LETTERS IN BIOMATHEMATICS
An International Journal

RESEARCH ARTICLE

OPEN ACCESS

Risk Adaptive Translational Data Continuity for Biomarker Assay Comparability and Clinical Database Lock

Archanaben Maisuria

Manager, Clinical Data Management, Taiho Oncology, Inc., Brentwood, CA, USA
archanamaisuria@ieee.org

ABSTRACT

Any change made to the biomarker assays might have an impact on the interpretation of longitudinal measures and clinical data. RA-TDC combines six pillars of continuity, changes detection in assays, provenance checking with versions, pre-lock assessment. In the oncology retrospective assessment, there were 1,200 patient-samples-assays-versions combinations and 240 defects records associated with the lock. The sensitivity of RA-TDC was 85.0%, while that of conventional quality control was 70.0%, for the identical 10.0% false positive rate, missing defects reduced two-fold. The accuracy rose from 86.0% to 89.0%. The index of Translational Data Continuity (TDCI) showed AUROC 0.86 compared to 0.75 in metadata completeness (the difference was 0.110; 95% CI 0.063 – 0.157). Changes sensitivity to assays failed rose from 65.0% to 85.0% while complete-and-correct provenance recover ability from 74.0% to 90.0%. There is a connection between unbridged changes and high level of classification disagreement (OR adj 2.65; 95% CI 1.50 – 4.68). Provided evidence supports the use of continuity assessment in the given population. Partial scoring and validation decrease reproducibility and generalizability.

ARTICLE HISTORY

Received October 15, 2026

Accepted November 19, 2026

KEYWORDS

Biomarker Continuity, Assay Comparability, Clinical Data Management, Data Provenance, Database Lock, Minimal Residual Disease, Provenance.

1. Introduction

The continuity of a biomarker is guaranteed if the measure remains valid upon the change of the laboratory performing the measurement, of the technology platform or reagents used for the measure, of the classification threshold, or of the definition of the endpoint itself. Despite the fact that some measures will be still kept in the database even after the loss of evidence necessary for comparison with earlier measures, TDC supports connection of scientific notions, history of analytics, and use of the measure during its lifecycle.

Clinical Data Management (CDM) is supposed to do more than just collect data, resolve inquiries, and perform quality reviews. The biomarker-driven development allows CDM teams to manage the stewardship of assay identifier, assay version, specimen metadata, transformation evidence, and endpoint derivation during laboratory work, clinical operations, biostatistical, and regulatory phases. This expansion allows CDM to identify the lack of traceability and to provide auditable record of the transition from original biomarker measure to its analysis and submission datasets. RA-TDC refers to the mentioned expanded roles as continuity evidence and pre-lock review, respectively, not presuming that the data manager team performs exclusively analytic and clinical comparability.

This problem becomes particularly important in the fields of oncology and minimal residual disease (MRD). The guidelines of European LeukemiaNet emphasize the standardization of MRD assessment [1], whereas ctDNA evaluation and validation ask for analytical validation and transparent reporting [2], [3]. The ESMO recommendations differentiate the applicability of supported applications and cases when additional evidence is necessary [4].

There are several examples of biomarker-guided management and postoperative residual disease analyses which demonstrate the context of clinical use of such measurements [5], [6]. In addition, the case of the lung cancer consensus guidance and research of implementation of liquid biopsy assays demonstrates the connection of assay performance and clinical context [7], [8]. Altogether, this evidence motivates to maintain the context of the measurement along the biomarker lifecycle.

CONTACT Archanaben Maisuria: archanamaisuria@ieee.org

RA-TDC uses the approach based on six-pillar index, risk-adaptive change assessment of the assay, and provenance graph verifiability to solve the problem. The former provides the retention of evidence, latter solves the problem of comparability and the former checks that the result and its derived endpoints have the history which could be recovered. This way, RA-TDC links continuity assessment to the identification of the gaps which have to be solved prior to database locking.

The main contribution of the work is integration of workflow for the management of evidence of the decision to modify assay or to lock the database. Evaluation answers five questions: is continuity score able to distinguish the problems, is risk-adaptive assessment of the changes of assay able to detect the assay failure, do bridging changes correlate with discordance, can verifiable history be recovered in case of graph verifiability, and does pre-lock detection improve under similar false positive rate of problems? The endpoints include measurement context, operational detection, and evidence recovery. The operational oncology case study inspires the architecture and is independent from quantitative population.

2. Related Work and Scientific Foundation

A. Assay Comparability and Clinical Interpretation

Different diseases and contexts require different technologies for MRD analysis [9]. EuroMRD guides include quality assurance elements [10]. This discussion indicates that assay identity, definitions of classifications, and application in a clinical study are separate issues, and the criteria for assay comparability are not limited to analytic sensitivity.

Cross-platform ctDNA analyses using sequencing facilities provide measurement performance characteristics and the analytics context [2], [11]. RA-TDC incorporates the consideration of history and bridging of assays for the interpretation. The governance score is not the surrogate for analytic validation.

B. Provenance and Reproducibility

The provenance research of biomedical applications is the source of data about origin, transformations, and transferring of data [12]. FAIR software guidelines, the emphasis on reproducibility, metadata, FAIR cookbook, and the guideline of FAIR workflow support the process of transparent computational artifacts generation and management of their transformation history [13]–[16]. Federated approaches to governance involve data ownership by domain and standardized interfaces in distributed data environments [36].

RO-Crate packages research artifacts with metadata machine-readable format [17]. The FAIR data workflow supports traceable workflow processes [18]; there are other provenance applications discussed in biomedical literature [19]. There are adequate frameworks from these literatures for linking specimen, assay version, results, derivations, and endpoints.

C. Continuity through Biomarker Lifecycle

There are six pillars of continuity that are: Scientific Traceability, Assay Traceability, Metadata Preservation, Version Governance, Data Lineage Management, and Regulatory Defensibility. These pillars represent the scientific purpose of a measurement, assay identity, preserved context, versioning, derivation, and defensible evidence. Example of implementation for the oncology includes a new endpoint of MRD introduced by protocol amendment after the recruitment stage that necessitates integration and resolution of past analysis before the interim one.

Analytic validation literature discusses the issue of measurement performance [2], [3]; RO-Crate and provenance workflow preserve the history of artifact generation and data transformations [17], [18]; RA-TDC combines both topics into the topic of continuity and pre-lock review. These works give conceptual framework and are not benchmarked software tools. Quantitative comparators are metadata completeness, fixed rules for assay changes, document tracing, and conventional QC).

3. Framework and Methodology

A. Study design and scope of evidence

In the study, the retrospective RA-TDC was evaluated using application of RA-TDC approach to the longitudinal oncology records with MRD, genomic biomarker, and IHC data. The cohort included 1,200 records labeled as held out. Each record consisted of one observation involving one patient, one specimen and one version of the assay; therefore, the denominator represents records, not patients. The protocol and amendments description served as operational context and do not make up a cohort.

Three non-additive subsets include three different problems: 400 assay change records for detection of errors, 400 bridging comparisons for classification difference, and 600 recoveries. The RA-TDC and conventional QC evaluations were performed on 1,200 records, while the graph-based and document tracing evaluations were done on 600 recoveries. Different numbers of unique entities,

repetitions and overlaps require description of the cohort to determine the dependence between records. The limitations of the evaluation should be interpreted against the background of the principles of prediction model reporting and evaluation [20] – [23]; held out status does not prove external validation.

TABLE I: SIX CONTINUITY PILLARS AND RETAINED EVIDENCE

Pillar	Evidence retained
Scientific Traceability	Biological rationale, discovery methods, intended use, biomarker definitions
Assay Traceability	Assay identity, version, platform, reagents, validation, laboratory
Metadata Preservation	Units, specimen conditions, storage, transformations, derived variables
Version Governance	Cutoff and algorithm changes, endpoint revisions, approvals
Data Lineage Management	Sources, transfers, transformations, endpoints, submission datasets
Regulatory Defensibility	Verifiable assay history, change rationale, comparability evidence

B. Continuity index and assay-change assessment

For record i, the value of each of continuity pillars S_{ij} is between 0 and 1, where larger values correspond to higher continuity. Index $TDCI_i = (S_{i1} + S_{i2} + S_{i3} + S_{i4} + S_{i5} + S_{i6})/6$. The contribution of each of the pillars to the index is the same and equals one sixth, thus the change in one of the pillars by 0.10 will change the TDCI value approximately by 0.0167 with other pillars' scores being constant. The evidence that remained in each of the domains is shown in Table I.

Equal weighting implies equality of the final weightings of each of the domains. It does not mean equal weights in the field-level analysis, as it is a function of the within-pillars weighting approach. Field coding, evidence sufficiency, and dealing with missingness, inconsistencies, and not applicable status are separate issues that should be defined. These definitions, the rationale behind the weights, and the example of scored record are all to be provided for reproduction (Section V-C).

Since the TDCI index and predicted defect risk are of the opposite nature – the higher the continuity is the more evidence exists, but the higher the defect risk is, the greater the concern about the evidence is – the AUROC of a predictor of defects necessarily requires specification of the scoring direction or risk transformation function. The resulting AUROCs are presented in the study results, while the unreported orientation is a reproducibility limitation.

The risk assessment considered clinically material comparability failure in terms of assay-change properties, bridging evidence, and clinical/laboratory covariates. The logistic regression with regularization and a gradient-boosted tree as a comparison group were defined in the study specification. Pathways considered routine verification, enhanced comparability review, and expert adjudication. The assessed strict, default, and permissive operating configurations and low, medium, and high-risk strata are categorical settings; the actual estimator used, configuration, and decision boundaries are to be defined.

Defect continuity meant a disruption of scientific, assay, metadata, version, lineage, or regulatory evidence. Lock-relevant defect had to be solved before database locking. Comparability failure implied a significant impact of the assay change on classification; discordance meant a disagreement of the classifications being compared. Provenance recovery implied a complete and correct provenance reconstruction. Such definitions allowed separating the evaluation tasks. Analytical interchangeability requires continuous assay agreement/equivalence evidence.

C. Provenance and database-lock review

The directed and versioned provenance graph G = (V, E) described specimens, assays, assay versions, laboratories, results, transformations, clinical endpoints, and submission datasets. The edges of the graph were generation, modification, derivation, validation, and transfer. Verification entailed missing dependencies, conflicting histories, invalid transformations, and incomplete endpoint ancestry. The connected path did not imply a complete and correct recovery endpoint.

The framework saved the original measurements separate from the versioned transformations, used standardized terminology and units, related specimens to assay history, assessed continuity and bridging evidence, and did endpoint dependency checking. Material

discrepancies not solved by the framework were sent to expert review. The comparator tests were associated with the endpoints: metadata completeness checked discrimination, rule-based testing was for assay-change failure detection, document traceability was for history recovery, and conventional QC was for pre-lock defects. The comparator rules and constraints on the graph are some of the limitations of the implementation process that have been summarized in Section V-C.

The processing sequence involved original measurements and versions preservation, continuity evidence assessment and aggregation, assay-change and bridging evidence assessment, endpoint dependencies traversal, discrepancies detection, and material discrepancies send to expert review. The inputs included record, assay history, bridging evidence, metadata, transformations, and endpoint; the outputs were continuity evaluation and discrepancy report. The above sequence constitutes the architecture as it is documented in this project. Executable scoring, graph constraints, and numerical routing rules required the records of implementation presented in Section V-C.

TABLE II: OBJECTIVES HYPOTHESES AND EVALUATION ENDPOINTS

Objective	Hypothesis	Endpoint and comparison
O1 Continuity measurement	H1	TDCI versus metadata-only AUROC
O2 Assay-change detection	H2	Sensitivity at matched false-positive rate
O3 Classification stability	H3	Unbridged versus bridged adjusted discordance odds
O4 Provenance reconstruction	H4	Graph versus document complete-and-correct recovery
O5 Database-lock integrity	H5	Defect sensitivity and false-alert noninferiority

D. Validation and Statistical Assessment

Reporting of statistics distinguishes descriptive measures recalculated from counts from the retained model-based estimates. Grouped nested cross-validation and external validation were specified in the protocol, but specific partition sizes, folds, and cluster IDs were not available. They are not considered as confirmed safeguards. No prospective sample size calculation is noted. Guidelines for modeling and validation [20]-[27] are used in interpreting these validation thresholds.

Each comparison corresponds to an estimand. H1 compares the difference in AUROC between TDCI and metadata completeness. H2 compares the difference in sensitivity of assay change failures. H3 compares the adjusted odds ratio of discordance of unbridged versus bridged tests. H4 compares the difference in paired recovery rates. H5 compares pre-lock sensitivity difference and assesses non-inferiority of false-positive rate. It is impossible to consider distinct outcomes and data subsets as repeated independent confirmation of one effect.

Joint outcomes define paired covariance, while repeated patient, specimen, or transition measures require a clear clustering scheme. Primary results are described as cluster-aware normal-approximation analysis, but their covariance calculations and clustering scheme cannot be replicated. All primary analyses estimates are retained under this limitation. Exploratory independent-cases sensitivity analysis of provenance recovery is additionally computed during manuscript revision based on complete paired data counts; this analysis was not a part of the original Holm family and cannot be used instead of the primary analysis.

Multiplicity adjustment involves Holm correction for five superiority comparisons. Non-inferiority H5 analysis uses RA-TDC minus conventional-QC false-positive rate with upper margin +3 percentage points. Margin justification, selection point, and significance level were not specified. The result is provided against the stated criterion, while prospective margin selection cannot be claimed.

Sensitivity = TP/(TP + FN), specificity = TN/(TN + FP), precision = TP/(TP + FP), accuracy = (TP + TN)/N, and F1 = 2TP/(2TP + FP + FN). Negative predictive value = TN/(TN + FN), and balanced accuracy = (sensitivity + specificity)/2. Matthews correlation coefficient (MCC) = $(TP \times TN - FP \times FN)/\sqrt{[(TP + FP)(TP + FN)(TN + FP)(TN + FN)]}$. Absolute rate differences use percentage points; relative reductions use the comparator count as the denominator. For the exploratory recovery analysis, let b and c denote graph-only and document-only successes. The paired difference is $d = (b - c)/n$ and its independent-case standard error is $\sqrt{[(b + c - nd^2)/(n(n - 1))]}$. A

normal-approximation interval uses d ± 1.96SE. The exact conditional test uses a two-sided binomial test with b successes among b + c discordant pairs under probability 0.5, calculated with SciPy 1.17.0. Calibration is interpreted separately from discrimination [28], with attention to leakage, implementation transparency, and class imbalance [29]–[33].

4. Results and Findings

A. Evaluation population and continuity scores

The evaluation population was composed of 600 MRD data (50.0%), 360 genomic biomarker data (30.0%) and 240 IHC data (20.0%). Lock-relevant defects were 240 while defect-free data were 960 resulting to a defect prevalence of 20.0%. The low-, medium-, and high-risk categories had 480, 480, and 240 data, respectively. The assay change category had 80 failed and 320 non-failed data, and the bridging analysis category had 200 bridged and 200 unbridged data (Table III).

TABLE III: REPORTED EVALUATION COMPOSITION AND ANALYSIS SUBSETS

Characteristic	Reported count	Percentage / description
Total held-out records	1,200	100.0%
MRD biomarker records	600	50.0%
Genomic biomarker records	360	30.0%
IHC biomarker records	240	20.0%
Confirmed lock-relevant defects	240	20.0%
Records without confirmed defects	960	80.0%
Low-risk records	480	40.0%
Medium-risk records	480	40.0%
High-risk records	240	20.0%
Assay-change prediction subset	400	33.3% of held-out records
Comparability failures within subset	80	20.0% of subset
Paired bridging-comparison records	400	200 bridged, 200 unbridged
Provenance reconstruction cases	600	50.0% of held-out records

The TDCI mean score was 0.80 (standard deviation 0.12) and median 0.83. The highest mean score in any pillar was for Metadata Preservation at 0.86, while Data Lineage Management (0.74) and Version Governance (0.76) scored the lowest. The average of the mean scores of all pillars is 0.80, rounding off as specified in the aggregation methodology. The mean for predicted defect risk is 0.20 (0.23). Table IV shows the distributions.

TABLE IV: CONTINUITY CONSTRUCT AND PREDICTED RISK DISTRIBUTIONS

Construct	N	Mean	SD	Median	Min–Max	IQR
Scientific Traceability	1,200	0.84	0.12	0.87	0.35–1.00	0.78–0.94
Assay	1,	0.7	0.	0.83	0.20–	0.71–

Traceability	200	9	15		1.00	0.91
Metadata Preservation	1,200	0.86	0.11	0.89	0.40–1.00	0.81–0.95
Version Governance	1,200	0.76	0.17	0.80	0.15–1.00	0.66–0.89
Data Lineage Management	1,200	0.74	0.18	0.78	0.10–1.00	0.63–0.88
Regulatory Defensibility	1,200	0.81	0.14	0.85	0.25–1.00	0.74–0.92
Equal-weighted TDCI	1,200	0.80	0.12	0.83	0.31–0.98	0.73–0.89
Predicted defect risk	1,200	0.20	0.23	0.10	0.01–0.96	0.04–0.28

B. Operational detection and failure of comparability

RA-TDC detected 204 out of 240 lock-related defects versus 168 by standard QC. Sensitivity raised from 70.0% to 85.0% with the same 10.0% rate of false positives. RA-TDC and conventional QC yielded 96 false positives and 864 true negatives. Unidentified defects dropped from 72 to 36, which is 50.0% fewer. Accuracy improved from 86.0% to 89.0%, precision from 63.64% to 68.00%, and F1 from 0.6667 to 0.7556 (Table V; Figure 1).

Count-based summaries for diagnostics helped make this comparison. Balanced accuracy rose from 80.0% to 87.5%, negative predictive value from 92.3% to 96.0%, and MCC from 0.5794 to 0.6928. The total number of misclassifications reduced from 168 to 132, which is 21.4% fewer. All these metrics correspond to the same test data and don't provide additional validation data or hypotheses.

TABLE V: OPERATIONAL PERFORMANCE ON THE SAME HELD-OUT RECORDS

Metric	RA-TDC	Conventional QC
Evaluation records	1,200	1,200
True positives (TP)	204	168
False negatives (FN)	36	72
False positives (FP)	96	96
True negatives (TN)	864	864
Accuracy	89.00%	86.00%
Precision	68.00%	63.64%
Sensitivity / recall	85.00%	70.00%
Specificity	90.00%	90.00%
F1-score	0.7556	0.6667
False-positive rate	10.00%	10.00%
False-negative rate	15.00%	30.00%
Total misclassifications	132	168
Error rate	11.00%	14.00%
Total flagged records	300	264

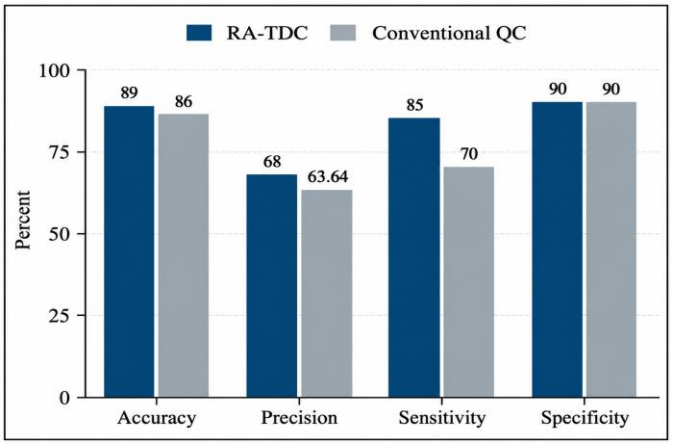


Fig. 1. Operational defect detection at a common false-positive rate. Values correspond to Table V.

For assay change failures, risk-adaptive assessment detected 68/80 (85.0%) instances, while the fixed rule approach identified 52/80 (65.0%). Both approaches had 32 false positives on 320 non-failures. The sensitivity difference is 20.0 percentage points (95% CI 8.2–31.8; p = 0.0009). The number of missed failures declined from 28 to 12, which was a relative reduction of 57.1%. This measure is about comparability failures in the assay change subgroup.

C. Comparison of hypothesis tests and lineage restoration

The TDCI method attained an AUROC of 0.86 (95% CI 0.83–0.89), whereas metadata completeness got 0.75 (95% CI 0.71–0.79). The difference between the two approaches was 0.110 (SE 0.024; z = 4.58; 95% CI 0.063–0.157). Table VI contains the five main hypothesis tests with raw and adjusted probabilities by Holm. Their reported confidence intervals did not include null hypotheses within Section III-D boundaries.

TABLE VI: REPORTED PRIMARY HYPOTHESIS TESTS AND UNCERTAINTY

H	Estimate	SE	z	95% CI	Raw p	Holm p
H1	ΔAUROC = +0.110	0.024	4.58	[0.063, 0.157]	<0.00001	<0.0001
H2	Δ = +0.200	0.060	3.33	[0.082, 0.318]	0.0009	0.0016
H3	OR 2.65; β 0.975	0.290	3.36	OR [1.50, 4.68]	0.0008	0.0016
H4	Δ = +0.160	0.024	6.67	[0.113, 0.207]	<0.000001	<0.000001
H5	Δ = +0.150	0.035	4.29	[0.081, 0.219]	<0.0001	<0.0001

The H5 sensitivity difference was 0.150 (SE 0.035; z = 4.29; 95% CI 0.081–0.219; p < 0.0001). The false-positive-rate difference was 0.0 percentage points (SE 1.0 percentage point; 95% CI −2.0 to +2.0 percentage points). The upper bound remained below the stated +3-percentage-point margin, and the one-sided test yielded z = 3.00 and p = 0.0013. The comparison met the stated noninferiority criterion; its prospective specification was not established.

Graph Verification obtained 540 out of 600 histories complete and correct (90.0%), compared to 444 out of 600 (74.0%) by Document Tracing. Both methods succeeded in 420 cases; Graph Verification alone succeeded in 120 cases; Document Tracing alone in 24; and neither in 36. The difference in recovery rate was 16.0 percentage points (SE 0.024; z = 6.67; 95% CI 11.3-20.7; p < 0.001; see Figure 2). Unrecovered histories were reduced from 156 to 60 (61.5% reduction).

Independent-case sensitivity analysis for exploration yielded exactly the same difference in recovery rate of 16.0 percentage points; SE 1.89 percentage points; approximate 95% CI 12.3-19.7 percentage points. Exact conditional paired test of 144 discordant pairs of results yielded $p = 1.49 \times 10^{-16}$. This statistic is consistent with the direction of the difference in recovery rates under independence of the cases. This is aggregate data analysis and not external validation or confirmation of the original cluster adjusted uncertainty.

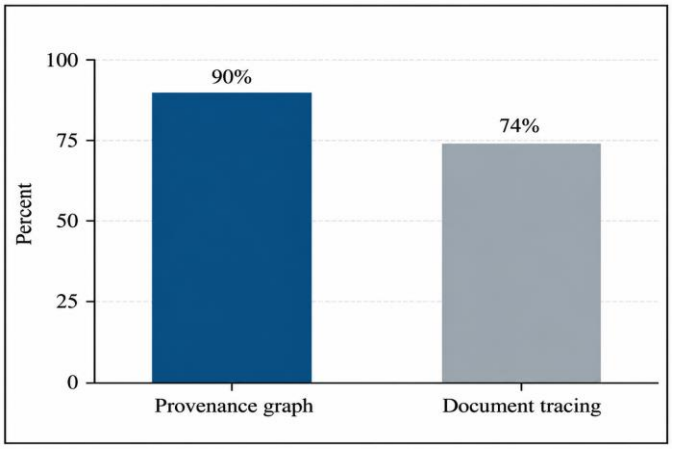


Fig. 2. Complete-and-correct provenance recovery across 600 cases. Graph verification recovered 540 paths and document tracing 444.

D. Biomarker subgroups and assay bridging

Sensitivity was 85.00% for MRD (102/120 defects), 84.72% for genomic biomarkers (61/72), and 85.42% for IHC (41/48), with 18, 11, and 7 missed defects. The observed sensitivity range was less than one percentage point. The subgroup-versus-overall comparisons in Table VII are descriptive because the overall estimate contains each subgroup; they do not establish equivalence or independent interaction effects.

TABLE VII: BIOMARKER DETECTION AND BRIDGING DISCORDANCE

Group	Records	Events	Rate	Reported comparison
MRD	600	102/120 detected	85.00 %	Δ 0.00 pp; CI [−6.4, 6.4]; p 1.00
Genomic	360	61/72 detected	84.72 %	Δ −0.28 pp; CI [−8.6, 8.0]; p 0.95
IHC	240	41/48 detected	85.42 %	Δ +0.42 pp; CI [−9.7, 10.5]; p 0.94
Successfully bridged	200	20 discordant	10.0 %	Reference
Unbridged	200	50 discordant	25.0 %	Adjusted OR 2.65; CI [1.50, 4.68]

Discordance occurred in 20/200 bridged comparisons (10.0%) and 50/200 unbridged comparisons (25.0%). The absolute difference was 15.0 percentage points. The adjusted odds ratio for unbridged changes was 2.65 (95% CI 1.50–4.68; p = 0.0008; Holm-adjusted p = 0.0016), with β = 0.975 and SE = 0.290 on the log-odds scale. This adjusted association does not establish a causal bridging effect (Figures 3 and 4).

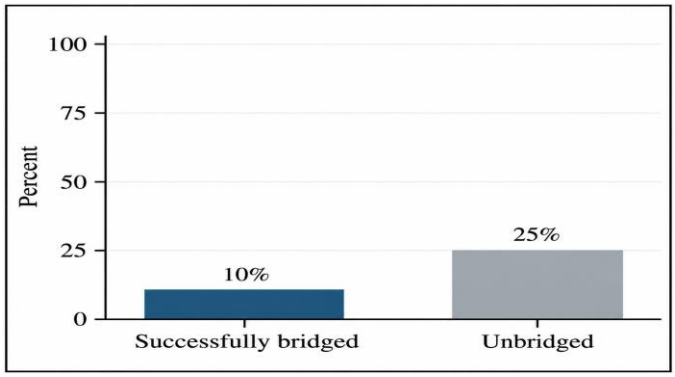


Fig. 3. Classification discordance by bridging status. Each group contains 200 comparisons.

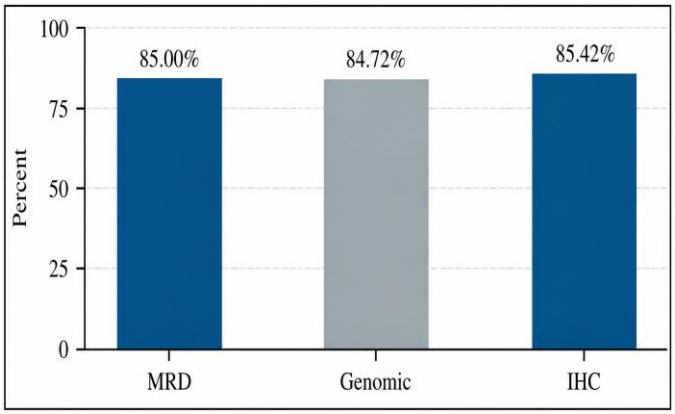


Fig. 4. Detection sensitivity by biomarker class. Counts are 102/120 for MRD, 61/72 for genomic biomarkers, and 41/48 for IHC.

E. Component ablation

All ablation configurations used 240 positive and 960 negative records, with 96 false positives. Removing graph verification reduced sensitivity by 12.50 percentage points, Data Lineage Management by 10.83 points, and risk-adaptive prediction by 10.00 points. Removing Scientific Traceability produced the smallest reduction, at 2.50 points. Table VIII and Figure 5 show the component comparisons.

TABLE VIII: ABLATION PERFORMANCE AT A COMMON FALSE-POSITIVE RATE

Framework configuration	TP / 240	FN	Sensitivity	Δ versus full	F1-score
Full RA-TDC	204	36	85.00%	0.00 pp	**0.7556**
Without Scientific Traceability	198	42	82.50%	−2.50 pp	**0.7416**
Without Assay Traceability	186	54	77.50%	−7.50 pp	**0.7126**
Without Metadata Preservation	194	46	80.83%	−4.17 pp	**0.7321**
Without Version Governance	190	50	79.17%	−5.83 pp	**0.7224**
Without Data Lineage Management	178	62	74.17%	−10.83 pp	**0.6926**
Without Regulatory Defensibility	196	44	81.67%	−3.33 pp	**0.7368**

Without risk-adaptive prediction	180	60	75.00%	−10.00 pp	**0.6977**
Without provenance-graph verification	174	66	72.50%	−12.50 pp	**0.6824**
Conventional QC baseline	168	72	70.00%	−15.00 pp	**0.6667**

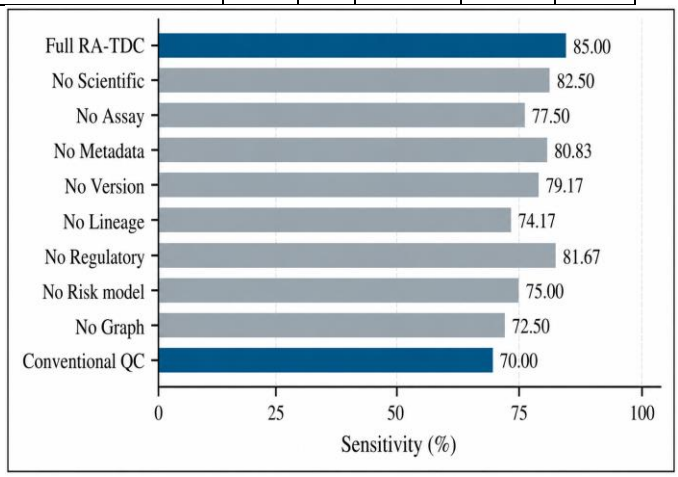


Fig. 5. Sensitivity after component removal. Every configuration uses 240 positive and 960 negative records with a 10% false-positive rate.

Graph verification had the largest observed sensitivity decrease upon removal. Component rankings remain descriptive because retraining, threshold retuning, and paired uncertainty were not documented. The reductions cannot be added to estimate a combined effect or interpreted as independent causal contributions.

F. Operating configurations and risk strata

Strict, default, and permissive configurations achieved sensitivities of 75.0%, 85.0%, and 90.0%, respectively, with false-positive rates of 5.0%, 10.0%, and 15.0%. Precision decreased from 78.95% to 68.00% to 60.00% as review became more permissive (Table IX; Figure 6).

Moving from strict to default increased detected defects by 24 and false positives by 48, raising total alerts from 228 to 300. Moving from default to permissive added 12 detected defects and 48 false positives, raising alerts to 360. The additional detection yield per added alert therefore decreased across these configurations. These marginal totals describe operating trade-offs without identifying an optimal threshold or paired record transitions.

TABLE IX: OPERATING CONFIGURATIONS AND DEFAULT-THRESHOLD RISK STRATA

Condition	Evaluation N	TP	FP	Sensitivity	False-positive rate	Precision
Strict threshold	1,200	180	48	75.00%	5.00%	78.95%
Default threshold	1,200	204	96	85.00%	10.00%	68.00%
Permissive threshold	1,200	216	144	90.00%	15.00%	60.00%
Low-risk stratum, default threshold	480	38	43	79.17%	9.95%	46.91%
Medium-risk stratum, default threshold	480	82	38	85.42%	9.90%	68.33%
High-risk stratum, default threshold	240	84	15	87.50%	10.42%	84.85%

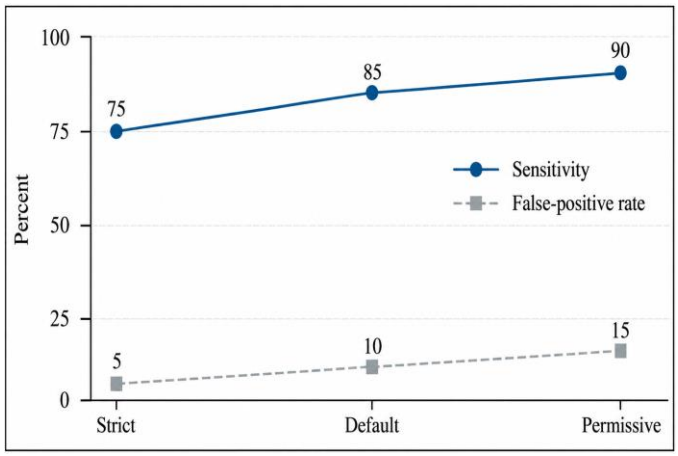


Fig. 6. Sensitivity and false-positive rate across categorical operating configurations. Numerical probability cutoffs are not supplied.

At the default configuration, low-, medium-, and high-risk sensitivities were 79.17%, 85.42%, and 87.50%; precisions were 46.91%, 68.33%, and 84.85%. Stratum counts reconciled with the full-population true-positive and false-positive totals (Table IX).

The corresponding defect prevalences were 10.0% (48/480), 20.0% (96/480), and 40.0% (96/240). This prevalence gradient helps explain the increasing precision despite broadly similar sensitivity and false-positive rates. Higher precision in the high-risk stratum therefore does not alone demonstrate better intrinsic classifier performance.

G. Reliability and diagnostic summaries

Cronbach's alpha was 0.84 and absolute-agreement ICC was 0.82. Confirmatory-factor analysis yielded a CFI of 0.95 and RMSEA of 0.050. The predicted risks had a Brier score of 0.108 and calibration slope of 0.97 (Table X). These statistics summarize internal consistency, rater agreement, factor fit, and probabilistic performance, respectively.

The pillars capture distinct operational evidence domains. Their interpretation as a reflective scale or a formative governance index affects the meaning of internal consistency and factor fit. Alpha and CFI therefore do not by themselves establish pillar interchangeability. Calibration slope likewise requires its evaluation design and uncertainty before transportability can be assessed.

TABLE X: REPORTED RELIABILITY CALIBRATION AND VALIDATION DIAGNOSTICS

Validation domain	Test / statistic	Reported value	95% CI / criterion
TDCI internal consistency	Cronbach's alpha	0.84	Acceptance threshold ≥0.70
TDCI inter-rater reliability	ICC, absolute agreement	0.82	Threshold ≥0.75
TDCI construct validity	Confirmatory factor analysis, CFI	0.95	CFI ≥0.95
TDCI construct validity	RMSEA	0.050	RMSEA ≤0.06
TDCI discrimination	AUROC	0.86	Source-reported CI [0.83, 0.89]
Metadata-only	AUROC	0.75	Source-reported CI [0.71, 0.79]

discriminatio n			
Defect-risk calibration	Brier score	0.108	Lower is better; no universal cutoff
Defect-risk calibration	Calibration slope	0.97	Reference value 1.00
Clinical bridging association	Mixed-effects logistic regression	Adjusted OR 2.65	[1.50, 4.68]
Provenance recovery	Paired recovery comparison	+16.0 pp	[11.3, 20.7] pp
Database-lock sensitivity	Cluster-adjusted Wald test	$z = 4.29$	$p < 0.0001$
False-alert noninferiorit y	One-sided Wald test	$z = 3.00$	$p = 0.0013$; margin +3 pp

Population, subgroup, risk-stratum, and confusion-matrix totals reconciled, and the count-derived rates agreed after rounding. This consistency check supports arithmetic coherence. It does not reconstruct record-level score distributions or independently validate reliability, calibration, and clustered uncertainty.

5. Discussion

A. Definition and application in science and operations

The essential finding was the improvement in defect detection without gain in false positives. RA-TDC discovered 36 additional defects related to locks and halved the missed defects. The remaining 36 missed defects and 96 false positives indicate the need for clinical/operational judgment in the continuity review. Severity of the errors was not estimated, so no direct translation of the detection gain into patient benefit was provided.

The bridging finding establishes the connection between evidence of retained comparability and classification stability. Unbridged comparisons had even higher level of discordance after correction. This indicates the retention of assay histories and bridging evidence, leaving the causal question open. The guidance on MRD and ctDNA analysis separates the aspects of measurement performance, intended use, and clinical interpretation [1], [3], [4], [10].

The improved provenance recovery and the graph removal ablation are consistent with the usefulness of the lineage information in continuity reviews. Provenance can be recovered only if the sources, versions and transformations are known. Therefore, the concept of provenance is complementary to continuity scores [12], [18], [19]. Correctness of recovery is a different concept from assay interchangeability and patient outcome improvement.

B. Context of implementation and review effort

The protocol amendment story demonstrates the positioning of the continuity assessment task in the context of coordination between clinical operations, monitors, sites, laboratories, and data management. The architecture links the evidence retention with discrepancy accountability. The implementation needs consistent identifiers, version history and review process able to resolve material discrepancies. The healthcare data engineering research addresses the importance of the workflow coordination across clinical and administrative systems [37].

C. Expanded role of Clinical Data Management

The RA-TDC methodology extends the operational role of Clinical Data Management from checking the individual data fields to the management of the evidence needed to interpret biomarker results. CDM may maintain the version-controlled library of assays, platforms, cutoffs, laboratory methods and relevant protocol amendments; check the consistency of the specimen and result identifiers; and document the transformation history of derived clinical endpoints. In the event of change of method or threshold, CDM may coordinate the comparability evidence with laboratory scientists, clinicians and statisticians; document the decisions and discrepancies; and resolve

material problems prior to database lock. These activities provide for inspection-readiness and reproducibility of the downstream analyses without substituting for the scientific validation and expert adjudication.

The study does not measure the CDM effort, review time, and comparative effectiveness of particular organizational models. The proposed responsibilities of CDM represent the implementation implications of the continuity workflow and are not independently verified benefits. Prospective implementation will require definition of the evidence owners, change control checkpoints, escalation criteria, and audit trails across the functions involved.

Default RA-TDC generated 300 alerts vs 264 in conventional QC, so that the increase in review load was 13.6%. The net gain of 36 alerts matched the net gain in defect detection because the number of false positives remained unchanged. The operating configurations clearly demonstrate the tradeoff, but the alert counts do not quantify the review effort and time required to fix the problems. The claims about faster database lock or lower costs need such operational measurements.

D. Boundaries of validation

The reproduction requires the completion of the cohort and the implementation record. The missing details of the cohort include the source identifiers, study dates, laboratories, eligibility criteria, extraction, exclusions, sampling strategy, number of patients/specimens/transitions, and the overlap between subsets. The missing details of the scoring include the field dictionary, encodings, aggregation within the pillar, rules for missing/conflicting evidence, justification for the weights, the real worked example. The final estimator, predictor set, preprocessing, fitted settings, software environment, numerical thresholds, graph identifiers and constraints, comparator rules, and adjudication procedures remain undocumented. All these limitations preclude the assessment of the selection bias, independence of reference labels, and implementation fidelity.

The validation record should establish the partition size, fold assignment, decision-time predictors availability, and fold-specific fitting. The cluster units and sizes, paired covariance, orientation of H1 score, H3 adjustment variables and random effects, and H5 margin selection remain undocumented. The reliability requires rater/sample details and factor specifications; the calibration requires fitting and evaluation details. The external validation, alternative weight and missing metadata robustness, precision-recall analysis, and the clinical severity of errors were not established. The exploratory paired case recovery calculation assumes independent cases and cannot account for the unrecorded patient or transition clustering. So, the evidence does not validate the subgroup equivalence, analytical interchangeability, bridging benefit, or generalizability.

E. Ethics and research transparency

Ethics approval or exemption, consent or waiver status, funding, competing interests, and data/code access arrangement require the author confirmed declarations. The status cannot be inferred from the lack of documentation. The reproducibility package should keep the scoring dictionary, anonymized worked record where applicable, model and threshold configuration, graph rules, software environment, and analysis scripts under access procedure that protects participant confidentiality. CONSORT-AI and SPIRIT-AI [34], [35] are applicable to future interventional evaluation but not to the current retrospective study.

6. Conclusion

RA-TDC combined continuity assessment, assay change assessment, and provenance assessment in the same workflow process before lock. It was found that there is improved defect detection at the same rate of false positives as well as an improvement in history completeness and correctness recovery. It was noted that changes that do not have any bridges show more discrepancies with regard to their classifications. This study favors the combined continuity assessment of the records analyzed in oncology.

References

M. Heuser, S. D. Freeman, G. J. Ossenkoppele et al., “2021 Update on MRD in acute myeloid leukemia: a consensus document from the European LeukemiaNet MRD Working Party,” Blood, vol. 138, no. 26, pp. 2753-2767, 2021, doi: 10.1182/blood.2021013626.

I. W. Deveson, B. Gong, K. Lai et al., “Evaluating the analytical validity of circulating tumor DNA sequencing assays for precision oncology,” Nature Biotechnology, vol. 39, no. 9, pp. 1115-1128, 2021, doi: 10.1038/s41587-021-00857-z.

C. M. Lockwood, L. Borsu, M. Cankovic et al., “Recommendations for Cell-Free DNA Assay Validations,” The Journal of Molecular Diagnostics, vol. 25, no. 12, pp. 876-897, 2023, doi: 10.1016/j.jmoldx.2023.09.004.

J. Pascual, G. Attard, F.-C. Bidard et al., “ESMO recommendations on the use of circulating tumour DNA assays for patients with cancer: a report from the ESMO Precision Medicine Working Group,” Annals of Oncology, vol. 33, no. 8, pp. 750-768, 2022, doi: 10.1016/j.annonc.2022.05.520.

J. Tie, J. D. Cohen, K. Lahouel et al., “Circulating Tumor DNA Analysis Guiding Adjuvant Therapy in Stage II Colon Cancer,” New England Journal of Medicine, vol. 386, no. 24, pp. 2261-2272, 2022, doi: 10.1056/NEJMoa2200075.

Y. Nakamura, J. Watanabe, N. Akazawa et al., “ctDNA-based molecular residual disease and survival in resectable colorectal cancer,” Nature Medicine, vol. 30, no. 11, pp. 3272-3283, 2024, doi: 10.1038/s41591-024-03254-6.

C. Rolfo, P. Mack, G. V. Scagliotti et al., “Liquid Biopsy for Advanced NSCLC: A Consensus Statement From the International Association for the Study of Lung Cancer,” Journal of Thoracic Oncology, vol. 16, no. 10, pp. 1647-1662, 2021, doi: 10.1016/j.jtho.2021.06.017.

M. Ignatiadis, G. W. Sledge, S. S. Jeffrey, “Liquid biopsy enters the clinic — implementation issues and future challenges,” Nature Reviews Clinical Oncology, vol. 18, no. 5, pp. 297-312, 2021, doi: 10.1038/s41571-020-00457-x.

J. Chen, R. P. Gale, Y. Hu et al., “Measurable residual disease (MRD)-testing in haematological and solid cancers,” Leukemia, vol. 38, no. 6, pp. 1202-1212, 2024, doi: 10.1038/s41375-024-02252-4.

V. H. J. van der Velden, I. Dombrink, J. Alten et al., “Analysis of measurable residual disease by IG/TR gene rearrangements: quality assurance and updated EuroMRD guidelines,” Leukemia, vol. 38, no. 6, pp. 1315-1322, 2024, doi: 10.1038/s41375-024-02272-0.

B. Gong, I. W. Deveson, T. Mercer et al., “Ultra-deep sequencing data from a liquid biopsy proficiency study demonstrating analytic validity,” Scientific Data, vol. 9, no. 1, Art. no. 170, 2022, doi: 10.1038/s41597-022-01276-8.

M. Johns, T. Meurers, F. N. Wirth et al., “Data Provenance in Biomedical Research: Scoping Review,” Journal of Medical Internet Research, vol. 25, Art. no. e42289, 2023, doi: 10.2196/42289.

M. Barker, N. P. Chue Hong, D. S. Katz et al., “Introducing the FAIR Principles for research software,” Scientific Data, vol. 9, no. 1, Art. no. 622, 2022, doi: 10.1038/s41597-022-01710-x.

J. Leipzig, D. Nüst, C. T. Hoyt et al., “The role of metadata in reproducible computational research,” Patterns, vol. 2, no. 9, Art. no. 100322, 2021, doi: 10.1016/j.patter.2021.100322.

P. Rocca-Serra, W. Gu, V. Ioannidis et al., “The FAIR Cookbook - the essential resource for and by FAIR doers,” Scientific Data, vol. 10, no. 1, Art. no. 292, 2023, doi: 10.1038/s41597-023-02166-3.

C. Goble, S. Cohen-Boulakia, S. Soiland-Reyes et al., “FAIR Computational Workflows,” Data Intelligence, vol. 2, no. 1-2, pp. 108-121, 2020, doi: 10.1162/dint_a_00033.

S. Soiland-Reyes, P. Sefton, M. Crosas et al., “Packaging research artefacts with RO-Crate,” Data Science, vol. 5, no. 2, pp. 97-138, 2022, doi: 10.3233/DS-210053.

S. N. Mitchell, A. Lahiff, N. Cummings et al., “FAIR data pipeline: provenance-driven data management for traceable scientific workflows,” Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 380, no. 2233, Art. no. 20210300, 2022, doi: 10.1098/rsta.2021.0300.

K. Gierend, F. Krüger, S. Genehr et al., “Provenance Information for Biomedical Data and Workflows: Scoping Review,” Journal of Medical Internet Research, vol. 26, Art. no. e51297, 2024, doi: 10.2196/51297.

G. S. Collins, K. G. M. Moons, P. Dhiman et al., “TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods,” BMJ, vol. 385, Art. no. e078378, 2024, doi: 10.1136/bmj-2023-078378.

K. G. M. Moons, J. A. A. Damen, T. Kaul et al., “PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods,” BMJ, vol. 388, Art. no. e082505, 2025, doi: 10.1136/bmj-2024-082505.

G. S. Collins, P. Dhiman, J. Ma et al., “Evaluation of clinical prediction models (part 1): from development to external validation,” BMJ, vol. 384, Art. no. e074819, 2024, doi: 10.1136/bmj-2023-074819.

R. D. Riley, L. Archer, K. I. E. Snell et al., “Evaluation of clinical prediction models (part 2): how to undertake an external validation study,” BMJ, vol. 384, Art. no. e074820, 2024, doi: 10.1136/bmj-2023-074820.

R. D. Riley, K. I. E. Snell, L. Archer et al., “Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study,” BMJ, vol. 384, Art. no. e074821, 2024, doi: 10.1136/bmj-2023-074821.

R. D. Riley, J. Ensor, K. I. E. Snell et al., “Calculating the sample size required for developing a clinical prediction model,” BMJ, vol. 368, Art. no. m441, 2020, doi: 10.1136/bmj.m441.

R. D. Riley, T. P. A. Debray, G. S. Collins et al., “Minimum sample size for external validation of a clinical prediction model with a binary outcome,” Statistics in Medicine, vol. 40, no. 19, pp. 4230-4251, 2021, doi: 10.1002/sim.9025.

O. Efthimiou, M. Seo, K. Chalkou et al., “Developing clinical prediction models: a step-by-step guide,” BMJ, vol. 386, Art. no. e078276, 2024, doi: 10.1136/bmj-2023-078276.

Y. Huang, W. Li, F. Macheret et al., “A tutorial on calibration measurements and calibration models for clinical prediction models,” Journal of the American Medical Informatics Association, vol. 27, no. 4, pp. 621-633, 2020, doi: 10.1093/jamia/ocz228.

G. Varoquaux, V. Cheplygina, “Machine learning for medical imaging: methodological failures and recommendations for the future,” npj Digital Medicine, vol. 5, no. 1, Art. no. 48, 2022, doi: 10.1038/s41746-022-00592-y.

S. Kapoor, A. Narayanan, “Leakage and the reproducibility crisis in machine-learning-based science,” Patterns, vol. 4, no. 9, Art. no. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

D. Chicco, G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” BMC Genomics, vol. 21, no. 1, Art. no. 6, 2020, doi: 10.1186/s12864-019-6413-7.

D. Chicco, N. Tötsch, G. Jurman, “The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation,” BioData Mining, vol. 14, no. 1, Art. no. 13, 2021, doi: 10.1186/s13040-021-00244-z.

I. Walsh, D. Fishman, D. Garcia-Gasulla et al., “DOME: recommendations for supervised machine learning validation in biology,” Nature Methods, vol. 18, no. 10, pp. 1122-1127, 2021, doi: 10.1038/s41592-021-01205-4.

X. Liu, S. C. Rivera, D. Moher et al., “Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension,” BMJ, vol. 370, Art. no. m3164, 2020, doi: 10.1136/bmj.m3164.

S. C. Rivera, X. Liu, A. W. Chan et al., “Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension,” BMJ, vol. 370, Art. no. m3210, 2020, doi: 10.1136/bmj.m3210.

A. V. Supramanian, “Data Mesh Architecture: Revolutionizing Enterprise Data Management through Decentralization,” Int. J. Sci. Res. Comput. Sci., Eng. Inf. Technol., vol. 11, no. 2, pp. 63–71, 2025, doi: 10.32628/CSEIT251112387.

R. R. Pai and J. P. Pendyala, “Optimizing healthcare pipelines for patient benefit: A data engineering perspective on preauthorization delays and denials,” in Software and Data Engineering, Springer, 2025, pp. 40–52, doi: 10.1007/978-3-032-08649-5_3.