ISSN:2584-2706 DOI: https://doi.org/10.5281/zenodo.22177411

R-DAIR the Open Revenue Data Assurance, Integrity & Reliability Standard

Adetoun Adeleke

Abstract

R-DAIR is an open, self-administered revenue-data assurance standard designed for small enterprises, credit unions, community banks, cooperatives, and other institutions that need continuous visibility into record integrity but cannot rely on large data-governance teams. The standard combines four maturity dimensions, five capability levels, measurable data-quality indicators, robust anomaly detection, owner-routed alerting, and explicit detection-latency accounting. This white paper extends the original R-DAIR design with a 90-day, eight-site pilot comprising 356 staff, 177,400 customer/master records, and 1,039,810 observed transactions. Across the pilot, record-weighted completeness was 96.15%, duplicate rate 2.32%, staleness 15.19%, validity 98.12%, and consistency 97.23%. Field anomaly detection produced 357 true positives, 48 false positives, and 29 false negatives, corresponding to overall precision of 0.881, recall of 0.925, and F1 of 0.903. The median detection latency was 29 seconds (IQR 16-48 seconds; 95th percentile 94 seconds), while median alert dispatch latency was 8 seconds. Average maturity rose from 2.22 to 3.47 across the four dimensions, with the largest gain in anomaly detection and revenue assurance. Sensitivity analysis showed that the default amount-outlier multiplier k=3.0 delivered the highest F1 among tested settings. The results support R-DAIR as a practical continuous-assurance architecture while also identifying important validation needs, especially inter-rater reliability, site-level D2 construct validation, usability measurement, and broader external replication.

Keywords:

Data Quality; Anomaly Detection; Revenue Assurance; Community Financial Institutions; Maturity Model; Continuous Monitoring; Low-Latency Alerting

Executive Summary

Revenue operations depend on customer, master, and transaction records that are continuously created, modified, duplicated, and consumed. Conventional data-quality reviews are often periodic, while billing, renewal, payment, forecasting, and automated decision processes operate continuously. R-DAIR addresses that mismatch by making integrity measurement and anomaly response persistent system functions rather than occasional audit events. The standard is intentionally lightweight: indicators are computable from an institution's own records, the reference implementation is designed around standard-library Python, and maturity scoring remains transparent enough for operational staff to reproduce and challenge.

The empirical extension reported here moves the framework beyond seeded engineering verification. The eight-site pilot spans credit unions, community banks, cooperatives, a regional distributor, and a small enterprise. It shows that a common assurance structure can be applied across heterogeneous institutions while preserving local threshold calibration. The central empirical finding is not that all sites reached the same maturity level; rather, it is that every site moved forward on the four-

dimensional profile, while detection latency remained in seconds rather than the hours or days implied by periodic review regimes.

Key empirical findings

- Pilot scale: 8 institutions, 356 staff, 177,400 master records, 1,039,810 transactions, common 90-day window.
- Record-weighted data quality: completeness 96.15%, duplicate rate 2.32%, staleness 15.19%, validity 98.12%, consistency 97.23%.
- Level-5 D1 tolerance status: 3 sites passed, 3 were partial, and 2 failed the default tolerance set.
- Field detection: precision 0.881, recall 0.925, F1 0.903, and 0.046 false alerts per 1,000 transactions.
- Latency: median detection 29 seconds, 95th percentile 94 seconds, median dispatch 8 seconds, and 0.5% of dispatches above 300 seconds.
- Maturity: mean four-dimension level increased from 2.22 to 3.47; D2 increased by 2.00 levels on average.
- Threshold sensitivity: k=3.0 produced the highest amount-outlier F1 (0.908) among k values from 1.5 to 3.5.

The paper also makes an important correction to a common audit-lag argument. A periodic review interval is not the minimum detection lag. Under uniformly distributed event times, the theoretical minimum approaches zero, the expected lag is approximately half the review interval, and the maximum approaches the full interval. R-DAIR therefore compares its measured latency with an explicit lag distribution rather than with an overstated minimum. This distinction strengthens the framework because its principal advantage is measurable: it shifts detection from a calendar property to an observed system metric.

1. Introduction: The Quiet Cost of Decaying Revenue Data

Revenue operations run on records: customers, contracts, invoices, payments, renewals, and account histories. Those records rarely fail in a single dramatic event. More often, integrity decays through duplicated identities, stale contact fields, missing values, invalid formats, contradictory dates, repeated transactions, and behavioral anomalies that remain individually small while accumulating operational consequences. Data-quality research has long treated quality as multidimensional and fitness-for-use oriented rather than reducible to accuracy alone (Wang & Strong, 1996; Pipino et al., 2002; Batini et al., 2009). R-DAIR adopts that principle but narrows it to the operational data most closely tied to revenue and customer service.

The difficulty for small institutions is not merely recognizing that poor data matter. It is sustaining a control system that can measure integrity continuously, detect transaction-level exceptions, route findings to someone who can act, and do so without an enterprise data-management function. Large governance frameworks are valuable reference points, but they can be difficult to operationalize in organizations where the same employee may own billing, reconciliation, customer service, and exception handling. ISO 8000-61 defines data-quality management processes and connects them to process capability and maturity, while DAMA-DMBOK provides a broad enterprise data-management body of knowledge (DAMA International, 2017; International Organization for Standardization, 2016). R-DAIR is narrower by design: it converts a small set of high-value integrity behaviors into a self-administered maturity profile and executable checks.

The design is also informed by work on data-constrained monitoring. Adeyekun, Orekoya, and Abiola (2023) show how heterogeneous evidence can be integrated into structured monitoring in a resource-limited context, while Adeyekun (2024) emphasizes predictive architectures that remain usable when data are sparse or uneven. Those studies concern coastal and environmental systems rather than revenue operations, so their domain findings are not transferred directly. Their methodological lesson is relevant, however: carefully engineered indicators, local baselines, transparent uncertainty, and

continuous observation can extract operational value from imperfect datasets without requiring an ideal data environment.

Automation raises the stakes. Batch billing, scoring, workflow automation, and AI-assisted systems can propagate record defects rapidly. Olaseinde and Azeez (2022) demonstrate why automated outputs in internal-control settings require verification rather than passive trust, and the NIST AI Risk Management Framework similarly emphasizes measurement, monitoring, and governance of risks created by automated systems (National Institute of Standards and Technology, 2023). R-DAIR therefore treats integrity scores as potential gating inputs for downstream automation. The framework does not require AI to function; rather, it establishes the data-control layer that makes any subsequent automation safer and more auditable.

2. Problem Definition and Design Principles

2.1 Six recurring modes of revenue-data degradation

R-DAIR defines six recurring degradation mechanisms. Duplication occurs when the same real-world entity or transaction is captured more than once. Staleness occurs when previously valid information is no longer current. Incompleteness is the absence of required information. Invalidity describes values that violate format or domain rules. Inconsistency occurs when fields contradict one another within or across records. Behavioral anomaly refers to transaction patterns that depart materially from expected sequences or entity-specific history. The first five mechanisms are primarily record-quality problems; the sixth is a stream-behavior problem. Together, they establish the scope of D1 and D2.

The choice of these dimensions follows established data-quality work but keeps the indicator set operationally small. Batini et al. (2009) review systematic data-quality assessment methods, and Pipino et al. (2002) demonstrate the value of combining objective metrics with context-sensitive assessment. R-DAIR follows the same logic: objective rates are computed from records, while maturity levels capture whether the institution can operationalize those rates through monitoring, workflow integration, and staff response.

2.2 Why periodic review is structurally insufficient

A periodic review does not necessarily fail because it is inaccurate. It fails as a continuous assurance mechanism because the event-to-review delay is governed by the review schedule. If anomalous events are uniformly distributed between reviews, the expected lag is approximately one-half of the review interval, with a maximum approaching the full interval. A daily review therefore implies an expected delay near 12 hours; a weekly review about 3.5 days; a 30-day review about 15 days. These are theoretical comparators, not claims about any specific audit team. Their purpose is to make the time cost of periodicity explicit.

This focus on response time is consistent with financial-fraud research. Olaseinde (2024) argues that instant-payment environments require low-latency detection because post hoc controls lose value as transaction velocity increases. R-DAIR applies the same temporal principle to a broader set of revenue-data anomalies: detection quality is incomplete without a timestamped measure of how long the system took to detect and dispatch the finding.

2.3 Governance must remain interpretable

Maturity frameworks can become opaque when numerous weighted variables collapse into a single score. Ashimolowo and Oginni (2022) emphasize the role of coordinated quality governance and assurance in maintaining performance stability, while Ashimolowo (2024) demonstrates how computational composite scoring can transform multiple governance indicators into decision support. R-DAIR uses that literature selectively. It calculates composite measures where they help describe data quality, but it retains the four-dimensional maturity vector as the primary governance output. A site that is strong in data quality but weak in workforce capability should not be allowed to hide that weakness inside a favorable overall average.

3. The R-DAIR Maturity Framework

R-DAIR expresses institutional readiness as four dimensions scored from Level 1 to Level 5. The dimensions are intentionally distinct. D1 asks whether the underlying records are measurable and within defined integrity tolerances. D2 asks whether anomalies are detected with acceptable accuracy

and latency. D3 asks whether controls are embedded inside operational workflows rather than performed beside them. D4 asks whether staff can interpret, verify, respond to, and improve the system. The profile is reported as a vector such as D1:4, D2:3, D3:2, D4:3. The vector drives remediation more effectively than a single maturity number because it identifies where capability is lagging.

Table 1. R-DAIR maturity dimensions

Dimension	What it measures
D1 - Data Quality & Record Integrity	Completeness, uniqueness/duplicate rate, validity, staleness, and consistency across customer, transaction, and supplier/vendor records.
D2 - Anomaly Detection & Revenue Assurance	Capacity to detect amount outliers, duplicate transactions, irregular sequences, orphan transactions, and related behavioral anomalies; includes detection latency.
D3 - Workflow & Systems Integration	Extent to which validation, deduplication, alert routing, and remediation are embedded inside operational workflows.
D4 - Workforce Capability	Ability of operational staff to interpret scores, respond to alerts, verify automated outputs, and evaluate technology against data conditions.

Source: R-DAIR framework specification in the author-provided draft.

Table 2. Five R-DAIR maturity levels

Level	Name	Characteristics
1	Ad Hoc	No defined ownership; reactive correction; little or no anomaly visibility.
2	Documented	Standards and ownership exist; quality is checked manually or occasionally; anomalies often surface through complaint or reconciliation.
3	Audited	Periodic review with known error rates; event-to-review lag is governed by the audit interval; between-review decay remains only partly visible.
4	Monitored	Automated scoring on defined indicators; scheduled dashboards and batch alerts; latency measured in hours or days.
5	Continuous	Near-real-time scoring and prioritized owner-routed alerts; remediation workflows are triggered and downstream automation can consume integrity scores as a gating input.

Source: R-DAIR framework specification, with the Level 3 latency wording corrected to reflect the review-lag distribution.

4. Formal Integrity Indicators and Detection Logic

4.1 Dimension 1 indicators

Let R denote the relevant record set, F_req the set of required fields, K the identity-key field set, and tau_s the staleness tolerance in days. R-DAIR uses normalized identity keys before duplicate comparison and excludes empty identity keys from duplicate matching so that missing identifiers are treated as completeness problems rather than as mutual duplicates.

Completeness: C(r) = number of populated required fields / number of required fields
Uniqueness: U(r) = 1 when the normalized identity key occurs once in R; otherwise 0
Staleness: S(r) = 1 when age(last_updated) > tau_s; otherwise 0
Validity: V(r) = passed applicable format checks / applicable format checks
Consistency: X(r) = 1 when all declared cross-field rules hold; otherwise 0

At the record level, the default composite is the equal-weight mean of the five good-direction components, with staleness inverted when necessary. Institutions may change the weights, but the component indicators remain visible. For cross-site descriptive visualization in this paper, an aggregate Data-Quality Index (DQI) is calculated as the equal-weight mean of completeness, 100 minus duplicate rate, 100 minus staleness rate, validity, and consistency. This DQI is not a replacement for the primary indicators or for the D1 maturity level; it is a compact visualization aid.

DQI = [Completeness + (100 - Duplicate) + (100 - Staleness) + Validity + Consistency] / 5

The default Level-5 D1 tolerance set requires duplicate rate <=2%, completeness >=95%, and staleness <=20%, in addition to continuous-scoring capability. The pilot dataset reports a tolerance status of pass, partial, or fail against these default data-quality conditions. Validity and consistency remain reported because they influence the integrity picture even where the maturity threshold is defined through the three stated defaults.

4.2 Dimension 2 anomaly classes and robust baselines

R-DAIR evaluates four transaction-stream classes: amount outliers, duplicate transactions, irregular sequences, and orphan transactions. Amount outliers use interquartile-range fences because revenue values can be skewed and because robust quartile-based methods reduce sensitivity to extreme observations (Tukey, 1977). The default multiplier is k=3.0. Global fences are available to all entities; an entity-specific fence requires at least eight historical transactions, after which a deviation from the entity's own baseline can receive higher severity.

Amount-outlier fence = [Q1 - k x IQR, Q3 + k x IQR], where IQR = Q3 - Q1 and default k = 3.0

Sequence logic operates on each customer's time-ordered stream. A transaction repeating the preceding amount within a one-day window is treated as a candidate duplicate transaction; a transaction arriving within 30 seconds of its predecessor is treated as an irregular sequence under the default configuration. Orphan transactions reference customer identifiers that do not exist in the declared master set. These rules are configuration values rather than hard-coded universal truths, which is important because transaction cadence and amount distributions differ by institution.

Evaluation uses precision, recall, and F1 rather than accuracy because anomalies are typically rare relative to the transaction population. Saito and Rehmsmeier (2015) show why precision-recall measures are especially informative under class imbalance. The framework also reports false alerts per 1,000 transactions, a workload-oriented measure that is easier for small operational teams to interpret.

Precision = TP / (TP + FP) Recall = TP / (TP + FN) F1 = 2 x Precision x Recall / (Precision + Recall)

4.3 Latency as a first-class assurance measure

Each anomaly carries an occurrence timestamp and a detection timestamp. The alert layer adds a dispatch timestamp. These timestamps separate two operational questions: how long the detector took to recognize the event, and how long the orchestration layer took to place the finding in front of the designated owner. This separation is important because a fast detector can still fail operationally if alert dispatch is slow or misrouted.

Detection latency $L_d = t_{detected} - t_{occurred}$

Dispatch latency $L_a = t_{dispatched} - t_{detected}$

5. Continuous Assurance Architecture and Operating Model

The R-DAIR architecture is deliberately modular. Source data are ingested and normalized, D1 indicators are calculated, D2 detectors evaluate the transaction stream, a priority engine assigns severity, the routing layer sends findings to named operational owners, and the remediation/audit-log layer records what happened next. D3 determines how deeply those controls are embedded into workflows, while D4 determines whether staff can interpret and challenge the system. The architecture therefore treats monitoring, workflow, and human capability as one operating loop rather than as separate technology and governance projects.

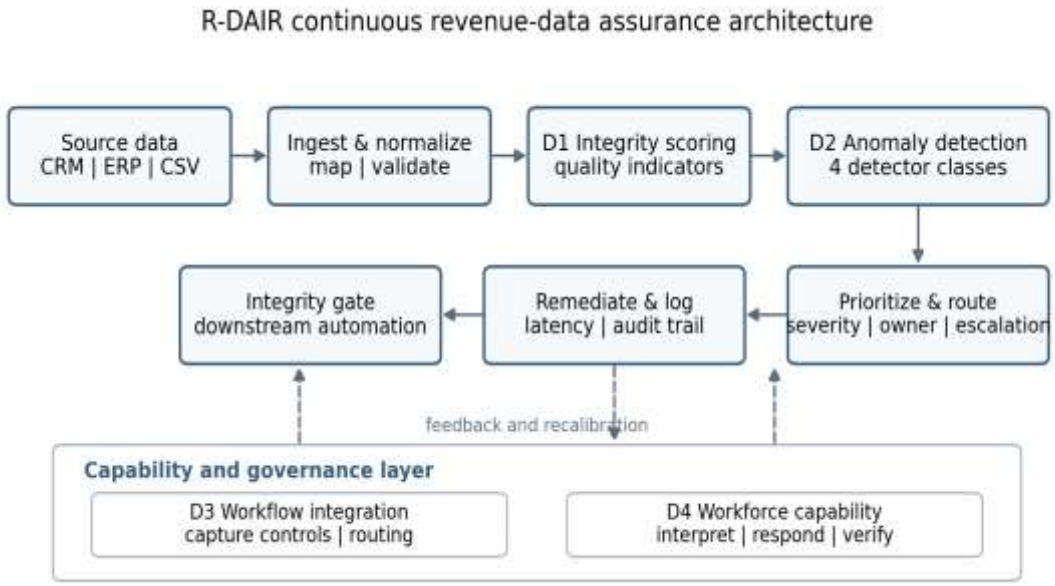

Figure 1. R-DAIR continuous revenue-data assurance architecture and capability feedback loop. Source: Author synthesis from the R-DAIR draft.

Three orchestration rules make the architecture practical for small organizations. First, findings are prioritized by severity rather than presented as an undifferentiated exception list. Second, findings are routed to named record or process owners, with an escalation owner for critical conditions. Third, every dispatch is timestamped and tested against the configured latency threshold, which is 300 seconds by default. The reference implementation described in the draft uses console output and JSON-lines audit logging, with a webhook interface available for integration with email or messaging systems.

The architecture is intentionally compatible with both conventional and AI-assisted workflows. Where AI is used, R-DAIR's role is not to judge model quality directly. It provides a control boundary around the records that feed automated systems. This aligns with the broader need for verification and human oversight identified by Olaseinde and Azeez (2022) and with the measurement and governance functions of the NIST AI RMF (National Institute of Standards and Technology, 2023).

6. Pilot Design and Evaluation Method

6.1 Pilot scope and observation window

The supplied empirical dataset describes an eight-site pilot conducted over a common 90-day observation window. The sites represent two credit unions, two community banks, two cooperatives, one regional distributor, and one small enterprise. Together they account for 356 staff, 177,400 customer/master records, and 1,039,810 observed transactions. Site identifiers are anonymized as I01 through I08, and the source materials do not provide geographic locations. The pilot therefore supports cross-institution operational analysis but not geographic generalization.

Table 3. Institutional profile for the eight-site pilot

Site	Institution type	Staff	Customer/master records	Transactions observed	Observation window
I01	Credit union	34	18,420	102,660	90 days
I02	Community bank	73	41,250	248,310	90 days
I03	Cooperative	22	9,870	54,980	90 days
I04	Regional distributor	48	14,610	89,440	90 days
I05	Credit union	29	12,980	76,550	90 days

I06	Community bank	91	56,400	331,720	90 days
I07	Cooperative	18	7,540	39,280	90 days
I08	Small enterprise	41	16,330	96,870	90 days
Total	**8 institutions**	**356**	**177,400**	**1,039,810**	**Common 90-day window**

Source: Author-provided empirical R-DAIR pilot dataset.

6.2 Measurement sequence

The evaluation sequence is reconstructed directly from the draft and the supplied pilot tables. First, each site's record extract is mapped to the required fields and evaluated for completeness, duplication, staleness, validity, and consistency. Second, the transaction stream is evaluated for the four D2 anomaly classes. Third, detector results are compared with the independently adjudicated ground-truth counts reported in the supplied data. Fourth, occurrence, detection, and dispatch timestamps are used to calculate latency. Fifth, each site is assessed on the four maturity dimensions at baseline and after 90 days. Finally, the amount-outlier multiplier k is varied from 1.5 to 3.5 to examine the precision-recall tradeoff.

The source dataset labels the field benchmark as independently adjudicated ground truth but does not provide the adjudication protocol, number of adjudicators, blinding procedure, or agreement statistic. This white paper therefore reports the supplied confusion counts and derived performance measures without claiming inter-rater reliability. Likewise, no site-level D2 confusion matrices are supplied, so D2 construct validation can be discussed only at the aggregate anomaly-class level.

6.3 Two-stage verification logic

The original draft also describes a deterministic synthetic dataset containing 206 customer records and 1,223 transactions with known seeded defects. That engineering test is retained conceptually as Stage 1 verification: it checks whether the implementation detects defects it was explicitly designed to detect. The 90-day pilot is Stage 2 field evaluation: it tests the same logic on operational data where false-positive burden and latency become observable. Separating these stages is important. Seeded data establish detector correctness under controlled ground truth; field data establish practical performance under messier conditions.

7. Empirical Results

7.1 Data-quality results

Across all 177,400 customer/master records, record-weighted completeness was 96.15%, duplicate rate 2.32%, staleness 15.19%, validity 98.12%, and consistency 97.23%. Three of eight sites met the default Level-5 D1 tolerance status, three were partial, and two failed. The strongest site-level profile was I06, which combined 98.4% completeness with a 0.9% duplicate rate and 8.7% staleness. I03 and I07 were the two failures, driven by lower completeness and higher duplicate and staleness rates.

Table 4. Dimension 1 data-quality results by institution

Site	Completeness %	Duplicate rate %	Staleness %	Validity %	Consistency %	Level-5 tolerance status
I01	97.2	1.6	12.1	98.7	97.9	Pass
I02	94.1	3.8	18.7	97.4	96.1	Partial
I03	91.8	5.6	27.5	95.2	93.8	Fail
I04	96.3	2.4	16.2	98.1	96.8	Partial
I05	95.6	1.9	21.6	97.9	97.2	Partial
I06	98.4	0.9	8.7	99.2	98.8	Pass
I07	92.7	4.9	24.8	96.0	94.6	Fail
I08	96.9	1.4	14.3	98.5	97.6	Pass

Source: Author-provided empirical R-DAIR pilot dataset. Level-5 tolerance status uses the default D1 conditions described in Section 4.1.

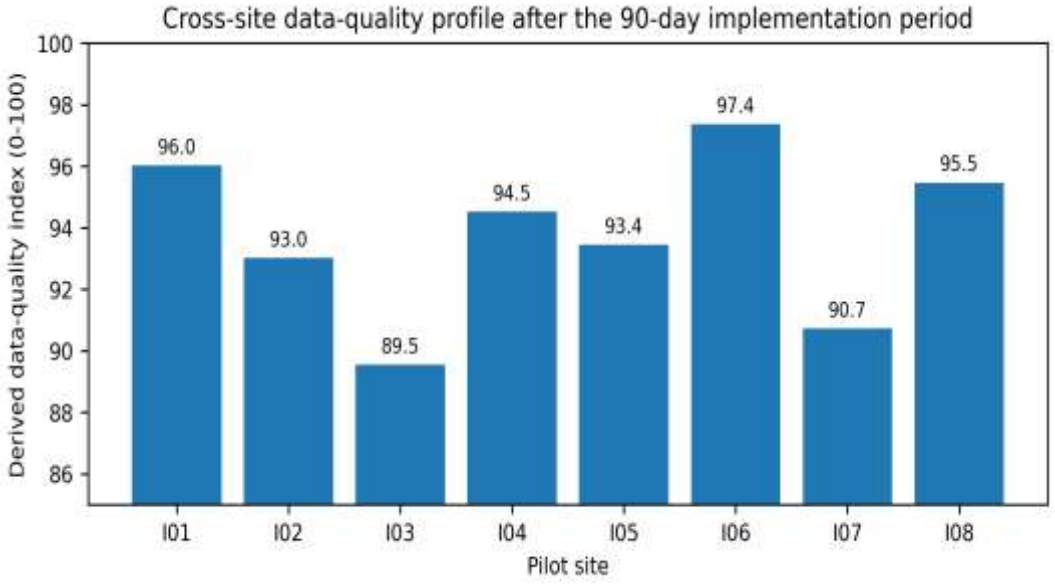

Figure 2. Derived cross-site Data-Quality Index after 90 days. The index is used for visualization only and does not replace the D1 indicator vector. Source: Analysis of the author-provided pilot dataset.

The derived DQI ranged from 89.54 at I03 to 97.36 at I06, with a record-weighted value of 94.80. DQI moved monotonically with the 90-day D1 maturity grouping in this small sample (Spearman rho approximately 0.85). That association is descriptive rather than independent validation because the D1 maturity rules themselves incorporate measured quality conditions, especially at Level 5.

7.2 Field anomaly-detection performance

The field evaluation contains 386 adjudicated anomaly instances: 357 were detected and 29 were missed. The detector also produced 48 false positives. Overall precision was 0.881, recall 0.925, and F1 0.903. Orphan transactions were the strongest class by F1 (0.963) and recall (0.975), while irregular sequences were the most difficult class, with precision 0.779 and F1 0.822. This pattern is operationally plausible because sequence anomalies depend more heavily on context and threshold choice than referential-integrity failures such as orphans.

Table 5. Field anomaly-detection performance against adjudicated ground truth

Anomaly class	TP	FP	FN	Precision	Recall	F1	False alerts / 1,000 txns
Amount outlier	148	18	12	0.892	0.925	0.908	0.017
Duplicate transaction	96	7	5	0.932	0.950	0.941	0.007
Irregular sequence	74	21	11	0.779	0.871	0.822	0.020
Orphan transaction	39	2	1	0.951	0.975	0.963	0.002
Overall	**357**	**48**	**29**	**0.881**	**0.925**	**0.903**	**0.046**

Source: Author-provided empirical R-DAIR pilot dataset.

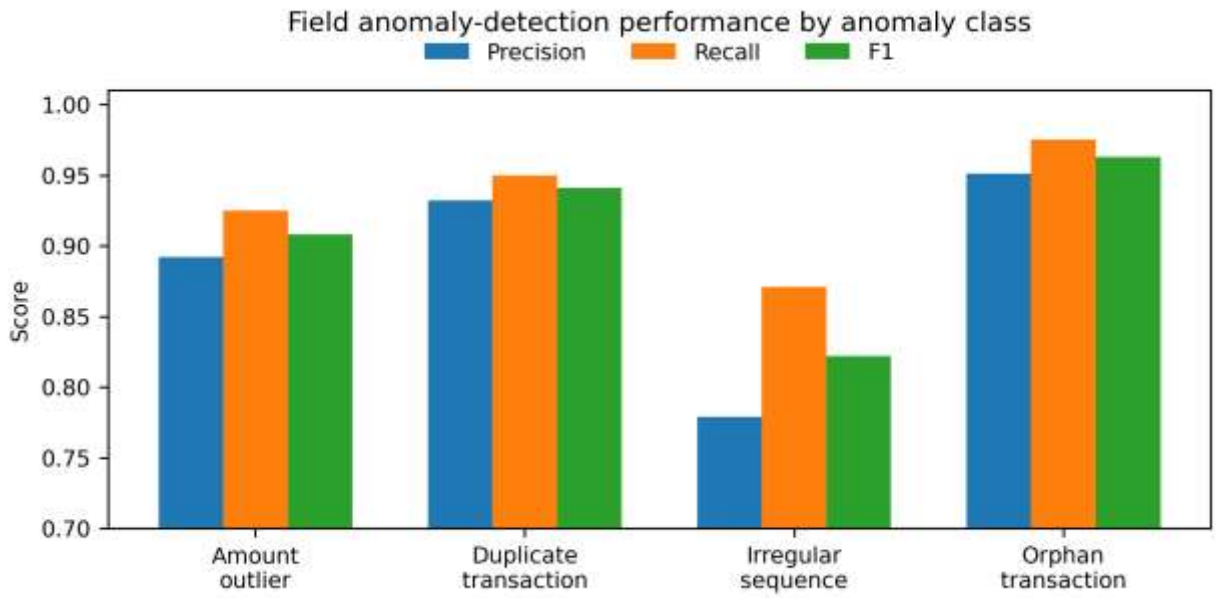

Figure 3. Precision, recall, and F1 by anomaly class. Source: Author-provided pilot results.

The overall false-alert rate of 0.046 per 1,000 transactions corresponds closely to the 48 supplied false positives over 1,039,810 observed transactions. This is useful because it translates statistical performance into an operational review burden. A detector with strong recall but a high false-alert rate can still overwhelm a small team. The pilot results therefore support keeping both classifier metrics and workload metrics in the R-DAIR reporting standard.

7.3 Detection and alert latency

Latency performance was measured only over successfully detected events, which avoids assigning arbitrary delays to false negatives. Across 357 detected events, the median detection latency was 29 seconds, the IQR was 16-48 seconds, and the 95th percentile was 94 seconds. Median dispatch latency was 8 seconds, and only 0.5% of dispatches exceeded the configured 300-second threshold. Irregular sequences were detected fastest at a median of 19 seconds; amount outliers were slowest at 38 seconds. Orphan transactions had the highest median dispatch latency at 11 seconds.

Table 6. Detection and alert-latency results

Anomaly class	Detected n	Median detection latency (s)	IQR (s)	95th percentile (s)	Median dispatch latency (s)	Dispatch >300 s (%)
Amount outlier	148	38	22-61	118	9	0.7
Duplicate transaction	96	24	15-42	79	7	0.0
Irregular sequence	74	19	12-31	66	8	0.0
Orphan transaction	39	31	18-49	95	11	1.1
All detected events	**357**	**29**	**16-48**	**94**	**8**	**0.5**

Source: Author-provided empirical R-DAIR pilot dataset. Detection latency is reported over detected events only.

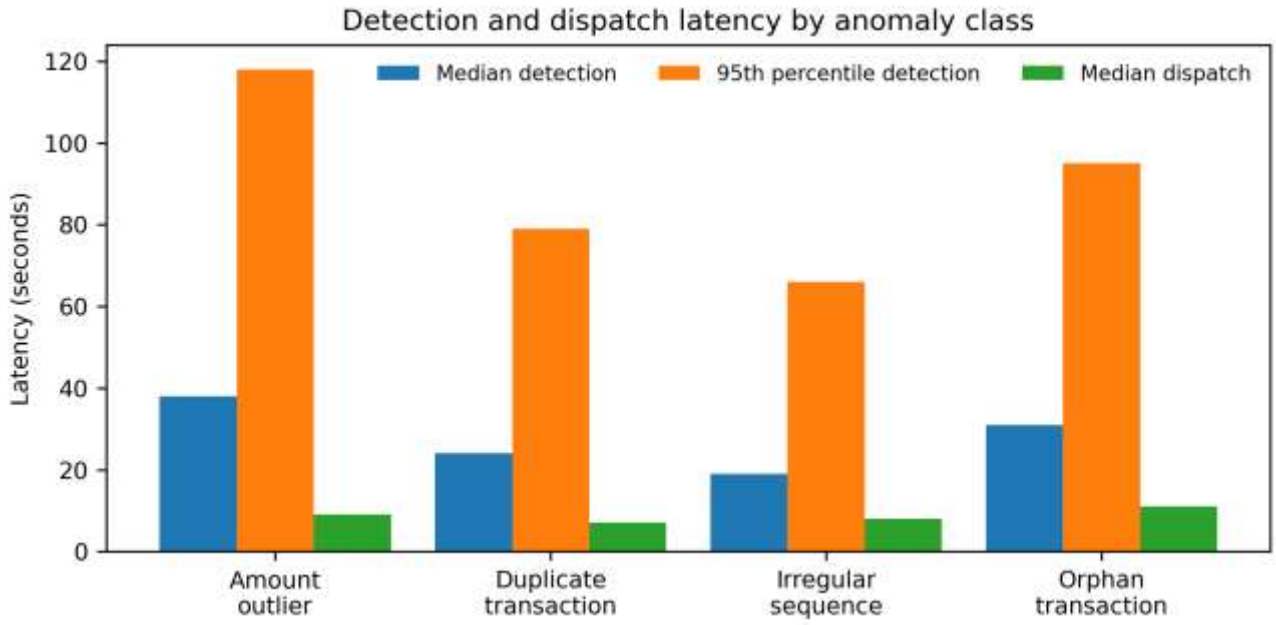

Figure 4. Detection and dispatch latency by anomaly class. Source: Author-provided pilot results.

7.4 Maturity movement over 90 days

The mean maturity profile changed from D1=2.88, D2=1.88, D3=1.88, D4=2.25 at baseline to D1=4, D2=3.88, D3=2.88, D4=3.12 at 90 days. The largest average movement was in D2, which increased by 2.0 levels. D1 increased by 1.12, D3 by 1.0, and D4 by 0.875. Across all 32 site-dimension observations, the mean level increased from 2.22 to 3.47. The pattern suggests that automation and alerting can improve more quickly than workflow redesign and workforce capability, which typically require procedural and behavioral change in addition to technical configuration.

Table 7. R-DAIR maturity profiles before and after the 90-day implementation period

Site	Baseline profile (D1/D2/D3/D4)	90-day profile (D1/D2/D3/D4)	Principal observed change
I01	3/2/2/2	4/4/3/3	Automated D1 scoring and owner-routed anomaly alerts.
I02	3/2/2/3	4/4/3/3	Reduced manual checks; scheduled monitoring and routing.
I03	2/1/1/2	3/3/2/3	Introduced periodic metrics and basic automated anomaly review.
I04	3/2/2/2	4/4/3/3	Embedded validation at capture and alert ownership.
I05	3/2/2/3	4/4/3/4	Staff training combined with automated scoring.
I06	4/3/3/3	5/5/4/4	Near-real-time monitoring and strong D1 indicator performance.
I07	2/1/1/1	3/3/2/2	Named ownership and routine review established.
I08	3/2/2/2	5/4/3/3	D1 indicator tolerances achieved with automated scoring.

Source: Author-provided empirical R-DAIR pilot dataset.

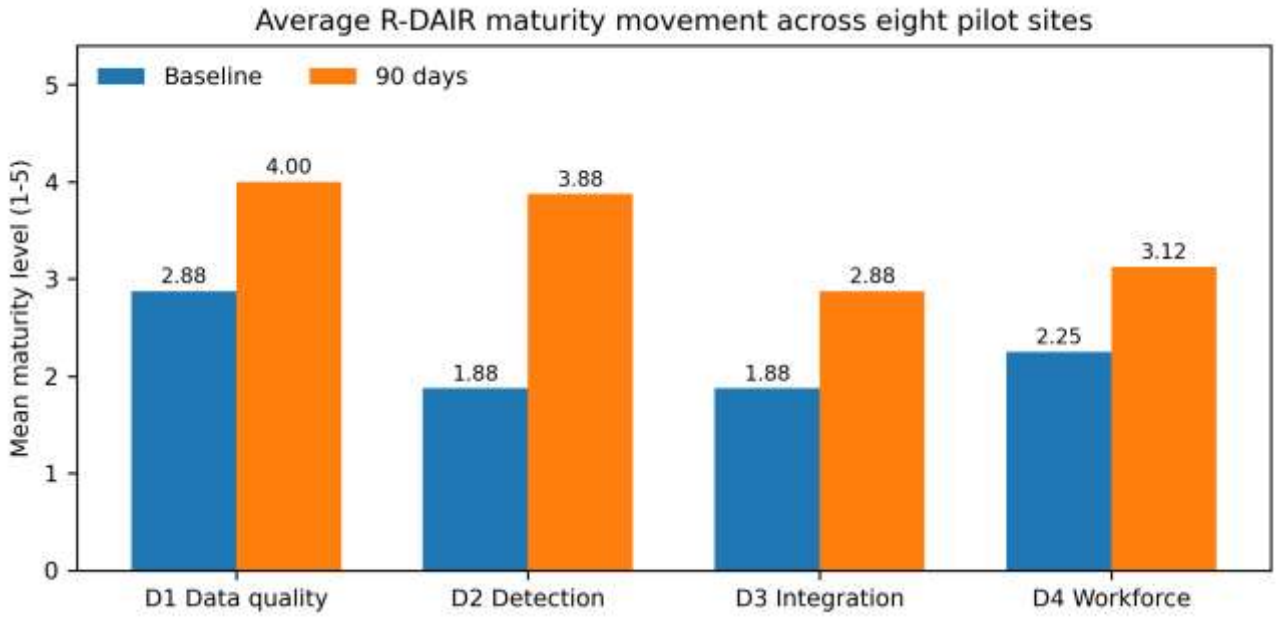

Figure 5. Mean R-DAIR maturity levels at baseline and after 90 days. Source: Analysis of the author-provided pilot dataset.

7.5 Amount-outlier threshold sensitivity

Threshold sensitivity shows the expected tradeoff. At k=1.5, recall was 0.981 but precision fell to 0.620, implying a much heavier review burden. Precision increased steadily as k widened, reaching 0.931 at k=3.5, but recall declined to 0.856. Among the five tested values, k=3.0 maximized F1 at 0.908. This result supports the draft default as a reasonable pilot starting point, not as a universal constant. Institutions with very high costs of missed anomalies may prefer a smaller k, while teams with severe review-capacity constraints may accept a larger k.

Table 8. Sensitivity analysis for the amount-outlier IQR multiplier k

k value	Precision	Recall	F1	Interpretation
1.5	0.620	0.981	0.760	Very sensitive; high review burden from false positives.
2.0	0.748	0.963	0.842	Strong recall, but still operationally noisy.
2.5	0.839	0.944	0.888	Balanced performance for institutions tolerating moderate review volume.
3.0	0.892	0.925	0.908	Best F1 among tested values; supports the draft default, subject to local calibration.
3.5	0.931	0.856	0.892	Lower alert burden at the cost of additional missed anomalies.

Source: Author-provided empirical R-DAIR pilot dataset.

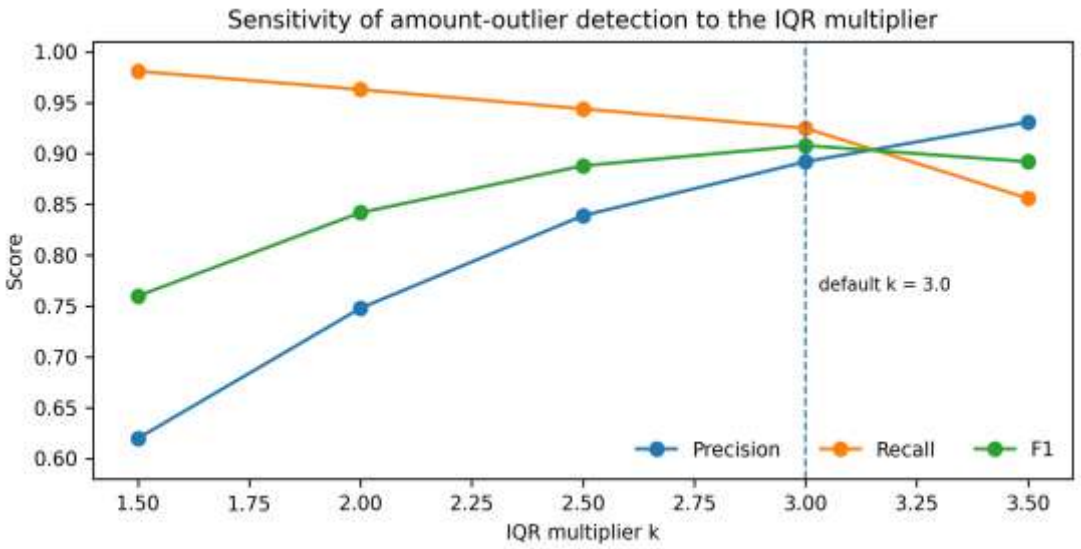

Figure 6. Precision-recall tradeoff across tested IQR multipliers. The dashed line marks the default k=3.0. Source: Author-provided pilot results.

8. Discussion: What the Pilot Establishes

8.1 Continuous assurance is empirically distinguishable from periodic review

The strongest result is the direct measurement of time. The pilot median of 29 seconds is not merely a statement that monitoring is continuous; it is an observed event-to-detection delay. Under a uniform timing assumption, the expected lag of daily, weekly, monthly, and annual periodic review is approximately 12 hours, 3.5 days, 15 days, and 182.5 days respectively. These are theoretical benchmarks, and real teams may review exceptions more or less frequently. Even so, they show why latency should be treated as a first-class assurance metric.

Table 9. Theoretical review-lag benchmark under uniformly distributed event times

Review regime	Approximate minimum lag	Expected lag	Approximate maximum lag	Interpretive use
Continuous R-DAIR	System-dependent	Measured from logs	System-dependent	Report empirical median, IQR, and 95th percentile.
Daily review	Approximately 0	Approximately 12 hours	Approximately 24 hours	Reference comparator when daily manual review is plausible.
Weekly review	Approximately 0	Approximately 3.5 days	Approximately 7 days	Useful comparator for small operational teams.
Monthly review (30 d)	Approximately 0	Approximately 15 days	Approximately 30 days	Illustrates the latency cost of infrequent review.
Annual review (365 d)	Approximately 0	Approximately 182.5 days	Approximately 365 days	Upper-bound comparison for periodic assurance.

Source: Author-provided R-DAIR empirical notes; periodic-review values are theoretical under uniformly distributed event timing.

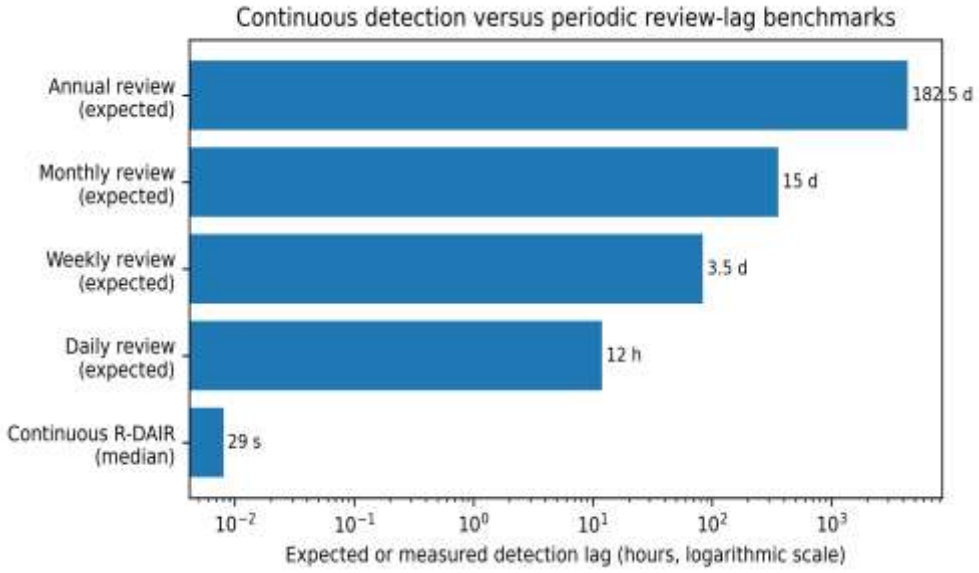

Figure 7. Measured R-DAIR median detection latency compared with expected periodic-review lags on a logarithmic scale. Source: Pilot data and theoretical benchmarks in Table 9.

On that theoretical basis, the 29-second median is roughly 1,500 times shorter than a 12-hour expected daily-review lag and more than 10,000 times shorter than the expected weekly-review lag. These ratios should not be interpreted as comparative effectiveness trials because no manual-review control arm was observed. They simply quantify the temporal difference between a measured continuous detector and a scheduled review model.

8.2 Performance must be read jointly with review burden

The pilot demonstrates why recall alone is inadequate. The amount-outlier detector could reach 0.981 recall at k=1.5, but its precision would drop to 0.620. For a small team, that difference is not statistical trivia; it determines how much human attention the control system consumes. Conversely, the orphan detector achieved precision of 0.951 and recall of 0.975 because referential-integrity violations are comparatively well defined. R-DAIR should therefore preserve class-specific thresholds and workload measures rather than forcing every anomaly type into one generic risk rule.

This conclusion is consistent with the anomaly-detection literature, which emphasizes the contextual nature of anomalies and the difficulty of balancing sensitivity against false positives (Chandola et al., 2009). It is also consistent with Olaseinde (2024), whose real-time fraud framework treats latency and operational constraints as part of detection design rather than as afterthoughts. R-DAIR extends that logic to revenue-data integrity in institutions whose primary constraint is often review capacity rather than model sophistication.

8.3 Maturity profiles reveal where technology does not solve the problem

D2 improved faster than D3 and D4. That is an important result because it prevents the white paper from reducing the intervention to a detector deployment. Detection can be automated relatively quickly; embedded workflow controls and workforce capability take longer. The gap is exactly why R-DAIR retains four dimensions. Ashimolowo and Oginni (2022) argue for integrated assurance rather than fragmented control mechanisms, while Ashimolowo (2024) shows the value of translating governance indicators into operational decision support. The pilot results reinforce both ideas: technical visibility produces value only when ownership, workflow, and human interpretation improve with it.

8.4 Data quality remains the gate for trustworthy automation

The D1 results also show that automated decision systems cannot assume clean upstream data. Two pilot sites failed the default tolerance set, and three were only partial. A downstream AI or scoring model can be technically sophisticated while still consuming duplicated, stale, or incomplete records.

Adeyekun (2024) emphasizes the need for interpretable, uncertainty-aware analysis when data are limited, and Olaseinde and Azeez (2022) show why human reviewers should verify automated conclusions rather than accept them uncritically. R-DAIR operationalizes that principle by making data integrity observable before records are passed into downstream automated processes.

9. Deployment Guidance for Resource-Constrained Institutions

R-DAIR is designed to be adopted incrementally. An institution does not need to reach Level 5 before obtaining value. The most practical sequence is to establish ownership and a minimal data contract, calculate D1 indicators on an existing extract, enable the least ambiguous D2 detectors, start owner-routed alerting, and then embed prevention controls at capture. This progression follows the maturity model rather than a large system-replacement program.

- Start with the existing extract. Declare required fields, identity keys, timestamp fields, and cross-field rules before adding new technology.
- Measure before automating. Establish baseline completeness, duplicate rate, staleness, validity, and consistency so that subsequent change is observable.
- Use conservative anomaly classes first. Orphan and duplicate-transaction checks are often easier to explain and adjudicate than broad behavioral outlier models.
- Route findings to named owners. A detector without ownership creates a backlog, not assurance.
- Track latency separately for detection and dispatch. This reveals whether delay is caused by analytics, orchestration, or response workflow.
- Calibrate thresholds locally. The pilot supports k=3.0 as a reasonable default for amount outliers, but the correct operating point depends on review capacity and the cost of missed events.
- Gate high-impact automation on integrity. Where billing, scoring, or AI workflows can amplify bad records, expose D1/D2 status as a control input rather than as a retrospective dashboard.

This implementation strategy parallels the broader principle of data-driven monitoring in constrained settings: begin with the information that already exists, engineer interpretable indicators, then add prediction or automation only where the data support it (Adeyekun et al., 2023; Adeyekun, 2024). It also follows strategic assurance logic by tying each metric to an owner and intervention rather than treating measurement as an end in itself (Ashimolowo & Oginni, 2022).

10. Limitations and Validation Agenda

The pilot materially strengthens the R-DAIR evidence base, but it does not complete validation. First, eight sites are too few to establish universal thresholds across institution types. Second, the supplied data do not include site geography, selection criteria, or detailed ground-truth adjudication procedures. Third, D2 performance is aggregated by anomaly class rather than reported per site, which prevents a direct site-level test of whether higher D2 maturity corresponds to stronger detection performance and shorter latency. Fourth, the self-assessment's inter-rater reliability and completion burden have not yet been reported. Fifth, the reference implementation remains intentionally lightweight and is not equivalent to a production-grade streaming platform.

The next validation phase should therefore separate content validity, scoring reliability, construct validity, usability, and external replication. De Bruin et al. (2005) emphasize staged development and validation of maturity assessment models; R-DAIR should follow that discipline rather than treating a successful pilot as final standardization. The table below converts the outstanding validation questions into measurable evidence requirements.

Table 10. Recommended empirical validation plan for the R-DAIR maturity assessment

Validation question	Data to collect	Recommended statistic	Expected evidential pattern
Do different assessors assign the same level?	Two independent assessors on the same institutions.	Weighted Cohen's kappa or ICC.	Substantial agreement; investigate systematic disagreement by dimension.

Does D1 maturity correspond to measured record quality?	D1 level plus completeness, duplicate, staleness, validity, and consistency.	Spearman rank correlation / monotonic trend.	Higher D1 levels should coincide with better measured quality; account for threshold rules that partly define the level.
Does D2 maturity correspond to operational detection?	Site-level D2 level, recall, false-alert rate, and detection latency.	Trend tests / rank correlation.	Higher D2 should correspond to shorter latency and stronger operational detection, not merely more tooling.
Is the self-assessment usable by small institutions?	Completion minutes, assistance requests, and missing responses.	Median/IQR and completion rate.	Most institutions complete within one working day without specialist support.
Are assessment items content-valid?	Structured expert ratings of relevance and clarity.	Item- and scale-level content-validity indices.	Weak or ambiguous items are revised before standard release.

Source: Author-provided R-DAIR empirical validation plan, expanded for interpretive clarity.

Threshold calibration should also be expanded beyond the single amount-outlier parameter. Future pilots should report precision-recall curves by institution type, alert workload per staff member, time-to-resolution, recurrence after remediation, and downstream business impact. Where AI-assisted controls are added, the evaluation should include model drift, false-confidence risks, traceability of automated recommendations, and explicit human verification, consistent with the concerns raised by Olaseinde and Azeez (2022) and the NIST AI RMF (National Institute of Standards and Technology, 2023).

11. Conclusion

R-DAIR addresses a practical assurance gap: small institutions need the same visibility into record decay and transaction anomalies that larger organizations obtain through specialized teams and enterprise platforms, but they need it in a form they can operate themselves. The framework combines five-level maturity assessment with executable integrity indicators, robust anomaly detection, owner-routed alerting, and explicit latency measurement. The eight-site pilot provides evidence that this combination can operate at meaningful scale: more than one million transactions were observed, overall anomaly-detection F1 reached 0.903, median detection latency was 29 seconds, median dispatch latency was 8 seconds, and the average maturity profile improved across every dimension over 90 days.

The results also clarify what R-DAIR should not claim. A successful pilot does not make the default thresholds universal, does not establish inter-rater reliability, and does not remove the need for human judgment. Instead, it demonstrates a measurable operating model for continuous revenue-data assurance and defines the evidence required for the next stage of standardization. That is the central contribution: integrity becomes a continuously observed property of the revenue system, with transparent indicators, traceable alerts, and a maturity path that small institutions can act on.

References

- Adeyekun, A. A. (2024). Artificial intelligence framework for predictive sediment dynamics and environmental vulnerability assessment in data-limited coastal systems. Journal of Computational Analysis and Applications, 33(2), 1133-1158.
- Adeyekun, A. A., Orekoya, O. H., & Abiola, O. A. (2023). Data-driven monitoring and modelling of coastal and sedimentary systems for climate resilience. International Journal of Data Science and Engineering, 1(1), 10-42. https://doi.org/10.34218/IJDSE_01_01_002
- Ashimolowo, M. B. (2024). A computational model for risk-based quality governance assessment in high-risk infrastructure projects: A weighted composite scoring approach. Journal of Computational Analysis and Applications, 33(8), 8352-8374.
- Ashimolowo, M. B., & Oginni, A. G. (2022). Quality governance and performance stability in high-risk projects: A framework for strategic assurance. International Journal of Scientific Research and Modern Technology, 1(2), 16-40. https://doi.org/10.38124/ijsrmt.v1i2.1282

- Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM Computing Surveys, 41(3), Article 16. https://doi.org/10.1145/1541880.1541883
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), Article 15. https://doi.org/10.1145/1541880.1541882
- DAMA International. (2017). DAMA-DMBOK: Data management body of knowledge (2nd ed.). Technics Publications.
- de Bruin, T., Freeze, R., Kulkarni, U., & Rosemann, M. (2005). Understanding the main phases of developing a maturity assessment model. In Proceedings of the 16th Australasian Conference on Information Systems (ACIS 2005) (Paper 109).
- International Organization for Standardization. (2016). ISO 8000-61:2016 Data quality - Part 61: Data quality management: Process reference model. ISO.
- National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1
- Olaseinde, A. J. (2024). The speed of theft: Developing real-time detection protocols for authorized push payment (APP) fraud in instant payment systems. Journal of Computational Analysis and Applications, 33(6), 4034-4059.
- Olaseinde, A. J., & Azeez, B. B. (2022). The trust paradox: Analyzing auditor reliance on hallucinating generative AI models in internal control testing. International Journal of Artificial Intelligence Engineering and Transformation, 3(1), 38-49. https://doi.org/10.54660/IJAIET.2022.3.1.38-49
- Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45(4), 211-218. https://doi.org/10.1145/505248.506010
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432
- Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley.
- Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. JournalofManagementInformationSystems,12(4),5-33. https://doi.org/10.1080/07421222.1996.11518099

Appendix A. Seeded Engineering Verification from the Reference Implementation

The original R-DAIR draft describes a deterministic seeded dataset used to verify that the reference implementation recognizes known defect classes before field deployment. This controlled test is not a substitute for the pilot results in Section 7; it is an engineering verification layer. The supplied draft reports complete detection for the separable seeded classes listed below.

Table A1. Seeded defect detection reported in the reference-implementation draft

Seeded defect class	Seeded	Detected
Duplicate identity records	12	12
Missing required fields	8	8
Invalid email format	6	6
Per-customer amount outliers	5	5
Duplicate transactions (second of pair)	4	4
Irregular sequences	3	3
Orphan transactions	4	4

The draft further notes that staleness and date-inversion checks produced additional rule-valid findings because the generator itself created naturally stale records and side effects during stale-date forcing. That behavior illustrates why field validation is necessary: a detector can be correct against its declared rule while the evaluation generator creates cases that complicate a simple seeded-versus-detected count. The field pilot therefore provides the more informative basis for false-positive burden, missed detections, threshold fitness, and latency.