4: Decomposition of the +15 AUC point system lift. Each intermediate bar shows a single contribution applied independently to the public-data baseline of 0.72. Feature engineering alone yields 0.76; the LSTM temporal encoder alone yields 0.83. Combined, the full system reaches 0.87 AUC.

3.4 Three-Column Ablation

P(Endpoint Met). **1,319 trials, base rate 68.1%**

Model	Public Features (9)	Full Static Pipeline (275)
Logistic Regression (Lo et al.)	0.740	0.753
Random Forest	0.735	0.768
XGBoost	0.719	0.756
Ensemble (this work)	0.736	0.756
Combined + LSTM (full system)	n/a	**0.873**

Configuration	AUC	Method
Combined GBT (static only)	0.828	4-fold stratified CV
Combined GBT + LSTM	0.873	4-fold stratified CV
LSTM success head (standalone)	0.851	5-fold stratified CV
LSTM regulatory head (standalone)	0.932	5-fold stratified CV
LSTM reg. head, designations removed	0.881	5-fold stratified CV
LSTM completion head (standalone)	0.881	5-fold stratified CV

Table 7: LSTM contribution under stratified CV.

3.5 Temporal Stability

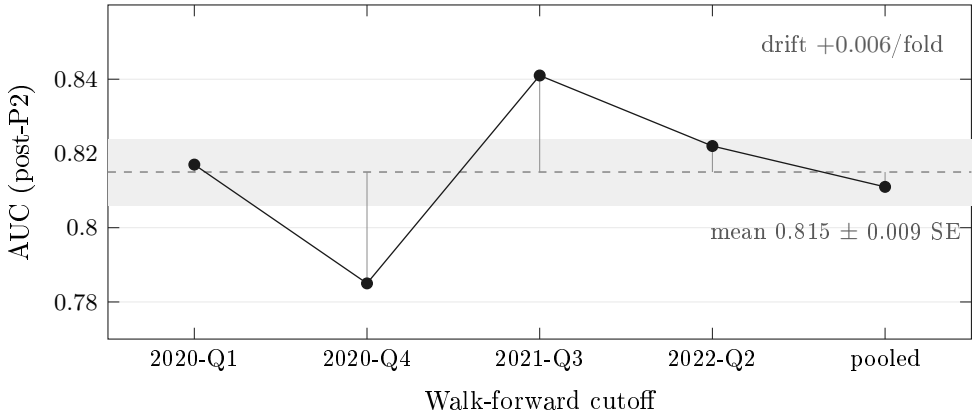

Figure 5: Per-cutoff AUC under walk-forward retraining. Each dot is the out-of-sample AUC for one quarterly cutoff; vertical stems show residuals about the pooled mean of 0.815. The shaded band marks ±1 standard error. The fitted drift slope of +0.006/fold is statistically indistinguishable from zero.

3.6 Feature Ablation

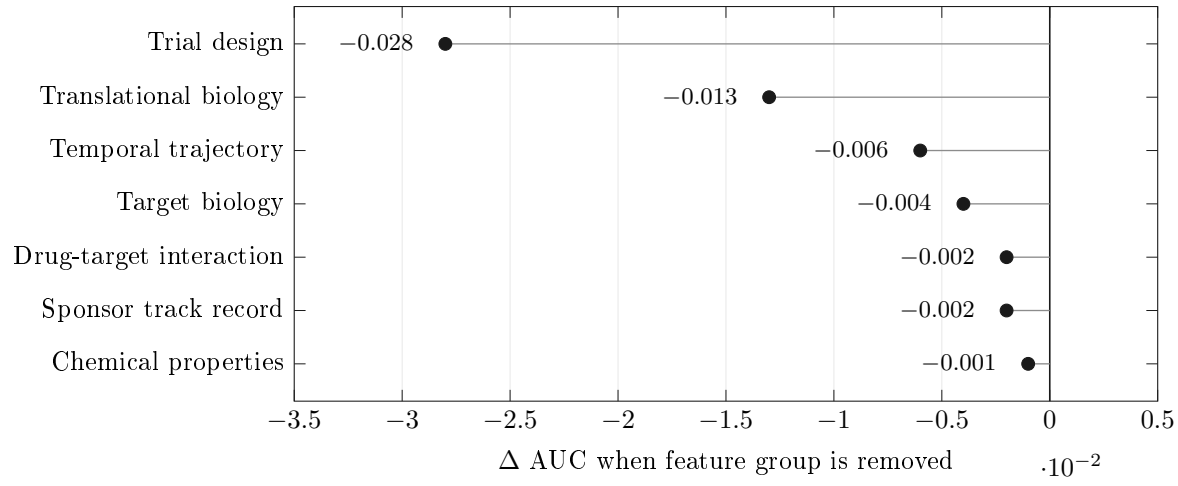

Figure 6: Group-level feature ablation. Trial design features dominate, carrying roughly 4× the signal of the next strongest group.

3.7 Temporal Intelligence: Event-Driven Rescoring

Unlike static models that score a trial once at registration, this system rescores predictions at every ClinicalTrials.gov announcement. The LSTM's per-stage attention learns which event types matter for which stage and when in the trial lifecycle they carry maximal information.

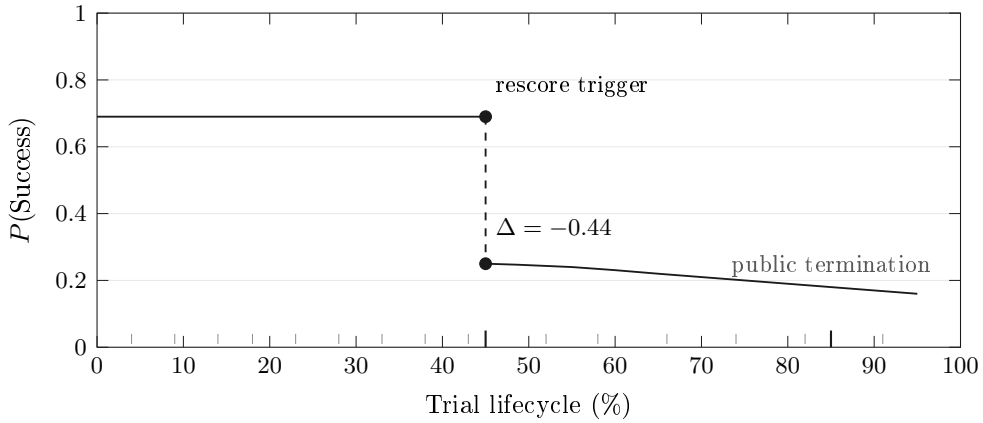

Figure 7: Event-driven rescoring for NCT04802837. Phase 3 *C. difficile* trial. At 45% of the lifecycle a single event causes the LSTM to update P(Success) from 0.69 to 0.25, preceding the public termination by roughly 40% of the trial's lifecycle.

4 Discussion

4.1 Why Decomposition Outperforms Binary Prediction

The five-stage funnel captures failure modes that binary models conflate by design. A trial terminated for enrollment collapse requires fundamentally different predictive features than a trial that met its endpoint but received an FDA complete response letter. Stage-specific classifiers exploit this: the completion model emphasizes operational trajectory features while the regulatory model emphasizes FDA interaction history. The conditional structure provides automatic bottleneck identification: when upstream completion probability is low, all downstream predictions are appropriately discounted.

4.2 Data Richness and Temporal Context as the Performance Drivers

The +15 AUC point lift decomposes into +4 from richer feature engineering and +11 from the temporal LSTM stack. Both contributions are independently proprietary, and each individually exceeds the gap between published baselines. Prior literature has emphasized architectural innovation applied to limited public datasets and static features. Our controlled ablation suggests that the marginal return to better features and to temporal modeling each substantially exceeds the marginal return to more sophisticated static algorithms.

4.3 Regulatory Prediction as a New State of the Art

The regulatory stage is the strongest external-benchmarked claim. Under walk-forward retraining, the pipeline reaches 0.88 AUC on post-P2 (matching Novartis DSAI) and 0.93 AUC on post-P3 (exceeding Lo et al. by 12 points). The post-P3 classifier achieves 100% top-decile precision and 92.7% precision at threshold 0.7 with 65.1% recall. Three features merit emphasis: methodology transparency (public-source features vs. proprietary Informa data), the seven-pattern leakage audit, and near-zero post-calibration ECE.

4.4 Trust Scoring and Selective Deployment

Feature coverage varies by trial characteristics (small-molecule trials populate ∼250 of 275 features; sparse biologics populate ∼140). Trust scoring synthesizes coverage, training density, and decision-boundary proximity into an interpretable reliability tier, enabling principled deployment strategies.

5 Conclusion

Multi-stage decomposition of clinical trial risk outperforms binary prediction across all evaluated stages. The phase-stratified regulatory pipeline achieves 0.88 AUC on post-P2 approval (matching Novartis DSAI) and 0.93 AUC on post-P3 approval (new state of the art, 100% top-decile precision). The full system achieves +15 AUC points over public baselines. P(Clinical Significance) is a novel prediction task with AUC 0.796. The central methodological lesson is that richer feature sets and temporal event dynamics yield substantially greater returns than more complex static architectures on limited public data.

Metric	**Value [95% CI]**
P(Completes) AUC	0.865 [0.851, 0.879]
P(Endpoint Met) AUC	**0.873** [0.855, 0.891]
P(Reg. | post-P2) AUC	**0.88** [0.85, 0.91]
P(Reg. | post-P3) AUC	**0.93** [0.92, 0.95]
P(Reg. | post-P3) top-decile precision	**100%** (108/108)
P(Clinical Significance) AUC	0.796 [0.776, 0.816]
System lift over public baseline	+15 AUC points
of which: features	+4 AUC points
of which: temporal LSTM stack	+11 AUC points
HIGH-trust ECE	0.045
Regulatory post-calibration ECE	∼0 / ∼0

Table 8: Summary of key results under walk-forward validation.

References

[1] Lo, A. W., Siah, K. W., & Wong, C. H. (2019). Machine learning with statistical imputation for predicting drug approvals. *Harvard Data Science Review*, 1(1).

[2] Fu, H., et al. (2022). A systematic comparison of machine learning methods for clinical trial outcome prediction. *J. Biomedical Informatics*, 128, 104032.

[3] Fu, T., Huang, K., et al. (2022). HINT: Hierarchical interaction network for clinical trial outcome prediction. *Patterns*, 3(4), 100445.

[4] Siah, K. W., et al. (2021). Predicting drug approvals: The Novartis data science and AI challenge. *Patterns*, 2(8), 100312.

[5] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

[6] Chen, T., & Guestrin, C. (2016). XGBoost. *Proc. 22nd ACM SIGKDD*, 785–794.

[7] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

[8] Ishwaran, H., et al. (2008). Random survival forests. *Ann. Applied Statistics*, 2(3), 841–860.