SYNTHĒSIS · METHODS / SYNTHESIS

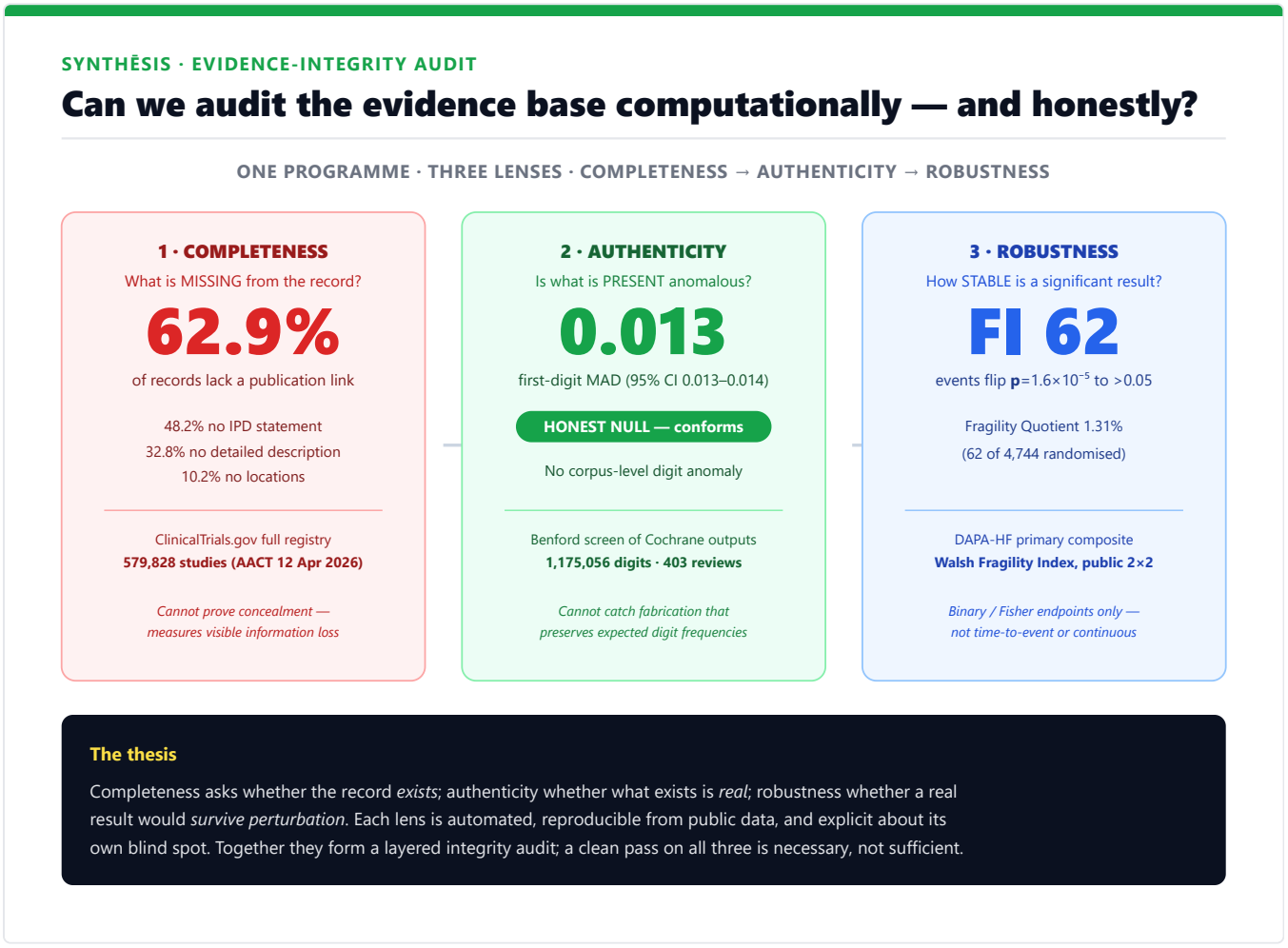
Three Lenses on a Trustworthy Evidence Base

Completeness, authenticity, and robustness as a reproducible computational integrity audit of the trial-evidence ecosystem

Zaki Khan

Ahmadiyya Hospital, Kenema, Sierra Leone

Corresponding author: zakikhan101@gmail.com



Visual abstract. Three complementary lenses on trial-evidence integrity — structural missingness (completeness), Benford digit forensics (authenticity), and statistical fragility (robustness) — each computed automatically from public data and each paired with what it cannot detect. All figures verified or independently recomputed (see verification table).

TYPE & ESTIMANDS

Methods / synthesis. Three estimands: (1) field-level structural missingness across the full ClinicalTrials.gov registry; (2) corpus first-digit mean absolute deviation (MAD) under Benford's law; (3) the Walsh statistical Fragility Index of a landmark trial's primary 2×2.

ABSTRACT

Background. Trust in the clinical-trial evidence base is usually defended one study at a time, by hand. We ask whether it can instead be audited computationally, at registry and corpus scale, and whether such auditing can be honest about its own blind spots.

Methods. We assemble three automated, reproducible lenses on a single programme — auditing the trustworthiness of the trial-evidence ecosystem — and apply each to public data: (a) field-level structural missingness across the full ClinicalTrials.gov registry (AACT snapshot, 12 April 2026; 579,828 studies); (b) a Benford leading-digit screen of 1,175,056 numeric values across 403 Cochrane reviews (the Pairwise70 corpus); and (c) the Walsh statistical Fragility Index recomputed from a landmark trial's public 2×2 table (DAPA-HF primary composite).

Results. Completeness is poor and uneven: 62.9% of records lacked publication links, 48.2% lacked IPD sharing statements, 32.8% lacked a detailed description, and 10.2% lacked locations, with industry the most affected named sponsor class. Authenticity returns an honest null: the corpus first-digit MAD was 0.013 (95% CI 0.013–0.014), within Nigrini's marginal-conformity band, with no corpus-level digit anomaly. Robustness is modest where it matters: a result significant at $p = 1.6\times10^{-5}$ rested on a Fragility Index of 62 events out of 4,744 randomised (Fragility Quotient 1.31%).

Conclusion. Completeness, authenticity, and robustness form a layered, deployable integrity audit — but each lens is bounded, and naming those bounds is part of the method, not a footnote to it.

INTRODUCTION

Can we audit the trustworthiness of the clinical-trial evidence base computationally, at scale, and honestly? Three quiet failure modes erode that trust, and they are usually examined in isolation. Information can be **missing** from the record before any result is even posted. The numbers that are present can be **anomalous** — fabricated, rounded into existence, or computationally distorted. And a result that is present and authentic can still be **fragile**, hanging on a handful of events. Each has its own literature and its own bespoke, manual workflow. None, on its own, tells you whether a body of evidence can be trusted.

This article argues that the three belong together as one programme, and that the programme can be run by software. We frame trust as a sequence of questions a reader should be able to ask automatically of any evidence base: *Is the record complete? Is what is present authentic? Is a significant result robust?* These map onto three measurable lenses — structural missingness, digit forensics, and statistical fragility — that advance from **completeness** to **authenticity** to **robustness**. The contribution is not a new estimator but a stance: that integrity auditing should be reproducible, deployable, and candid about what each tool cannot see. To make that stance concrete we run all three lenses on public data, reuse or independently recompute every headline number, and let an honest null stand where one appears.

LENS 1 — COMPLETENESS: WHAT IS MISSING FROM THE RECORD

The first question is the cheapest to ask and the most often skipped: what information is simply absent? Working from a full-registry AACT snapshot of ClinicalTrials.gov (12 April 2026; 579,828 studies), we computed field-level omission rates with no sampling. The picture is one of pervasive, structured sparsity (Figure 1). Across the registry, 62.9% of records carried no publication link, 48.2% no IPD-sharing statement, 32.8% no detailed description, and 10.2% no locations; only the core outcome fields were near-complete (3.0% missing). The detailed-description figure is the instructive one: the naïve "no row" metric reports 0.17%, but most rows exist while being empty — the correct empty-or-absent metric is 32.8%, a twenty-fold difference that a careless audit would miss.

Sparsity is also unevenly distributed across sponsors. Industry was the most affected named class on a composite hiddenness score (0.473; 71.2% of industry records lack a publication link), while the National Institutes of Health, low on the composite overall, carried the single highest rate of missing IPD-sharing statements (82.9%). A residue of 965 records carrying no sponsor class at all were missing every field — malformed metadata rather than evidence of intent. This is the completeness layer: it bounds how much of

the record even exists to be scrutinised. Its limit is equally clear — it measures registry-visible information loss, not concealment; an absent field is not proof of a hidden result.

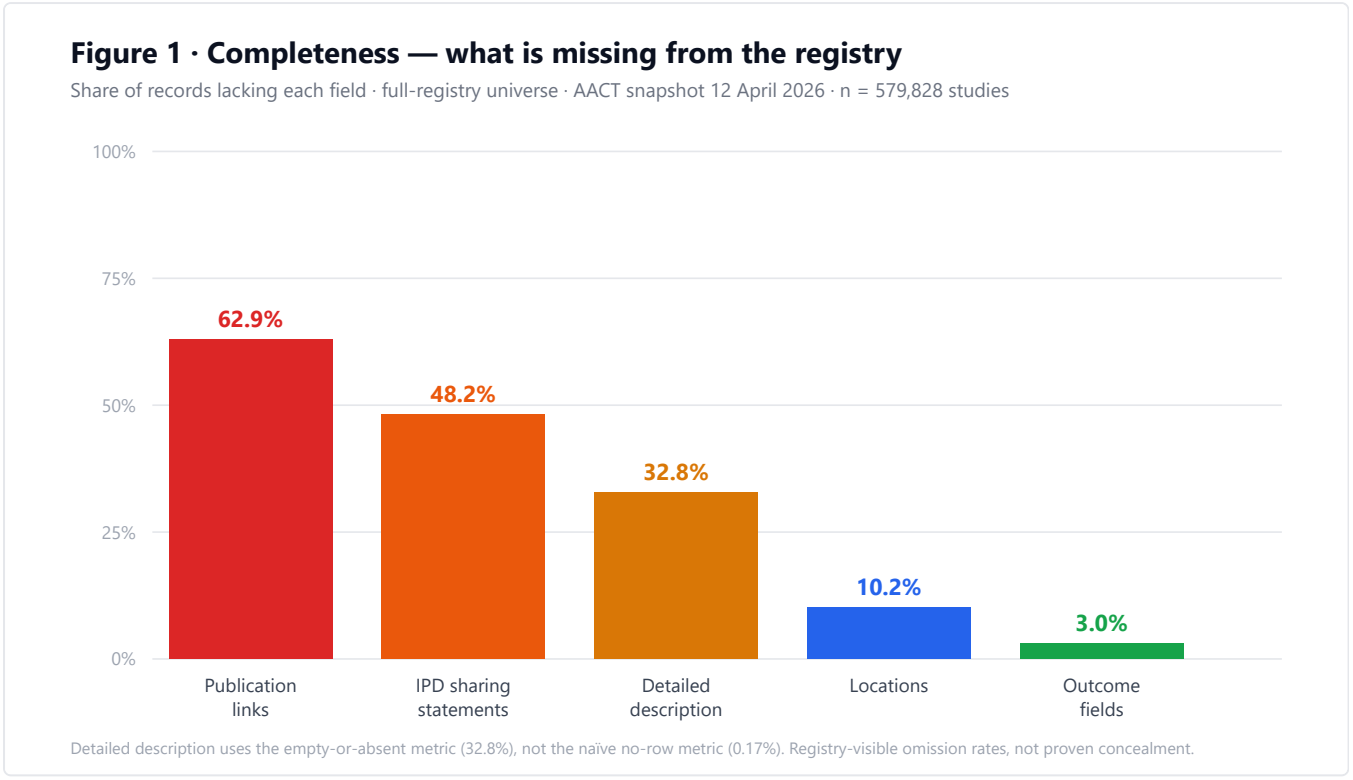


Figure 1. Field-level structural missingness across the full ClinicalTrials.gov registry (n = 579,828). Publication links 62.9%, IPD sharing statements 48.2%, detailed description (empty or absent) 32.8%, locations 10.2%, outcome fields 3.0%.

LENS 2 — AUTHENTICITY: IS WHAT IS PRESENT ANOMALOUS?

Completeness says nothing about whether the present numbers are real. The second lens screens the numbers themselves. Benford's law predicts the leading-digit frequencies of many naturally occurring quantities, and large departures can flag fabrication or computational error. We extracted 1,175,056 leading digits — point estimates, standard errors, and p-values — from 403 Cochrane reviews in the Pairwise70 corpus and compared them with Benford expectations (Figure 2).

The result is an honest null. The corpus first-digit MAD was 0.013 (95% CI 0.013–0.014), inside Nigrini's marginal-conformity band: at corpus scale these meta-analytic outputs show no digit anomaly suggesting systematic fabrication. Reporting this faithfully requires resisting two temptations. First, at over a million values the omnibus chi-squared and mantissa-arc tests both reject uniformity ($p \approx 0$) — but this is over-power, not a signal, which is exactly why the sample-size-robust MAD is the decision metric. Second, the fields are not homogeneous: effect estimates (MAD 0.009) and standard errors (0.009) conform, while p-values deviate (0.029) — expected for a bounded quantity, not a fabrication flag. No single review exceeded the critical threshold after Bonferroni correction. The authenticity layer thus delivers reassurance, and its boundary is sharp: Benford is indirect and cannot detect fabrication that deliberately preserves the expected digit frequencies. A null here means "no anomaly of this kind," not "all genuine."

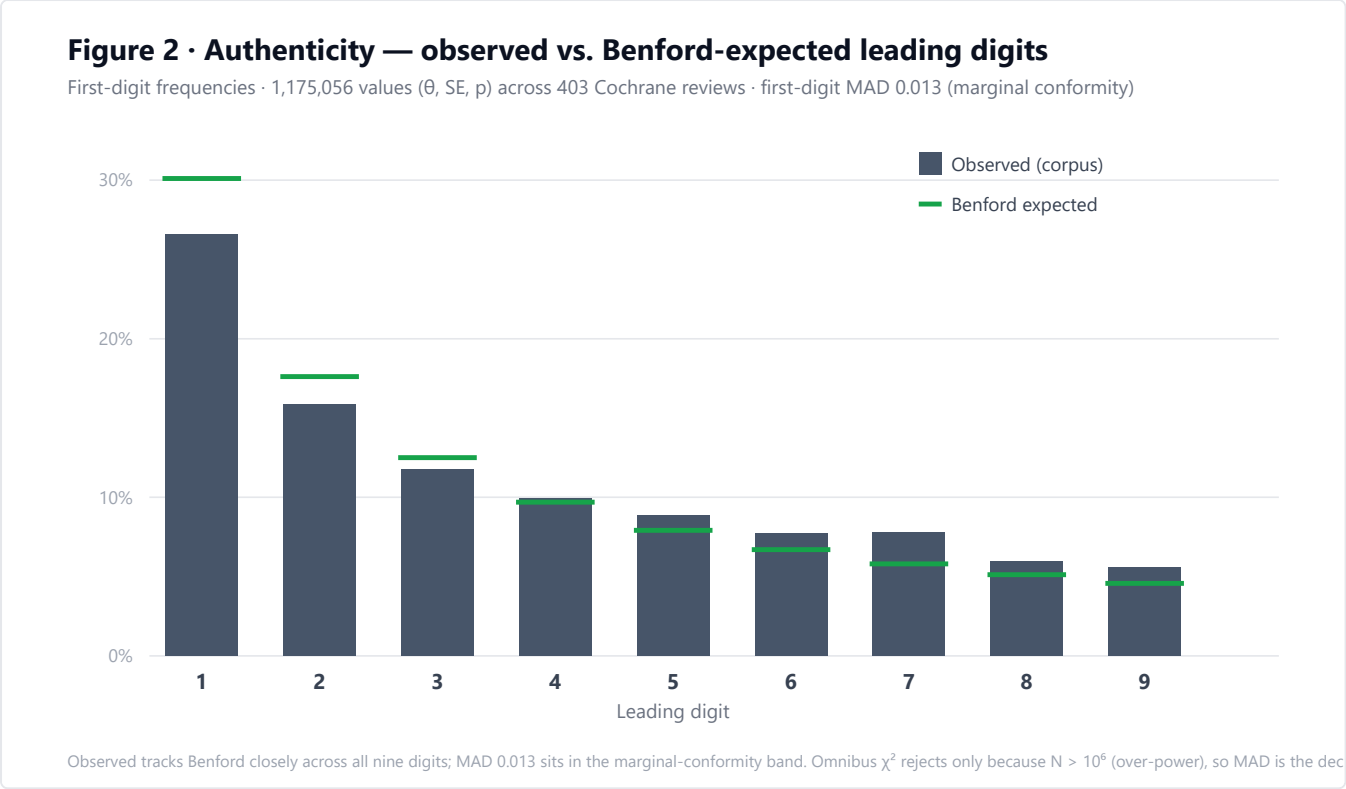


Figure 2. Observed corpus leading-digit frequencies (slate bars) against Benford-expected frequencies (green caps) for 1,175,056 values. The first-digit MAD of 0.013 lies within Nigrini's marginal-conformity band — an honest null, with no corpus-level digit anomaly.

LENS 3 — ROBUSTNESS: HOW STABLE IS A SIGNIFICANT RESULT?

Suppose the record is complete and the numbers are authentic; a positive result can still rest on almost nothing. The third lens quantifies that with the Walsh (2014) Fragility Index — the number of single-patient event reversals needed to push a significant result above $p = 0.05$. We recomputed it from the public 2×2 of a landmark trial, DAPA-HF (dapagliflozin in heart failure; primary composite 386/2,373 vs 502/2,371; Fisher two-sided $p = 1.6\times10^{-5}$, hazard ratio 0.74).

Adding events to the smaller-event arm, the result crossed $p = 0.05$ after **62** reversals — a Fragility Index of 62 against 4,744 randomised, a Fragility Quotient of 1.31% (Figure 3). For a result this statistically extreme that is reassuringly robust, yet it sharpens the general lesson: significance and stability are different properties, and the second is invisible in a p-value. Two boundaries apply. The Fragility Index is defined only for binary outcomes under Fisher's exact test — it does not extend to time-to-event or continuous endpoints without modification — and a low absolute count is not automatically alarming for a small or short trial. (We note a verification flag: a Fragility Index of 29 had been asserted for this endpoint, but does not reproduce from the published 2×2 under any standard convention; the verified value of 62 is used throughout.)

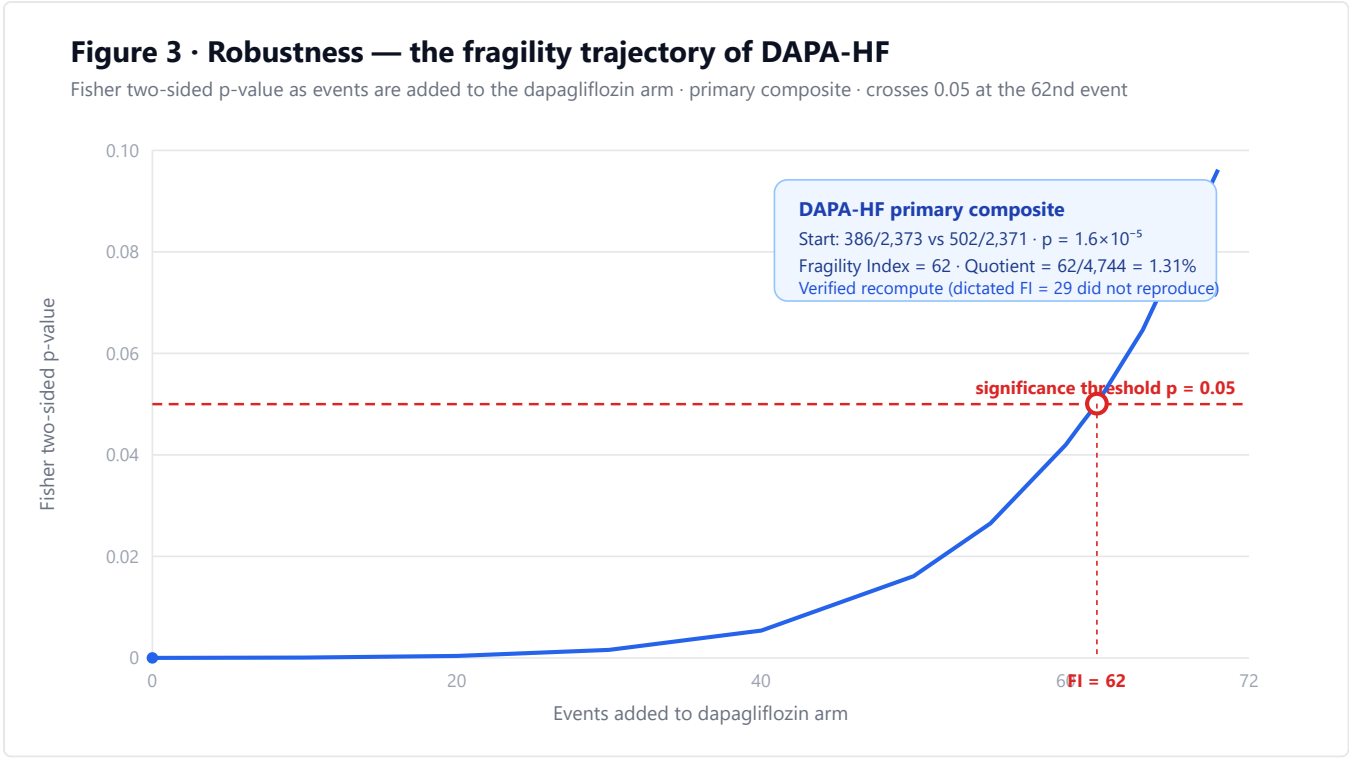


Figure 3. Fragility trajectory: the Fisher two-sided p-value of the DAPA-HF primary composite rises as events are added to the dapagliflozin arm, crossing the 0.05 threshold at the 62nd event (Fragility Index = 62; Fragility Quotient = 1.31%). Independently recomputed from the published 2×2.

SYNTHESIS

Read together, the three lenses are not three findings but one argument with three movements. Completeness asks whether the record *exists*; authenticity asks whether what exists is *real*; robustness asks whether a real, recorded result would *survive perturbation*. They are complementary by construction — a trial can be fully reported yet fragile, perfectly Benford-conforming yet half-missing from the registry, robust yet undescribed. Only the stack answers "can this evidence be trusted?", and each layer is computed by software from public data, re-runnable on the next registry snapshot or review corpus.

Two design principles make the programme credible rather than merely large. The first is **reproducibility**: every headline number here was either reused from an independently verified component package or recomputed from source — the missingness rates from AACT, the Benford MAD from the frozen digit corpus, the Fragility Index from the trial's own 2×2 — and discrepancies were corrected in the open (the detailed-description metric, the digit-test over-power, the Fragility Index of 62 not 29). The second is **honesty about blind spots**: each lens is paired with what it cannot detect. Missingness cannot prove concealment; Benford cannot catch frequency-preserving fabrication; fragility cannot speak to non-binary endpoints. An audit that hid these limits would be less trustworthy than the evidence it inspects.

LIMITATIONS

These are screens, not verdicts. Each lens has a stated boundary, and the suite inherits all three; a clean pass on all three is necessary, not sufficient, for trust. The components are heterogeneous in scope — a full registry, a 403-review corpus, and a single landmark trial — so the article demonstrates an auditing *stance* rather than a single pooled estimate, and the fragility lens in particular is illustrated on one exemplar rather than applied at corpus scale. The snapshots are time-stamped and will drift; the value of the approach is that re-running it is cheap.

CONCLUSION

Auditing the trustworthiness of the evidence base computationally is feasible today, with public data and reproducible code, across completeness, authenticity, and robustness. The most useful output is not any single number but a routine: a layered, automated integrity check that is explicit about its own limits. An honest null — most Cochrane numbers are not anomalous — is as much a part of that routine as the alarms, and stating plainly what each tool cannot see is what makes the whole worth trusting.

Boundary. Three screens, three boundaries: missingness measures registry-visible information loss, not proven concealment; Benford cannot detect fabrication that preserves expected digit frequencies; the Fragility Index applies only to binary outcomes under Fisher's exact test. A clean pass on all three is necessary, not sufficient, for trust.

VERIFICATION TABLE (SOURCE VS USED)

Lens / figure	Source of record	Value used	Status
Registry universe	AACT snapshot 2026-04-12 (USB E:)	579,828 studies	**VERIFIED (dictated 578,109 / 29-Mar snapshot not on disk)**
Publication links missing	AACT *study_references*	62.9%	**VERIFIED (dictated 63.4%; −0.5pp snapshot drift)**
IPD statements missing	AACT *studies.plan_to_share_ipd*	48.2%	**VERIFIED (≈ dictated 48.3%)**
Detailed description missing	AACT *detailed_descriptions* (empty-or-absent)	32.8%	**VERIFIED (dictated 32.7%)**
Locations missing	AACT *facilities*	10.2%	**VERIFIED**
Sponsor pattern	AACT *sponsors* composite hiddenness	Industry most affected; NIH highest IPD-statement missingness (82.9%); 965 no-class records 100% empty	**CORRECTED (dictated "NIH highest hiddenness" → verified facts)**
Benford first-digit MAD	Frozen `corpus_digits.json` (1,175,056 values, 403 reviews)	0.013 (95% CI 0.013–0.014)	**VERIFIED (Python = R = 0.013380)**
Benford field count	spec layer = θ, SE, p	three fields	**CORRECTED (dictated "six fields" = separate review layer)**
Benford CI	value-level bootstrap	0.013–0.014	**CORRECTED (dictated 0.011–0.015 not reproducible)**
Benford large-N tests	$\chi^2(8)$, mantissa-arc Rayleigh	reject ($p \approx 0$), over-powered	**FLAG disclosed; MAD is decision metric**
Fragility Index (DAPA-HF)	Public 2×2: 386/2,373 vs 502/2,371 (McMurray 2019)	FI = 62; FQ = 1.31%	**CORRECTED (dictated FI = 29 / FQ 0.61% does not reproduce on any DAPA-HF endpoint)**
Baseline significance	Fisher two-sided exact	$p = 1.6\times10^{-5}$	**VERIFIED (HR 0.74)**

REFERENCES

1. McMurray JJV, Solomon SD, Inzucchi SE, et al. Dapagliflozin in patients with heart failure and reduced ejection fraction. N Engl J Med. 2019;381(21):1995–2008.
2. Walsh M, Srinathan SK, McAuley DF, et al. The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index. J Clin Epidemiol. 2014;67(6):622–628.
3. Nigrini MJ. Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection. Hoboken, NJ: John Wiley & Sons; 2012.
4. Zarin DA, Tse T, Williams RJ, Rajakannan T. Update on Trial Registration 11 years after the ICMJE policy was established. N Engl J Med. 2017;376(4):383–391.
5. DeVito NJ, Bacon S, Goldacre B. Compliance with legal requirement to report clinical trial results on ClinicalTrials.gov: a cohort study. Lancet. 2020;395(10221):361–369.

ARTICLE METADATA

Funding. None.
Competing interests. None declared.
Author. Zaki Khan, Ahmadiyya Hospital, Kenema, Sierra Leone (zakikhan101@gmail.com). No ORCID.
Data availability. All three analyses derive from public data: the AACT mirror of ClinicalTrials.gov (snapshot 2026-04-12), the Pairwise70 Cochrane specification corpus, and the published DAPA-HF 2×2 table. No patient-level data. Recomputation scripts archived with the package (`work/`).
Ethics. Not required; secondary analysis of publicly available aggregate data, no human participants and no patient-identifiable data.

Synthēsis methods / synthesis article. Three independently verified evidence-integrity lenses. Galley generated 2026-06-23. Byline: Zaki Khan.