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

Raminderpal Singh

raminderpal@hitchhikersai.org

http://raminderpalsingh.com | http://scienceclaw.ai

v0.2 (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 architecture is motivated by and leverages the OpenClaw multi-agent framework (openclaw.ai), but has been architected to mitigate the security challenges that OpenClaw's permissive agent execution model poses: all data collection, scoring, and structured output are handled by deterministic Python scripts rather than by the agent directly, confining the language model to reasoning tasks where its outputs are validated before action. This revised version extends the original methods framework in three directions: (1) a composite anomaly scoring formula with empirically derived weights, (2) a discovery layer for autonomous detection of emerging therapeutic trends, cross-trial convergence, and sponsor portfolio shifts, and (3) integration of additional public data sources including the EU Clinical Trials Information System (CTIS) and oncology conference abstracts via PubMed. We also describe a feedback mechanism for prospective validation of detection quality.

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

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. These choices are made by individual sponsors but collectively shape the evidentiary landscape for regulators, clinicians, and patients. A trial designed around progression-free survival (PFS) as its primary endpoint generates fundamentally different evidence than one designed around overall survival (OS), even when testing the same drug in the same indication.

In oncology, the volume and complexity of trial design decisions is particularly acute. As of March 2026, the ClinicalTrials.gov registry contains approximately 12,950 active Phase 2 and Phase 3 interventional trials in oncology—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 non-small cell lung cancer (NSCLC) immuno-oncology trials overwhelmingly use PFS as the primary endpoint, while melanoma trials show a more balanced distribution across PFS, objective response rate (ORR), and relapse-free survival (RFS). When a new trial deviates from these conventions, the deviation may signal a regulatory shift, a competitive response, a novel mechanistic hypothesis, or simply an idiosyncratic sponsor decision.

Currently, detection of such deviations relies on manual review by clinical development professionals, periodic landscape reports from commercial intelligence services, or retrospective academic analyses. No system continuously monitors the registry for design anomalies in real time, investigates potential explanations, and reports only those deviations that survive a confidence filter.

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 or updated 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. This revised paper extends the original framework with three additions: a quantitative scoring formula that prioritises the most anomalous trials for investigation, a discovery layer that autonomously identifies emerging trends and cross-trial convergence patterns, and integration of additional public data sources including the EU Clinical Trials Information System and oncology conference abstracts.

2. Design Principles

The framework is guided by five principles that distinguish it from conventional landscape analysis or registry dashboards.

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. Most trials match their baseline and are filed without further analysis. Only deviations receive investigation. On days when no anomalies are detected, the system produces no output—silence is a valid result indicating the field is behaving as expected.

Narrow findings with confidence. Each reported finding concerns a specific trial, compound, or design choice—not a broad trend. A finding must pass three checks before reporting: the deviation must be confirmed as real (not a data artefact), verified as novel (not previously reported), and triangulated against at least one independent source. This conservative filtering prioritises precision over recall.

Investigation, not just detection. Flagging a deviation is not sufficient. The system investigates each anomaly by searching for regulatory guidance changes, competitor readouts, published validation studies, or safety signals that might explain the deviation. Findings are reported as either "explained" (with the identified cause) or "unexplained" (investigation found no cause—often the most interesting category).

Evidence thresholds for pattern claims. The system does not assert field-level trends (e.g., "the field is shifting toward composite endpoints") unless supported by data from at least five independent trials over a 30-day accumulation window. Below this threshold, observations are noted as "possible signals" rather than confirmed patterns.

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: interpreting clinical context, following adaptive investigation chains, and synthesising findings into human-readable reports. This separation emerged from empirical observation that large language models are unreliable for schema-adherent structured output and file operations, but excel at clinical interpretation when

given pre-computed data.

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. The data sources serve distinct roles in the baseline construction, anomaly detection, investigation, and discovery stages.

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 coverage	None
ChEMBL	REST API	Compound enrichment: mechanism, molecular properties, ADMET	None
Open Targets	GraphQL API	Target enrichment: genetic evidence, disease associations, tractability	None
bioRxiv / medRxiv	Content API	Literature: preprints on trial methodology, biomarker validation	None
openFDA	REST API	Regulatory: approvals, label changes, adverse event signals	Free key
PubMed E-utilities	REST API	Literature: protocols, reviews, conference abstracts (ASCO/ESMO/ASH)	Free key
WHO ICTRP	Web service	Global registry aggregator: non-US/EU trials (gated access)	Application

Table 1. Data sources integrated into the anomaly detection framework. Sources above the line are fully integrated; CTIS and ICTRP are proposed additions. The ClinicalTrials.gov v2 API and ChEMBL REST API require no authentication; openFDA and PubMed offer free API keys that increase rate limits.

The ClinicalTrials.gov v2 API serves as the primary data source. It provides structured fields for study phase, conditions, interventions, primary and secondary outcome measures, eligibility criteria, sponsor, enrolment count, study design (allocation, masking, intervention model), and registration/update dates. The API supports filtering by condition, phase, study type, and update date, enabling targeted queries for newly registered or recently modified trials.

3.1 ClinicalTrials.gov v2 API: Non-Obvious Behaviours

Several non-obvious behaviours of the ClinicalTrials.gov v2 API were discovered during implementation and are documented here for reproducibility. Arms data resides at `protocolSection.armsInterventionsModule.armGroups` (not `.arms`). The `fields` query parameter omits arms data; the full record must be requested by omitting the `fields` parameter entirely 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; it is not returned in standard search responses. The API rate limit is 3 requests per second with no authentication required; scripts use a 0.4-second sleep between paginated requests.

3.2 EU Clinical Trials Information System (CTIS)

The EU Clinical Trials Information System 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) and `/retrieve/{EUCT_ID}` (GET), plus an RSS feed for recently updated trials. The API requires no authentication. CTIS contains over 7,000 trials

and is growing, with particular strength in EU-initiated academic trials and early-stage European biotech programmes that may not be cross-registered on ClinicalTrials.gov. Integration requires a mapping layer between CTIS and ClinicalTrials.gov data models, and deduplication against NCT IDs for cross-registered trials. Rate limits are undocumented; conservative pacing (1 request per second) is recommended.

3.3 Conference Abstracts via PubMed

Oncology conference abstracts represent a critical intelligence source: they contain clinical results data 12–18 months before full journal publication. ASCO Annual Meeting abstracts are published as supplements to the Journal of Clinical Oncology and indexed in PubMed with publication type "Meeting Abstract". ESMO abstracts appear in Annals of Oncology supplements. ASH (American Society of Hematology) abstracts appear in Blood supplements. All are searchable via PubMed E-utilities using the existing API integration. PubMed indexing of conference abstracts lags by weeks to months, so this source does not provide same-week coverage but is valuable for historical pattern detection and investigation enrichment.

3.4 Mechanistic and Literature Sources

ChEMBL and Open Targets provide the mechanistic context that registry data lacks. When a trial tests a novel compound, ChEMBL supplies the mechanism of action, target, molecular properties, and drug-likeness assessment. Open Targets supplies the genetic evidence linking the target to the disease indication. bioRxiv, PubMed, and openFDA serve the investigation stage: when a deviation is detected, these sources are queried for potential explanations including recently published biomarker validation studies, new FDA guidance documents, adverse event signals, or competitor readouts.

4. Baseline Construction

4.1 Scope and Filtering

Baselines are constructed for each indication–modality–line-of-therapy triplet within oncology. An "indication" is a specific tumour type (e.g., NSCLC, melanoma, DLBCL). A "modality" is the therapeutic approach (e.g., checkpoint inhibitor, ADC, bispecific antibody). A "line of therapy" is the treatment setting (e.g., first-line, second-line-plus, adjuvant, neoadjuvant, maintenance). The three-level granularity is essential because design conventions vary substantially by line: first-line NSCLC checkpoint inhibitor trials overwhelmingly use PFS as the primary endpoint, second-line-plus trials favour ORR, adjuvant trials use DFS, neoadjuvant trials use pCR, and maintenance trials use PFS. A single combined baseline would obscure these clinically distinct distributions.

The registry query scope is limited to Phase 2 and Phase 3 interventional trials in oncology indications. Phase 1 trials are excluded because their designs are near-formulaic (dose-escalation, 3+3 or Bayesian) and offer minimal design variation. Phase 1b/2 hybrid trials are also excluded because the ClinicalTrials.gov record does not reliably distinguish the dose-escalation portion from the expansion cohort.

4.2 Line-of-Therapy Classification

Line-of-therapy classification uses a section-aware regex classifier applied to four text fields from the registry record: title, brief summary, arms/interventions description, and eligibility criteria. The classifier assigns section-dependent weights: title and summary text receive full weight (1.0×), while eligibility criteria text receives reduced weight (0.5×) because eligibility criteria often mention prior treatment lines as exclusion criteria rather than as the trial's own setting.

A prior-treatment context suppression rule handles cases where eligibility criteria mention "prior adjuvant" or "progressed after neoadjuvant": in these cases, the adjuvant/neoadjuvant/maintenance scores are recalculated from high-priority text only, and the second-line-plus category receives a +5 boost. A compound pattern rule handles maintenance-after-first-line designs: when "maintenance" and "following/after first-line" co-occur, the first-line score is capped below the maintenance score. The classifier passes 28 of 28 test cases (23 synthetic text, 5 live ClinicalTrials.gov records).

Trials that cannot be confidently classified into a specific line are assigned to an "all-lines" bucket and scored against the combined baseline for their indication–modality pair. This affects approximately 16–22% of trials per baseline. Combined baselines have wider distributions, making anomaly detection less sensitive for unclassifiable trials.

4.3 Baseline Structure

Each baseline is a structured JSON object capturing the empirical distribution of design choices across all trials in its indication–modality–line triplet. The six baseline dimensions are: primary endpoint distribution (modal endpoint and frequency of alternatives), comparator strategy (active, placebo, none), enrichment biomarkers (selection and stratification factors), sample size range (interquartile range for Phase 2 and Phase 3 separately), design architecture (randomised, single-arm, adaptive, basket, umbrella, platform), and sponsor composition (set of active sponsors with trial counts). Additionally, each baseline stores an intervention names dictionary mapping drug names to frequency counts, enabling first-in detection.

4.4 Minimum Baseline Size

A baseline is considered usable for anomaly detection when it contains at least 15 trials. Below this threshold, the empirical distributions are too sparse to distinguish genuine anomalies from normal variation. Baselines with fewer than 15 trials are marked as "insufficient" and excluded from the detection phase. This threshold was chosen to balance detection sensitivity against false positive rate; the optimal threshold may vary by indication and should be empirically calibrated.

4.5 Operational Baseline Data

Table 2 summarises the 17 indication–modality pairs for which baselines have been constructed from ClinicalTrials.gov data as of March 2026. Each combined baseline was split into line-specific sub-baselines where sufficient trial volume exists.

Indication × Modality	Combined 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 as of March 2026: 17 indication–modality pairs, 21 combined baselines, and 54 line-specific sub-baselines, built from 4,000+ real trials on ClinicalTrials.gov. Modal Endpoint reflects the most frequent primary endpoint across all lines in the combined baseline. Abbreviations: 1L = first-line, 2L+ = second-line-plus, adj = adjuvant, neo = neoadjuvant, maint = maintenance; other = no single endpoint exceeds 50% frequency.

4.6 Baseline Refresh

Baselines are not static. The framework includes a monthly refresh cycle that reconstructs the top three baselines with the highest false positive rates, plus any baseline manually flagged as stale. This prevents the system from flagging normal field evolution as anomalous—if the field genuinely shifts (e.g., from PFS to a composite endpoint), the baseline should eventually reflect this shift.

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 (primary endpoint not in the baseline's top three), comparator deviation (comparator type diverges from the baseline modal), enrichment deviation (biomarker not present in baseline or standard biomarker omitted), sample size deviation (enrolment target outside the baseline p25–p75 range), design deviation (architecture not represented in baseline's top three), and new entrant deviation (sponsor or modality not previously seen in this indication).

5.2 Composite Anomaly Scoring

The original framework described deviation dimensions qualitatively. This revision introduces a weighted composite anomaly score that ranks candidates by severity. The score is the sum of per-dimension contributions, each scaled by a dimension-specific weight reflecting clinical importance:

```
endpoint: 0 if in baseline top-3 endpoints, else (1 - pct/100) × 3.0 comparator: 0 if modal comparator, else (1 - pct/100) × 2.0 biomarker: 0 if modal biomarker, else (1 - pct/100) × 1.5 sample_size: 0 if within p25-p75, else |log2(observed/median)| × 1.0 design: 0 if in baseline top-3 designs, else 2.0 first_in: 0 (computed separately; adds +3.0 boost if detected)
```

The weights reflect a clinical judgment hierarchy: endpoint choice has the largest impact on the evidence a trial will generate (weight 3.0), followed by comparator strategy (2.0) and design architecture (2.0). Biomarker enrichment (1.5) and sample size (1.0) are lower-weighted because they are more commonly explained by legitimate design variation. Within each dimension, the score is proportional to rarity: a trial using an endpoint present in 40% of the baseline scores lower than one using an endpoint present in 2% of the baseline. The sample size dimension uses a logarithmic scale to handle the wide range of enrolment targets across Phase 2 and Phase 3 trials.

Trials scoring below a minimum threshold (default: 3.0) are excluded from the candidate list. Table 3 shows the impact of line-specific baselines and composite scoring on candidate volume.

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 3. Impact of progressive refinements on anomaly detection volume, measured over a 30-day retrospective scan. Line-of-therapy splitting reduced the flagging rate from 61% to 44% by aligning baselines with clinically distinct trial populations. Composite scoring with a 3.0 threshold further reduced the rate to 30%, concentrating investigation on the most anomalous trials.

5.3 First-In Detection

A separate first-in detection stage checks the top 20 daily candidates for three patterns that indicate novel competitive entry. First, *first_sponsor*: the trial's lead sponsor has no prior trials in this indication–modality pair's baseline. Second, *first_phase3*: the trial's drug is entering Phase 3 for the first time in this indication, verified by querying the ClinicalTrials.gov API directly with intervention type filtering (DRUG/BIOLOGICAL only) to eliminate false positives from diagnostic procedures and radiation therapy entries. Third, *first_combination*: the trial combines two agents that have not been previously tested together in this indication. When any first-in pattern is detected, the trial's composite score receives a +3.0 boost, ensuring it appears in the investigation queue regardless of its baseline deviation score.

5.4 Investigation

Each day, the top five anomaly candidates are investigated by an autonomous agent. The investigation priority ordering is: (A) watchlist status changes with high priority (TERMINATED, WITHDRAWN), (B) first-in findings, (C) highest composite anomaly score. The agent follows deviation-specific investigation chains: endpoint deviations trigger searches for recent FDA guidance documents and methodology publications; novel compounds trigger ChEMBL mechanism lookups and Open Targets genetic evidence queries; trial terminations trigger safety signal searches in openFDA and status change checks in related trials. Investigations are bounded by a decision budget of 10–20 API calls per candidate.

5.5 Three-Part Confidence Filter

After investigation, each candidate must pass three sequential checks before reporting. Check 1: Is the deviation real? Confirms the deviation is not a data entry artefact, a classification mismatch, or an anomaly against a baseline with fewer than 15 trials. Check 2: Is it novel? Verifies the deviation has not been previously reported by the system and is not simply a single sponsor's programme producing multiple similar trials. Check 3: Can it be triangulated? Requires at least one corroborating source beyond the registry entry itself. For "explained" anomalies, the explanation must reference a specific document or event. For "unexplained" anomalies, the investigation must have covered at least three source types.

5.6 Watchlist Monitoring

Trials scoring above the anomaly threshold are automatically added to a persistent watchlist. A daily deterministic script checks all watchlisted trials for status changes: RECRUITING to TERMINATED (high priority), RECRUITING to WITHDRAWN (high), RECRUITING to SUSPENDED (medium), and RECRUITING to ACTIVE_NOT_RECRUITING (low). The script also detects endpoint amendments and enrolment changes exceeding 20% of the original target. Pruning removes COMPLETED trials after 90 days and TERMINATED/WITHDRAWN trials after 30 days. As of mid-March 2026, the watchlist tracks 93 active trials.

6. Autonomous Direction Discovery

The detection framework described above operates on fixed baselines and scores individual trials in isolation. It does not autonomously identify emerging therapeutic areas, detect convergent trends across trials, or recognise strategic shifts in sponsor behaviour. This section describes four discovery mechanisms that layer on top of the daily detection pipeline to address these gaps. All four are designed as deterministic weekly scripts that read accumulated daily candidate data, with optional monthly LLM interpretation sessions.

6.1 Unmatched Trial Analysis

The daily registry scan discards trials that do not match any of the 17 existing indication–modality baselines. These unmatched trials represent a direct signal of blind spots in baseline coverage. A weekly script collects all unmatched trials from the preceding 30 days, normalises their condition terms (using a regex-based synonym map, not LLM inference), and groups them by condition and intervention type. When three or more unmatched trials cluster around the same condition–modality pair within the 30-day window, the cluster is flagged as a potential new baseline candidate and included in the weekly discovery digest. This mechanism is designed to detect emerging therapeutic areas (e.g., bispecific antibodies in pancreatic cancer) before they accumulate enough trial volume to register through other channels.

6.2 Temporal Trend Detection

A weekly script reads 30–60 days of accumulated candidate JSON files and computes directional trends within each baseline: endpoint drift (is a previously rare endpoint appearing more frequently?), comparator shift (are more trials switching from placebo to active comparator?), biomarker adoption (is a new biomarker appearing in eligibility criteria?), sample size trends (is median enrolment growing or shrinking?), and phase distribution changes (is the Phase 2 to Phase 3 ratio shifting?). A trend is flagged when a previously sub-10% design element appears in more than 25% of new trials in the window. This mechanism detects field-level evolution—for example, the emergence of ctDNA-based endpoints in colorectal trials—and may also trigger baseline refresh when the shift is sustained.

Limitation: Statistical power is weak at small sample sizes. A baseline with 5 new trials per month where 2 use a rare endpoint is not a trend—it is two trials. Windows of 60–90 days may be needed for meaningful signals, introducing detection latency.

6.3 Cross-Trial Convergence Detection

This mechanism detects when multiple independent sponsors make the same unusual design choice within a time window. A weekly script reads 7–30 days of candidates and groups them by unusual shared deviations:

same non-modal endpoint, same non-modal biomarker, same novel drug target (resolved via ChEMBL mechanism data), or same first-in-indication entry. When two or more sponsors independently deviate in the same direction, the convergence is flagged as a high-value signal. This is arguably the highest-value intelligence pattern: convergent independent decisions by different sponsors suggest genuine biological or regulatory consensus is forming.

Limitation: Defining "same direction" deterministically is challenging. Same endpoint is straightforward; same drug target requires ChEMBL lookups that may not resolve for all trials. The overlap rate at 93 daily candidates is expected to be low—most days will have zero convergent pairs. The mechanism only works for unusual shared deviations; two sponsors both using ORR in melanoma 2L+ is the modal design, not a convergence signal.

6.4 Sponsor Portfolio Tracking

For the top 25 oncology sponsors, the system builds portfolio profiles from all their active trials across baselines: which indications, which modalities, which phases, which lines of therapy. A daily deviation detector then flags four types of strategic moves: new indication entry (sponsor has no prior trials in this baseline), modality pivot (sponsor historically in one modality, now opening a trial in another), registration surge (two or more Phase 3 trials opened in the same indication within 90 days), and withdrawal cluster (two or more trials terminated or withdrawn in the same indication within 90 days). Portfolio profiles are rebuilt monthly from full registry data.

Limitation: Sponsor names in ClinicalTrials.gov are inconsistent ("Merck Sharp & Dohme LLC", "MSD", "Merck & Co., Inc." are the same entity). A manually curated synonym map for the tracked sponsors is required. Portfolio profiles become stale if not rebuilt periodically; monthly reconstruction from full registry data is recommended, requiring 3,000+ API calls.

7. Validation and Feedback

The framework currently has no mechanism for validating whether reported anomalies are genuinely interesting to a clinical development professional. This section describes a feedback loop designed to close this gap.

7.1 Feedback Mechanism

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: anomaly score, score breakdown, baseline indication and modality, first-in types, investigation source, classification, and deviation types. Feedback is stored as an append-only JSON file. Unlabelled trials in a partially responded email remain unlabelled (not assumed to be NOISE).

7.2 Precision Analysis

A weekly analysis script computes precision (USEFUL / total labelled) across multiple dimensions: overall, by investigation source (watchlist / first-in / score), by baseline indication, by deviation type, by anomaly score range, and by first-in type. Score threshold analysis computes the precision–recall tradeoff at each candidate threshold (3.0, 3.5, 4.0, 4.5, 5.0), showing how many USEFUL findings would have been lost at higher thresholds. Results are reported in a weekly [FEEDBACK SUMMARY] email.

7.3 Limitations of Expert Feedback

Several biases constrain the feedback loop's reliability. Self-reinforcing bias: findings are labelled "USEFUL" based on the reviewer's existing understanding of competitive importance, which may optimise the system toward confirming priors rather than surfacing genuine novelty. Labelling inconsistency: the reviewer's attention varies day to day, introducing noise that is indistinguishable from signal quality variation. Small-sample instability: with 50 data points split across 17 baselines, most baselines will have 2–4 labelled findings, too few for reliable per-baseline analysis. Meaningful per-baseline precision requires 10+ findings

per baseline, which may take months to accumulate.

8. System Architecture

The framework is implemented as an autonomous agent pipeline running on a dedicated server (Apple Silicon MacBook Pro M1, 16 GB RAM) using the OpenClaw multi-agent framework for orchestration. The language model backend is GLM-5:cloud (256K context window) accessed via Ollama. All external data sources are accessed through thin Python wrapper scripts (stdlib only, no external pip dependencies) that each make a single API call and return structured JSON. The agent decides which scripts to call, in what order, and with what parameters.

8.1 Daily Cycle

Table 4 summarises the full daily detection cycle. All times are UK (Europe/London). Each stage runs as an isolated cron job with file-based handoffs between stages.

Time (UK)	Stage	Script / Agent	What Happens
11:45	Watchlist check	retrospective_scan.py (deterministic)	Checks tracked trials for status changes (TERMINATED, WITHDRAWN, SUSPENDED), endpoint amendments, and enrolment changes >20%. Produces watchlist alerts.
12:00	Registry scan	retrospective_scan.py (deterministic)	Fetches new/updated trials, scores against baselines across 6 dimensions using weighted composite formula. Filters by --min-score 3.0. Writes candidates JSON.
12:15	First-in detection	first_in_detector.py (deterministic)	Checks top 20 for novel sponsor/phase/combination. Adds +3.0 score boost to any first-in detection. Updates candidates JSON.
12:30	Sponsor deviation check	sponsor_deviation_detector.py (deterministic)	Checks daily candidates against sponsor portfolio profiles. Detects new indication entries, modality pivots, registration surges, and withdrawal clusters.
13:30	Investigation	trials-oncology agent (GLM-5 agentic)	Investigates top 5 candidates with adaptive investigation chains across ChEMBL, Open Targets, bioRxiv, openFDA, PubMed. Budget: 10–20 API calls per candidate. Writes investigated JSON.
14:30	Email alert	trial_anomaly_emailer.py (deterministic)	Reads investigated JSON. Formats and sends HTML email with score bars, badges, deviation tables, and investigation narratives. Silent if no confirmed findings.
14:35	Report publication	publish_anomaly_report.py (deterministic)	Generates standalone HTML report page. Updates reports.json manifest and findings.json flat table. Archives reports >30 days. Deploys to Vercel via CLI. Skips if no confirmed findings.

Table 4. Daily anomaly detection cycle. Each stage runs as an isolated cron job with file-based handoffs. Total daily processing time: 75–150 minutes. The pipeline runs 12:00–14:35 UK time, staggered to avoid LLM contention with a separate daily research thread (Thread 1, 02:00–13:00).

8.2 Memory Architecture

The agent operates with tiered memory to maintain constant token consumption across sessions. A "hot" tier (last 14 days) contains full findings and is loaded into every session (~12K tokens). A "warm" tier (days 15–60) contains weekly summaries (~5K tokens). A "cold" tier (60+ days) is archived and accessible only by file search. Weekly rotation summarises hot entries into warm entries and ages warm entries to cold. Total per-session memory budget is approximately 17–20K tokens, leaving the majority of the 256K context window available for the day's data and reasoning.

9. Illustrative Anomaly Scenarios

The following scenarios are drawn from the first week of operational deployment (13–20 March 2026) to illustrate the types of anomalies the system detects in practice.

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 (AstraZeneca had no prior trials in the bladder × checkpoint-inhibitor baseline) and an enrichment deviation (omitting biomarker selection in a PARP inhibitor trial is unusual). Investigation via ChEMBL and PubMed found no published rationale for biomarker-unselected PARP inhibition in bladder cancer, classifying the finding as "unexplained"—the most interesting category.

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, reflecting the clinical significance of a new mechanism class reaching late-stage development.

Scenario C: Convergent signals across sponsors. In the same week, Roche and Merck both opened trials with mRNA-based components in bladder cancer. Neither company had prior bladder cancer mRNA trials. The system detected both as individual *first_sponsor* anomalies but did not automatically link them. This convergence was identified through manual review, highlighting the need for the cross-trial convergence detection mechanism described in Section 6.3.

10. Output Format

The system produces anomaly alerts emailed only when the daily cycle produces findings that pass the three-part confidence filter. Each alert contains, for each finding: the trial NCT ID and sponsor, the composite anomaly score and per-dimension breakdown, first-in badges (if applicable), the deviation type, the baseline expectation, the observed deviation, the baseline size, a 2–4 sentence investigation summary, and a significance assessment.

On days with no reportable anomalies, no email is sent. Each confirmed finding is additionally published as a standalone HTML report page, deployed to Vercel, and indexed in a sortable flat findings table publicly accessible at https://scienceclaw.ai/findings.html. The table is filterable by date, indication, phase, score, and classification, and links directly to each day's full report page. On-demand queries are available via email: the user sends a question tagged with [TRIAL DESIGN] and the system produces a focused answer drawn exclusively from the accumulated knowledge base.

11. Limitations

Several significant limitations constrain the framework's applicability.

Registry data quality. ClinicalTrials.gov registrations are sponsor-submitted and intentionally vague on design rationale. Dosing logic, enrichment justification, and statistical assumptions are typically absent until the protocol paper publishes. The framework detects deviations in what is registered, not in the full protocol design.

Baseline sensitivity. Baselines are only as good as the registry data they are built from. Misclassified indications, inconsistent modality labelling, and incomplete endpoint descriptions introduce noise. The "all-lines" bucket (16–22% of trials) hits combined baselines with wider distributions, reducing sensitivity for unclassifiable trials. Small baselines (DLBCL × bispecific with 46 trials) produce less reliable distributions.

Scoring weight calibration. The dimension weights (endpoint 3.0, comparator 2.0, biomarker 1.5, sample size 1.0, design 2.0, first-in 3.0) and threshold (3.0) were set based on clinical judgment, not empirical optimisation. The feedback mechanism described in Section 7 is intended to enable data-driven calibration, but requires sustained engagement over months to accumulate sufficient labelled data.

False positive risk. The system will flag deviations that are normal variation, particularly in indication–modality pairs with heterogeneous baselines. The three-part confidence filter reduces but does not eliminate false positives. The monthly baseline refresh introduces a lag: if the field shifts rapidly, the baseline

may briefly be out of date, producing a burst of false positives.

LLM reliability. The investigation and confidence-filtering stages depend on a language model's ability to compare structured data, follow multi-step investigation chains, and make judgment calls about deviation significance. Errors in this reasoning degrade output quality in ways that are difficult to detect without manual review.

Detection versus assessment. The framework detects anomalous design choices. It does not assess whether those choices are good or bad. A trial using a novel endpoint may be innovative or misguided; a trial with an unusually small sample size may be elegantly designed or underpowered. Clinical judgment is required to interpret the significance of any finding.

Design intent versus clinical results. The system operates almost entirely in the "design intent" space (trial registrations describe what sponsors plan to do). The highest-value competitive intelligence is in the "what happened" space (conference abstracts and publications contain actual results). Conference abstract integration via PubMed (Section 3.3) partially bridges this gap but with a publication lag of weeks to months.

No prospective validation. Prospective evaluation metrics—false positive rate, detection sensitivity, time-to-detection versus manual review, and the proportion of "unexplained" anomalies that are later confirmed as meaningful—require a pilot deployment period of at least 12 weeks with sustained expert feedback.

12. Discussion

This framework represents a shift from landscape analysis (what is happening in oncology trials?) to anomaly detection (what is happening that shouldn't be?). The distinction matters because the volume of oncology trial registrations—approximately 12,950 active Phase 2/3 interventional studies—makes comprehensive monitoring impractical for any individual analyst.

The composite scoring formula introduced in this revision addresses a key weakness of the original qualitative framework: without quantitative ranking, all deviations received equal investigation priority regardless of severity. The weighted formula reduced daily candidates from 118 to 93 while preserving the most anomalous trials. The first-in detection mechanism with its +3.0 boost addresses a second weakness: genuinely novel competitive entries (a new sponsor in an established indication, a new mechanism reaching Phase 3) may have moderate baseline deviation scores but represent the highest-value intelligence signals.

The line-of-therapy split was the single most impactful refinement, reducing the flagging rate from 61% to 44%. This validates the clinical intuition that trial design conventions vary substantially by treatment setting within the same tumour type. First-line and second-line-plus trials in the same indication often have different modal endpoints, different comparator strategies, and different sample size expectations. A combined baseline treats these as a single population and flags legitimate second-line designs as anomalous against first-line norms.

The autonomous discovery layer (Section 6) addresses the framework's most fundamental limitation: it detects anomalies within fixed baselines but cannot discover new areas of investigation. The four proposed mechanisms—unmatched trial analysis, temporal trend detection, cross-trial convergence, and sponsor portfolio tracking—are designed to surface emerging signals that no individual trial-level anomaly would reveal. However, all four are proposed and untested. Their value depends on the volume and quality of accumulated candidate data, which increases over time.

A contrary viewpoint deserves explicit acknowledgment: an experienced oncology analyst scanning ClinicalTrials.gov once a week would likely identify the same high-impact findings faster than an automated system that processes 93 candidates per day to surface 5. The system's advantage is consistency (it runs every day, never forgets) and coverage (it checks all 17 baselines simultaneously). Its disadvantage is that it cannot tell the difference between a statistically unusual trial and a strategically important one. That

distinction requires domain judgment that neither a language model nor deterministic scoring can reliably provide. The feedback mechanism (Section 7) is designed to gradually close this gap, but closing it fully may not be achievable.

The reliance on free, public APIs is both a strength and a constraint. It makes the framework reproducible without commercial data subscriptions. However, undocumented APIs (such as CTIS) can change without notice. The WHO ICTRP crawling service is already "currently not available." Each new external dependency increases operational fragility. The framework's core detection capability depends only on the ClinicalTrials.gov v2 API, which is stable, documented, and operated by a US government agency; the enrichment and discovery sources are valuable but not load-bearing.

13. Data and Code Availability

All data presented in this paper was retrieved from publicly available APIs (ClinicalTrials.gov v2 API, ChEMBL REST API, Open Targets Platform GraphQL API) 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, including the full agent configuration, skill files, utility script specifications, and baseline JSON schemas, is available from the corresponding author on request.

References

1. 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.

2. Califf RM, Zarin DA, Kramer JM, Sherman RE, Aberle LH, Tasneem A. Characteristics of Clinical Trials Registered in ClinicalTrials.gov, 2007–2010. *JAMA.* 2012;307(17):1838–1847.

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

4. Ochoa D, Hercules A, Mouber BM, et al. The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Res.* 2023;51(D1):D1353–D1359.

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

6. Katz DH, Levin K, Gersing K. openFDA: An Innovative Platform Providing Access to a Wealth of FDA's Publicly Available Data. *J Am Med Inform Assoc.* 2016;23(3):596–600.

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

8. Chen EY, Raghunathan V, Prasad V. An Overview of Cancer Drugs Approved by the US FDA Based on the Surrogate End Point of Response Rate. *JAMA Intern Med.* 2019;179(7):915–921.

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

10. Gyawali B, Hey SP, Kesselheim AS. Assessment of the Clinical Benefit of Cancer Drugs Receiving Accelerated Approval. *JAMA Intern Med.* 2019;179(7):906–913.

11. Namiot ED, Smirnovova D, Sokolov AV, et al. The international clinical trials registry platform (ICTRP): data integrity and trends in clinical trials, diseases, and drugs. *Front Pharmacol.* 2023;14:1228148.

12. Clinical Trials Regulation (EU) No 536/2014 of the European Parliament and of the Council. *Official Journal of the European Union.* 2014;L158:1–76.