ReadoutBench: a prospective benchmark for forecasting clinical trial readouts

Design and Round 1 protocol

Lucas Heinrich

lucas@readoutbench.com https://readoutbench.com

1 November 2026

Version 1, written before any result of the round is known

Abstract

Benchmarks with an answer key can leak into training data, and most medical benchmarks are exams with published answers. ReadoutBench asks questions whose answers do not yet exist: the results of randomized clinical trials that have not reported. Round 1 holds 653 company-run phase 2 and 3 trials from every therapeutic area whose registered primary completion falls between July 2026 and March 2027. For each of the 548 trials with an efficacy or equivalence comparison, an entry gives the probability that the trial meets its primary objective and the probability that a result comes out, and for most it also gives the control-arm value and the effect size as 80% ranges; all 653 carry questions on trial operations. Submissions close on 15 November 2026. Two independent labelling agents date and read each trial's first public statement, and any trial with a result public before the close is dropped. Probabilities are scored with the Brier score and ranked in tiers against a logistic regression on seven registry fields; numbers are scored with a percentile loss and ranked against a frontier-model reference. This paper fixes the design before any result arrives. Results follow at 31 March, 30 June and 31 December 2027.

1 Introduction

A static benchmark measures what a model knows about a fixed set of questions. Once the questions and answers circulate, it also measures whether the model has seen them. Medical benchmarks are mostly exam questions with published answers (MedQA and its relatives), which makes contamination hard to rule out and harder to measure. A forecast of an event that has not happened cannot be looked up. ForecastBench applied this idea to general forecasting questions [1], and CT Open to clinical trial outcomes [2].

Trial readouts suit a prospective benchmark. They are frequent and dated: hundreds of company trials report each quarter. They are consequential: a readout decides whether a drug moves on, and while a trial runs, its arms are offered to patients whose odds nobody knows. And people already forecast them, with documented and modest skill. In a randomized study, 148 cancer specialists forecast whether 25 ongoing trials would meet their primary outcomes; their mean Brier score was 0.21 to 0.22, better than the 0.25 of an uninformative 50%, with an area under the ROC curve near 0.7 [3]. Principal investigators forecasting their own trials did no better than 50% on primary outcomes, and worse on recruitment and timelines, with consistent over-optimism [4]. These are the human numbers a model has to beat before its trial forecasts are worth much.

The obvious guess is also strong. Among the earlier company trials on which Round 1's registry model was fitted (Section 7), 63.8% met their primary endpoint within the round's reporting window. A

forecaster who gives every trial that probability already scores well. The question ReadoutBench asks is whether anything does better, and by how much.

This paper describes Round 1 and fixes its design before any of its results are known: how the trials were chosen (Section 3), how the answers were kept unknown (Section 4), how results are read (Section 5) and scored (Section 6), and the reference entries an entry is compared with (Section 7). The full rules, the code, every pre-launch check and the build inputs are public.

2 Related work

Trial outcome prediction. HINT predicts trial success from drug molecules, disease codes and eligibility criteria; it was trained and tested on several thousand historical trials and scored with F1 and related metrics [5]. Such datasets are retrospective: the outcome was known when the test set was built. CT Open runs quarterly challenges on trials expected to report in the following three months, labels whether each outcome measure reached statistical significance with an automated model-based pipeline, scores with macro-F1 and balanced accuracy, and reports that simple machine-learning baselines remain competitive with large language models [2].

Forecasting benchmarks. ForecastBench poses questions from prediction markets and public datasets that resolve after submission, compares models with superforecasters and the public, and scores with the Brier score [1, 6]. ReadoutBench borrows the scoring logic of forecasting tournaments and the interval scoring used by the COVID-19 forecast hubs [7].

Base rates. Large registry analyses give success rates by phase, disease and sponsor [8]. ReadoutBench's registry model is a small, transparent version of that idea, refitted on recent company trials and labelled the way the round labels them.

What differs here. Table 1 sets ReadoutBench beside its closest relative. The main differences are the scoring (proper scores on probabilities and ranges, not classification metrics), what counts as a result (the first public statement of any kind, since a press release often comes months before posted registry results), and the bar (a published registry model rather than other entries).

Table 1: ReadoutBench and CT Open [2].

	CT Open	ReadoutBench
What is forecast	A class per outcome measure: did it reach statistical significance	A probability per trial that it meets its registered primary objective; control-arm value and effect as 80% ranges; subgroups, safety, timing and operations
Scoring	Macro-F1, balanced accuracy	Brier score and percentile loss (proper scores); calibration shown
Bar	Language models beside classical models	Tiers against a published registry model, with the base rate beside it; three frontier-model references, hashed before opening and published at the close
Results read from	An automated pipeline over reported results	Two independent labelling agents reading press releases, filings, congress abstracts, papers and posted results; organiser adjudication; labels published with sources
Entries	Any method	Two boards: one model with fixed instructions, or anything

3 Round 1

3.1 The trials

The candidates were the 1,085 company-run trials on ClinicalTrials.gov registered as randomized phase 2 or 3 with a registered primary completion between July 2026 and March 2027. Of these, 432 were removed before launch: stale records (still recruiting after their own completion date), trials already stopped, trials whose primary result or stop was found public by the checks in Section 4, trials with an individual or non-profit sponsor, and a few others. Every step, with its NCT numbers, is listed in the published funnel. Company trials were chosen because companies tend to announce readouts quickly; academic trials can take years to publish, which would leave most of a round unscored. Table 2 shows what the remaining 653 ask, and Figure 1 when they complete.

Table 2: The 653 trials of Round 1 and the questions each asks.

Trials	*n*	`p_met`	Numbers	Result out?	Operations
Superiority and non-inferiority	536	yes	509 (subgroups on 275)	yes	yes
Equivalence (biosimilars, bioequivalence, lot consistency)	12	equivalence	no	yes	yes
No efficacy goal (safety only, pilot, descriptive, uncontrolled)	105	no	no	no	yes

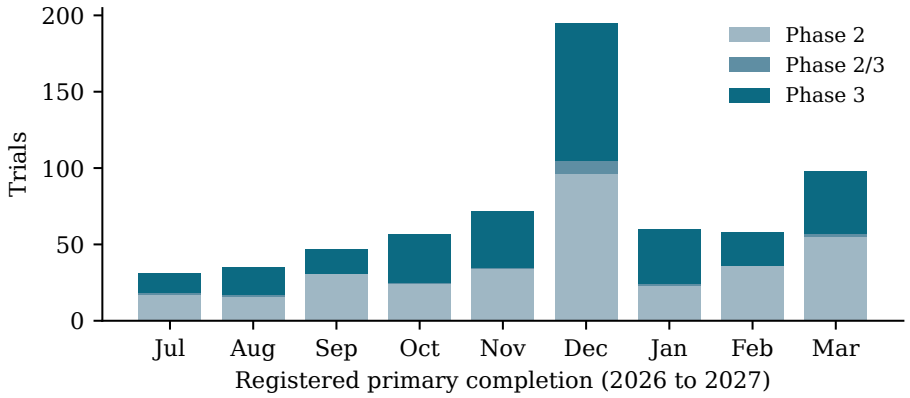


Figure 1: Registered primary completion of the 653 trials, by phase (332 phase 2, 16 phase 2/3, 305 phase 3). December 2026 holds the most (195), reflecting year-end dates in the registry.

3.2 What a forecaster sees

Each trial comes with a packet (its ClinicalTrials.gov record as pulled on 29 October 2026, with no result), a question card and an operations row. The card fixes what the questions refer to: the primary endpoint, the experimental and control arms with their planned sizes, the design (superiority, non-inferiority or equivalence) and significance level, the effect measure (a hazard ratio, a difference in means or in response rates) and the control-arm measure with its unit. Cards were written from the record by language-model agents; equivalence cards were written by hand; independent agents that did not write the cards checked 150 of them against the records, and the errors found were fixed. Every card is public. The round's prompt for the Models board is public too.

3.3 The questions

`p_met` is the probability that the trial meets its primary objective, given that a result is public by 30 June 2027. Optional versions for 31 March and 31 December 2027 serve the early and final passes. Most efficacy trials also ask for the control-arm value and the effect, each as a median with a 10th and 90th percentile; the probability that the key secondary endpoint is met; the probability of the next development step; adverse-event shares; and, for phase 3 trials, subgroup effects. The same 548 trials ask whether a result will be public by each deadline (`p_public`). All 653, including the 105 with no efficacy goal, ask about operations, read from each record on 31 August 2027: whether the trial ran late, stopped, posted results or changed its primary outcomes.

3.4 Boards and entries

The Models board takes one model answering from the packet alone with the round's instructions unchanged; it compares models. The Open board takes anything else, declared with the entry: any harness, tools, web access, several models, people in the loop. It shows how far a system can go. A forecast is required for every trial with a `p_met` question. An account may hold up to three entries; entries may be uploaded again until the close, the last upload counts, and the first is scored beside it, so the value of the extra weeks can be seen.

4 Keeping the answers unknown

Before launch every trial was searched for a public result, an announced stop or any statement on its primary outcome. The checks used PubMed by NCT number; web searches in the sponsor's language (Chinese, Japanese or Korean for trials run only in East Asia); congress abstract books; sponsors' pipeline documents and third-quarter reports; and stock-exchange filings. A trial stayed in only if a check met a stated bar: at least two distinct searches, at least one opened page that is not a registry listing or a copy of one, and a check date on or after 29 October 2026. Before launch, the full package was reviewed repeatedly by AI review agents run independently of the build, which searched samples of the round for missed results; every miss they found was removed, and several prompted wider sweeps (for example, of every large sponsor's pipeline documents and quarterly reports). No human expert reviewed the trial list. Every check row, with its queries, pages and decisions, is published.

The checks cannot prove a negative, so the scores do not depend on them. Labellers record the date of each trial's first public statement on its primary result. Any trial whose first statement came before the round closed, whether the checks missed it or it appeared during the two open weeks, is dropped from scoring and listed. Only results that appear after the close are scored. A missed leak therefore costs a trial, not the validity of the score.

Eighty-five trials could not be searched to the full bar, mostly because sponsors' pages and filings in China were hard to open, and were kept, marked *below the bar*. They are not a random part of the round: 56% of them run only in China, against 20% of the rest. The leaderboard shows each entry's Brier score, skill and tier also without them.

5 Resolution

Labels come from public sources: press releases, regulatory and stock-exchange filings, congress abstracts and presentations, papers, and posted registry results. Each trial is labelled by one of the organiser's labelling agents and checked by a second, independent one; neither sees any forecast. The organiser settles disagreements. The agreement rate, the labels, their sources and the models used are published, and corrections are made in the open.

A trial meets its objective when it achieves the primary objective as it defined it: superiority at its own significance level, or non-inferiority against its own margin, on the card's primary outcome within the trial's own testing order. A numerical improvement without significance, a subgroup win or a post-hoc analysis is not met. A statement that the trial was "positive" or "encouraging" with nothing on the primary endpoint keeps the question open; if nothing clearer comes by the deadline, the trial counts as not met in that pass. That rule is a choice, on the view that a sponsor who talks about a trial without claiming its primary endpoint has more often missed it, not a measurement, so these trials are marked and every score is also shown without them. A stop for futility or harm is not met; a stop for a reason unrelated to efficacy with no efficacy result drops the trial. The rules list the remaining cases, including several primaries, several arms, interim analyses and equivalence.

6 Scoring

6.1 The ranking

The headline is the Brier score [6] of `p_met` over the trials that report after the close and by 30 June 2027,

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2, \tag{1}$$

where p_i is the forecast and $y_i \in \{0,1\}$ the outcome. Because `p_met` is conditional on reporting, knowing that trials which report early tend to succeed is part of the forecast. Each entry is placed in one of three tiers against the registry model: *clearly better* (Brier score lower by more than 0.005), *no clear difference*, or *clearly worse* (higher by more than 0.005). The test is a paired comparison whose standard deviation comes from a bootstrap that resamples whole sister groups (trials of the same experimental drug, which tend to succeed or fail together); the margin sits in each null hypothesis. For *clearly better* the test is of $H_0\colon \mathrm{BS}_{\text{entry}} \geq \mathrm{BS}_{\text{B1}} - 0.005$ against $H_1\colon \mathrm{BS}_{\text{entry}} < \mathrm{BS}_{\text{B1}} - 0.005$; for *clearly worse*, of $H_0\colon \mathrm{BS}_{\text{entry}} \leq \mathrm{BS}_{\text{B1}} + 0.005$ against $H_1\colon \mathrm{BS}_{\text{entry}} > \mathrm{BS}_{\text{B1}} + 0.005$. Both tests are fixed in advance; their p-values come from a normal approximation, and the smaller one is doubled, so no direction is chosen after seeing the data, and compared with $\alpha = 0.05$. An entry whose true gap is exactly at the margin is therefore placed on the wrong side of it at most 2.5% of the time in each direction. Each entry is tested on its own, so no other entry can move its tier.

How large a lead a tier takes was simulated with the round's own sister groups (Table 3). With all 548 trials scored, an entry needs a Brier score about 0.022 lower than the registry model's to be placed clearly better half the time. In the simulation, an entry exactly at the margin is wrongly placed in either direction about 2–3% of the time, close to the nominal 2.5%. With smaller and more shared errors the required lead is about a third lower; both simulations are published.

There is no correction across entries. An account may hold up to three entries, each a different harness and each tested separately, so for an account whose three entries all sit exactly at the margin, the chance that at least one is wrongly placed clearly better is up to about 7% ($1 - 0.975^3$), and less when the entries are correlated, as entries from one account usually are. The leaderboard lists every entry with the account that uploaded it and shows the number of entries tested.

6.2 The number score

Control-arm values and effects are given as 10th, 50th and 90th percentiles and scored with the percentile (pinball) loss, which equals the weighted interval score of the COVID-19 forecast hubs up to a constant [7]: each resolved answer scores the width of its range, plus ten times any miss outside it, plus five times the distance from the median to the result. Ratios and times are scored on the log scale. To compare questions with different units, each score is divided by the width of the Opus reference entry's range for the same

Table 3: Lead over the registry model needed for the tier "clearly better", and false calls at the margin, by number of scored trials. Simulation with the round's sister groups (1,000 rounds per size); rates have a standard error of about 0.5 points.

Trials scored	Lead, 50% of the time	Lead, 80% of the time	False "better"	False "worse"
100	0.044	0.061	2.6%	3.1%
200	0.033	0.045	1.7%	3.0%
400	0.025	0.033	1.9%	2.2%
548	0.022	0.029	2.7%	2.8%

question. A floor keeps a near-zero reference range from inflating one question. For time-to-event and responder questions it is the 10th percentile of the reference's widths over the round's questions of the same part (control or effect), endpoint kind and scale. Continuous questions do not share a unit, so each trial gets a size: the absolute value of the reference's control-arm median plus the width of the reference's control-arm range. The same size serves the trial's control and effect questions, and the floor is that size times the 10th percentile, over the round's continuous questions of the same part, of the reference's width divided by its trial's size. The floors depend only on the frozen reference and the cards, so they were fixed before any result. One edge case remains: on a change-from-baseline endpoint whose reference control median is near zero and whose control range is narrow, the size, and with it the floor, is small, so that question carries more weight. The score stays proper; the median and the trimmed mean shown beside the number score show how much single questions move it. The number score is the plain mean of these scaled scores, the mean of a proper score, so honest ranges do best; a range as wide as the reference's with its median on the result scores 1, unless the floor applies. Number answers come as one block: an entry that gives them must give the control value and the effect for all 509 number questions, or none (the upload check refuses anything in between), so no entry can pick the questions it is sure of. The number score is the second ranking, in tiers against the Opus reference with a margin of 0.05, decided by the same test as the `p_met` tiers: the paired difference in number score over the resolved number questions, with the same bootstrap over sister groups, both one-sided tests with the margin in the null, and $\alpha = 0.05$.

6.3 Shown, not ranked

Next to each entry the leaderboard shows the skill against the base rate and the registry model with intervals; whether the entry adds information to the registry model in a cross-fitted regression; AUROC and the Brier decomposition into reliability and resolution; log loss and calibration slope; the share of results inside the 80% ranges; subgroup, safety, key-secondary, next-step, `p_public` and operations scores; and, for entries that state assumptions, the share of a checked sample that their sources support.

7 Reference entries

B0, the base rate. Each trial gets the share met among earlier company trials followed over that trial's own reporting window, shrunk towards 63.8% with a weight of 20 trials. It ranges from about 0.58 to 0.71.

B1, the registry model. B1 moves B0 on the logit scale by an L2-regularised logistic regression on seven fields: phase 3, placebo control, enrolment (log), cancer, large sponsor, China-only, and a time-to-event endpoint. Its training table holds 218 earlier company trials, completed in 2026, whose results were labelled the way the round labels them, after an audit of the labels; the regression was fitted on the 105 of them followed over the median round trial's reporting window, of which 67 (63.8%) were met. (These 105 training trials are unrelated to the 105 round trials without an efficacy goal.) Its training table and script are published so anyone can refit it exactly. Its forecasts for the 548 trials range from 0.25 to 0.90, with a mean of 0.62 (Figure 2). B1 leans on the China-only field: 0.76 on average for the 147 China-only

trials, 0.58 for the rest. That coefficient may reflect how often such sponsors report vaguely as much as how often they succeed, so skill against B1 is also shown separately for those trials.

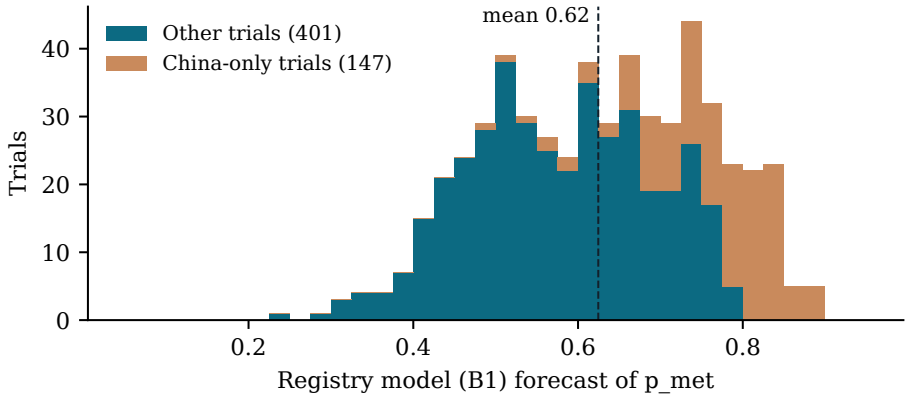


Figure 2: The registry model's forecasts for the 548 trials with a `p_met` question. China-only trials (147) sit higher.

Frontier models. Three models from Anthropic, `claude-opus-5-5`, `claude-sonnet-5-5` and `claude-haiku-5-5`, were run on 29 October 2026 on the Models board's rules, with every question: Claude Code subagents with default settings, 72 batches of up to 10 trials, one agent per batch, the packet and card only and no web access. Opus needed 7 batch reruns, Sonnet and Haiku 6 each. Their forecast files are fixed by the SHA-256 values in Table 4 but are not public while the round is open, so that no entry can copy the Opus reference that scales the number score; only their metadata and hashes are. They are published at the close of submissions, and anyone can then check them against the table. The Opus entry is the reference that scales and ranks the number score. A batch was rerun only when the checker refused an answer or the run log showed a tool other than reading the instructions and writing its own answers; reruns and their reasons are in the entries' metadata. One trial was removed before launch because Opus would not answer it in any run; removing the trial rather than the entry is a selection, and it is disclosed.

8 Limitations

The trials are a selected cohort: company-run, randomized, completing in one nine-month window, and only those that report by a deadline are scored for `p_met`. The scores describe forecasting this cohort's readouts, not trial forecasting in general. Reporting is not random: trials that report quickly lean towards success, which the conditional framing of `p_met` builds in but does not remove.

The labels depend on company statements, which vary in candour, and the rule that a vague statement counts as not met is an assumption. The labellers are language-model agents with organiser adjudication, not clinicians. Some results that were public before launch will have been missed; they are dropped when found. The pre-launch reviews were run by AI agents, not by human experts. Each account may hold three entries without a correction across them (Section 6). The 85 below-the-bar trials are concentrated in China. The question cards were written by agents and checked in samples, so some will be imperfect. The questions added between 29 and 31 October 2026 (the three deadlines, the number ranking, and the safety, key-secondary, subgroup, next-step and amendment questions) were checked by independent agents and their mechanics by the outside AI reviews, but their design had less review than the core. Round 1 has no human-expert arm, so the human numbers in the introduction come from other studies and other trials. The round has a single organiser.

ReadoutBench is a research benchmark. It forecasts trial-level outcomes from public information. It is not medical advice and not a tool for decisions about individual patients, and nothing in it is investment advice.

9 Timeline and availability

The round opened on 1 November 2026 and closes at 23:59:59 UTC on 15 November 2026. An early look scores results public by 31 March 2027, the ranked result those public by 30 June 2027, and a final, unranked pass those public by 31 December 2027; operations are read from the registry on 31 August 2027. The trial files, the rules, the code (MIT licence), every pre-launch check and the build inputs are linked from `https://readoutbench.com`. A second version of this paper will report the results.

Fingerprints. Table 4 gives the SHA-256 of the files that fix the round: the manifest, which holds the SHA-256 of every packet, of the question cards, of the operations rows, of the sister groups and of the forecasting instructions; the registry of reference entries; the forecast file of each reference entry; and the labelling instructions. Anyone can recompute them from the published files. This version of the paper was deposited on Zenodo on 1 November 2026, the day the round opened, and the deposit's date is the independent timestamp: a file whose SHA-256 differs from the table was changed after this paper. The public build package reproduces the manifest exactly; its own dry run registers only B0 and B1 and so gives a different registry fingerprint (`bb8c14e5...`), while the registry below is the organiser's build with all five reference entries.

Table 4: SHA-256 fingerprints of Round 1, computed on 31 October 2026, before the round opened.

File	SHA-256
`manifest.json`	`3df6536cd92ccd8e654242887647420fe09b23e0802a0ace6b5e986f9a8df910`
`registry.json`	`44eda736835cdf69101b27841cc6258d01fb4a4151f72e15110b0f0e231025d9`
B0 forecasts	`30f818f28c3f8c9947f0997d6877d61803309f59999e6313a1247f1e56a384e8`
B1 forecasts	`7e2115bd78ac20f02a68241a065f8a94059ded0eeb93b5eac2b30548f022e3b1`
Opus 5.5 forecasts	`b327da3d9ad0d5f3f46f80405801424e99657a4fd54fcad4a7afc6d73a363c76`
Sonnet 5.5 forecasts	`b54374184abee348401a9763fab90b158c92f64a653ea591b2fc5935870c336e`
Haiku 5.5 forecasts	`e910e8ae6cffddd123f33ec479be08808b2bce45b5c4364a95ef141d735cfcb0`
`label_instructions_r1.md`	`80246ddb90faf891b68320c12604263af31044398f1770a652dd8ae988048850`
`amend_label_instructions.md`	`2c9775060823740810f349223d08735c6e0b7fac18d4881177fd71ca8eb20dbb`
`next_step_label_instructions.md`	`57ca0bdc5e1ea9a5995cb328caada6e952fe34b7060e4dc77b7613bf06ae7671`

References

[1] Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2409.19839.

[2] Jianyou Wang, Youze Zheng, Longtian Bao, Hanyuan Zhang, Qirui Zheng, Yuhan Chen, Yang Zhang, Matthew Feng, Maxim Khan, Aditya K. Sehgal, Christopher D. Rosin, Ramamohan Paturi, Umber Dube, and Leon Bergen. CT Open: An open-access, uncontaminated, live platform for the open challenge of clinical trial outcome prediction. arXiv:2604.16742, 2026.

[3] Adélaïde Doussau, Patrick Kane, Jeffrey Peppercorn, Aden C. Feustel, Sylviya Ganeshamoorthy, Natasha Kekre, Daniel M. Benjamin, and Jonathan Kimmelman. The impact of feedback training on prediction of cancer clinical trial results. *Clinical Trials*, 21(2):143–151, 2024. doi: 10.1177/17407745231203375.

[4] Daniel M. Benjamin, Spencer P. Hey, Amanda MacPherson, Yasmina Hachem, Kara S. Smith, Sean X. Zhang, Sandy Wong, Samantha Dolter, David R. Mandel, and Jonathan Kimmelman. Principal investigators over-optimistically forecast scientific and operational outcomes for clinical trials. *PLoS ONE*, 17(2):e0262862, 2022. doi: 10.1371/journal.pone.0262862.

[5] Tianfan Fu, Kexin Huang, Cao Xiao, Lucas M. Glass, and Jimeng Sun. HINT: Hierarchical interaction network for clinical-trial-outcome predictions. *Patterns*, 3(4):100445, 2022. doi: 10.1016/j.patter.2022.100445.

[6] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[7] Johannes Bracher, Evan L. Ray, Tilmann Gneiting, and Nicholas G. Reich. Evaluating epidemic forecasts in an interval format. *PLoS Computational Biology*, 17(2):e1008618, 2021. doi: 10.1371/journal.pcbi.1008618.

[8] Chi Heem Wong, Kien Wei Siah, and Andrew W. Lo. Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20(2):273–286, 2019. doi: 10.1093/biostatistics/kxx069.