Supplementary Information

Clinical trial success prediction from open registry text using classical machine learning

Michael R. Doane

Iambic Therapeutics, San Diego, CA, USA

Draft v1.0 | 11 August 2026

Contents

Supplementary Methods

Human-reference challenge-set validation

The manually annotated CTO 2020 to 2024 set was treated as an untouched post-fit challenge set. The official release describes the set as difficult cases remaining after easier rule-based classifications were removed. NCT identifiers were normalized and matched to saved all-to-all predictions. Duplicate human records and conflicting predictions were audited explicitly. The prespecified primary model was random forest, evaluated separately by phase. Metrics were computed against human labels without retraining, model selection, threshold selection, exclusion-rule changes, or recalibration. Confidence intervals used 2,000 unique-NCT bootstrap replicates with seed 42. Calibration intercept and slope were diagnostic only.

Duplicate-NCT sensitivity

Duplicate NCT groups were audited within each phase cohort. Groups had identical final model text, rounded binary labels, phase, cutoff side, eligibility, and neuroscience membership. Some groups had different continuous CTO probabilities. The sensitivity selected one canonical source row per NCT and refit the nine random-forest configurations. Standard metrics were compared across the original row-weighted test, a unique-NCT test using the frozen model, and a unique-NCT test using the deduplicated training model. Paired comparisons used identical test NCTs and bootstrap indices.

Organizational-field ablation

Three deduplicated all-indication random forests were refit after removing the dedicated Sponsor, Collaborators, and Locations fields. All other predictors, hyperparameters, exclusions, temporal rules, seeds, and evaluation procedures were unchanged. The ablated vocabularies were fitted from the modified pre-2019 text. Full and ablated predictions were compared on identical unique-NCT temporal test sets and on the human challenge set.

Residual-language sensitivity

The feature audit identified six clear retrospective contexts linked to two training NCTs and two held-out NCTs. The sensitivity excluded the two training NCTs, removed only the smallest adjudicated retrospective clauses from the two test NCTs, and refit the nine affected Phase 3 configurations. Ambiguous and prospective contexts were not changed. Effects of test exclusion, test-clause correction, training exclusion, and combined remediation were estimated separately.

Feature interpretation

Coefficients and impurity-based feature importances were extracted from the saved estimators and matched to the fitted TF-IDF vocabularies. The top 200 features per run were mapped to source fields using the fitted tokenizer and assigned to a reproducible category system. Logistic coefficients were interpreted by sign; tree importances were treated as nondirectional. Feature stability was summarized by Jaccard overlap and ranked-list similarity. No SHAP analysis or outcome-conditioned feature selection was performed.

Neuroscience cohort definition

The approved doctoral praxis Appendix B contained 772 source entries grouped into 491 concepts and 17 categories. Case-insensitive literal substring matching against the ClinicalTrials.gov Conditions field reproduced all 12,909 Phase 1, 12,088 Phase 2, and 7,109 Phase 3 neuroscience records. Missing Conditions values were treated as nonmatches. The historical filter retained broad terms, spelling variants, historical names, and 101 short uppercase abbreviations. It was designed for broad capture and was not clinically validated for sensitivity or specificity. The complete dictionary and matching audit are provided as a separate Supplementary Table S1 workbook.

Supplementary Table S2. Human-reference challenge-set validation

Phase	Matched NCTs	Success	Failure	Prevalence	RF ROC-AUC	95% CI	Mean probability	Calibration intercept	Calibration slope
1	2,287	604	1683	0.264	0.668	0.642 to 0.692	0.772	-2.268	0.879
2	3,004	1224	1780	0.407	0.669	0.650 to 0.689	0.772	-2.340	1.520
3	1,969	1195	774	0.607	0.686	0.664 to 0.710	0.833	-1.968	1.461

The challenge set was enriched for difficult cases and is not representative of the full trial population. Raw probabilities were not recalibrated with human outcomes.

Supplementary Table S3. One-NCT random-forest sensitivity

Phase	Experiment	Original AUC	Unique-test AUC	Deduplicated AUC	Original Brier	Unique-test Brier	Deduplicated Brier
1	All to all	0.7384	0.7455	0.7463	0.1209	0.1196	0.1194
1	All to neuro	0.7100	0.7209	0.7273	0.1453	0.1430	0.1421
1	Neuro to neuro	0.7016	0.7101	0.7127	0.1460	0.1441	0.1439
2	All to all	0.6185	0.6175	0.6207	0.1590	0.1600	0.1597
2	All to neuro	0.6256	0.6326	0.6270	0.1700	0.1667	0.1671
2	Neuro to neuro	0.6070	0.6155	0.6191	0.1715	0.1682	0.1679
3	All to all	0.6770	0.6718	0.6772	0.1187	0.1234	0.1234
3	All to neuro	0.6735	0.6695	0.6742	0.1318	0.1347	0.1347
3	Neuro to neuro	0.6577	0.6553	0.6533	0.1330	0.1357	0.1361

Supplementary Table S4. Organizational-field ablation

Phase	Train NCTs	Test NCTs	Full AUC	Ablated AUC	Delta AUC	95% CI	Full Brier	Ablated Brier	Human AUC full/ablated
1	23,946	12,364	0.7463	0.7367	-0.0096	-0.0132 to -0.0058	0.1194	0.1206	0.6669 / 0.6569
2	24,788	9,719	0.6207	0.6035	-0.0172	-0.0244 to -0.0102	0.1597	0.1616	0.6736 / 0.6796
3	16,952	5,379	0.6772	0.6546	-0.0226	-0.0359 to -0.0103	0.1234	0.1244	0.6900 / 0.7060

Ablated models excluded the dedicated Sponsor, Collaborators, and Locations fields. Organizational names could remain in other retained narrative fields.

Supplementary Table S5. Residual-language sensitivity

Remediation	Affected configurations	ROC-AUC change range	PR-AUC change range	Brier change range	Log-loss change range
Test-NCT exclusion	9 Phase 3	-0.0013 to +0.0007	+0.0000 to +0.0006	-0.0008 to +0.0002	-0.0021 to +0.0004
Test-clause correction	9 Phase 3	0 to +0.0002	about 0 to +0.0000	-0.0000 to +0.0000	-0.0001 to 0
Training exclusion and refit	9 Phase 3	-0.0060 to +0.0001	-0.0021 to +0.0009	-0.0000 to +0.0005	-0.0003 to +0.0016
Combined remediation	9 Phase 3	-0.0054 to +0.0001	-0.0021 to +0.0009	-0.0000 to +0.0004	-0.0003 to +0.0014

All combined-remediation paired 95% confidence intervals included zero. The sensitivity addressed two training NCTs and two held-out NCTs and does not establish that all retained registry text was historically prospective.

Supplementary Table S6. Feature-interpretability summary

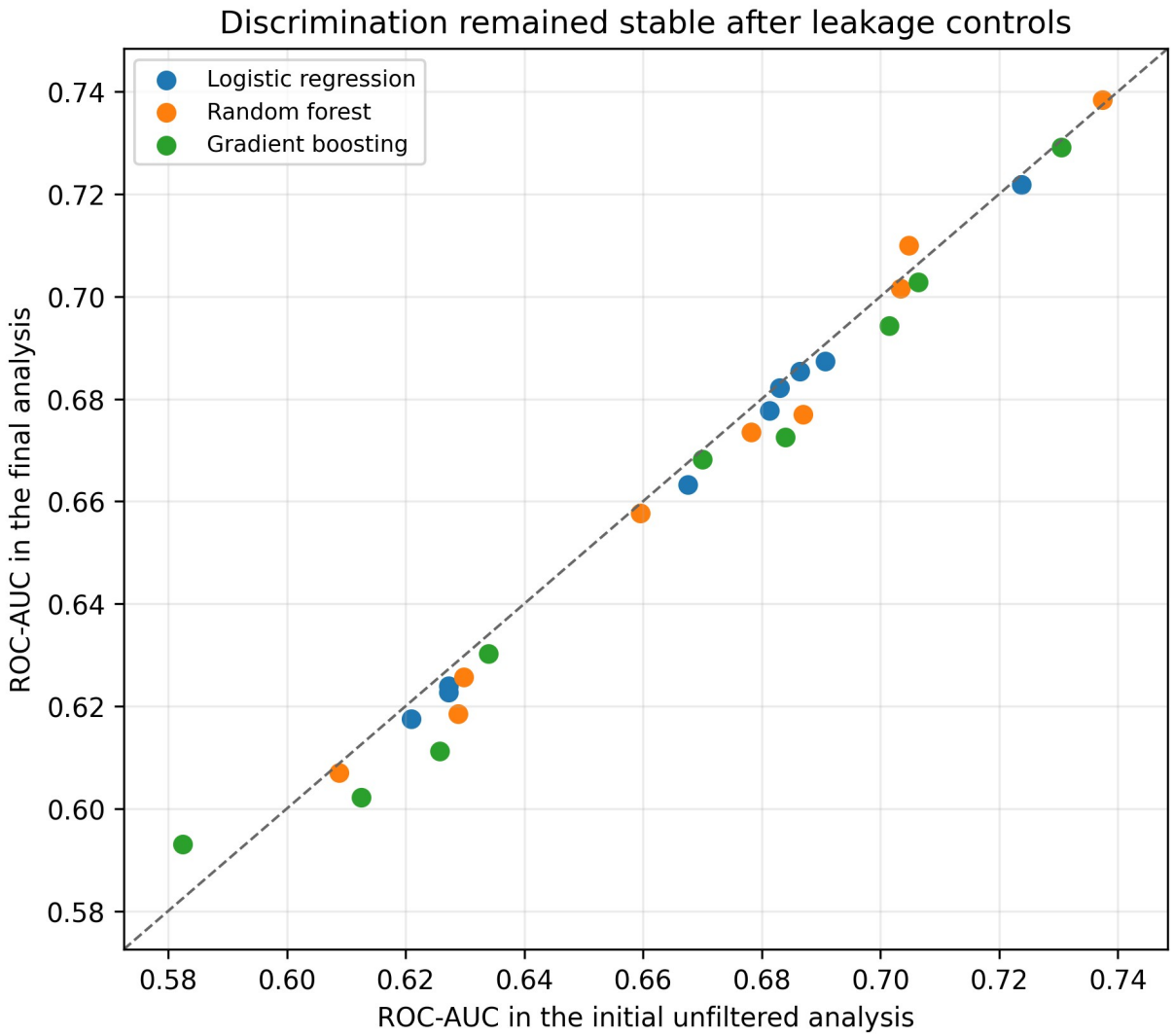
Model	Median top-100 share	Features for 50%	Organizational/ geographic terms in 300 top-100 entries	Cross-phase top-100 Jaccard	Interpretation
Logistic regression	0.074	1,148	15	0.042	Highly diffuse signed coefficients; low cross-phase stability.
Random forest	0.222	507	21	0.550	Moderately diffuse and the most stable across phases and domains.
Gradient boosting	0.722	43	12	0.070	Concentrated importance with low overlap with random forest.

Tree importance is nondirectional and should not be interpreted causally. Category counts refer to automated, auditable feature classifications and retain an unclear or mixed category when assignment was uncertain.

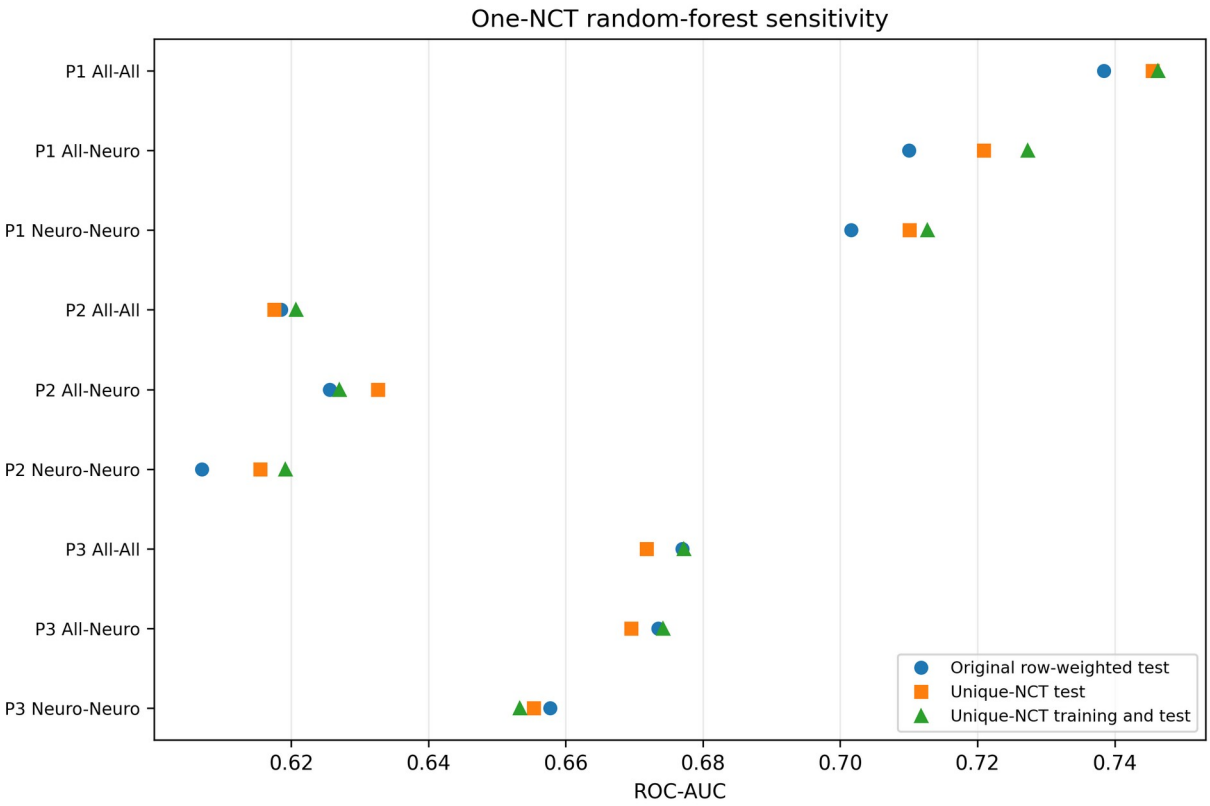
Supplementary Table S7. CTO label-source overlap audit

CTO label source	Direct model input	Possible indirect overlap	Mitigation	Residual concern
Publications	No	Intervention and disease text may correlate with publication availability	No publication text used	Low
Phase progression and	No direct progression or	Phase and intervention	Separate models by	Moderate

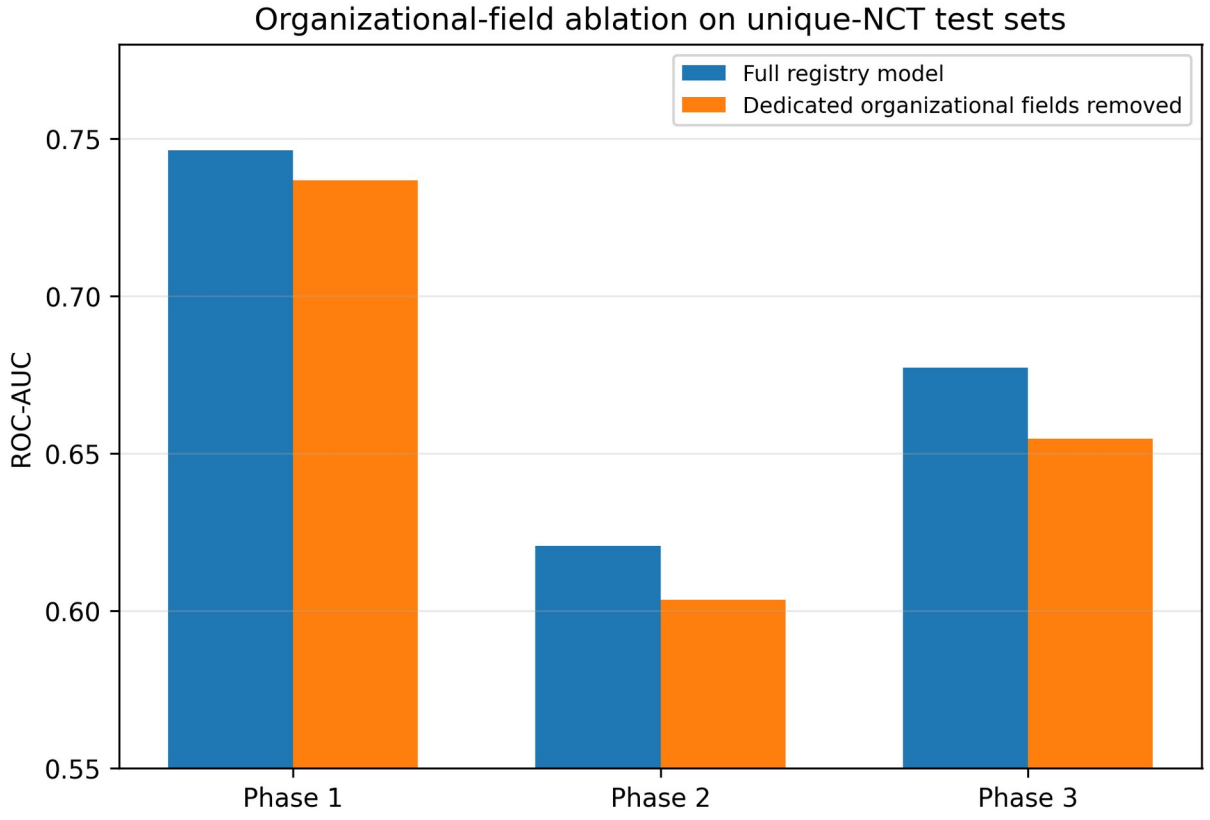
CTO label source	Direct model input	Possible indirect overlap	Mitigation	Residual concern
FDA linkage	FDA feature	identify development context	phase; human-reference validation	
News	No	Sponsor and intervention identity may correlate with news coverage	Organizational-field ablation	Moderate
Stock movement	No	Sponsor identity may indicate public-company status	Organizational-field ablation	Moderate
Study status and result metadata	No	None after explicit field blocking	Status, results, dates, enrollment, and duration blocked	Low
Prospective registry fields	Yes, selected fields	Shared trial-design construct	Human-reference validation; leakage audit	Moderate



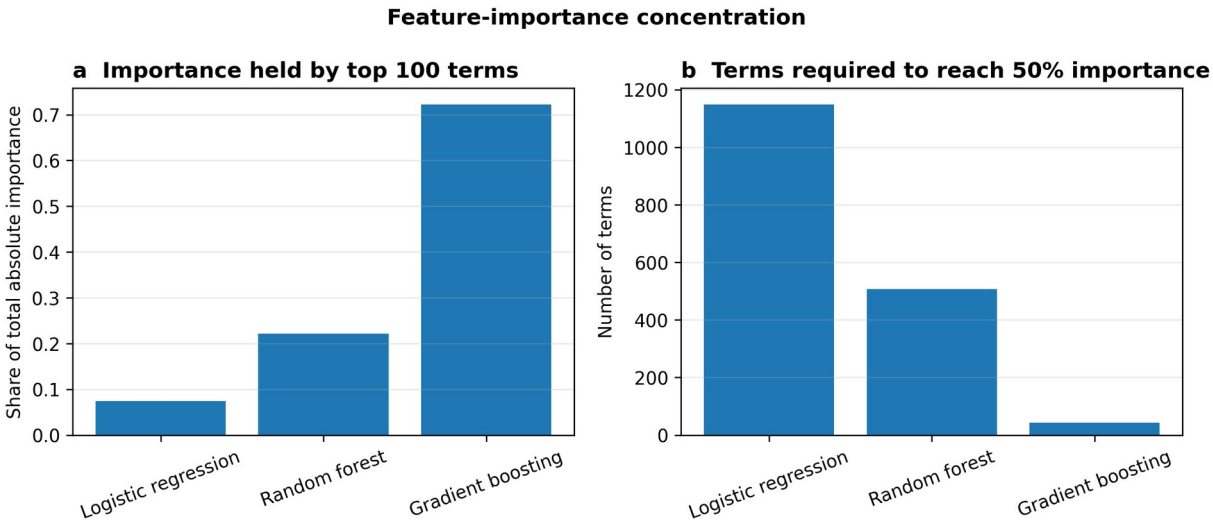
Supplementary Figure S1. Stability across leakage-control profiles. Final ROC-AUC is plotted against the corresponding estimate from the initial unfiltered analysis for all 27 configurations.



Supplementary Figure S2. One-NCT random-forest sensitivity. The three estimates separate the effect of duplicate weighting during evaluation from the effect of duplicate weighting during training.



Supplementary Figure S3. Organizational-field ablation. Random-forest ROC-AUC is shown before and after removing the dedicated Sponsor, Collaborators, and Locations fields from the unique-NCT all-indication models.



Supplementary Figure S4. Feature-importance concentration. Logistic-regression values use absolute coefficients; tree-model values use normalized impurity-based importance.

Supplementary data files

- Supplementary Table S1: complete neuroscience dictionary workbook (XLSX) and concept-level CSV.
- Complete 27-run primary results and paired cross-domain comparisons.
- Human-reference match audit, metrics, bootstrap summaries, and calibration bins.
- One-NCT deduplication manifests and three-way comparisons.
- Organizational-field ablation predictions, paired comparisons, ranking-overlap summaries, and vocabulary audit.
- Residual-language correction manifest and targeted sensitivity tables.
- Complete feature rankings, category rules, source mappings, stability tables, and contextual audit outputs.
- Run provenance, source hashes, and machine-readable validation records for each analysis package.