

Evidence Quality in Assurance-Oriented Benchmark Construction

Byungsik Seo

seo307@gmail.com

Research Article

Keywords: software quality, data quality, software assurance, evidence provenance, incident reconstruction, Model Context Protocol

Posted Date: September 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10839605/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Assurance studies built from public software incidents depend on evidence whose fitness for reconstruction and downstream assurance analysis is rarely characterized at the level of the semantic claims the analysis consumes. We study this evidence-quality problem in a frozen 40-case frame of public tool-using-agent and Model Context Protocol incidents, separating reconstructability, replay/source resolution, oracle feasibility, pairing, releaseability, and element-level provenance of authored assurance semantics. Reproducible-vulnerability benchmarks and vulnerability-data-quality studies already establish disclosure filtering and intrinsic data-quality assessment; our contribution is narrower and task-relative. Although 39/40 cases exposed a patch or corrected state, only 20/40 met the strict paired criteria. Among 15 prefix survivors that were not strict-eligible, nonexclusive blockers were reconstruction (14), releaseability (9), replay/source resolution (4), pairing (3), and oracle feasibility (1); C38 was technically reconstructable but excluded solely by Conditional releaseability. In a ten-case deep audit, 190 representation rows were provenance-classified, and all 52 analyzer-relevant semantic attributes were source-informed rather than independent ground truth. C4 showed conclusion-relevant uncertainty: unresolved enforcement activity produced different fixed-state conclusions under fail-open versus fail-closed assignments. We additionally release AIRR-40, a 40-case, 49-field registry with a 20-case exact-upstream strict manifest and reproducibility artifacts. The study does not estimate population readiness or detector accuracy; it characterizes evidence quality for assurance-oriented benchmark construction.

1. Introduction

Assurance-oriented benchmarks built from public vulnerability advisories, repository issues, and incident reports inherit an evidence-quality problem before any analyzer is evaluated. A patch or corrected state may be visible while the affected state, deployment reconstruction, replay decision, oracle, release conditions, or assurance-relevant semantics remain incomplete or ambiguous. For tool-using agents, those missing semantics can include identity/context fields, binding relationships, transformations, and enforcement activity that determine whether a static assurance question is even well posed.

Prior work already establishes important neighboring facts. Vul4J filtered 912 real-world Java vulnerabilities and 1,803 fix commits through explicit source, test, reproducibility, and patch-isolation criteria to produce 79 reproducible vulnerabilities (Bui et al. 2022). Croft et al. (2023) showed that vulnerability datasets can contain substantial problems in accuracy, uniqueness, consistency, completeness, and currentness, with downstream effects. ReposVul and MegaVul further improve patch quality, repository context, code representations, and scale (Wang et al. 2024; Ni et al. 2024). We therefore do not claim novelty for filtering public vulnerabilities, for showing that a fix is insufficient for reproducibility, or for treating vulnerability-data quality as a research object.

The narrower gap studied here is task-relative evidence quality for software assurance over a reconstructed system representation. Even after an affected/fixed pair is available, an analyst may need

to author source and consequence boundaries, required capabilities, context fields, bindings, transformations, and deployment activity that public records do not state directly. The evidentiary status of those authored semantics is distinct from ordinary source availability or label correctness, and public release may be constrained even when the technical pair is reconstructable. This study therefore combines a frame-level qualification analysis with a ten-case element-level provenance and uncertainty audit.

The study makes four contributions:

1. An empirical characterization of assurance-oriented evidence quality, including the 39/40 patch-versus-20/40 strict gap, a nonexclusive blocker decomposition, and C38, where favorable technical evidence was insufficient because Conditional releaseability alone prevented strict inclusion.
2. An element-level provenance audit of a secondary evidence-bounded representation contract, separating analyzer-relative semantics (Γ), nominal provenance (P), and unresolved alternatives (U); the study does not claim a complete/minimal Γ or an ordinal evidence-strength scale.
3. C4 as a conclusion-relevant uncertainty boundary: an unresolved enforcement-activity field yields different fixed-state conclusions under explicit fail-open versus fail-closed assignments, showing why unresolved alternatives should remain first-class rather than be forced into a negative finding.
4. AIRR-40 (Agent Incident Reconstructability Registry), released as a reusable companion resource: a frozen 40-case, 49-field registry, a 20-case exact-upstream strict manifest, public evidence pointers, case-status metadata, release/license notes, and reproducibility artifacts.

2. Background and Related Work

2.1 Reproducible vulnerability benchmarks

Defects4J and BugSwarm demonstrate the value of frozen faulty/fixed artifacts for controlled studies (Just et al. 2014; Tomassi et al. 2019), and crash-reproduction research emphasizes explicit reproduction conditions (Soltani et al. 2020). Vul4J is the closest precedent for the qualification logic: it requires Java-source involvement, an appropriate test/PoV, publicly accessible and buildable revisions whose test status changes across the patch, and an isolated security patch (Bui et al. 2022). Starting from 912 vulnerabilities and 1,803 fix commits, it retained 79 reproducible vulnerabilities from 51 projects. Thus, large disclosure-to-benchmark attrition is prior knowledge, not a novelty claim of AIRR-40.

ReposVul addresses tangled patches, repository-level context, and outdated patches (Wang et al. 2024), while MegaVul expands scale and code representations (Ni et al. 2024). AIRR-40 is intentionally smaller and is not a detector-training corpus; its purpose is to expose reconstructability, release constraints, and assurance-evidence status for public agent/MCP incidents.

2.2 Vulnerability-data quality

Croft et al. (2023) provide the closest vulnerability-data quality precedent. They operationalize accuracy, consistency, completeness, and currentness from the ISO/IEC 25012 data quality model and add uniqueness for vulnerability datasets. ISO/IEC 25012 defines a general model for structured-data quality and supports the specification, measurement, and evaluation of data quality under stated conditions (ISO/IEC 2008). The present constructs are not proposed as new ISO/IEC 25012 characteristics or as a validated extension of that model. Rather, reconstructability, releaseability, and element-level semantic provenance are task-relative admissibility concerns that arise when public incident records are reused to construct an assurance representation. AIRR-40 therefore complements intrinsic data-quality analysis without claiming to replace it.

2.3 Architecture reconstruction and assurance semantics

Software architecture reconstruction treats higher-level system views as derived artifacts rather than direct readings of source code (Ducasse and Pollet 2009; Tamburri and Kazman 2018). Recovery approaches can differ in recovered characteristics and accuracy (Schneider et al. 2025), while model-to-analysis traceability can itself be uncertain (Trubiani et al. 2017). For agent incidents, causal understanding can require logs, documentation, system access, and tool information that public reports omit (Ezell et al. 2025). Peer-reviewed or accepted MCP studies characterize protocol threats, attack vectors, and vulnerability datasets (Hou et al. 2026; Song et al. 2026; He et al. 2026). Our contribution is upstream of detector accuracy: a reusable evidence registry and an element-level audit of the semantics used to ask an assurance question.

2.4 Explicit novelty boundary

Prior work	Established contribution	Narrower contribution here
Vul4J (2022)	Filtering + executable reproducibility for real vulnerabilities	Agent/MCP reconstruction, releaseability, and assurance-oriented evidence fields; AIRR-40 resource.
Croft et al. (ICSE 2023)	Four ISO/IEC 25012-derived data-quality characteristics plus uniqueness; vulnerability-dataset quality and downstream impact	Task-relative reconstructability/releaseability and element-level assurance-semantic provenance/uncertainty; not a new ISO/IEC 25012 model.
ReposVul / MegaVul (2024)	Higher-quality collection, repository context, representations, scale	Small audited reconstructability registry; no training-corpus claim.
Architecture reconstruction	Derived system views and uncertain traceability	Evidence status of reconstructed assurance semantics and unresolved alternatives.

3. Study Design

We report an exploratory empirical reconstruction study following software-engineering case-study transparency principles (Runeson and Host 2009). The unit of analysis is an incident-derived

affected/fixed case. AIRR-40 is a constructed challenge frame, not a probability sample; all rates are descriptive of this frame.

RQ1 – Qualification and blockers

What evidence dimensions and blockers separate public incident disclosure from the frozen strict paired subset?

RQ2 – Residual semantic provenance

After source-pair acquisition in the deep subset, what provenance supports representation elements and analyzer-relevant semantic attributes?

RQ3 – Conclusion-relevant uncertainty

Does any explicitly unresolved representation assumption alter the static assurance conclusion under alternative admissible assignments?

3.1 AIRR-40 and qualification protocol

AIRR-40 contains 49 frozen fields spanning public evidence pointers, timestamps, source stratum, affected-version knowledge, pre-fix source availability, patch/corrected-state availability, deployment reconstructability, replay-triage status, oracle feasibility, releaseability, paired-control availability, fidelity, ethics/license notes, and other descriptors. The fields are not a single scale and are not all nested gates.

The primary status is the frozen strict definition; a pre-specified lenient definition admits configured Partial or Conditional states where allowed. In this frame, lenient eligibility coincided exactly with replay-decidability: both identified the same 31 cases. Replay triage is a decisionability screen, not a claim that an exploit was reproduced, and uses a hard 30-minute time box. Cases lacking an exact reproducibly acquirable pair are recorded as source-resolution-required rather than replay failures. Blockers are nonexclusive.

Online Resource 1 releases AIRR-40 and the 20-case strict manifest. All 20 strict rows carry the frozen exact-upstream fidelity label. The resource preserves public pointers, case-status fields, version/mechanism notes, and release/license notes; upstream source trees are not redistributed.

3.2 Deep evidence-bounded representation audit

Before deep outcomes, eight strict primary cases (C2, C3, C4, C5, C8, C11, C16, C19) and two lenient-only stress cases (C1, C17) were frozen. Gate2A independently reacquired affected/fixed artifacts and used normalized fingerprints, changed production files, and fixed-only anchors. Pre-outcome corrections changed the C4 affected reference, normalized C19 extraction roots, and strengthened mechanism

localization. A post-result audit later identified an inert source-diversity selector; the cohort was preserved to avoid outcome-informed reselection.

For the deep subset we document $R=(G, \Gamma, P, U)$, where $G=(V,E,s,t)$ is the represented source-to-consequence graph and $\Gamma=(C_{\text{req}}, X_{\text{req}}, B, T, A)$ contains required capability, required context, bindings, transformations, and activity. Γ is analyzer-relative; no completeness, minimality, or universality claim is made. P maps elements/attributes to nominal provenance categories, not an ordinal evidence ladder. U records unresolved propositions and alternatives.

Online Resource 1 includes a post-result structural conformance contract and machine-readable graph schema. These check well-formedness and provenance-record presence only. Let $R_{\Phi}(E)$ denote the set of representations consistent with evidence E and the frozen derivation/conformance rules Φ . This study does not enumerate that set or estimate its size; RQ3 asks only whether the frozen audit contains a conclusion-relevant unresolved alternative.

3.3 Analyzer role and claim discipline

A0 checks path/capability mediation; A1 additionally consumes context binding/preservation. For a common modeled path set with resolved activity, A0 Witness implies A1 Witness by construction. Thus the direction of A0/A1 disagreement is definition-constrained and the 4/8 count is not a performance result. The case-level comparison remains in Online Resource 1 only as implementation context. The main text uses analyzer output only for C4, where both analyzers agree about the effect of alternative activity assumptions.

4. Results

4.1 RQ1 — Qualification and blocker profile

Five of 40 AIRR-40 candidates failed the prefix evidence screen, leaving 35 survivors. Thirty-one reached a valid replay-feasibility decision. Four survivors (C27, C28, C29, C32) remained source-resolution-required and are not replay failures. Twenty cases met the strict paired definition and all 20 were exact-upstream; 31 met the lenient definition.

Patch availability was substantially less restrictive than strict inclusion: 39/40 cases had a patch or corrected state, while 20/40 entered the strict subset. Thirty-six had a known affected version, 29 had fully available pre-fix source, 21 were fully deployment-reconstructable, 34 had a fully feasible atomic oracle, 26 were fully releasable, and 32 had paired control available. The point is not that attrition itself is novel; rather, the registry exposes assurance-specific reconstruction and release dimensions after public fixes are already visible.

Evidence dimension	n	% of frame
Affected version known	36	90.0
Pre-fix source fully available	29	72.5
Patch/corrected state available	39	97.5
Deployment fully reconstructable	21	52.5
Valid replay decision	31	77.5
Atomic oracle fully feasible	34	85.0
Fully releasable	26	65.0
Paired control available	32	80.0
Strict exact-upstream subset	20	50.0
Lenient eligible	31	77.5

Among the 15 prefix survivors outside the strict subset, reconstruction was recorded in 14, releaseability in 9, replay/source resolution in 4, pairing in 3, and oracle feasibility in 1. Counts overlap. Eleven cases were lenient-eligible but not strict-eligible: ten were partial-upstream reconstruction cases; C38 was exact-upstream and failed strict inclusion only because releaseability was Conditional. C38 therefore separates technical reconstructability from releaseability under the study criteria.

Nonexclusive blocker	n	% of 15
Reconstruction	14	93.3
Releaseability	9	60.0
Replay/source resolution	4	26.7
Pairing	3	20.0
Oracle	1	6.7

4.2 RQ2 — Element-level provenance after source-pair acquisition

The ten-case post-result audit contains 190 graph/assumption rows: 12 DIRECT_ANCHOR_SUPPORTED, 20 DOCUMENTED_ABSTRACTION, 43 COARSE_FILE_LEVEL_SUPPORT, 49 INFERRED_EXPLICIT, and 66 SYNTHESIS_EXPLICIT. These are nominal categories, not five validated levels of evidence strength.

Nominal provenance class	Rows
DIRECT_ANCHOR_SUPPORTED	12
DOCUMENTED_ABSTRACTION	20
COARSE_FILE_LEVEL_SUPPORT	43
INFERRED_EXPLICIT	49
SYNTHESIS_EXPLICIT	66
Total	190

The audit separately records 52 analyzer-relevant semantic attributes: 20 binding, 14 capability, 14 activity, and four transformation attributes. All 52 are explicitly marked as not independent ground truth. This is not an error rate; it is the central construct-validity observation of the deep audit. Exact or partial source-pair recovery does not automatically provide independent ground truth for the semantic fields consumed by a downstream assurance abstraction.

Every declared consequence sink in the deep representations was inferred. A direct/documented-only ablation therefore removes the source-to-consequence question in every deep state. We do not treat the all-Unmodeled output as 20 independent findings; it follows mechanically once inferred sinks are removed. The supported boundary is that source acquisition alone did not provide a direct/documented consequence boundary in this selected subset.

4.3 RQ3 – C4 and conclusion-relevant underdetermination

C4 is the strongest non-tautological representation boundary. In the fixed representation an Origin-validation enforcement node is represented in code, but its deployment/activity state remains ambiguous. R0 therefore yields Witness → Unmodeled. R3_fail_open treats the ambiguous activity as inactive and yields Witness → Witness; R3_fail_closed treats it as active and yields Witness → No-Witness-under-Abstraction. A0 and A1 follow the same reversal.

Variant	Activity interpretation	A0/A1 affected → fixed
R0	unresolved	Witness → Unmodeled
R3_fail_open	ambiguous treated inactive	Witness → Witness
R3_fail_closed	ambiguous treated active	Witness → No-Witness-under-Abstraction

C4 shows that U is not merely documentation overhead: one unresolved activity proposition is conclusion-relevant under the frozen representation. It also illustrates the underdetermination captured by $R_{\Phi}(E)$. The study does not quantify the R-versus-system gap, estimate how often such alternatives occur, or enumerate all admissible representations.

5. Discussion

5.1 What the registry adds beyond disclosure filtering

Vul4J already establishes that public vulnerabilities need technical filtering and reproducibility checks. AIRR-40 should not be read as rediscovering that fact at smaller scale. Its task-specific contribution is to make reconstruction, oracle, pairing, release, and downstream semantic-evidence deficits visible for assurance-oriented agent/MCP studies. The blocker categories and row-level metadata are more transferable than the frame-specific 20/40 ratio.

C38 makes releaseability concrete: the case is exact-upstream and favorable on the major technical evidence fields, yet Conditional releaseability keeps it outside the strict subset. This does not establish a universal benchmark rule; it shows that release can be an explicit study-quality constraint rather than a packaging footnote.

5.2 Why P and U are separate

P asks what type of support is recorded for an element or semantic attribute; U asks whether unresolved alternatives remain. These questions are not equivalent. A source-informed synthesis can be explicit in P without a currently recorded alternative, while C4 shows that a directly represented enforcement node can still have an unresolved activity assignment that changes the conclusion. The distinction is operational in the released audit, but it is not a validated quality model.

The provenance categories are intentionally non-ordinal. Direct code, documentation, coarse file support, explicit inference, and synthesis differ in kind and relevance to particular claims; the study provides no empirical basis for a universal total order or numeric score.

5.3 Representation limits and falsifiability

Γ is analyzer-relative: its five semantic components were retained because A0/A1 consume them. The study does not show that they are sufficient or minimal for tool-using-agent assurance. Trust boundaries, temporal ordering, concurrency, multiplicity, or other semantics may be needed elsewhere. The representation should be revised if independent outcome-blind derivations repeatedly disagree, if direct evidence contradicts authored semantics, or if held-out cases require additional field types to pose the assurance question.

$R \neq S$ remains an epistemic warning rather than a measured distance. C4 is one local probe. A stronger representation study would freeze Φ , use at least one independent derivation, report field-level disagreement, and study the size or structure of $R_\Phi(E)$. Those tests are outside the evidence of Paper A.

5.4 Resource use

AIRR-40 can be reused without accepting the paper's strict criteria as universal. Researchers can choose alternative admissibility policies over the frozen 49 fields, compare strict/lenient or new criteria, reacquire the 20 exact-upstream strict cases through public pointers, and retain release/license notes when deciding what can be redistributed. The resource deliberately preserves excluded and unresolved cases rather than publishing only a cleaned final benchmark.

6. Threats to Validity

Novelty and construct boundary. Qualification funnels and vulnerability-data quality are established by Vul4J, Croft et al., and related dataset work. The paper claims novelty only for the narrower assurance-oriented combination of reconstructability/releaseability resource metadata and semantic provenance/uncertainty auditing. Strict, lenient, and blocker categories are study-specific operational constructs.

Selection validity. AIRR-40 is a constructed challenge frame, not a probability sample. The deep cohort is not source-stratified because an intended source-diversity selector was inert; the pre-outcome cohort was preserved to avoid outcome-informed reselection.

Derivation validity. The same author who studied the incident evidence authored the deep representation semantics. All 52 analyzer-relevant attributes are marked as non-independent ground truth, and no inter-derivation reliability statistic is possible. The post-result conformance contract checks structure only and does not validate the construct.

Internal validity and chronology. Recovered Gate1/Gate2A/Gate2B receipts support the chronology of freezing and pre-outcome corrections and the integrity of the computational rerun. They do not independently validate source/sink or semantic choices.

External and runtime validity. Deep provenance and C4 findings cannot be extrapolated to all 40 cases or unseen systems. Replay triage is a feasibility decision, not full exploit reproduction. Witness is existential under R; No-Witness-under-Abstraction is not a safety certificate, and Unmodeled is not a safe state.

7. Reproducibility, Resource, and Data Availability

Online Resource 1 contains AIRR-40, the frozen 40-row/49-field registry, case-status file, a 20-case exact-upstream strict-pair manifest, data dictionary/resource card, original Gate1 closeout archive, and scripts that recompute the qualification/evidence/blocker results. It also contains Gate2A correction records and the slim Gate2B post-result archive with frozen graphs, policy, analyzer code, minimal evidence freeze, outputs, and provenance audit.

For v0.6.2, the resource validator confirms 40 rows $\times$ 49 fields and exact correspondence between the 20-case pair manifest and frozen strict_s8 status. The Gate1 counts recompute from row-level records;

the representation validator checks structural conformance and the presence of all 52 semantic provenance rows; and the Gate2B analyzer rerun reproduces six designated outputs byte-for-byte. These establish resource consistency and computational reproducibility, not independent representation-derivation replicability.

Full upstream source trees are not redistributed. Public pointers, version/mechanism notes, release/license notes, fingerprints, and the computational core are retained. The authored AIRR-40 registry/status metadata, data dictionary, and resource documentation in Online Resource 1 are released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Original study/package scripts and code authored by the author are released under the MIT License. These licenses do not relicense upstream projects or third-party artifacts, which remain governed by their original terms.

8. Conclusion

Assurance-oriented benchmark construction is not exhausted by identifying a public vulnerability and its fix. In the constructed AIRR-40 frame, 39/40 cases exposed a patch or corrected state but only 20/40 met the frozen strict paired criteria; reconstruction and releaseability were the most frequent recorded blockers among strict-ineligible prefix survivors. The deeper evidence-quality issue appears after source acquisition: across 190 audited representation rows, all 52 analyzer-relevant semantic attributes were source-informed rather than independent ground truth. C4 shows why uncertainty must remain explicit, because one unresolved enforcement-activity assignment changes the fixed-state conclusion under opposite admissible assumptions. These findings do not define a universal software-quality metric or estimate population readiness. They identify task-relative quality conditions for assurance-oriented reuse of public incident evidence—reconstructability, releaseability, semantic provenance, and conclusion-relevant uncertainty—and release AIRR-40 and its 20-case exact-upstream strict manifest so that those conditions can be inspected, recomputed, and replaced by alternative admissibility policies.

Declarations

Competing interests. The author declares no competing interests.

Funding. No specific external funding was received for this work.

Data availability. AIRR-40 and the accompanying reproducibility/evidence package are provided as Online Resource 1. The authored AIRR-40 registry/status metadata and documentation are licensed under CC BY 4.0, and original study/package scripts and code authored by the author are licensed under the MIT License. Full upstream source trees are not redistributed and remain governed by their original licenses; public source pointers, versions, fingerprints, frozen status records, and the computational core used for the reported recomputations are included.

Ethics approval. Not applicable. The study analyzes public software-security records and software artifacts and does not involve human participants or private personal data.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

Author contributions. The author conceived the study, designed the methodology, conducted and reviewed the analysis, curated the study records, interpreted the results, and prepared the manuscript.

AI-assisted research disclosure. Generative AI tools were used during research planning, code drafting, manuscript drafting, language editing, structural consistency checking, and limited execution support. The author independently reviewed the research question, public source evidence, frozen study data, calculations, interpretations, references, and final manuscript. AI tools were not treated as independent validators and did not alter the frozen research records.

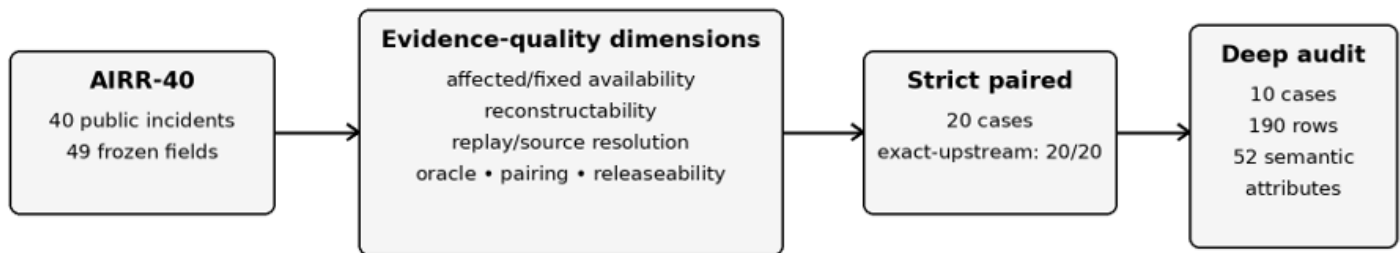
References

1. Bui, Q. C., Scandariato, R., & Díaz Ferreyra, N. E. (2022). Vul4J: A Dataset of Reproducible Java Vulnerabilities Geared Towards the Study of Program Repair Techniques. *Proceedings of MSR*, 464–468. <https://doi.org/10.1145/3524842.3528482>
2. Croft, R., Babar, M. A., & Kholoosi, M. M. (2023). Data Quality for Software Vulnerability Datasets. *ICSE 2023*, 121–133. <https://doi.org/10.1109/ICSE48619.2023.00022>
3. Ducasse, S., & Pollet, D. (2009). Software Architecture Reconstruction: A Process-Oriented Taxonomy. *IEEE TSE*, 35(4), 573–591. <https://doi.org/10.1109/TSE.2009.19>
4. Ezell, C., Roberts-Gaal, X., & Chan, A. (2025). Incident Analysis for AI Agents. *AAAI/ACM AIES*, 8(1), 865–878. <https://doi.org/10.1609/aies.v8i1.36596>
5. He, X., Patel, S., & Wang, S. (2026). Tracing MCP Security Vulnerabilities: Constructing and Analyzing an MCP Security Vulnerability Dataset. *IEEE SPW*, 381–389. <https://doi.org/10.1109/SPW72489.2026.00047>
6. Hou, X., Zhao, Y., Wang, S., & Wang, H. (2026). Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions. *ACM TOSEM*. <https://doi.org/10.1145/3796519>
7. ISO/IEC. (2008). *ISO/IEC 25012:2008, Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model*. International Organization for Standardization and International Electrotechnical Commission.
8. Just, R., Jalali, D., & Ernst, M. D. (2014). *Defects4J: A database of existing faults to enable controlled testing studies for Java programs* (pp. 437–440). *ISSTA*. <https://doi.org/10.1145/2610384.2628055>
9. Ni, C., Shen, L., Yang, X., Zhu, Y., & Wang, S. (2024). *MegaVul: A C/C++ Vulnerability Dataset with Comprehensive Code Representations* (pp. 738–742). *MSR*. <https://doi.org/10.1145/3643991.3644886>
10. Runeson, P., & Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. *EMSE*, 14(2), 131–164. <https://doi.org/10.1007/s10664-008-9102-8>

11. Schneider, S., et al. (2025). Comparison of static analysis architecture recovery tools for microservice applications. *EMSE*, 30, 128. <https://doi.org/10.1007/s10664-025-10686-2>
12. Soltani, M., Derakhshanfar, P., Devroey, X., & van Deursen, A. (2020). A benchmark-based evaluation of search-based crash reproduction. *EMSE*, 25, 96–138. <https://doi.org/10.1007/s10664-019-09762-1>
13. Song, H., et al. (2026). Beyond the Protocol: Unveiling Attack Vectors in the Model Context Protocol Ecosystem. *IEEE TSE*, 52(8), 2410–2426. <https://doi.org/10.1109/TSE.2026.3694876>
14. Tamburri, D. A., & Kazman, R. (2018). General methods for software architecture recovery: a potential approach and its evaluation. *EMSE*, 23(3), 1457–1489. <https://doi.org/10.1007/s10664-017-9543-z>
15. Tomassi, D. A., et al. (2019). *BugSwarm: Mining and continuously growing a dataset of reproducible failures and fixes* (pp. 339–349). ICSE. <https://doi.org/10.1109/ICSE.2019.00048>
16. Trubiani, C., Ghabi, A., & Egyed, A. (2017). Exploiting traceability uncertainty between software architectural models and extra-functional results. *JSS*, 125, 15–34. <https://doi.org/10.1016/j.jss.2016.11.032>
17. Wang, X., Hu, R., Gao, C., Wen, X. C., Chen, Y., & Liao, Q. (2024). ReposVul: A Repository-Level High-Quality Vulnerability Dataset. *ICSE Companion*, 472–483. <https://doi.org/10.1145/3639478.3647634>.

Figures

Source stratum	Cases
NVD/GHSA	6
GitHub issue/advisory	27
Vendor	4
OWASP	2
Incident catalog	1



Patch/corrected state: 39/40

Prefix survivors: 35/40; 5 excluded for prefix evidence insufficiency

Replay-decidable: 31/40; 4 source-pair resolutions remained unresolved

In this frame, lenient eligibility coincided exactly with replay-decidability (31 cases).

Strict inclusion adds the frozen reconstruction/oracle/pairing/release conditions; C38 was excluded solely by Conditional releaseability.

Counts describe the constructed challenge frame and are not population prevalence estimates.

Figure 1

Evidence-quality structure of AIRR-40. The lenient definition coincided exactly with replay-decidability (31 cases) in this frame; strict inclusion required the additional frozen criteria. Counts describe the constructed challenge frame and are not prevalence estimates.