Baseline-Deviation Analysis for Automated Detection of Anomalous Design Choices in Oncology Clinical Trials

Raminderpal Singh
raminderpal@hitchhikersai.org
raminderpalsingh.com | scienceclaw.ai

v0.4 (March 2026)

Abstract

Clinical trial design decisions — endpoint selection, comparator strategy, biomarker-driven enrichment, and adaptive elements — determine what evidence the oncology field will generate years before results are available. Despite this, no systematic method exists for continuously detecting when a newly registered trial's design deviates meaningfully from established norms in its indication and therapeutic modality. We describe a computational framework for automated anomaly detection in oncology clinical trial design. The system constructs empirical baselines at the indication–modality–line-of-therapy level from ClinicalTrials.gov registry data, scores each new trial against its baseline using a weighted composite anomaly formula, and investigates high-scoring deviations through automated cross-referencing of six public biomedical APIs. We report operational results from the first deployment: 17 indication–modality pairs with 75 baseline files (21 combined, 54 line-specific) built from 4,000+ real trials, a composite scoring threshold that reduced daily candidates from 61% to 30% of matched trials, and first-in detection covering novel sponsors, first Phase 3 entries, and first combination patterns. The system is implemented using the OpenClaw multi-agent framework; all data collection, scoring, and structured output are handled by deterministic Python scripts, confining the language model to reasoning tasks. This version (v0.4) extends the v0.3 framework with: updated server specification (Apple Silicon M4 48 GB), and documentation of Thread 3 (Oncology Trial Insights), a companion KB-building and synthesis cycle that consumes confirmed anomaly findings as Tier 1 evidence and produces weekly synthesised intelligence at scienceclaw.ai/oncology-insights.html.

Keywords:
clinical trial design, anomaly detection, oncology, ClinicalTrials.gov, registry surveillance, endpoint selection, baseline-deviation, composite scoring, autonomous discovery, OpenClaw, EU CTIS

1. Introduction

The design of a clinical trial — its choice of primary endpoint, comparator arm, enrichment strategy, and statistical framework — is among the most consequential decisions in drug development. In oncology, as of March 2026, the ClinicalTrials.gov registry contains approximately 12,950 active Phase 2 and Phase 3 interventional trials spanning checkpoint inhibitors, antibody–drug conjugates (ADCs), bispecific antibodies, cell therapies, targeted small molecules, and radiopharmaceuticals. Within this landscape, individual indications exhibit strong design conventions: first-line NSCLC trials overwhelmingly use PFS as the primary

endpoint, while melanoma trials show a more balanced distribution across PFS, ORR, and RFS.

We describe a methods framework for automated, continuous anomaly detection in oncology clinical trial design. The system builds empirical baselines of design norms from structured registry data, scores each new trial against the relevant baseline using a weighted composite formula, investigates flagged deviations using a toolkit of biomedical data APIs, and applies a three-part confidence filter before reporting.

2. Design Principles

Anomaly detection, not landscape summarisation. The system does not attempt to comprehensively track all oncology trials. It builds baselines of what is normal, then scans for deviations. On days when no anomalies are detected, the system produces no output — silence is a valid result.

Narrow findings with confidence. Each reported finding concerns a specific trial, compound, or design choice. A finding must pass three checks: the deviation must be confirmed as real, verified as novel, and triangulated against at least one independent source.

Investigation, not just detection. Flagging a deviation is not sufficient. The system investigates each anomaly through regulatory guidance changes, competitor readouts, published validation studies, or safety signals. Findings are reported as 'explained' or 'unexplained'.

Evidence thresholds for pattern claims. The system does not assert field-level trends unless supported by data from at least five independent trials over a 30-day window.

Deterministic computation, agentic reasoning. All data collection, counting, scoring, and structured output are handled by deterministic Python scripts. The language model is used exclusively for reasoning tasks.

3. Data Sources and Integration

The framework integrates eight structured data sources, each accessed via free public REST APIs or GraphQL endpoints. No commercial data subscriptions are required.

Source	API	Role in Framework	Auth
ClinicalTrials.gov v2	REST API	Primary registry: trial search, endpoints, eligibility, status	None
EU CTIS	REST API (public)	EU-only trials; supplements ClinicalTrials.gov. Fully integrated in v0.3.	None
PubMed E-utilities	REST API	Literature: protocols, reviews, conference abstracts. Integrated in v0.3.	Free key
ChEMBL	REST API	Compound enrichment: mechanism, molecular properties, ATC	None
Open Targets	GraphQL API	Target enrichment: genetic evidence, disease associations	None
bioRxiv / medRxiv	Content API	Preprints on trial methodology, biomarker validation	None
openFDA	REST API	Regulatory: approvals, label changes, adverse event signals	Free key
WHO ICTRP	Web service	Global registry aggregator (access currently unavailable)	Application

Table 1. Data sources integrated into the anomaly detection framework.

3.1 ClinicalTrials.gov v2 API: Non-Obvious Behaviours

- Arms data resides at protocolSection.armsInterventionsModule.armGroups. The fields query parameter omits arms data; the full record must be requested when comparator classification is needed.
- Phase filtering uses the syntax AREA[Phase](PHASE2 OR PHASE3) combined with AREA[StudyType]INTERVENTIONAL.

- Date filtering uses AREA[LastUpdatePostDate]RANGE[{date},MAX]. The countTotal=true parameter is required for total counts.
- The API rate limit is 3 requests per second; scripts use a 0.4-second sleep between paginated requests.

3.2 EU Clinical Trials Information System (CTIS)

The EU CTIS has been mandatory for all clinical trial applications in the EU and EEA since January 2023. It provides a public REST API at euclinicaltrials.eu/ctis-public-api/search (POST). CTIS contains over 7,000 trials. Confirmed API quirks: the containAll field does not support OR syntax — a query such as 'cancer OR lymphoma' returns zero results. Individual search terms must be looped separately and results deduplicated by EUCT trial number. The containAny field also returns zero results. Date fields use DD/MM/YYYY format. Conservative pacing (1 request per second) is recommended.

3.3 Conference Abstract Monitoring via PubMed

A daily deterministic script (pubmed_conference_bridge.py) queries PubMed E-utilities for publications citing watchlisted and recently confirmed trial NCT IDs, using the secondary identifier field ([si]) as the search anchor. Nine key oncology journals are monitored: Journal of Clinical Oncology, Annals of Oncology, Blood, Lancet Oncology, NEJM, Lancet, Nature Medicine, Cancer Cell, and JAMA Oncology. Known limitation: the [si] field is not universally populated, so coverage is incomplete.

4. Baseline Construction

4.1 Scope and Filtering

Baselines are constructed for each indication–modality–line-of-therapy triplet within oncology. The three-level granularity is essential because design conventions vary substantially by line: first-line NSCLC trials overwhelmingly use PFS, second-line-plus favour ORR, adjuvant trials use DFS, neoadjuvant use pCR, and maintenance use PFS. The registry query scope is limited to Phase 2 and Phase 3 interventional trials.

4.2 Operational Baseline Data

Indication × Modality	Trials	Modal Endpoint	Line Sub-Baselines
NSCLC × checkpoint-inhibitor	619	PFS	1L(139), 2L+(215), adj(44), neo(45), maint(75)
NSCLC × combination-IO	603	PFS	1L(133), 2L+(206), adj(41), neo(38), maint(73)
Melanoma × checkpoint-inhibitor	368	other	1L(43), 2L+(135), adj(42), neo(22), maint(26)
RCC × checkpoint-inhibitor	181	ORR	1L(29), 2L+(60)
Bladder × checkpoint-inhibitor	165	ORR	1L(15), 2L+(46), neo(39)
Head & neck × checkpoint-inhibitor	373	ORR	1L(39), 2L+(151), adj(31), neo(37), maint(17)
Breast × ADC	166	PFS	2L+(68), adj(17), neo(34)
Breast × targeted-CDK	205	other	1L(26), 2L+(68), adj(23), neo(41)
Colorectal × targeted	642	PFS	1L(165), 2L+(176), adj(40), neo(49), maint(36)
Gastric × checkpoint-inhibitor	179	ORR	1L(31), 2L+(55), neo(20)
HCC × checkpoint-inhibitor	235	ORR	1L(37), 2L+(61), neo(15)
Ovarian × targeted-PARP	112	PFS	2L+(48), maint(20)
DLBCL × cell-therapy	75	ORR	2L+(33)
DLBCL × bispecific	46	other	2L+(20)
Multiple myeloma × bispecific	76	other	2L+(37)
AML × targeted	228	other	1L(46), 2L+(89), maint(70)
CLL × targeted	191	other	1L(52), 2L+(94), maint(16)

Table 2. Operational baselines (March 2026): 17 indication–modality pairs, 75 baseline files, 4,000+ trials.

5. Anomaly Detection and Scoring

5.1 Deviation Dimensions

Each newly registered or recently updated trial is compared against its relevant baseline across six dimensions: endpoint deviation, comparator deviation, enrichment deviation, sample size deviation, design deviation, and new entrant deviation.

5.2 Composite Anomaly Scoring

The weighted composite anomaly score sums per-dimension contributions. Weights reflect clinical importance: endpoint choice (3.0), comparator strategy (2.0), design architecture (2.0), biomarker enrichment (1.5), sample size (1.0). Within each dimension, the score is proportional to rarity. Trials scoring below the minimum threshold (default 3.0) are excluded.

Dimension	Weight	Score Formula
Endpoint	3.0	0 if in baseline top-3 endpoints, else (1 - pct/100) × 3.0
Comparator	2.0	0 if modal comparator, else (1 - pct/100) × 2.0
Biomarker	1.5	0 if modal biomarker, else (1 - pct/100) × 1.5
Sample size	1.0	0 if within p25–p75, else |log2(observed/median)| × 1.0
Design	2.0	0 if in baseline top-3 designs, else 2.0
First-in	+3.0 boost	Computed separately; +3.0 added if first_sponsor / first_phase3 / first_combination

Table 3. Composite anomaly scoring formula.

Metric	Original (combined baselines)	After line splitting	After scoring (threshold 3.0)
Matched trials	1,312	1,313	1,313
Trials flagged	~805 (61%)	573 (44%)	395 (30%)
Daily candidates (1-day scan)	~118	118	93

Table 4. Impact of progressive refinements on anomaly detection volume (30-day retrospective scan).

5.3 First-In Detection

A separate first-in detection stage checks the top 20 daily candidates for three patterns: first_sponsor (no prior trials in this indication–modality pair), first_phase3 (first Phase 3 entry for this drug in this indication), and first_combination (agents not previously combined in this indication). Any first-in pattern adds a +3.0 boost to the composite score.

5.4 Investigation

Each day, the top five anomaly candidates are investigated by an autonomous agent. Investigation priority ordering: (A) watchlist status changes with high priority (TERMINATED, WITHDRAWN); (B) first-in findings; (C) highest composite anomaly score. The investigation toolkit includes: ct_details.py, chembl_compound.py, ot_search.py, pubmed_search.py, biorxiv_search.py, and conference_search.py. Investigation is bounded by a decision budget of 10–20 API calls per candidate. Each investigated trial is classified as CONFIRMED_UNEXPLAINED, CONFIRMED_EXPLAINED, FALSE_POSITIVE, or DEFERRED.

6. Autonomous Direction Discovery

Four discovery mechanisms layer on top of the daily detection pipeline. All four are implemented as deterministic weekly scripts.

6.1 Unmatched Trial Analysis

A weekly script collects all unmatched trials from the preceding 30 days and groups them by condition and intervention type. When three or more cluster around the same indication–modality pair within 30 days, the cluster is flagged as a potential new baseline candidate.

6.2 Temporal Trend Detection

A weekly script reads 30–60 days of accumulated candidate data and computes directional trends within each baseline: endpoint drift, comparator shift, biomarker adoption, sample size trends. A trend is flagged when a

previously sub-10% design element appears in more than 25% of new trials in the window.

6.3 Cross-Trial Convergence Detection

A weekly script reads 14 days of candidates and groups them by unusual shared deviations. A minimum of three independent sponsors making the same unusual choice is required for a signal. Convergent independent decisions by different sponsors suggest genuine biological or regulatory consensus is forming.

6.4 Sponsor Portfolio Tracking

For the top 25 oncology sponsors, the system builds portfolio profiles from all active trials. A daily deviation detector flags four types of strategic moves: new indication entry, modality pivot, registration surge (two or more Phase 3 trials in the same indication within 90 days), and withdrawal cluster. Portfolio profiles are rebuilt monthly.

7. Validation and Feedback

The user replies to anomaly alert emails with structured labels: USEFUL: NCTxxxxxxxx or NOISE: NCTxxxxxxxx. The webhook server parses these replies and records each feedback item alongside the trial's metadata. A weekly analysis script computes precision (USEFUL / total labelled) across multiple dimensions: overall, by baseline indication, by deviation type, by anomaly score range, and by first-in type.

Deferred items: Thread 3 integration into the feedback loop (using insight record quality to calibrate synthesis prompts) is gated on at least 3 months of USEFUL/NOISE data accumulation from Thread 2 anomaly email feedback.

8. System Architecture

The framework runs on a dedicated server (Apple Silicon MacBook Pro M4, 48 GB RAM) using the OpenClaw multi-agent framework for orchestration. The language model backend is GLM-5:cloud (256K context window) accessed via Ollama Cloud.

8.1 Daily Cycle

Time (UK)	Stage	Script / Agent	What Happens
11:45	Watchlist check	watchlist_checker.py	Checks tracked trials for status changes, endpoint amendments, enr
12:00	Registry scan	retrospective_scan.py	Fetches new/updated NCT trials; scores against baselines; filters by
12:15	First-in detection	first_in_detector.py	Checks top 20 NCT candidates for novel sponsor/phase/combination
12:20	EU CTIS scan	ctis_scan.py	Fetches EU CTIS oncology trials updated in last 1 day; matches to c
12:30	Sponsor deviation check	sponsor_deviation_detector.py	Checks daily candidates against sponsor portfolio profiles
13:30	Investigation	trials-oncology agent (GLM-5)	Investigates top 5 with adaptive chains across ChEMBL, Open Targe
14:20	PubMed readout bridge	pubmed_conference_bridge.py	Queries PubMed [si] field for watchlisted and recent CONFIRMED_U
14:30	Email alert	trial_anomaly_emailer.py	Formats and sends HTML email. Silent if no confirmed findings.
14:35	Report publication	publish_anomaly_report.py	Generates HTML report; updates findings.json; deploys to Vercel

Table 5. Daily anomaly detection cycle (v0.4). Thread 3 runs 04:45–08:00 before Thread 2 begins.

8.2 Thread 3: Oncology Trial Insights (Companion System)

Thread 3 (Oncology Trial Insights) runs on the same server as a companion KB-building and synthesis cycle. It consumes confirmed Thread 2 anomaly findings (findings.json) as Tier 1 evidence, ingesting CONFIRMED_UNEXPLAINED and CONFIRMED_EXPLAINED classifications daily at 05:00 London. Thread 3 runs 04:45–08:00, completing before Thread 2 begins at 11:45, with GLM-5 calls staggered to prevent concurrent requests.

Time (UK)	Stage	Script	What Happens
04:45	KB maintenance	oncology_kb_maintenance.py	Hot/warm/cold rotation; log rotation
05:00	Trial findings ingest	ingest_trial_findings.py	Reads Thread 2 findings.json; writes trial-finding KB entries t
05:15	News scan	oncology_news_scan.py	Brave Search across 5 oncology topics; 204 articles/day typic
06:00	Topic analysis	oncology_topic_analysis.py	Two-pass GLM-5 analysis (Pass 1: direct findings; Pass 2: cr
07:00	Cross-synthesis	oncology_cross_synthesis.py	GLM-5 synthesis: themes, trial-news connections, gaps, qual
07:30	Insight generation	generate_insight_records.py	Deterministic extraction + focused GLM-5 per theme; Tier 1/2
08:00	Retrospective	oncology_retrospective.py	Quality review; term discovery (max 3 proposals/day, 2 signa

Table 6. Thread 3 daily cycle. findings.json from Thread 2 is the Tier 1 evidence source for Thread 3.

Thread 3 weekly output: every Friday at 09:00 London, the weekly page generator publishes insight records to scienceclaw.ai/oncology-insights.html. At 09:30, the email distribution script runs Kimi K2.5 verification on the content before sending. Quality score gate: 0.7+ sustained for 4+ consecutive retrospective runs before external distribution is enabled.

8.3 Technology Stack

Component	Demo Instance	Role
Agent framework	OpenClaw	Multi-agent orchestration, cron, session isolation
Analysis LLM	GLM-5:cloud (Ollama Cloud)	Investigation reasoning, synthesis, insight extraction
Verification LLM	Kimi K2.5 (Ollama Cloud)	Cross-model compliance verification (Thread 3 email gate)
Email I/O	AgentMail (scienceclaw@agentmail.to)	Webhook + REST API for inbound and outbound
Networking	Tailscale Funnel	Stable public HTTPS tunnel, survives reboots
Server	macOS, Apple Silicon M4 48 GB	Dedicated 24/7 server

Table 7. Technology stack for the demo instance.

8.4 Output: Findings and Insights

Thread 2 produces two complementary outputs:

- **Findings (2a)**: Raw anomaly detection output. Each confirmed finding is published as a standalone HTML report and indexed in a sortable flat findings table at scienceclaw.ai/findings.html, filterable by classification (Confirmed unexplained, Confirmed explained, False positive, Deferred). Email alert sent only when confirmed findings exist.
- **Insights (2b)**: Thread 3 synthesised intelligence, produced by the companion KB-building cycle. Where Findings shows what happened (deterministic, scored from registry data), Insights shows why it matters (synthesised from news + trial signals, reasoned by GLM-5, verified by Kimi K2.5). Published weekly at scienceclaw.ai/oncology-insights.html.

9. Illustrative Anomaly Scenarios

The following scenarios are drawn from the first two weeks of operational deployment (13–27 March 2026).

Scenario A: First-in sponsor in bladder IO

AstraZeneca registered a Phase 3 trial of olaparib in bladder cancer without BRCA or HRR biomarker enrichment. The system flagged this as a first_sponsor deviation and an enrichment deviation. Investigation via ChEMBL and PubMed found no published rationale for biomarker-unselected PARP inhibition in bladder cancer: CONFIRMED_UNEXPLAINED.

Scenario B: First Phase 3 for a novel mechanism

BioNTech's acasunlimab (a 4-1BB agonist) reached Phase 3 in NSCLC. The system detected a first_phase3 pattern: no 4-1BB agonist had previously entered Phase 3 in NSCLC. The +3.0 first-in boost elevated this trial above several higher-scoring baseline deviations.

Scenario C: Convergent mRNA signals in bladder cancer

In the same week, Roche (NCT06534983) and Merck (NCT06833073) both opened trials with mRNA-based components in bladder cancer. Both were classified as CONFIRMED_EXPLAINED by Thread 2. Thread 3's cross-synthesis identified the two trials as a convergence signal, producing a Modality expansion insight record published to the Insights tab.

Scenario D: PubMed readout alert

On 27 March 2026, the PubMed readout bridge identified a JAMA Oncology publication (PMID 36416836) for NCT02506153 (SWOG S1404 pembrolizumab melanoma trial), demonstrating the registration-to-results bridge operational.

10. Output Format

Anomaly alerts are emailed only when the daily cycle produces findings that pass the three-part confidence filter. Each alert contains: trial NCT ID or EUCT ID and sponsor, composite anomaly score and per-dimension breakdown, first-in badges (if applicable), investigation summary, significance assessment, and a Readout Alerts section when PubMed publications have been found for tracked trials.

Each confirmed finding is also published as a standalone HTML report at scienceclaw.ai, indexed in the sortable flat findings table (scienceclaw.ai/findings.html) filterable by classification via a five-button filter bar. Thread 3 insight records derived from these findings are published weekly at scienceclaw.ai/oncology-insights.html.

11. Limitations

Registry data quality. ClinicalTrials.gov registrations are sponsor-submitted and intentionally vague on design rationale.

Baseline sensitivity. The 'all-lines' bucket (16–22% of trials) hits combined baselines with wider distributions, reducing sensitivity.

CTIS baseline coverage. CTIS candidates are scored against combined baselines only — no line-of-therapy splitting, as CTIS records do not reliably encode treatment line.

Scoring weight calibration. The dimension weights and threshold (3.0) were set based on clinical judgment, not empirical optimisation. The feedback mechanism enables data-driven calibration but requires months of sustained engagement.

Thread 3 evidence mismatch risk. On the first production run, Kimi K2.5 identified NCT03698019 (a melanoma trial) being cited for NSCLC efficacy claims in Thread 3 insight records. This is a cold-start KB accuracy issue: with only news snippets and no topic-specific trial context, GLM-5 pulled nearby NCT IDs into themes they don't belong to. Expected to reduce as the news KB accumulates topic-specific context.

LLM reliability. The investigation stage depends on GLM-5's ability to follow multi-step investigation chains. Errors in this reasoning degrade output quality.

No prospective validation. Prospective evaluation metrics require a pilot deployment of at least 12 weeks with sustained expert feedback.

12. Discussion

This framework represents a shift from landscape analysis to anomaly detection. The composite scoring formula reduced daily candidates from 118 to 93 while preserving the most anomalous trials. The line-of-therapy split was the single most impactful refinement, reducing the flagging rate from 61% to 44%. CTIS integration extends registry coverage to EU-initiated trials not cross-registered on ClinicalTrials.gov.

Thread 3 adds a second intelligence layer: where Thread 2 detects what is anomalous, Thread 3 synthesises why it matters. The two threads are complementary rather than redundant: Thread 2 provides the Tier 1 evidence that anchors Thread 3's synthesis, and Thread 3 provides the contextual analysis that gives Thread 2's findings strategic significance.

A contrary viewpoint deserves explicit acknowledgment: an experienced oncology analyst scanning ClinicalTrials.gov once a week would likely identify the same high-impact findings. The system's advantage is consistency and coverage. Its disadvantage is that it cannot distinguish a statistically unusual trial from a strategically important one.

13. Data and Code Availability

All data was retrieved from publicly available APIs in March 2026. No proprietary data sources were used. The system is implemented using the OpenClaw multi-agent framework (open source) with GLM-5:cloud as the language model backend. The implementation guide is available from the corresponding author on request. The demo instance is deployed at scienceclaw.ai.

References

1. Singh R. Autonomous Market Intelligence for Science R&D;: A Multi-Agent Configurable Web Service with Cross-Model Verification. v1.1 (March 2026). [Companion paper describing the KB-building framework used by Thread 3.]

2. Zarin DA, Tse T, Williams RJ, Carr S. Trial Reporting in ClinicalTrials.gov — The Final Rule. N Engl J Med. 2016;375(20):1998–2004.

3. Califf RM, Zarin DA, Kramer JM, et al. Characteristics of Clinical Trials Registered in ClinicalTrials.gov, 2007–2010. JAMA. 2012;307(17):1838–1847.

4. Mendez D, Gaulton A, Bento AP, et al. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res. 2019;47(D1):D930–D940.

5. Ochoa D, Hercules A, et al. The next-generation Open Targets Platform. Nucleic Acids Res. 2023;51(D1):D1353–D1359.

6. Sever R, Roeder T, et al. bioRxiv: the preprint server for biology. Cold Spring Harb Perspect Biol. 2019;11(10):a035063.

7. Katz DH, Levin K, Gersing K. openFDA: An Innovative Platform. J Am Med Inform Assoc. 2016;23(3):596–600.

8. FDA. Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics: Guidance for Industry. 2018.

9. Beaver JA, Howie LJ, et al. A 25-Year Experience of US FDA Accelerated Approval of Oncology Products. J Clin Oncol. 2018;36(11):1126–1133.

10. Clinical Trials Regulation (EU) No 536/2014 of the European Parliament and of the Council. 2014.