IntuitionLabs

INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

TrialGPT Clinical Trial Matching: Pilot Metrics and Review

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-11-23 · trialgpt clinical trial matching · trialgpt 2.0 · nih-trialbench · ai clinical trial matching validation · patient-to-trial matching metrics · hit rate precision at k · human in the loop eligibility screening · automated prescreening · clinical trial pilot acceptance criteria

Original article: https://intuitionlabs.ai/articles/trialgpt-2-sponsor-pilots-metrics-human-review



In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes: What the TrialGPT 2.0 Evidence Adds
4. Eligible, Recommended and Enrolled: Three Different Outcomes
5. Ground-Truth Protocol and Acceptance-Test Matrix
6. Data Analysis and Evidence
7. Implementation Considerations and Process Changes
8. Implications and Future Directions
9. Frequently Asked Questions (FAQs)
10. Conclusion

Executive Summary

TrialGPT clinical trial matching is best evaluated as a chain of separate outcomes (trial retrieval, criterion assessment, clinician recommendation, reviewer workload and downstream enrollment), not as a single accuracy number. The 1 September 2026 TrialGPT 2.0 preprint, revised as version 2 on 28 October 2026, is a useful worked example because its authors state that **eligibility is necessary but not sufficient for recommendation** (arxiv.org). The arXiv record lists no journal reference, so every figure below is author-reported preprint evidence, not an independently reproduced benchmark.

The preprint reports several distinct evaluations with different denominators. Its headline hit rate of about **91%** was computed on 188 of the 288 retrospective cases, where clinicians reviewed 440 patient-trial pairs; the abstract places the figure beside its 288-case count, but the full text's top-ranked review covers only those 188 cases, and the other 100 cases are the exhaustively reviewed NCI cohort (arxiv.org). Its screening-time reduction of **55.0%** (129.8 to 58.4 seconds per patient) comes from a separate 100-case cohort and excludes patient-context review and report generation (arxiv.org). Its prospective result of **90.9%** is a rise from 11 to 21 cases (out of 27) with at least one final recommendation (arxiv.org), and the authors state that downstream recruitment outcomes were not assessed (arxiv.org). These numbers answer different questions, and none of them is an enrollment rate.

Enrollment context explains why the distinction matters. A pooled analysis found that no trial was available at the treating institution for 55.6% of patients (pmc.ncbi.nlm.nih.gov), and in a genomic-testing program 11% of patients with actionable mutations went on to genotype-matched trials (pmc.ncbi.nlm.nih.gov). A matching system can raise the number of reviewed options without moving enrollment, so a pilot should log enrollment separately and never infer it from recommendations.

This report supplies an acceptance-test workbook: an evidence-denominator table, an eight-class error taxonomy with proposed synthetic cases, a metric sheet with formulas (hit rate@K, recommended precision@K, severe reversals, override, unresolved-information rate, review-time change), a source and version manifest, and a go/hold/stop decision record. Thresholds are deliberately left blank for the sponsor's clinical owners. Regulatory framing is generic: FDA's January 2026 clinical decision support guidance conditions non-device status on all four criteria, including that clinicians can independently review the basis of

recommendations (www.fda.gov), and clinicians retain eligibility and treatment decisions throughout. Version changes to the model, prompt, corpus or retrieval should trigger local revalidation, a position the preprint itself takes (arxiv.org).

Introduction and Background

Sponsor clinical-innovation leaders, oncology recruitment operations teams and clinical software owners increasingly receive proposals to pilot large language model (LLM) systems that read a patient summary and propose candidate trials. The question this report addresses is narrow: is such a system useful and safe enough to assist expert trial review in a defined local workflow, and what stopping and escalation criteria should be written down before the pilot starts? It is not a vendor ranking and does not endorse any product.

The original TrialGPT was developed at the US National Library of Medicine and has three modules, **TrialGPT-Retrieval**, **TrialGPT-Matching** and **TrialGPT-Ranking** (www.ncbi.nlm.nih.gov). The Matching module assigns each inclusion criterion one of four labels (Included, Not included, Not enough information, Not applicable) (www.ncbi.nlm.nih.gov). In the 2024 Nature Communications paper it reached a criterion-level accuracy of 0.873, close to the expert range (www.ncbi.nlm.nih.gov), on 1,015 patient-criterion pairs (www.ncbi.nlm.nih.gov). Its pilot user study, however, covered six oncology trials (www.ncbi.nlm.nih.gov), and the paper reports a time saving of about 42.6% (www.ncbi.nlm.nih.gov) while noting that accuracy differences between assisted and unassisted screening were not statistically significant (www.ncbi.nlm.nih.gov).

TrialGPT 2.0 is described in a preprint titled "Towards AI-Assisted Clinical Trial Matching: Practical Considerations, Multicenter Evaluation, and Real-World Deployment." The NLM project page lists "TrialGPT 2.0" as an award headline (www.ncbi.nlm.nih.gov); we did not find a public release page, repository or license for version 2.0 itself. For context, the original TrialGPT repository carries a notice that the software is a "United States Government Work" (github.com), and its README states that it is not intended for medical decision-making without review and oversight (github.com). Readers should confirm the status of any 2.0 code or data directly with the authors or the vendor.

IntuitionLabs is an AI and life-sciences advisory firm; it does not sell trial-matching software and is not one of the options discussed here. Its perspective in this report is limited to pilot design: define the decision, separate the outcomes, freeze the evidence, and write the stop rules before the first case is reviewed.

Synthetic cases proposed for local testing are labeled as examples, and all worked calculations use reported counts or hypothetical inputs.

Key Changes: What the TrialGPT 2.0 Evidence Adds

Change 1: From criterion accuracy to clinician recommendation

The 2024 evaluation asked whether the system's criterion-level labels matched expert labels. The 2026 preprint adds a second layer: whether a technically eligible trial is one a clinician would actually advise considering. It defines "Recommended" as meaning the trial is clinically advisable for consideration (arxiv.org),

and "Eligible but Not Recommended" as meaning the trial is not clinically advisable or not a priority in the current clinical context (arxiv.org). This is the central design point for a sponsor pilot, because a system can be right about eligibility and still produce a list that wastes reviewer time.

The prospective data show how large the gap can be. Of 339 system-generated candidate trials in the tumor-board cohort, UIC clinicians labeled 45 (13.3%) as Recommended and 282 (83.2%) as Eligible but Not Recommended (arxiv.org). A precision figure that counted both classes as correct (327 of 339, or 96.5%, by our arithmetic) would look far better than one that counted only clinician-recommended trials, and the two should be reported side by side.

Change 2: Several evaluation designs with different denominators

The preprint combines retrospective, prospective, synthetic-benchmark, public-benchmark and user-interface evaluations. *Table 1* below lists each with its population, reference label, outcome and main limitation, using figures reported in the preprint (version 2) unless noted.

Evaluation design	Population and denominator	Reference label	Outcome reported	Limitation to carry forward
Retrospective, top-ranked review	188 de-identified cases across four settings, 440 reviewed pairs	Clinician label on up to ten trials the system itself rated Highly Recommended	Hit rate about 0.91 from K=3 to K=10 (arxiv.org)	Review was restricted to model-selected trials, so hit rate is not recall over the full corpus (arxiv.org)
Retrospective, exhaustive (NCI)	100 cases against nine trials, 900 pairs (arxiv.org)	Three annotations per pair, discordance settled by a fourth clinician (arxiv.org)	System exact agreement 89.0% (arxiv.org); mean screening time 129.8 to 58.4 seconds	Two clinicians each reviewed pairs with and without assistance on opposite 50-case halves; timing annotators included study coauthors
Prospective tumor board (UIC)	27 cases, 339 pairs (arxiv.org); one referred case excluded because the patient had died	Final tumor-board recommendation list	Cases with at least one final recommendation rose from 11 to 21	Single workflow, limited period; routine and system lists were merged and reviewed together
NIH-TrialBench (synthetic)	126 vignettes, 990 reviewed pairs (arxiv.org); 1,373 candidate trials	One prespecified target trial per vignette, authored by 24 NIH investigators	Macro-averaged Recall@10 of 70% versus 54% for the original framework	Macro-average weights each institute equally; synthetic vignettes are not live patients
Public benchmarks	58 SIGIR cases with 3,835 judged pairs, plus TREC 2021 and 2022 cohorts	Pooled relevance judgments	Precision@10 up 4.5% relative to TrialGPT 1.0	Relative changes averaged across benchmarks; unjudged trials count as not relevant
Interface and survey	Item-level survey ratings (104 ratings; the preprint does not state a respondent count)	Likert ratings	Lowest items were perceived efficiency and decision confidence, at 3.85 and 3.92	Respondent count not reported, so the sample size cannot be judged; not an outcome study

The table shows that the preprint contains no single population on which every claim holds. The 91% hit rate, the 55% time reduction and the 90.9% increase come from three different cohorts of 188, 100 and 27 cases. Ask which row any vendor figure belongs to.

Change 3: Hit rate replaces recall as the primary retrospective metric

The preprint defines hit rate@K at case level as the fraction of cases with at least one trial judged Recommended in the top K (arxiv.org). It reports a hit rate of 0.91 at K=3 that stays at 0.91 through K=10, with eligible and recommended precision at K=10 of 0.87 and 0.83. The authors chose this metric because it reflects the intended workflow, and warn that it is not recall over the full trial corpus. For the synthetic benchmark the metric differs: Recall@10 is the share of vignettes whose known target trial appears among the top ten.

NIST's rationale for counting only eligible trials in precision-type metrics is that real users are extremely dissatisfied when shown trials they are explicitly excluded from (trec.nist.gov). Hit rate and ranking metrics have different jobs. The TREC Clinical Trials track, run by the US National Institute of Standards and Technology (NIST), used normalized discounted cumulative gain (nDCG) at ranks 5 and 10 as its main metric (trec.nist.gov) and in 2022 reported four primary metrics including NDCG@10 and P@10 (trec.nist.gov). NDCG applies a rank discount so that top-ranked documents count more. A pilot that cares whether a coordinator sees one good option should use hit rate; one that cares about how much poor material sits above the good option should add a rank-sensitive measure.

Change 4: The system is treated as a versioned artifact

The preprint states that decoding settings, code version, prompt and backbone model were fixed across evaluations (arxiv.org), that all main analyses used GPT-4.1, and that a change to the backbone, prompt, retrieval, corpus or policy should be treated as a new system version requiring local revalidation before clinical use. It also records the trial-corpus snapshot used for each prospective run for auditability. This is the right posture for a pilot, and it converts directly into the source and version manifest described later in this report.

Retrieval is activated only for local trial lists longer than 1,500 (arxiv.org), while retrospective local lists ranged from 9 to 1,871 candidate trials, so a sponsor with a small portfolio may never exercise the retrieval path. Local matching policies were also refined using clinician feedback from other cases, so ask how much tuning a local deployment needs.

Eligible, Recommended and Enrolled: Three Different Outcomes

A pilot scorecard that mixes these three outcomes will overstate value. The working glossary below is the one used throughout this report.

- **Eligible (technical):** the patient appears to meet the written inclusion and exclusion criteria on the available evidence. In TREC's 2022 judgments an Eligible trial is one the patient is eligible to enroll in (trec.nist.gov), and Eligible trials made up 11% of the 35,394 judged pairs, with 9% Excluded.
- **Recommended (clinical):** a clinician would advise considering the trial in the present context, which also reflects priorities, performance status, prior therapy and logistics.

- **Enrolled (operational):** the patient consented, passed screening and was registered. No evaluation in the preprint measures this.
- **Unresolved:** the record lacks the information needed to decide. FDA's decision support guidance recommends that output show knowns and unknowns, including missing or unexpected inputs (www.fda.gov), and the original TrialGPT carries an explicit "Not enough information" label (www.ncbi.nlm.nih.gov).

Why an eligible trial may be unsuitable to recommend

The preprint reports that the system occasionally applied eligibility criteria too literally. A trial can also be eligible on paper yet inappropriate because of travel burden, prior-therapy sequencing or a closed arm. The original TrialGPT benchmarks excluded factors such as geolocation and trial recruitment status (www.ncbi.nlm.nih.gov), so a user must check those separately. For a sponsor, this means the "Eligible but Not Recommended" class is expected output, not a defect, and the pilot should measure how much reviewer time it consumes.

Why recommended does not equal enrolled

Published oncology data show repeated attrition between the identification of an option and enrollment. In a meta-analysis of 13 studies, a trial was unavailable at the patient's institution 55.6% of the time (pmc.ncbi.nlm.nih.gov), 21.5% of patients with an available trial were ineligible (pmc.ncbi.nlm.nih.gov), and 14.8% of the remainder did not enroll (pmc.ncbi.nlm.nih.gov). The authors note that eligible patients may fail to enroll because they are not asked or because they decline (pmc.ncbi.nlm.nih.gov), yet when offered a trial they say yes more than half the time (pmc.ncbi.nlm.nih.gov).

Genomically matched programs show the same pattern. In a 2,000-patient testing protocol, 789 patients (39%) had at least one potentially actionable mutation (pmc.ncbi.nlm.nih.gov), but only 83 (11%) of those with actionable mutations went on genotype-matched trials (pmc.ncbi.nlm.nih.gov). In the NCI-MATCH interim analysis, 56 of 645 sequenced tumors (8.7%) had an actionable alteration for an open subprotocol (pmc.ncbi.nlm.nih.gov), only 33 (5.1%) of screened patients were eligible for assignment (pmc.ncbi.nlm.nih.gov), and 16 of those 33 (48.5%) enrolled (pmc.ncbi.nlm.nih.gov). Accrual shortfalls are a long-standing feature of oncology trials: among 238 phase III cooperative-group trials, 66% closed with sufficient accrual (pmc.ncbi.nlm.nih.gov). The eligibility criteria themselves are part of the problem. ASCO and Friends of Cancer Research stated that excessive or overly restrictive eligibility criteria can slow trial accrual (ascopubs.org), and FDA's 2020 oncology guidance notes some criteria are used as a template across trials without clear scientific or clinical rationale (www.fda.gov). Determining eligibility has long required a lengthy manual review of data for a single evaluation (pmc.ncbi.nlm.nih.gov). The enrollment figures above come from different eras and programs, so they are not benchmarks for TrialGPT. They show that a pilot should treat enrollment as a separate, slower measurement with its own denominators.

Ground-Truth Protocol and Acceptance-Test Matrix

Building reference labels

The reference label determines every metric downstream, so write the protocol first. The preprint's strongest labeling design is the NCI cohort, where each pair received three annotations, including one from an independent clinician blinded to system output, and a fourth clinician settled discordant pairs. For the four top-ranked cohorts, the text we read does not state the number of reviewers per pair, and its quality control is described as manual review of annotations. A sponsor pilot should specify the following before review begins.

- **Two independent reviewers** per case-trial pair, blinded to each other and, for at least a sample, blinded to system output.
- **Adjudication** by a third clinician for disagreements, with the adjudicator's identity and role logged.
- **An explicit unresolved label** for pairs where the chart lacks required evidence, reported as its own rate and never silently folded into "not eligible."
- **A temporal cutoff:** the patient record as of the decision date, not later information that was unavailable at the time.
- **Local trial availability:** whether the trial was open at the patient's treating site on that date.
- **Protocol and arm version:** the amendment in force, because a registry record can lag a protocol change.

Agreement should be reported as both percent agreement and kappa; the McHugh review recommends calculating both (pmc.ncbi.nlm.nih.gov), treats any kappa below 0.60 as inadequate agreement (pmc.ncbi.nlm.nih.gov), and suggests no fewer than 30 comparisons (pmc.ncbi.nlm.nih.gov). Manual screening is itself imperfect: a breast cancer pilot reported interrater reliability for manual screening of 0.60 to 0.77 (pmc.ncbi.nlm.nih.gov), and the SIGIR 2016 collection recorded 70% agreement on a control query (bevankoopman.github.io). The SIGIR collection's top grade is "highly likely to refer" (bevankoopman.github.io), a referral judgment rather than a formal eligibility determination, which is another reason to define labels locally. Labels from a human panel are a reference standard, not a truth.

Registry snapshots and status

Trial status is time-dependent. ClinicalTrials.gov data are refreshed daily Monday through Friday (clinicaltrials.gov), and federal rules allow the responsible party up to 30 calendar days to update overall recruitment status (www.ecfr.gov) and site status (www.ecfr.gov), with a similar window after an approved protocol amendment (www.ecfr.gov) (rule text current as of 19 November 2026 (www.ecfr.gov)). The registry's eligibility field is a limited free-text list intended to help potential participants find studies (clinicaltrials.gov), not a validated structured specification. Studies with unknown status are considered closed (clinicaltrials.gov), and a multi-site study's overall status is derived from site statuses (clinicaltrials.gov), so "Recruiting" at study level does not prove the patient's site is open.

The acceptance-test matrix

Table 2 below combines the error taxonomy, the evidence that motivates each class, the clinical consequence and the human action. The "proposed synthetic case" column contains illustrative cases we propose for local construction; they are not drawn from any study and are labeled as examples.

Error class	Proposed synthetic case (illustrative example)	Source evidence	Clinical consequence	Human action and pilot rule
1. Retrieval miss	A relevant trial for a rare tumor type is absent from the candidate list	Retrieval is only activated above 1,500 local trials; pooled public judgments mean unjudged trials count as not relevant (trec.nist.gov)	Patient never sees a suitable option	Seed known-positive trials; measure hit rate against a clinician-built reference list
2. Wrong biomarker or treatment-line match	Note says "second-line" after a regimen change; trial requires first-line	Temporal data feature in 40% of sampled criteria (escholarship.org); 85% had significant semantic complexity (escholarship.org)	Patient steered to an ineligible trial	Clinician checks line of therapy against chart; log as severe if recommended
3. Missing labs	Creatinine absent from the summary	Matching includes a Not enough information label (www.ncbi.nlm.nih.gov); unstructured notes were essential to 59% of CLL criteria in one study (pmc.ncbi.nlm.nih.gov)	False eligible or false ineligible	System must show the gap (www.fda.gov); unresolved-information rate tracked
4. Negation and chronology	"No prior immunotherapy" versus "prior immunotherapy stopped 2019"	Incorrect reasoning was the largest error type, at 30.7% of errors (www.ncbi.nlm.nih.gov)	Wrong exclusion or inclusion	Paired contrast cases in the test set; reviewer confirms in source note
5. Stale trial status	Trial closed to accrual at the local site last week	Registry updates may lag by up to 30 days (www.ecfr.gov); multi-site status can mask closed sites (clinicaltrials.gov)	Wasted referral	Coordinator verifies site status at the time of recommendation
6. Incorrect exclusion	A criterion applied too literally removes a trial a clinician would keep	Authors report occasionally literal reading; exclusion-criteria accuracy of 0.859 was below inclusion (www.ncbi.nlm.nih.gov)	Missed opportunity (false negative)	Review a sample of low-fit trials each week
7. Unsupported explanation	Rationale cites a sentence that is not in the note	Sentence-location was evaluated against experts (www.ncbi.nlm.nih.gov); FDA expects reviewable basis (www.fda.gov)	Clinician trusts a flawed rationale	Reviewer must be able to click through to source text; sample explanations
8. Confident output on unknown evidence	High fit score despite an unresolved biomarker	Fit score above 90 yields "Highly Recommended"; automation bias is a recognized risk (www.fda.gov)	Over-reliance	Cap category when critical evidence is unresolved; track override and reversal

The matrix gives each failure a test, a consequence and an owner. Classes 2, 4, 6 and 8 are the highest consequence because they can reverse a recommendation; the preprint's NCI analysis counted severe Highly Recommended to Low Fit reversals, with 1 pair under assisted review versus 14 unassisted and 8 for the system alone (arxiv.org), which is why "severe recommendation reversals" appears as its own line in the metric sheet below.

Data Analysis and Evidence

Reading the headline numbers correctly

Worked reading example (not a replication). The preprint's reported counts allow simple arithmetic that a reader can verify.

- **Prospective expansion.** Cases with at least one final recommendation rose from 11 to 21 of 27. The increase is (21 minus 11) divided by 11, which is 90.9%. As a share of all cases, the rise is 10 of 27 (37.0%), and the proportions are 40.7% (11 of 27) versus 77.8% (21 of 27). Our own Wilson 95% intervals on these counts are about 24.5% to 59.3% and 59.2% to 89.4%. They barely touch, at about 59%, and are marginal intervals on paired same-case proportions, so they are not a significance test (a paired test such as McNemar's would be the right comparison, and none was run here); they illustrate how wide uncertainty is at n=27; the Wilson interval is recommended for small samples (www.doi.org). We computed these intervals ourselves from the reported counts, and the preprint does not report them (it uses percentile bootstrap intervals over cases in figures).
- **Recommendation-level attribution.** The tumor board selected 54 final recommendations, of which the system identified 45 (83.3%) (arxiv.org), while the routine workflow identified 17. The abstract's 69% corresponds to the 37 recommendations identified by the system alone. For a sponsor, the useful numbers are the 10 cases and 37 recommendations added, not the percentage.
- **Wording note.** A pilot report should describe the 90.9% figure as an increase in the number of cases with at least one final recommendation, not as an enrollment or patient-access outcome.
- **Review time.** The preprint's definition of timing covers trial-level decision time and excludes patient-context review and automated PDF report generation. The relative change, (129.8 minus 58.4) divided by 129.8, is 55.0%. A separate runtime test excluded patient-context entry and manual clinician review. Neither is an end-to-end recruitment timing, and the original NIH news release gave a rounded 40% for the earlier study (www.nih.gov) while the paper reported 42.6% (www.ncbi.nlm.nih.gov); use the paper's figure and state the task definition.
- **Precision depends on the label definition.** At the tumor board, 13.3% of candidates were Recommended. Using Wilson intervals on 45 of 339, we obtain roughly 10.1% to 17.3%. Counting Eligible but Not Recommended as correct would produce 96.5%, a very different story from the same data.
- **Class-level agreement matters.** The system matched consensus on 89.0% of NCI pairs overall, but only 27 of 59 Possible Match pairs (arxiv.org) (45.8%, Wilson interval about 33.7% to 58.3%). Intermediate categories are where reviewer attention should concentrate.

Comparison with other published matching evaluations

Other published studies use different denominators and cannot be ranked against TrialGPT. A retrieval-augmented GPT-4 study analyzed 178 patient cases (pmc.ncbi.nlm.nih.gov) and reported 63.0% precision and 100.0% recall against tumor-board physician recommendations (pmc.ncbi.nlm.nih.gov). The RECTIFIER system for a heart-failure trial used 100 development, 282 validation and 1,894 test patients (medrxiv.org), and its authors recommend final clinician review before patient engagement (medrxiv.org). Judgments in TREC 2022 were pooled to depth 40 from 41 runs (trec.nist.gov), so precision and recall values are tied to the pool.

The metric sheet

Table 3 below is the local pilot scorecard. It defines each metric, its denominator and the reported reference value where one exists. The "local threshold" column is intentionally blank: acceptable values should be set with clinical owners and recorded before data collection, because no universal threshold is supported by the evidence reviewed here.

Metric	Definition and denominator	Reference point in the preprint	Local threshold (set by clinical owner)
Hit rate@K	Cases with at least one clinician-recommended trial in the top K divided by reviewed cases	About 0.91 retrospective, K=3 to 10	Blank: owner to define
Recommended precision@K	Clinician-recommended items divided by reviewed top-K items; state how cases with fewer than K candidates are handled	0.83 at K=10 retrospective; 13.3% of all prospective candidates	Blank
Eligible precision@K	Recommended plus Eligible but Not Recommended divided by reviewed top-K items	0.87 at K=10	Blank
Severe recommendation reversal	Pairs where the system rates a trial highly that adjudicated review rates ineligible or clearly unsuitable, per reviewed pair	8 system-alone pairs in the NCI cohort of 900	Blank, with a stop rule
Human override rate	Reviewer decisions that reverse a system rating, split by direction	Not reported as a single rate	Blank
Unresolved-information rate	Pairs labeled unresolved divided by reviewed pairs	Not reported	Blank
Review-time change	(baseline minus assisted) divided by baseline, same defined task only	55.0% in one 100-case cohort	Blank
Downstream enrollment	Consented and registered patients divided by recommended patients, tracked separately	Not assessed	Blank, long horizon

When a case has fewer than K reviewed candidates, the preprint averages precision over the smaller of K and the number of annotated Highly Recommended trials, so a local pilot must state its own rule. Hit rate is comparable across sites only if the unit of review is the same; showing ten trials versus three is a different task. Report hit rate with a confidence interval, because with fewer than about 30 cases the interval will be wide.

(Hypothetical Example) A site reviews 40 cases. In 31 cases a clinician recommends at least one trial in the top 5. Hit rate@5 is 31 divided by 40, or 77.5%. If the 40 cases produced 170 reviewed top-5 items and clinicians recommended 62, recommended precision@5 is 62 divided by 170, or 36.5%. If baseline review averaged 6.0 minutes per case on a defined task and assisted review 4.2 minutes, the review-time change is (6.0 minus 4.2) divided by 6.0, or 30%. These inputs are invented to show the arithmetic; they are not evidence about any system.

Implementation Considerations and Process Changes

Pilot design: frozen versus refreshed snapshots

Run two phases. A **frozen phase** fixes the trial corpus to a dated snapshot, so that any difference between runs is attributable to the system, not to registry drift. The preprint's retrospective cohorts used fixed pre-inference trial corpora, and the NIH-TrialBench search space was a predefined list of 1,373 trials. A **refreshed phase** then mirrors production: the prospective study used the live local trial list, recording the snapshot for each run. The refreshed phase is where stale-status errors appear, and it should have its own log. The original benchmark snapshot for TREC was a single ClinicalTrials.gov collection of 375,581 trial descriptions (trec.nist.gov), a reminder that benchmark corpora are fixed by design.

Source and version manifest

Keep a manifest in the pilot record with, at minimum, the following entries.

- **Model and configuration:** backbone model name and date, decoding settings, prompt version, and retrieval configuration, matching the preprint's practice of fixing them.
- **Code and data version:** repository commit or build identifier and license notice; for the original TrialGPT, the repository notice is a US Government Work statement (github.com).
- **Trial corpus:** registry source, snapshot timestamp, number of trials, and status-filter rules; the ClinicalTrials.gov API documentation tells users to check a data timestamp to confirm the most recent data (clinicaltrials.gov).
- **Patient context:** extraction method, note types included, date cutoff, de-identification method.
- **Label set:** reviewer roster by role, adjudicator, definitions of Recommended, Eligible but Not Recommended, Ineligible and Unresolved.
- **Change log:** every change to any entry above, with date and revalidation decision.

Human review, logging and escalation

Clinicians retain eligibility and treatment decisions; the preprint states that outputs were used to support clinician review, and Good Clinical Practice text states that the investigator retains ultimate responsibility when activities are delegated (database.ich.org). The same guideline asks that computerised systems be fit for purpose (database.ich.org) and that sponsors keep a record of important computerised systems used (database.ich.org). Logging should capture the input version, output, reviewer action, override reason and time stamps.

Automation bias is a documented risk. A systematic review retrieved 13,821 papers and included 74 on the topic (pmc.ncbi.nlm.nih.gov), and found that mitigators included training and emphasizing user accountability (pmc.ncbi.nlm.nih.gov). FDA's decision support guidance defines automation bias as over-reliance on an automated suggestion (www.fda.gov), and notes developers may need usability testing to show clinicians can independently review the basis (www.fda.gov). Pilots can insert known-wrong items to test whether reviewers catch them.

Governance and privacy controls (generic, not legal advice)

A sponsor should consult its privacy officer, institutional review board and counsel. Three generic reference points are worth noting. HHS states that both HIPAA de-identification methods, even when properly applied, retain some risk of identification (www.hhs.gov), and the Safe Harbor method addresses date elements directly related to an individual (www.hhs.gov), which matters for chronology-sensitive matching. HHS describes the minimum necessary standard as taking reasonable steps to limit use and disclosure of protected health information (www.hhs.gov). Ethics pathways differ by setting, so a sponsor should confirm the applicable oversight route for its own pilot.

Regulatory status is context-dependent, and it is not accurate to say all recruitment software is a regulated device. FDA's clinical decision support guidance, issued in January 2026, says a function is excluded from the device definition only if it meets all four criteria (www.fda.gov). FDA's January 2025 draft guidance on AI supporting regulatory decision-making is a draft distributed for comment (www.fda.gov); it defines the context of use (COU) as the specific role and scope of the AI model (www.fda.gov), expects rigor to be commensurate with model risk (www.fda.gov), and expects ongoing monitoring so the model remains fit for use (www.fda.gov). It also excludes some operational-efficiency uses that do not affect patient safety or the reliability of results (www.fda.gov). In the European Union, the AI Act requires high-risk systems to be designed for effective human oversight (eur-lex.europa.eu), with overseers aware of automation bias (eur-lex.europa.eu) and able to decide not to use the system (eur-lex.europa.eu); whether a given matching tool is high-risk is a legal determination outside this report. The voluntary NIST AI Risk Management Framework (www.nist.gov) calls for defined human oversight (nvlpubs.nist.gov) and post-deployment monitoring plans (nvlpubs.nist.gov).

Change triggers

Define in advance the events that force revalidation. The preprint's own position is that such changes make a new system version.

- **Model changes:** a backbone model change or deprecation, or a prompt or retrieval change.
- **Data changes:** a new registry schema, a new tumor type or site, or a shift in case mix. FDA's draft describes data drift as input that differs from the data on which the model was trained (www.fda.gov).
- **Performance signals:** a sustained rise in override or unresolved-information rates, or any severe reversal.

Go, hold and stop decision record

Write the record before the pilot and fill it at each gate. The items below are criteria types, not numeric values.

- **Go** (continue or expand): hit rate and recommended precision meet owner thresholds with intervals reported; severe reversals at or below the owner limit; override and unresolved rates stable; reviewers can trace explanations to source text; manifest complete and unchanged or revalidated.
- **Hold** (pause, fix, retest): any metric within the owner's warning band; a change trigger fired; agreement between reviewers below the pre-set level; stale-status errors above limit in the refreshed phase.
- **Stop** (end the pilot or the use case): a severe reversal that reached a patient decision; explanations found unsupported in a sample above the owner limit; inability to log or reproduce outputs; a privacy or governance requirement that cannot be met.
- **Escalate:** any single severe event goes to the clinical owner and quality or compliance lead the same day, with the case record and manifest attached.

Questions for the vendor or internal team

- Which evaluation row (retrospective, exhaustive, prospective, synthetic, public) supports each performance claim, and what is its denominator?
- How are cases with fewer than K candidates treated in precision@K, and are cases with zero Highly Recommended trials in the hit-rate denominator?

Implications and Future Directions

For sponsors, the practical implication is that a defensible pilot is built on denominators and definitions, not on a single percentage. The TrialGPT 2.0 preprint is unusually explicit about several of them, which makes it a good teaching example, but it is also author-reported. The authors identify limitations that apply to any transfer: a single prospective workflow over a limited period, US-centered trial corpora and workflows, and no assessment of downstream recruitment outcomes. A sponsor outside the United States, or one with a different case mix, should treat the reported figures as hypotheses for local testing.

Three developments would improve the evidence base. First, independent replication on a public benchmark with a posted code version would let evaluators separate system effects from reviewer effects. Synthetic cohorts (www.ncbi.nlm.nih.gov) cannot reveal real chart messiness. Second, prospective studies that follow recommended patients through contact, screening and enrollment would turn a recommendation metric into an operational one. Third, shared reporting conventions for hit rate, precision and unresolved labels would enable multi-site comparison.

IntuitionLabs' remit in such an engagement is pilot design, acceptance criteria and local clinical-data integration, with clinicians and the sponsor's governance bodies owning every eligibility and treatment decision. Final decisions about regulatory status, privacy basis and clinical use remain with the sponsor's qualified professionals.

Frequently Asked Questions (FAQs)

What is TrialGPT and how is clinical trial matching evaluated?

TrialGPT is an NIH-developed framework with retrieval, criterion-level matching and ranking modules (www.ncbi.nlm.nih.gov). Evaluation can occur at several levels: retrieval hit rate, criterion accuracy (0.873 in the 2024 paper (www.ncbi.nlm.nih.gov)), clinician-recommended precision, review time, and enrollment. The levels answer different questions and need separate denominators.

What metrics matter for AI clinical trial patient matching?

Use hit rate@K, recommended precision@K, severe reversals, override rate, unresolved-information rate and review-time change, with downstream enrollment tracked separately. For ranking quality, add NDCG or precision@10 as TREC does (trec.nist.gov).

How accurate is LLM patient-to-trial matching?

It depends on the task and denominator. The original paper reported 0.873 criterion-level accuracy (www.ncbi.nlm.nih.gov); the 2026 preprint reports 89.0% exact agreement with consensus on 900 NCI pairs but only 27 of 59 Possible Match pairs. These are author-reported results, not independently reproduced.

Does human-in-the-loop eligibility screening need special design?

Yes. Define reviewer roles, adjudication and override logging, and test for automation bias (pmc.ncbi.nlm.nih.gov). Clinicians keep the eligibility decision (database.ich.org).

What would sponsor pilot acceptance criteria look like?

The scorecard in *Table 3* with owner-set thresholds, a stop rule for severe reversals, a manifest and change triggers. No universal thresholds are supplied.

How do I validate AI eligibility criteria matching locally?

Use two reviewers with adjudication, a frozen dated snapshot, the matrix in *Table 2*, and intervals (www.doi.org).

Is TrialGPT open source?

The original TrialGPT code is published by NIH with a US Government Work notice (github.com). We did not find a public repository or license for TrialGPT 2.0, and the preprint says NIH-TrialBench is released to the research community. Confirm availability and terms before planning a deployment.

Conclusion

The central lesson of the TrialGPT 2.0 preprint for a pilot sponsor is methodological. It separates technical eligibility from clinical recommendation, reports several evaluations with different denominators, and states that downstream recruitment outcomes were not assessed. Reading it well means keeping the 188-case hit rate,

the 100-case timing result and the 27-case prospective result apart, and treating all of them as author-reported evidence from a preprint.

A local pilot should therefore be designed around five separate measurements (retrieval, criterion assessment, clinician recommendation, reviewer workload and enrollment), with reference labels from independent reviewers, a frozen and a refreshed snapshot phase, an explicit manifest, and a go/hold/stop record agreed before the first case. Thresholds belong to the clinical owners. Where the evidence is thin, as in the 27-case prospective cohort, the correct response is to report wide intervals and keep the decision with clinicians, not to assume transfer. The workbook here is a starting structure to be adapted by the sponsor's clinical, statistical, privacy and regulatory reviewers.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://arxiv.org/html/2609.01202v2
2. https://pmc.ncbi.nlm.nih.gov/articles/PMC6410951
3. https://pmc.ncbi.nlm.nih.gov/articles/PMC4550690
4. https://www.fda.gov/media/109618/download
5. https://www.ncbi.nlm.nih.gov/research/trialgpt
6. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11574183
7. https://github.com/ncbi-nlp/TrialGPT/raw/refs/heads/main/LICENSE
8. https://github.com/ncbi-nlp/TrialGPT
9. https://trec.nist.gov/pubs/trec30/papers/Overview-2021.pdf
10. https://trec.nist.gov/pubs/trec31/papers/Overview_trials.pdf
11. https://pmc.ncbi.nlm.nih.gov/articles/PMC7566320
12. https://pmc.ncbi.nlm.nih.gov/articles/PMC3977198
13. https://ascopubs.org/doi/10.1200/JCO.2017.73.7916
14. https://www.fda.gov/media/121317/download
15. https://pmc.ncbi.nlm.nih.gov/articles/PMC3500483
16. https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052
17. https://pmc.ncbi.nlm.nih.gov/articles/PMC8088869
18. https://bevankoopman.github.io/papers/sigir2016_clinicaltrials_collection.pdf
19. https://clinicaltrials.gov/data-api/api
20. https://www.ecfr.gov/current/title-42/chapter-I/subchapter-A/part-11/subpart-C/section-11.64
21. https://clinicaltrials.gov/policy/protocol-definitions
22. https://escholarship.org/uc/item/96k623g1
23. https://pmc.ncbi.nlm.nih.gov/articles/PMC4333685
24. https://www.doi.org/10.1214/SS/1009213286
25. https://www.nih.gov/news-events/news-releases/nih-developed-ai-algorithm-matches-potential-volunteers-clinical-trials

26. https://pmc.ncbi.nlm.nih.gov/articles/PMC11522650
27. https://medrxiv.org/content/10.1101/2024.02.08.24302376v1.full.pdf+html
28. https://database.ich.org/sites/default/files/ICH_E6%28R3%29_Step4_FinalGuideline_2025_0106.pdf
29. https://pmc.ncbi.nlm.nih.gov/articles/PMC3240751
30. https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html
31. https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/minimum-necessary-requirement/index.html
32. https://www.fda.gov/media/184830/download
33. https://eur-lex.europa.eu/legal-content/EN/TXT?uri=CELEX:32024R1689
34. https://www.nist.gov/itl/ai-risk-management-framework
35. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.