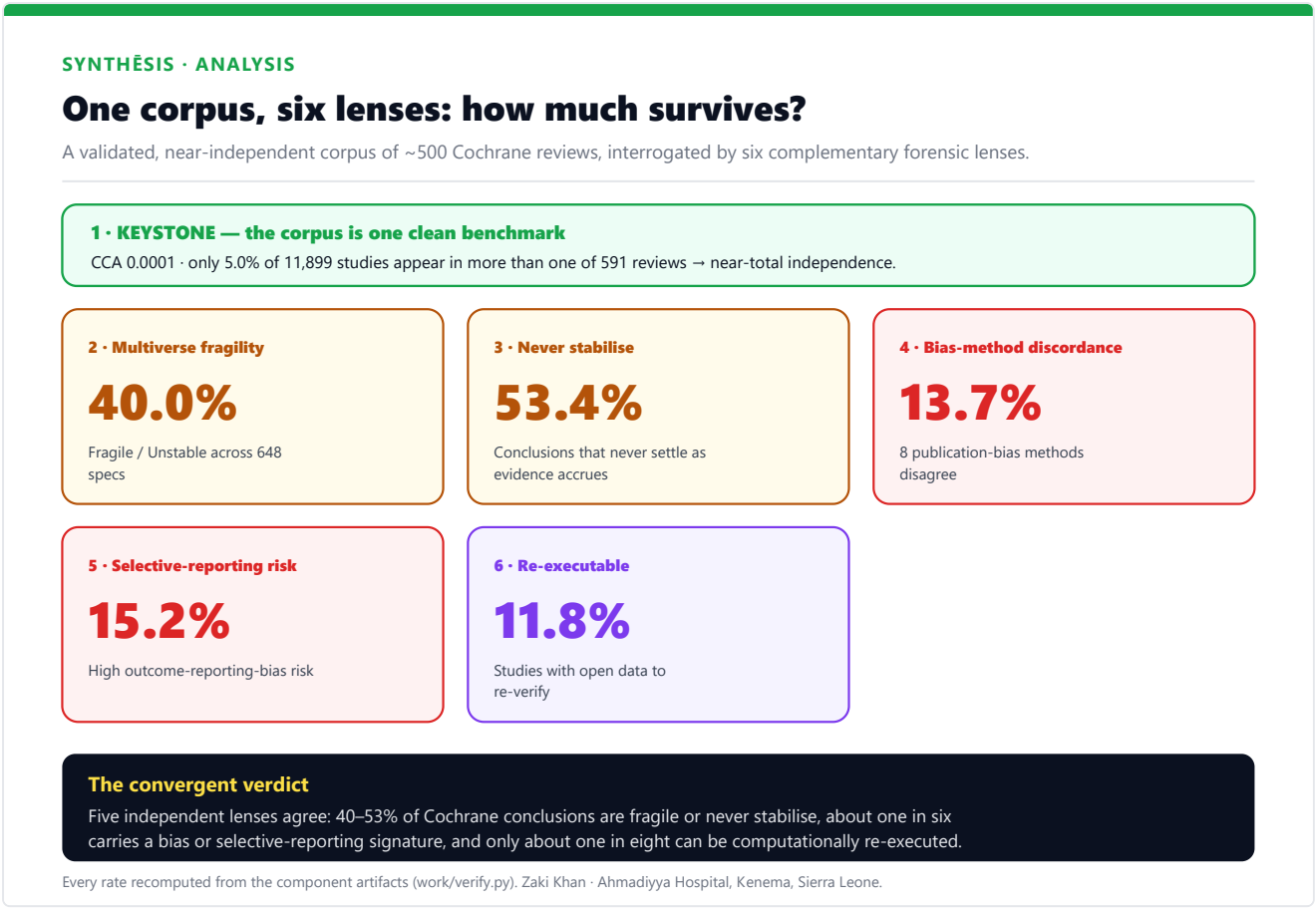
SYNTHĒSIS · ANALYSIS

One corpus, six lenses: how much of a Cochrane conclusion survives analyst choices, bias methods, and computational re-execution

Zaki Khan

Ahmadiyya Hospital, Kenema, Sierra Leone

Corresponding author: zakikhan101@gmail.com



Visual abstract. A near-independent corpus of ~500 Cochrane reviews (proven by the keystone, CCA 0.0001) read through five complementary forensic lenses, which converge: 40–53% of conclusions are fragile or never stabilise, ~1 in 6 carry a bias/reporting signature, only ~1 in 8 are re-executable.

ABSTRACT

Background. Whether a meta-analytic conclusion is robust depends on choices that are usually examined one review and one method at a time: which specification an analyst picks, which publication-bias method is trusted, how much evidence has accumulated, what was selectively reported, and whether the numbers can be re-derived at all. These properties are rarely measured together, and almost never on a single standardised corpus large enough to make the answer a property of the evidence base rather than of a chosen example.

Methods. We assembled one validated corpus — the Pairwise70 extract of roughly 500 Cochrane pairwise reviews — and first established its near-independence so it could serve as a single clean benchmark (the keystone), then interrogated it with five complementary forensic engines: multiverse specification search, publication-bias method discordance, cumulative analytic stability, outcome-reporting-bias excess significance, and computational re-execution from open-access sources. Denominators differ by each

engine's eligibility filter (307–591 reviews; 11,899–14,340 studies) and are reconciled explicitly. Every headline rate was reused from an independently verified component package or recomputed here from the row-level artifacts.

Results. The keystone holds: the corrected covered area was 0.0001, with only 5.0% of 11,899 studies appearing in more than one of 591 reviews — the corpus is near-independent. The forensic lenses then converge. Two robustness lenses put fragility at 40.0% of 403 reviews (Fragile or Unstable across 648 specifications) and 53.4% of 365 reviews never reaching analytic stability. Two bias lenses put the flagged share near one in six: 13.7% of 307 reviews were discordant across eight publication-bias methods, and 15.2% of 473 carried high outcome-reporting-bias risk. The re-execution lens found that only 11.8% of 14,340 studies (1,688) had the open data needed to re-verify them. Across the multiverse, bias correction was the single dominant source of specification variance ($\eta^2 = 0.929$ in 99.0% of reviews).

Conclusion. One near-independent corpus, read through five complementary lenses, returns a convergent verdict: a large share of Cochrane conclusions do not survive contact with defensible analyst choices, about one in six carry a bias or selective-reporting signature, and only about one in eight can even be re-executed. Robustness is measurable at corpus scale, reproducibly and honestly — and the most informative output is the convergence, not any single rate.

Central claim. One validated, near-independent Cochrane corpus, interrogated by five complementary forensic lenses, gives a convergent verdict on how much a meta-analytic conclusion survives analyst choices, bias-method selection, accumulating evidence, selective reporting, and computational re-execution — and the convergence, not any single rate, is the finding.

Introduction

A Cochrane meta-analysis is widely read as a settled answer. Yet the single number it reports is the product of dozens of defensible choices: which variance estimator and confidence-interval method to use, whether and how to correct for publication bias, which studies were available when the review was frozen, what the primary trials chose to report, and whether the underlying numbers can be re-derived at all. Each of these has a mature literature and a bespoke tool, and each is usually applied to one review with one method. None, on its own, tells you how an entire evidence base behaves when its conclusions are stress-tested.

This article asks that larger question. To make robustness a property of evidence rather than of a hand-picked example, two things are needed: a fixed, validated corpus, and complementary lenses that can be turned on the same reviews and read together. But a collection of reviews is only a clean benchmark if the reviews are themselves independent — if they do not silently share primary studies, double-counting the same evidence under different labels. So the first question is structural, not forensic: is the corpus independent enough to treat as one sample? With that established — the keystone of the exercise — five forensic lenses can interrogate the same corpus and their answers be compared. The contribution is not a new estimator but the convergence: five ways of asking “how much survives?” on one validated corpus, harmonised to stated denominators, every number traceable to source.

Methods

The corpus. All analyses use the Pairwise70 dataset, a standardised extract of Cochrane intervention reviews carrying pairwise meta-analyses, with per-study effect sizes, standard errors, and 2×2 tables. Depending on the engine and its filter, the corpus spans 307–591 reviews and 11,899–14,340 studies; the broadest review-level count is 591 and the largest study-level count is 14,340 records (11,899 unique after de-duplication). It is a single pull, not six different datasets.

The keystone: validating independence. Before treating the corpus as one benchmark, OverlapDetector mapped every primary study across all 591 reviews using normalised first-author–year keys, computing the corrected covered area (CCA), pairwise Jaccard similarity, and the study-frequency distribution across all 174,345 review pairs. This step licenses everything that follows: if reviews heavily shared studies, the downstream rates would describe the same handful of reviews counted repeatedly rather than the corpus as a whole.

Five forensic engines. (i) *Multiverse specification search* (MES) re-analysed each review across 648 or more specifications over six dimensions — heterogeneity estimator, CI method, bias correction, quality filter, design filter, leave-one-out — and classified each as Robust, Moderate, Fragile, or Unstable (403 reviews, k≥3). (ii) *BiasForensics* applied eight publication-bias methods (four detection, four correction) and classified the fingerprint as Clean, Suspected, Confirmed, or Discordant (307 reviews, k≥5). (iii) *EvidenceHalfLife* ordered studies by year and tracked cumulative robustness, defining stabilisation as sustained robustness above 70% (365 reviews, k≥5). (iv) *OutcomeReportingBias* (ORB) applied excess-significance testing plus three indicators to classify reviews Low, Moderate, or High risk (473 reviews, k≥3). (v) *MetaReproducer* re-extracted effect sizes from open-access source PDFs and re-pooled them against the Cochrane reference values (501 reviews, 14,340 studies).

Reconciling the denominators. The review counts differ because the engines impose nested eligibility filters on the same corpus, not because the data differ (Figure 3). The keystone and the re-execution audit use all reviews (591; and 501 reviews / 14,340 study records). The two k≥3 engines differ — 473 for ORB versus 403 for MES — because the multiverse needs studies usable across all six specification dimensions, stricter than merely having three primary studies. The two k≥5 engines differ similarly — 365 for EvidenceHalfLife versus 307 for BiasForensics — because computing a full eight-method bias fingerprint excludes reviews a cumulative-stability trajectory can still use. Study totals differ analogously: 11,899 unique studies after author–year de-duplication versus 14,340 study records before it. We therefore report each rate against its own denominator, and read the convergence across rates, not within a pooled sample.

Truth-first accounting. Every headline number was reused from an independently verified component package or recomputed here from the row-level artifacts (script archived with the package). Where a component had two legitimate runs we used the canonical one and disclosed the alternative: EvidenceHalfLife's committed 365-review result (53.4%) is used, with a stricter 307-review re-run (52.1%) noted; MetaReproducer's prevalence is the 11.8% pre-linkage figure, with a post-linkage 12.2% disclosed. Small discrepancies are preserved, not smoothed.

Results

The keystone holds. Overlap across the corpus was minimal: the CCA was 0.0001, and only 590 of 11,899 unique studies (5.0%) appeared in more than one review. Of 174,345 possible review pairs, just 875 shared any study, and the single most overlapping pair shared 43 studies (Jaccard 0.37). The Pairwise70 corpus is, for practical purposes, a collection of independent meta-analyses — a valid benchmark for the lenses that follow.

Robustness: 40–53% of conclusions do not hold up. Two independent lenses agree that a large fraction of conclusions are sensitive to defensible choices. In the multiverse, 161 of 403 reviews (40.0%) were Fragile or Unstable across 648 specifications, with a mean concordance of 0.763; and bias correction was the single dominant source of specification variance in 399 of 403 reviews (99.0%), with a mean dominant η^2 of 0.929 — nearly all of the analytic instability traces to whether and how one adjusts for publication bias. Cumulatively, 195 of 365 reviews (53.4%) never reached analytic stability even after all available studies were accumulated, with a median half-life of six studies among those that did stabilise.

Bias and selective reporting: about one in six. Two further lenses converge on a flagged share near one in six. Across eight publication-bias methods, 42 of 307 reviews (13.7%) were Discordant — the apparent direction of the conclusion depended on which method was chosen — while only 54 (17.6%) were unambiguously Clean; method-pair agreement ranged from 73% to 98%. Independently, an excess-significance screen classified 72 of 473 reviews (15.2%) as high outcome-reporting-bias risk, with a further 20.7% at moderate risk.

Re-execution: only about one in eight can be checked. The final lens is the most sobering. Of 14,340 studies, only 1,688 (11.8%) had the open-access source needed to re-extract and re-verify their effect sizes; the rest are computationally unverifiable not for any methodological reason but because the underlying data are not open. The barrier to reproducibility is infrastructural, and it caps how much of the evidence base any automated audit can re-execute.

Analysis

The five lenses are complementary, not redundant: each probes a different stage of the inferential pipeline — specification (MES), bias adjustment (BiasForensics), evidence accumulation (EvidenceHalfLife), selective reporting (ORB), and re-execution (MetaReproducer). That they converge is the finding. A single reported point estimate implies more solidity than the corpus carries: by two robustness lenses, four to five conclusions in ten are fragile or never settle; by two bias lenses, roughly one in six carries a detectable signature of bias or selective reporting; and for seven in eight reviews you cannot even check, because the source data are closed. The recurrence of bias correction as the dominant lever — it explains almost all multiverse variance and is the axis along which the bias methods most disagree — points to a specific culprit: the most consequential analyst choice is also the one on which the field has least consensus.

The keystone is what turns six findings into one argument. Because the corpus is near-independent (CCA 0.0001), the five rates describe a sample of largely non-overlapping reviews; the fragility is not an artefact of the same few shaky reviews being scored again and again by different tools. The convergence is therefore a property of the contemporary Cochrane evidence base as sampled here, not of the benchmark or of any one engine. Read together, the lenses answer the question a careful reader should ask of any synthesis — how much of this would survive a competent, sceptical re-analysis? — and the answer, uncomfortably, is: often less than the abstract implies.

Limitations

These are screens, not verdicts, and each has a boundary the suite inherits. The robustness and bias lenses (MES, ORB, BiasForensics) operate on pre-computed effect-and-standard-error grids rather than re-analysing raw trials; they detect statistical signatures of fragility and bias, not bias itself, and cannot separate publication bias from clinical or methodological heterogeneity. The re-execution lens is bounded by open-access availability — an infrastructural, not methodological, limit — and its rate is sensitive to source linkage (11.8% here, 12.2% after registry linkage). The keystone relies on first-author–year matching, which can produce false positives when distinct studies share a surname and year; that error would only inflate apparent overlap, so the true independence is at least as strong as reported. Finally, the denominators are heterogeneous by construction, so the convergence is a qualitative agreement across lenses, not a single pooled estimate. A clean pass on any one lens is necessary, not sufficient, for trust.

Conclusion

One validated, near-independent corpus, interrogated by five complementary forensic lenses, returns a convergent verdict: a Cochrane conclusion frequently does not survive contact with analyst choices (40–

53% fragile or never-stabilising), about one in six carries a bias or selective-reporting signature, and only about one in eight can be computationally re-executed at all. None of these rates is new on its own; their agreement, on one corpus whose independence is proven rather than assumed, is. The robustness of evidence synthesis is measurable at scale, reproducibly and candidly — and the most useful product is not any single number but the routine: validate the benchmark, turn complementary lenses on it, harmonise the denominators, and let five independent answers converge on the same uncomfortable truth.

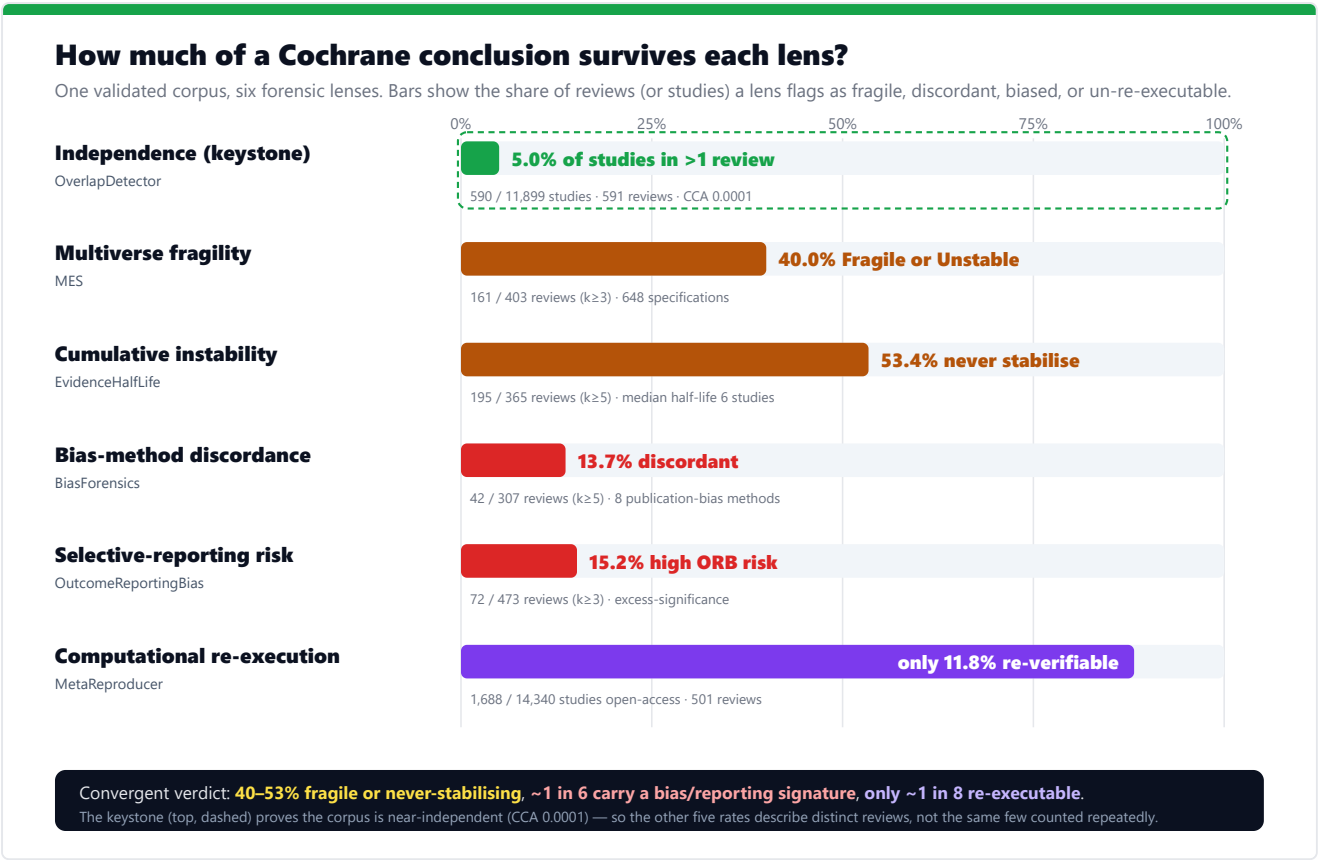


Figure 1. The six lenses on one axis. Bars show the share each lens flags (the re-execution bar shows the 88.2% that cannot be re-verified). Two robustness lenses (40.0%, 53.4%), two bias lenses (13.7%, 15.2%), and the re-execution lens (only 11.8% verifiable) converge; the keystone (top) certifies the corpus as near-independent.

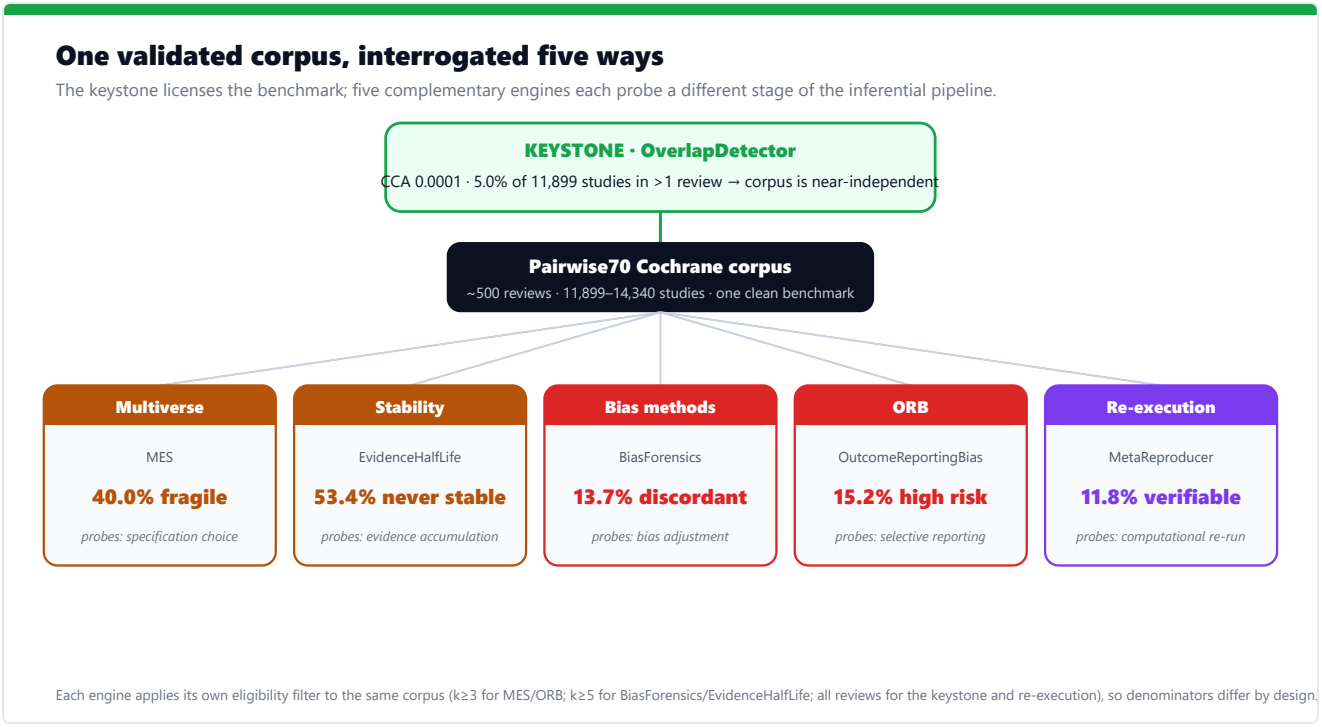


Figure 2. Architecture of the audit. The keystone (OverlapDetector) certifies the corpus as near-independent; five forensic engines then each probe a different stage of the inferential pipeline — specification, accumulation, bias adjustment, selective reporting, and re-execution.

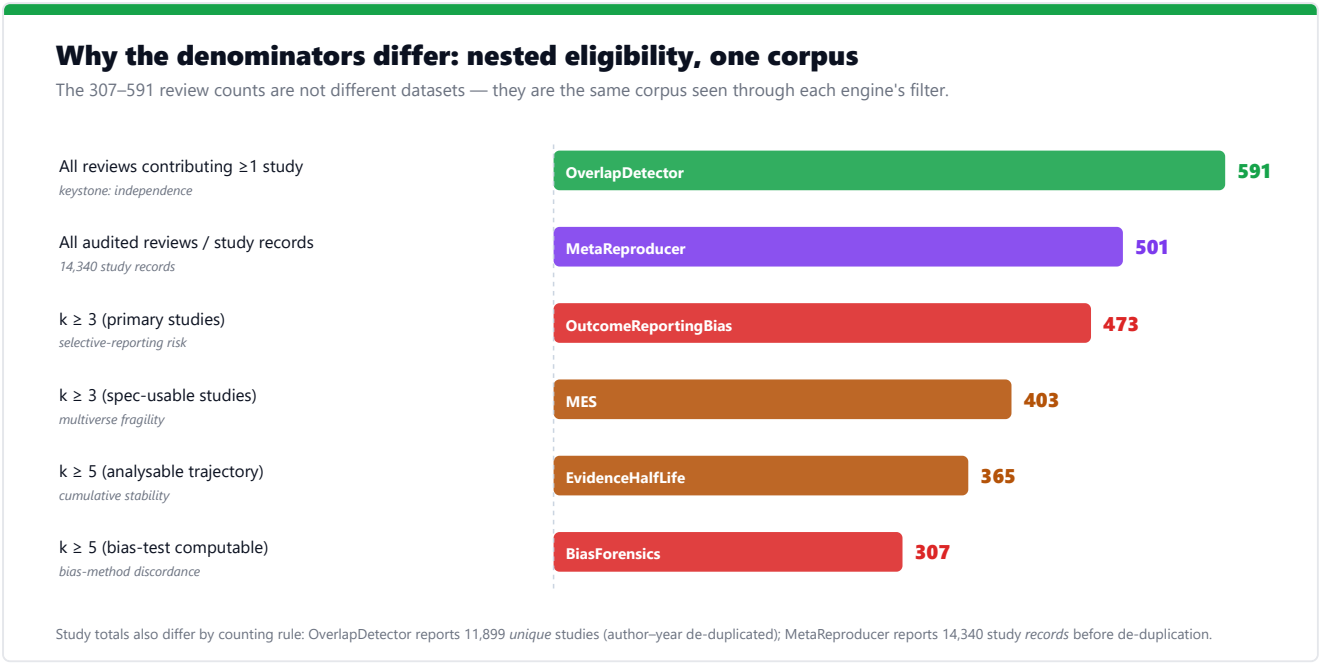


Figure 3. Denominator reconciliation. The 307–591 review counts are the same corpus seen through each engine's eligibility filter (k≥3 for MES/ORB, k≥5 for BiasForensics/EvidenceHalfLife, all reviews for the keystone and re-execution). Study totals differ as unique studies (11,899) versus study records (14,340).

Verification table (component: source versus used)

Component (entry)	Lens	Source value	Used	Status
OverlapDetector [127]	Keystone independence	CCA 0.0001; 590/11,899 (5.0%) multi-review; 591 reviews	same	**VERIFIED — recomputed from overlap_summary distribution (479+68+43=590)**
MES [87]	Multiverse fragility	161/403 (40.0%) Fragile/Unstable; η^2 0.929 in	same	**VERIFIED — recomputed from batch_results.csv (403 OK rows)**

		399/403		
EvidenceHalfLife [52]	Cumulative stability	195/365 (53.4%) never stabilise; median half-life 6	53.4% (365)	**VERIFIED — git-canonical summary; working-tree 307-run (52.1%) disclosed**
BiasForensics [12]	Bias-method discordance	42/307 (13.7%) discordant; mean shift 1.01	same	**VERIFIED — recomputed from bias_forensics_results.csv (307 rows)**
OutcomeReportingBias [126]	Selective-reporting risk	72/473 (15.2%) high; 98 (20.7%) moderate	15.2% (473)	**VERIFIED — orb_summary.json; reconciled vs rewrite 403/16.1% (not used)**
MetaReproducer [106]	Computational re-execution	1,688/14,340 (11.8%) open-access	11.8%	**VERIFIED — summed n_with_pdf over 501 reviews (pre-AACT); post-AACT 12.2% disclosed**

All rates recomputed in *work/verify.py* against the row-level component artifacts. Two reconciliations are disclosed rather than smoothed: EvidenceHalfLife uses the git-canonical 365-review run (53.4%), with a stricter 307-review re-run (52.1%) noted; MetaReproducer uses the 11.8% pre-linkage open-access prevalence, with a post-linkage 12.2% noted.

References

1. Pieper D, Antoine S-L, Mathes T, Neugebauer EAM, Eikermann M. Systematic review finds overlapping reviews were not mentioned in every other overview. J Clin Epidemiol. 2014;67(4):368–375.
2. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing transparency through a multiverse analysis. Perspect Psychol Sci. 2016;11(5):702–712.
3. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629–634.
4. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2000;56(2):455–463.
5. Stanley TD, Doucouliagos H. Meta-regression approximations to reduce publication selection bias. Res Synth Methods. 2014;5(1):60–78.
6. Ioannidis JPA, Trikalinos TA. An exploratory test for an excess of significant findings. Clin Trials. 2007;4(3):245–253.
7. Dwan K, Gamble C, Williamson PR, Kirkham JJ; Reporting Bias Group. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One. 2013;8(7):e66844.
8. Nosek BA, Hardwicke TE, Moshontz H, et al. Replicability, robustness, and reproducibility in psychological science. Annu Rev Psychol. 2022;73:719–748.

Article metadata

Author. Zaki Khan, Ahmadiyya Hospital, Kenema, Sierra Leone (zakikhan101@gmail.com). No ORCID.

Funding. None.
Competing interests. None declared.
Article type. Review / Analysis (synthesis of six computational meta-research analyses on one corpus).
Data availability. All six analyses derive from the public Pairwise70 Cochrane corpus; every headline rate was reused from an independently verified component package or recomputed from the row-level artifacts (recomputation script archived with this package). No patient-level data.
Ethics. Not required; secondary analysis of public, aggregate meta-analytic data with no human participants.

Synthēsis · Analysis. Body ≈1594 words. Galley generated 2026-06-23.