ELSEVIER

Check for updates

Feature

Estimating and predicting clinical trial outcomes in China

Chaoyi Zhao [1,*], Timothy Pang [2], Andrew W. Lo [1,3,4]

Around a quarter of drug development now has a Chinese origin, yet China-specific success benchmarks and forecasting tools remain scarce. We analyze all nongeneric drug trials registered on China's national drug clinical trial registry from 2013 to 2025, benchmark phase-transition probabilities of success (PoSs)—where success denotes a program advancing to the next phase or to approval—and forecast trial outcomes with machine learning using structured registry features and Chinese text. China's overall phase I-to-approval PoS is below 30% across most therapeutic areas, while phase II-to-phase III transition rates run far above global benchmarks. In out-of-sample evaluations, random forests reach areas under the receiver operating characteristic curve approaching 80% and logistic regression approaches 70%, with the sponsor's track record and text-derived signals as key drivers.

Keywords: China; clinical trials; probability of success (PoS); machine learning; feature importance; pipeline prioritization; portfolio prioritization

FEATURE

Introduction

Clinical trials are the central and most expensive bottleneck of drug development: they determine whether promising scientific ideas translate into therapies that regulators will approve and clinicians will use. However, clinical development remains highly uncertain. Historical evidence shows that a large fraction of programs fail somewhere along the phase I–III path, and even late-stage studies can end without drug approval.[(p1),(p2),(p3)] These attrition dynamics imply enormous economic and public health stakes. When a trial fails, sponsors lose years of time and substantial capital; when a weak program is advanced too far, scarce clinical resources are misallocated; and when a strong program is prematurely terminated, potential therapies are delayed or never reach patients. Clinical development is both capital-intensive and risk-dominated, making improved decision-making under uncertainty a first-order problem.[(p1),(p4),(p5),(p6),(p7),(p8),(p9)]

As a result, the ability to quantify and forecast clinical-development risk has become increasingly important for R&D portfolio planning, trial and pipeline prioritization, financing decisions and policy evaluation. The literature has developed along two related lines. One estimates probabilities of success (PoSs), either for individual phase transitions or for the overall path from phase I to approval; early studies established broad benchmarks from historical development portfolios, while later registry-based methods made censoring, delayed outcomes and unobserved transitions explicit.[(p2),(p3),(p10)] The other moves from aggregate benchmarks to the trial level and asks which specific ongoing trials are more likely to succeed or fail, with recent models drawing on inputs that range from drug and target properties to early-phase subject-level data, structured trial characteristics and trial text.[(p11),(p12),(p13),(p14),(p15)] The first line provides empirical benchmarks for development risk, while the second is more directly relevant for operational

https://doi.org/10.1016/j.drudis.2026.104802

decisions involving individual trials and projects. A fuller review of both research paths is provided in Appendix A in the supplementary material online.

China is a particularly important setting in which to study these issues. Over the past decade, China's clinical-development ecosystem has expanded rapidly, shaped by regulatory reforms, growing domestic biopharma capacity and deeper integration with global R&D pipelines.[p16],[p17],[p18] Approximately a quarter of active drug development now has Chinese origins.[p19] At the same time, building on a long history of clinical experimentation,[p20],[p21] China has developed a public-facing clinical trial registration and disclosure infrastructure for drug trials (Appendix B in the supplementary material online), creating a rare opportunity to study a large national trial system using a comprehensive, regulator-linked registry. This combination—fast growth plus systematic disclosure—makes China a natural laboratory for both PoS estimation and trial-outcome forecasting.

However, most large-scale predictive studies and many widely used PoS benchmarks rely on ClinicalTrials.gov and global aggregators,[p12],[p22] whose structure, reporting norms and development practices might not carry over to China. This gap matters for three decisions that increasingly depend on China-origin programs: valuing a China-origin asset or in-licensing deal, prioritizing which ongoing trials and programs to back or drop and benchmarking China-origin risk against global development. Each needs China-specific evidence: how high are PoSs in China, how do they vary across therapeutic areas and how do they compare with global benchmarks? And can models built on registry fields—including Chinese language text—forecast individual trial outcomes as well as models built on global data?

Using all nongeneric drug trial records from 2013—when China's registry platform was established—through 2025, this study estimates China-specific PoSs, compares these estimates with global benchmarks and evaluates trial-level outcome prediction using structured registry fields and Chinese language text. It therefore connects two related uses of clinical trial registries—aggregate PoS benchmarking and trial-level prediction—in a national setting that has been largely absent from prior work. To our knowledge, this is the first study to benchmark PoS and to build and evaluate trial-level outcome-prediction models at this scale using China's national registry. Four messages emerge.

First, China's clinical-development activity has expanded rapidly at both the project and trial levels, with oncology dominating the landscape.

Second, China's PoS profile is distinctive. The overall phase I-to-approval PoS is generally below 30% across most therapeutic areas. This is consistent with the closest prior work by Wang *et al.*[p23] who studied 1076 Chinese innovative drug development projects through 2019; our sample of 5648 projects through 2025 is roughly five times larger and extends the analysis to therapeutic area-level global comparison and trial-level prediction. In addition, China's phase-transition profile differs systematically from global benchmarks: China's PoS from phase II to phase III is substantially higher than global estimates across therapeutic areas. The other transitions are closer to the global benchmark, making this phase II to phase III gap the most distinctive feature of China's phase-transition profile.

Third, registry-inferred trial outcomes in China can be predicted. In out-of-sample, forward-looking evaluations, our best models achieve discrimination on par with leading studies on global trial registries: for the most recent cohorts, random forests reach AUC (area under the receiver operating characteristic curve) values approaching 80%, and logistic regression delivers AUC around 70%.[p12],[p14],[p15]

Fourth, the models are interpretable rather than black boxes. The sponsor's historical PoS is the single most influential predictor, followed by text-derived signals and time since disclosure—yielding a transparent map from registry disclosures to forecasted outcomes. We quantify the contribution of each feature block to these forecasts, including the models' performance for sponsors that are new to the registry.

China's clinical-development landscape

Our dataset was downloaded from the DXY Insight database, which includes all clinical trials registered on China's Drug Clinical Trial Registration and Information Disclosure Platform since its establishment in 2013. The original DXY database is documented in Chinese; we therefore provide an English–Chinese translation reference in Appendix E in the supplementary material online. Because the goal of this paper is to predict the outcomes of clinical trials for nongeneric drugs, we downloaded data only for these drugs; generics and biosimilars were excluded. The database provides two datasets: the dataset of drug development projects (the Project Dataset) and the dataset of clinical trials (the Trial Dataset). The Project Dataset includes 5648 drug development projects for nongeneric drugs from 2013 to 2025, and the Trial Dataset includes 14 769 clinical trials over the same period.

The Project Dataset and the Trial Dataset are closely related and complementary. Each drug development project represents a combination of all clinical trials across different phases, conducted by the same sponsor for the same drug in the same indication. In other words, the Project Dataset aggregates information at the project level, while the Trial Dataset provides trial-level details for each phase in that project. Of the 14 769 clinical trials in the Trial Dataset, 13 339 clinical trials appear in the Project Dataset, covering 90.32% of all trials. In the following analysis, we match the two datasets and focus only on the 13 339 trials that appear in both datasets.

Both drug development projects and clinical trials exhibit a clear upward trend over time, consistent with the rapid expansion of China's pharmaceutical and biotechnology sectors. At the project level, both chemical and biologic projects increase over time, with chemical projects remaining more numerous throughout the sample period and domestic projects increasing substantially to dominate project counts in later years. At the trial level, oncology accounts for the largest share of clinical trials in our sample, and phase I trials make up the largest share across most therapeutic areas. These patterns suggest sustained growth in clinical trial activity in China together with substantial heterogeneity across therapeutic areas, phases, sponsors and operational scale. We report these detailed descriptive results in Appendix B.

China-specific PoS benchmarks

In this section, we estimate the PoS for clinical trials. We first preprocess the trial records and classify each trial's registry-inferred development outcome as successful, failed, still in progress or not available. Following Wong *et al.*,[p3] 'success' here means that the associated program advances to a later phase or reaches approval, rather than that the individual trial meets its own primary endpoint. Specifically, a phase I trial is labeled successful if its project has any phase II or phase III trial or has been approved; a phase II trial is labeled successful if its project has any phase III trial or has been approved; and a phase III trial is labeled successful only if its project has been approved. Three labeling rules matter for reading the results: combined-phase trials (e.g. phase I/II) are assigned to the earlier phase; missing completion dates are imputed from the median duration of trials in the same therapeutic area; and a completed trial that has not yet advanced remains in progress for a grace period equal to half the relevant median trial duration before it is classified as failed. Transitions that remain in progress are kept as a separate category and enter the transition-specific PoS estimate only after their registry-inferred outcome is observed. Full methodological details, including data preprocessing, trial-outcome labeling, counts of successful, failed and still-in-progress transitions as well as formal definitions of the PoS estimators, are reported in Appendix C in the supplementary material online.

Table 1 compares the estimated path-by-path PoSs for Chinese versus global clinical development across therapeutic areas and phase transitions as of 2025. The global benchmarks are from Project ALPHA (2025Q4 release), which applies the same path-by-path estimator to Citeline drug-development data; the crosswalk between the two therapeutic area classifications is reported in Appendix C. The most salient difference between the Chinese and global PoS results is that China's PoS from phase II to phase III is substantially higher than the global benchmark across almost all therapeutic areas. At the pooled level, the estimate is 81.29% for China, nearly twice the global estimate of 43.77%. This pattern is robust to alternative outcome-labeling assumptions and to restricting the sample to trials first disclosed by 2019, for which nearly all transition outcomes are resolved by the end of the sample; in this more mature cohort, the pooled phase II-to-phase III PoS remains high at 77.26%. The pattern also holds within novelty strata: when trials are stratified by registry-derived proxies—global launch status, domestic versus imported origin, biologic versus chemical modality and, within oncology, PD-1/PD-L1 (programmed cell death protein 1/programmed death-ligand 1) versus other targets—the phase II-to-phase III transition remains elevated even among not-yet-launched and domestic programs (Appendix C). Differences in individual therapeutic areas with few trials are noisier and should be read with the standard errors in mind (Table 1).

The other transitions present a more mixed picture. China's phase I to phase II PoS is lower than the global benchmark in 17 of the 18 therapeutic areas, whereas the phase III to approval transition is broadly comparable (a pooled 61.51% vs. 66.74%). Because the large phase II to phase III advantage is only partly offset by the lower phase I to phase II transition, China's overall phase I to approval PoS is, if anything, modestly higher than the global benchmark in most areas.

More generally, for most therapeutic areas, the overall PoS remains below 30%. This result aligns with the PoS estimates for Chinese clinical trials reported in the literature.[p23] Another prominent pattern concerns oncology, where China's overall PoS is 20.11%, markedly higher than the global estimate (8.88%). This gap is driven primarily by the transition from phase II to phase III: China's phase II trials in oncology have a PoS of 86.68%, compared with 28.65% globally. These comparisons describe registry-inferred progression and should not be interpreted as direct evidence about trial efficacy, evidentiary standards or innovation quality.

The high phase II to phase III estimate might have several causes. First, imported, bridging or multinational programs can enter the Chinese pathway after evidence has already accumulated outside China; for these programs, the Chinese registry observes only part of the development history.[p24],[p25],[p26] Second, as shown earlier, China's phase I to phase II PoS is lower than the global benchmark in nearly all therapeutic areas, so more programs might drop out before they enter the observed phase II risk set, which might raise the transition rate from phase II to phase III among the projects that remain.[p27],[p28]

Third, part of the estimate reflects how the path-by-path estimator handles incomplete phase histories.[p3] This matters because phase II records are often missing among projects that later reach phase III or approval. Missing intermediate phases are not unique to the Chinese registry: the path-by-path method of Wong *et al.* was designed for global clinical-development databases in which a higher stage can be observed even when an intermediate phase is not.[p3] The missing-phase II share is 59.1% in our data, higher than the approximately 34.6% implied by Wong *et al.*'s global estimates. It is also higher than the missing-phase I (30.1%) and missing-phase III (14.0%) shares in the Chinese data (Appendix C).

This higher Chinese missing-phase II share does not appear to be driven primarily by incomplete coverage during the registry's early years. Using the same definition among the 1664 projects that reached phase III or approval, the proportions without an observed phase II record were 57.7%, 49.2% and 54.9% in the three more mature first-disclosure cohorts (2013–2015, 2016–2018 and 2019–2021), showing no sustained decline as registry coverage matured. The corresponding proportion was 72.3% for 2022–2025, although this latest cohort is selectively composed of programs that advanced rapidly enough to reach phase III or approval by the end of the sample. Separately, public company filings show that some programs proceeded directly from phase I to phase III without a phase II trial. These examples show that true skipping occurs, but the registry data cannot determine for each missing phase II record whether phase II was truly skipped or unobserved in the linked data. Appendix C provides the calculation of these shares, the cohort breakdown by first disclosure year and details on the four examples based on public filings.

Taken together, the markedly higher phase II-to-phase III transition should be treated as a hypothesis-generating feature

TABLE 1

Path-by-path PoS (%) for Chinese versus global clinical development by therapeutic area.

Therapeutic area	Phase I to phase II		Phase II to phase III		Phase III to approval		Overall	
	China	Global	China	Global	China	Global	China	Global
Oncology	54.36 (1.25)	**62.42** (0.34)	**86.68** (1.35)	28.65 (0.43)	**76.66** (2.50)	58.44 (0.94)	**20.11** (1.21)	8.88 (0.21)
Cardiovascular	60.64 (2.52)	**73.15** (0.83)	**78.16** (3.13)	56.36 (1.13)	59.76 (5.42)	**62.28** (1.55)	18.28 (2.36)	**23.52** (0.84)
Dermatology	74.43 (2.50)	**76.92** (0.88)	**81.50** (2.95)	44.11 (1.23)	63.01 (5.65)	**71.74** (1.78)	**25.14** (3.21)	22.05 (0.91)
Gastroenterology	61.62 (2.53)	**66.77** (0.68)	**67.47** (3.64)	46.05 (0.92)	**69.77** (4.95)	67.47 (1.35)	**21.28** (2.44)	18.56 (0.59)
Endocrinology	63.12 (2.25)	**66.77** (0.68)	**79.49** (2.64)	46.05 (0.92)	61.90 (4.33)	**67.47** (1.35)	**22.67** (2.26)	18.56 (0.59)
Ophthalmology	69.14 (3.63)	**82.58** (0.90)	**80.43** (4.14)	51.74 (1.39)	39.53 (7.46)	**69.62** (1.88)	15.32 (3.42)	**27.27** (1.14)
Hematology	73.15 (3.63)	**74.76** (1.07)	**87.50** (3.38)	64.04 (1.41)	67.92 (6.41)	**78.89** (1.56)	34.29 (4.63)	**35.54** (1.23)
Immunology	**67.90** (1.96)	65.99 (1.03)	**75.09** (2.56)	51.88 (1.40)	**77.78** (4.18)	73.75 (1.79)	21.88 (2.20)	**22.97** (0.96)
Infectious	66.79 (2.03)	**71.11** (0.61)	**77.08** (2.48)	59.64 (0.81)	48.47 (3.91)	**73.95** (0.98)	19.41 (1.96)	**29.31** (0.64)
Psychiatry	56.45 (3.64)	**67.07** (0.57)	**79.01** (4.52)	44.10 (0.78)	**62.00** (6.86)	60.80 (1.22)	**20.95** (3.34)	16.03 (0.47)
Otolaryngology	77.78 (5.24)	**82.58** (0.90)	**59.46** (8.07)	51.74 (1.39)	64.29 (12.81)	**69.62** (1.88)	20.93 (6.20)	**27.27** (1.14)
Musculoskeletal	64.57 (3.20)	**68.88** (1.09)	**81.74** (3.60)	56.87 (1.46)	67.24 (6.16)	**69.91** (1.91)	**24.68** (3.43)	24.65 (1.07)
Neurology	59.36 (2.54)	**67.07** (0.57)	**79.89** (3.04)	44.10 (0.78)	52.69 (5.18)	**60.80** (1.22)	**17.50** (2.27)	16.03 (0.47)
Other	66.17 (2.05)	**77.80** (1.88)	**82.61** (2.28)	55.29 (2.73)	66.67 (3.97)	**67.86** (3.60)	25.47 (2.27)	**26.82** (2.15)
Rare	62.68 (2.20)	**71.14** (0.37)	**85.20** (2.38)	42.29 (0.50)	**80.70** (3.70)	65.91 (0.83)	**28.05** (2.48)	17.56 (0.34)
Reproductive	63.16 (4.95)	**75.30** (1.09)	**81.25** (5.63)	53.92 (1.51)	66.67 (9.07)	**69.89** (1.98)	25.35 (5.16)	**26.33** (1.17)
Respiratory	67.42 (2.49)	**71.48** (0.94)	**79.39** (3.15)	39.52 (1.26)	**69.33** (5.32)	61.91 (2.06)	**23.21** (2.82)	16.17 (0.80)
Urology	74.87 (3.14)	**75.30** (1.09)	**68.81** (4.44)	53.92 (1.51)	67.35 (6.70)	**69.89** (1.98)	25.19 (3.79)	**26.33** (1.17)
All areas (pooled)	61.78 (0.74)	**67.64** (0.20)	**81.29** (0.86)	43.77 (0.27)	61.51 (1.57)	**66.74** (0.41)	**19.66** (0.72)	17.68 (0.17)

The all-area pooled estimate is in the final row. Each cell reports the estimate with its standard error in parentheses; for each therapeutic area and transition, the larger of the China and global estimates is shown in bold. Because several therapeutic areas have few trials, the standard errors in those rows are large, so bold differences there should be read with caution. China standard errors are Bernoulli; global estimates and their Bernoulli standard errors are from Project ALPHA (MIT Laboratory for Financial Engineering; https://projectalpha.mit.edu/pos/), 2025Q4 release. The therapeutic area crosswalk is reported in Appendix C.

of the registry. It does not, by itself, establish less-selective phase II evidence or weaker innovation quality. Decision-makers should examine the underlying endpoints, effect sizes, trial design, external evidence and complete development history rather than infer evidence quality from advancement alone. Additional trial-outcome summaries, full PoS tables and time-series PoS results are reported in Appendix C.

Trial-outcome prediction: performance and limits

Now we move from descriptive analysis to prediction using machine learning models to forecast clinical trial outcomes. For a given calendar year, for each trial, we compute various features using only information available up to the end of that year. Our feature set combines structured trial characteristics, such as phase, therapeutic area, modality, enrollment, scope and sponsor track record. In addition, we extract Chinese language textual features from key narrative fields using a pretrained Chinese BERT (bidirectional encoder representations from transformers) model. A more detailed description of feature construction is provided in Appendix D in the supplementary material online.

To predict trial outcomes, we focus on two representative classifiers: logistic regression, a linear baseline with strong interpretability, and random forest, a nonlinear benchmark that captures complex feature interactions. For each calendar year, we train only on trials whose outcomes are already known by the end of that year and then forecast the outcomes of trials that are still pending—a strictly forward-looking, out-of-sample design that guards against look-ahead bias. Because the pending trials' outcomes become known over the following years, we compare each forecast with the actual outcome; trials whose outcomes are still unknown at the end of 2025 have no actual outcome to compare against and are excluded from the evaluation. Appendix D details the training, validation and testing procedure and reports results for additional machine learning models.

Table 2 summarizes yearly out-of-sample performance for random forest and logistic regression at validation-selected classification thresholds. For each trial, both models output a predicted probability of eventual success. AUC measures 'discrimination'—the ability to rank trials by these predicted probabilities—which is the chance that a randomly chosen trial that eventually succeeds receives a higher predicted probability than a randomly chosen trial that eventually fails, so 50% corresponds to no ranking ability and 100% to a perfect ordering. Random forest generally leads on AUC and F1—consistent with its ability to capture nonlinearities and interactions—while logistic regression stays competitive. Beyond the metrics in Table 2, Appendix D reports, for every model and year, the area under the precision–recall curve (PR-AUC) and per-class precision and recall as well as decision-oriented metrics; both models exceed a no-skill baseline on discrimination (AUC and PR-AUC) and on decision-oriented metrics such as top-decile lift.

Performance also improves over time for both models, consistent with the growing number of labeled trials available for training: the AUC of the random forest rises from 55.87% in 2015 to 78.59% in 2024, and that of the logistic regression model from 54.04% to 70.80%. The 2024 values are based on smaller realized-outcome test sets, so they should be read with some caution (Appendix D). The overall magnitude is comparable to that reported in prior predictive-modeling studies on global clinical trial datasets,[p12],[p14],[p15] indicating that our setup yields

performance in a similar range despite differences in countries and data sources.

Discrimination asks only whether the predicted probabilities rank trials correctly—placing more-promising trials above less-promising ones; 'calibration' asks a different question—whether the predicted probabilities are numerically accurate: among trials assigned around 70%, do roughly 70% in fact succeed? Panel B of Table 2 reports two standard calibration measures, the Brier score and the expected calibration error (ECE; lower is better for both), alongside a no-skill benchmark that assigns every trial the historical success rate. Such a constant forecast can be reasonably well calibrated even though it cannot rank trials at all (its AUC is 50% by construction). On calibration, the models offer no advantage over this benchmark: random forest (Brier 0.259, ECE 0.169) is essentially on par with it (0.263, 0.164), and logistic regression (0.284, 0.218) is worse. Their advantage lies in discrimination, with weighted AUCs of 60.84% and 58.10%, respectively, against the benchmark's 50%. In practical terms, the predictions are most informative when used to compare trials—separating more-promising from less-promising ones—rather than as precise success probabilities for individual trials.

Panel B also quantifies the contribution of each feature block. Removing the sponsor-history block lowers weighted random forest AUC by 2.95 percentage points, compared with 0.51 points for removing the BERT text score, so the sponsor track record is the single largest contributor; for trials whose sponsors have no prior registry record, full-model AUC is 51.52% for random forest and 52.32% for logistic regression, so the scores are most informative for sponsors with a registry history (full ablation results are in Appendix D).

To understand which features are associated with predicted success, we examine the logistic regression coefficients (Table 3) alongside model-based feature importance. First, the sponsor's historical PoS (its success share among its own previously resolved trials) is strongly positive, whereas its trial volume is not—past performance matters more than past quantity. Second, the BERT text score is large and positive, so registry narratives add information beyond the structured variables. Third, timing, scope and design variables behave intuitively: days since first disclosure and combination therapy are positively associated with predicted success, whereas international scope, larger target domestic enrollment and older enrollment ages are negatively associated. We stress that these coefficients are predictive associations, not causal effects: the design variables, in particular, are confounded with therapeutic area, trial complexity, sponsor type and regulatory pathway and should not be read as levers—a trial would not become more

TABLE 2

Out-of-sample model performance by year for random forest (RF) and logistic regression (LR), at validation-selected classification thresholds (Panel A), with weighted averages across evaluation years, calibration and feature-ablation summaries (Panel B).

Panel A: performance by year

Year	Model	# Train	# Valid	# Test	Precision (%)	Recall (%)	F1 (%)	AUC (%)
2015	RF	328	83	9115	66.00	79.89	72.29	55.87
2015	LR	328	83	9115	65.15	88.93	75.20	54.04
2016	RF	496	124	8906	65.19	83.18	73.09	55.53
2016	LR	496	124	8906	64.90	81.47	72.25	53.67
2017	RF	713	179	8634	64.91	65.40	65.15	54.14
2017	LR	713	179	8634	64.60	65.11	64.85	52.92
2018	RF	981	246	8299	69.35	47.78	56.58	59.30
2018	LR	981	246	8299	67.18	64.52	65.82	57.21
2019	RF	1392	348	7786	64.18	87.73	74.13	61.24
2019	LR	1392	348	7786	64.91	74.34	69.30	57.43
2020	RF	1854	464	7208	63.78	87.63	73.83	62.40
2020	LR	1854	464	7208	66.54	67.31	66.92	60.16
2021	RF	2531	633	6362	63.04	86.97	73.10	64.77
2021	LR	2531	633	6362	63.22	78.16	69.90	61.32
2022	RF	3410	853	5263	67.42	75.19	71.09	70.01
2022	LR	3410	853	5263	63.08	78.97	70.13	66.18
2023	RF	4647	1162	3717	59.02	88.05	70.67	72.61
2023	LR	4647	1162	3717	59.33	75.95	66.62	67.06
2024	RF	6181	1546	1799	63.52	78.12	70.07	78.59
2024	LR	6181	1546	1799	53.50	79.38	63.92	70.80

Panel B: weighted averages across years—calibration and feature ablation

Model	AUC (%)	Brier	ECE	AUC loss (pp): sponsor/BERT	New-sponsor AUC (%)
RF	60.84	0.259	0.169	2.95/0.51	51.52
LR	58.10	0.284	0.218	2.04/0.20	52.32
No-skill	50.00	0.263	0.164	–	50.00

Precision, Recall and F1 refer to the success class; per-class metrics for both classes are reported in Appendix D. # Train, # Valid and # Test are the numbers of trials used for training, validation and out-of-sample evaluation; # Test counts the realized-outcome test set (Appendix D). Panel B reports averages across evaluation years, weighted by the number of realized outcomes. Brier is the mean squared error between the predicted success probability and the realized binary outcome; ECE is the expected calibration error across ten equal-width probability bins (Appendix D). AUC loss is the full-model AUC minus the AUC with the indicated feature block removed (sponsor history or the BERT text score), in percentage points (pp). New-sponsor AUC is the full-model AUC among trials whose sponsors have no prior registry record. The No-skill benchmark makes the same prediction for every pending trial in year *Y*: the overall success rate among trials whose outcomes were already known by the end of that year.

TABLE 3

Selected coefficients from the logistic regression model.

Feature name	Coefficient	*t*-statistics
Text: BERT PoS	4.54***	(18.64)
Sponsor: past PoS	6.44***	(35.19)
Sponsor: number of past sponsored trials	−0.00	(−0.05)
Information: days since first disclosure	0.21***	(7.22)
Information: international scope	−0.67***	(−6.05)
Design: target domestic enrollment	−0.08**	(−2.34)
Design: with combination therapy	0.37***	(4.10)
Design: minimum enrollment age	−0.06**	(−2.22)
Design: maximum enrollment age	−0.07**	(−2.53)
Therapeutic area: oncology	−0.32***	(−2.87)
Therapeutic area: gastroenterology	−0.21*	(−1.86)
Therapeutic area: neurology	−0.03	(−0.29)
Therapeutic area: psychiatry	−0.38**	(−2.41)
Therapeutic area: dermatology	0.38**	(2.42)
Therapeutic area: hematology	0.60***	(3.38)
Line: first	0.71***	(4.64)
Line: last	−0.43**	(−2.31)
Modality: CAR-T	0.68*	(1.85)
Modality: gene therapy	−1.25***	(−3.04)
Modality: antibody–drug conjugate	1.42***	(5.19)

BERT: bidirectional encoder representations from transformers; CAR-T: chimeric antigen receptor T cell therapy.

* Significance at 10% level.

** Significance at 5% level.

*** Significance at 1% level.

likely to succeed merely by being run domestically rather than internationally.

Appendix D reports the full logistic regression coefficients, feature-importance results for the other models, calibration measures for all models and the complete feature-ablation analyses. These results show that sponsor track record, text-derived signals and basic trial-design characteristics all contribute to prediction.

Implications and limitations

Our results turn China's registry disclosures into something that sponsors, investors and regulators can act on: China-specific PoS benchmarks across phases and therapeutic areas and trial-level forecasts built on the same data. We conclude by spelling out their implications.

The most actionable signal concerns phase II. China's phase II-to-phase III transition is above the global benchmark in every therapeutic area. The practical implication is not to discount China-originated phase II assets mechanically, but to interpret phase II-to-III advancement with China-specific benchmarks in mind. For sponsors and investors, the benchmark helps to place a program's progression in context when valuing assets or prioritizing portfolios. For regulators and policymakers, comparisons with global development can help monitor where rapid advancement is concentrated and whether the pattern changes as the ecosystem matures.

Our study has several limitations that point to natural directions for future work. First, because registry data do not provide clean outcome labels in real time, any registry-based success definition must make assumptions about censoring, missing transitions and reporting lags.[p3],[p12] We make several such assumptions in our outcome-labeling procedure (Appendix C). Second, our models rely mainly on registry-visible data, and the scores are better suited to comparing trials than to serving as absolute success probabilities for individual trials. Incorporating molecular descriptors, target biology and external evidence such as publications, patents or real-world utilization could improve forecasts and calibration and help separate scientific from operational drivers; the emergence of large language models could make such extensions increasingly feasible. This reliance binds hardest on new sponsors: with little registry history available, the models' discrimination for trials from sponsors with no prior record—new entrants and first-time licensors—falls to near chance (Appendix D). Third, the novelty stratifications reported in Appendix C show that the elevated phase II-to-phase III transition persists among not-yet-launched and domestic programs, but these registry-derived proxies are coarse and cannot identify the mechanism behind the gap. How PoS relates to innovativeness remains debated.[p29],[p30]

More broadly, our results underscore the value of national registries as scientific and policy infrastructure. China's biopharmaceutical industry is now large enough, and its registry is detailed enough to support benchmarking and prediction at scale and to allow direct comparison with the predominantly US-centric literature built on ClinicalTrials.gov and global aggregators. The deeper lesson is that transparent disclosure, when used well, can be turned into actionable learning about risk, outcomes and the drivers of success in a fast-moving clinical-development ecosystem.

CRediT authorship contribution statement

Chaoyi Zhao: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Timothy Pang:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Andrew W. Lo:** Writing – review & editing, Writing – original draft, Validation,

Supervision, Resources, Project administration, Methodology, Investigation, Conceptualization.

Acknowledgments

The research support from the MIT Laboratory for Financial Engineering is gratefully acknowledged. We thank Shomesh E. Chaudhuri for helpful comments. We thank Jayna Cummings for editorial support.

Declarations of interest

T. P. and C. Z. have nothing to declare.

A.W.L. reports the following competing interests: he is a director of Annual Reviews, BridgeBio Pharma, Global Coalition for Adaptive Research (GCAR), n-Lorem Foundation, TechTransferase, Uncommon Cures and Vesalius Therapeutics; an advisor to AACR Oncology Development Fund, Apricity Health, Aracari Bio, Aurora Therapeutics, AUTS2 Research Collaborative, Cancer Focus Fund, Ellison Medical Institute, Gondola Bio, Health at Scale, Khora Therapeutics, Quantile Health, Ride Therapeutics, Roivant Social Ventures, Sakura Health, Swiss Finance Institute, Thalēs, Think Therapeutics and xCures; a cofounder of QLS Technologies; a cofounder and chairman of QLS Advisors; and he has personal investments in ACISM, Apricity Health, Aracari Bio, BillionToOne, BridgeBio Pharma, Fidelity Select Biotech Fund, Khora Therapeutics, Ladder Tx, LS Polaris Innovation Fund, MPM Capital Fund, Nautilus Biotech, Nurture Genomics, Peer Collective, Polaris Innovation Fund II, Proactive Labs, PsiThera, Quantile Health, Ride Therapeutics, Sakura Health, Stratus Therapeutics, SURGE Therapeutics, Uncommon Cures and Waypoint Bio.

During the most recent 6-year period, A.W.L. has received speaking/consulting fees, honoraria or other forms of compensation from AbCellera, AlphaSimplex Group, Annual Reviews, Apricity Health, Aracari Bio, Atomwise, BridgeBio, CME, Enable Medicine, Journal of Investment Management, Lazard, MIT, New Frontier Advisors, Oppenheimer, Princeton University Press, Q Group, QLS Advisors, Quantile Health, Roivant, SalioGen, Stern Speakers, Swiss Finance Institute, Think Tx, Vesalius and W. W. Norton. In addition, the MIT Laboratory for Financial Engineering, for which A.W.L. serves as director, has received funding support from the Critical Path Institute, J.P. Morgan Asset & Wealth Management, Schmidt Futures and Wellcome Leap.

Data availability

The data that support the findings of this study were obtained from the DXY Insight database. Readers can subscribe to the DXY Insight database to access the data. The global probability-of-success benchmarks used for comparison are publicly available from Project ALPHA (https://projectalpha.mit.edu/pos/).

The code used to construct the variables, estimate the PoS and train and evaluate the predictive models is available from the corresponding author upon reasonable request.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Claude Opus 4.8, Claude Fable 5, ChatGPT 5.5 and ChatGPT 5.6 Sol to improve the language and readability of the manuscript, to help organize the text, tables and figures as well as to assist in writing and standardizing the code that generates the manuscript's tables and figures. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.drudis.2026.104802.

References

1. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov*. 2004;3:711–716. https://doi.org/10.1038/nrd1470.
2. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. *Nat Biotechnol*. 2014;32:40–51. https://doi.org/10.1038/nbt.2786.
3. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. *Biostatistics*. 2019;20:273–286. https://doi.org/10.1093/biostatistics/kxx069.
4. Grabowski H, Vernon J. A new look at the returns and risks to pharmaceutical R&D. *Manage Sci*. 1990;36:804–821. https://doi.org/10.1287/mnsc.36.7.804.
5. Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat Rev Drug Discov*. 2012;11:191–200. https://doi.org/10.1038/nrd3681.
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. https://doi.org/10.1016/j.jhealeco.2016.01.012.
7. Ringel MS, Scannell JW, Baedeker M, Schulze U. Breaking Eroom's law. *Nat Rev Drug Discov*. 2020;19:833–834. https://doi.org/10.1038/d41573-020-00059-3.
8. de Villemeur EB, Scannell JW, Versaevel B. Biopharmaceutical R&D outsourcing: short-term gain for long-term pain? *Drug Discov Today*. 2022;27, 103333. https://doi.org/10.1016/j.drudis.2022.08.001.
9. Lo AW, Thakor RT. Financing biomedical innovation. *Annu Rev Financ Econ*. 2022;14:231–270. https://doi.org/10.1146/annurev-financial-031721-081537.
10. DiMasi JA, Feldman L, Seckler A, Wilson A. Trends in risks associated with new drug development: success rates for investigational drugs. *Clin Pharmacol Ther*. 2010;87:272–277. https://doi.org/10.1038/clpt.2009.295.
11. Gayvert KM, Madhukar NS, Elemento O. A data-driven approach to predicting successes and failures of clinical trials. *Cell Chem Biol*. 2016;23:1294–1301. https://doi.org/10.1016/j.chembiol.2016.07.023.
12. Lo AW, Siah KW, Wong CH. Machine learning with statistical imputation for predicting drug approvals. *Harv Data Sci Rev*. 2019;1. https://doi.org/10.1162/99608f92.5c5f0525.
13. Qi Y, Tang Q. Predicting phase 3 clinical trial results by modeling phase 2 clinical trial subject level data using deep learning. In: *Machine Learning for Healthcare Conference*. PMLR; 2019:288–303.
14. Fu T, Huang K, Xiao C, Glass LM, Sun J. HINT: hierarchical interaction network for clinical-trial-outcome predictions. *Patterns*. 2022;3, 100445. https://doi.org/10.1016/j.patter.2022.100445.
15. Zheng W et al. LIFTED: multimodal mixture-of-experts for clinical trial outcome prediction. In: *ICML 2024 Workshop on Foundation Models in the Wild*:1–13.
16. Chakma J, Sun GH, Steinberg JD, Sammut SM, Jagsi R. Asia's ascent—global trends in biomedical R&D expenditures. *N Engl J Med*. 2014;370:3–6. https://doi.org/10.1056/NEJMp1311068.
17. Tannoury M, Attieh Z. The influence of emerging markets on the pharmaceutical industry. *Curr Ther Res*. 2017;86:19–22. https://doi.org/10.1016/j.curtheres.2017.04.005.
18. Zhou Q, Chen XY, Yang ZM, Wu YL. The changing landscape of clinical trial and approval processes in China. *Nat Rev Clin Oncol*. 2017;14:577–583. https://doi.org/10.1038/nrclinonc.2017.10.
19. Pang T, Chancellor D. The growing pains of Chinese biotech. In Vivo. 2023. Citeline, Global Vision. https://insights.citeline.com/IV148602/The-Growing-Pains-Of-Chinese-Biotech/.
20. Yu X, Chen Y, Podolsky SH. The history of controlled clinical trials in China. Part 1: from the advent of clinical trials through the conduct of

single-centre or small-sample RCTs. *J R Soc Med*. 2025;118:230–236. https://doi.org/10.1177/01410768251337693.
21. Yu X, Chen Y, Podolsky SH. The history of controlled clinical trials in China. Part 2: from the advent of large-scale multicentre randomised controlled trials through contemporary self-reflection. *J R Soc Med*. 2025;118:258–264. https://doi.org/10.1177/01410768251347259.
22. Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The ClinicalTrials.gov results database—update and key issues. *N Engl J Med*. 2011;364:852–860. https://doi.org/10.1056/NEJMsa1012065.
23. Wang QH, Lin J, Peng N, Wang QM, Li M, Liu GG. Clinical development success rates for innovative drugs in China. *Int J Pharmacol*. 2022;18:1137–1150. https://doi.org/10.3923/ijp.2022.1137.1150.
24. National Medical Products Administration. Clinical technical requirements for drugs marketed overseas but not marketed in China. Updated 18 November 2020. https://english.nmpa.gov.cn/2020-11/18/c_568155.htm.
25. Wang T, Cao X, He Y, Chen X. Innovation drug approvals based on a bridging study: from concept to practice. *Transl Breast Cancer Res*. 2022;3:2. https://doi.org/10.21037/tbcr-21-43.
26. Huang H et al. Accelerating the integration of China into the global development of innovative anticancer drugs. *Lancet Oncol*. 2022;23:e515–e520. https://doi.org/10.1016/S1470-2045(22)00483-1.
27. Wu SS, Fernando K, Allerton C, Jansen KU, Vincent MS, Dolsten M. Reviving an R&D pipeline: a step change in the phase II success rate. *Drug Discov Today*. 2021;26:308–314. https://doi.org/10.1016/j.drudis.2020.10.019.
28. Fernando K et al. Achieving end-to-end success in the clinic: Pfizer's learnings on R&D productivity. *Drug Discov Today*. 2022;27:697–704. https://doi.org/10.1016/j.drudis.2021.12.010.
29. Kesselheim A, Wang B, Avorn J. Defining "innovativeness" in drug development: a systematic review. *Clin Pharmacol Ther*. 2013;94:336–348. https://doi.org/10.1038/clpt.2013.115.
30. Pang T et al. *Study on the Impact of Mergers and Acquisitions on Innovation in the Pharmaceutical Sector*. Luxembourg: Publications Office of the European Union; 2020.

Chaoyi Zhao [1,*], Timothy Pang [2], Andrew W. Lo [1,3,4]

[1] MIT Sloan School of Management and Laboratory for Financial Engineering, 100 Main Street, Cambridge, MA 02142, USA
[2] Clarivate Consulting, Akasaka Park Building, 5-2-20 Akasaka, Minato-ku, Tokyo 107-6118, Japan
[3] MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, Cambridge, MA 02139, USA
[4] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

* *Corresponding author.*
cy_zhao@mit.edu (C. Zhao)