Computers and Artificial Intelligence; Vol.3, No.5; 2026
E-ISSN: 3006-3728 P-ISSN: 3006-371X
DOI: https://doi.org/10.70267/cai.v3n5.163170
Published by: Zeus Press

A Reproducible Evaluation Workflow for Quantitative Benchmarking of Medical Data Quality Tools

Xuanhao Shen*

Xinjiang University of Science and Technology, Korla, Xinjiang, China

**Corresponding author: Xuanhao Shen.*

Abstract

Propagating data defects in Electronic Health Records (EHRs) undermine the reliability of clinical AI models. However, the absence of ground-truth defect labels in real-world datasets hinders the standardized evaluation of data quality tools. This study establishes a robust evaluation workflow for a multimodal medical data quality assurance system [A], comprising controlled defect injection, real-dataset profiling, cross-tool benchmarking, and ablation studies. Experimental results demonstrate that the system achieves high detection efficacy, with a recall of 98.00% and precision of 91.91%. The system's discriminative power is validated by a monotonic decline in quality scores as defect ratios increase. Validation on the UCI Diabetes 130 dataset showed missing-rate statistics consistent with existing literature (maximum deviation < 0.05%). Crucially, comparative analysis reveals that general-purpose tools (Great Expectations, DeepChecks, ydata-profiling) lack native medical-semantic capabilities (e.g., unit consistency) and require clean baselines, whereas our system operates effectively without such dependencies. Ablation studies further identify completeness and accuracy as the primary drivers of detection performance. This work offers a reproducible, quantitative benchmarking framework for advancing medical data quality assessment.

Keywords

medical data quality, defect injection, clinical AI, benchmarking, data profiling

1. Introduction

Clinical AI research relies heavily on high-quality medical data. Defects such as missing values, coding errors, unit conflicts and temporal logic contradictions can distort feature statistics and affect model output [1-2]. However, real clinical data lacks record-by-record ground-truth defect labels, making it difficult to directly measure the actual detection capability of quality inspection tools. General-purpose open-source data quality tools are developed for general business scenarios, and their capability boundaries in medical settings have not been quantitatively evaluated.

Existing research mainly focuses on data quality assessment methodology, and fewer provide complete reproducible evaluation benchmarks [3-5]. Controlled defect injection is a feasible solution to the lack of ground-truth labels: known types of defects are artificially injected into clean datasets to obtain ground-truth labels. Combined with natural defect profiling on public datasets, cross-tool comparison and ablation experiments, the performance of quality inspection systems can be comprehensively characterized.

© 2026 The Author(s). Authors retain copyright in their articles.

This paper takes the multimodal medical quality assurance system developed as the system-under-test and conducts a full set of empirical evaluations. The research questions are: (1) What is the system's detection capability for various medical data defects (2) Can the quality score sensitively reflect data quality degradation? (3) What shortcomings do general-purpose open-source quality tools have in medical scenarios (4) What are the respective contributions of the four dimensions: completeness, consistency, accuracy and timeliness? The main work of this paper includes: constructing a controlled defect injection benchmark; conducting real public dataset profiling and comparing with literature results; performing controlled cross-tool comparison with three mainstream general-purpose tools; carrying out four-dimension ablation experiments; and truthfully disclosing the limitations of the evaluation framework itself.

2. Related Work

Data quality standards define core dimensions such as completeness, consistency, accuracy and timeliness. General-purpose open-source tools Great Expectations, DeepChecks and ydata-profiling are mature in general scenarios but lack built-in medical domain knowledge [6-8].

Medical data quality related literature points out that electronic health record defects propagate downstream [9]. Some studies use unsupervised outlier detection to identify suspicious records. Local outlier factor [10], isolation forest [11] and histogram-based outlier score [12] are representative methods that rank suspicious records based on how much a record deviates from the data, without requiring error labels. Controlled defect injection is widely used to evaluate data cleaning and quality inspection algorithms, artificially creating known defects to obtain ground truth and circumventing the difficulty of real data lacking annotations. MIMIC-III, UCI Diabetes 130 and PhysioNet series datasets are commonly used as benchmarks for medical data experiments [13-14].

Currently, the public literature lacks a complete and reproducible evaluation process oriented to medical scenarios. This paper integrates injection evaluation, real-data profiling, cross-tool comparison and ablation into a complete evaluation workflow. The system architecture and module implementation of the system-under-test are described in detail in reference.

3. Evaluation Framework and Experimental Setup

3.1 Overall Evaluation Scheme

The entire evaluation includes four groups of experiments: (1) Controlled defect injection experiment: 7 types of common medical defects are artificially injected into a clean dataset, and detection recall and precision are calculated. The discriminative power staircase of the score as a function of defect proportion is tested. (2) Real public dataset profiling: quality profiling is performed on multiple public medical datasets, and the missing rate of UCI Diabetes 130 is compared with published literature results to verify the correctness of statistical. (3) Cross-tool horizontal comparison: under a unified experimental protocol, the detection capability of this system is compared with Great Expectations, DeepChecks and ydata-profiling for various types of defects. (4) Four-dimension ablation experiment: all checks under one dimension are removed one by one, and recall and precision are re-measured to evaluate the contribution of each dimension to defect detection.

Defect detection adopts the clean-frame vs defect-frame differential approach: inherent alarms on clean datasets are deducted, and only newly added problem records after injection are counted as detections. The total score weighting rules are: completeness 0.35, consistency 0.25, accuracy 0.25, timeliness 0.15, consistent with the system implementation in reference.

3.2 Experimental Environment and Datasets

The experimental environment is Windows 11, Python 3.11.9, pandas 2.3.3, Spark 4.1.1, DeepChecks 0.19.1, Great Expectations 1.23.2, and ydata-profiling 4.18.4. Different tools are installed in separate virtual environments to avoid dependency conflicts. All injection experiments fix the random seed to ensure reproducibility.

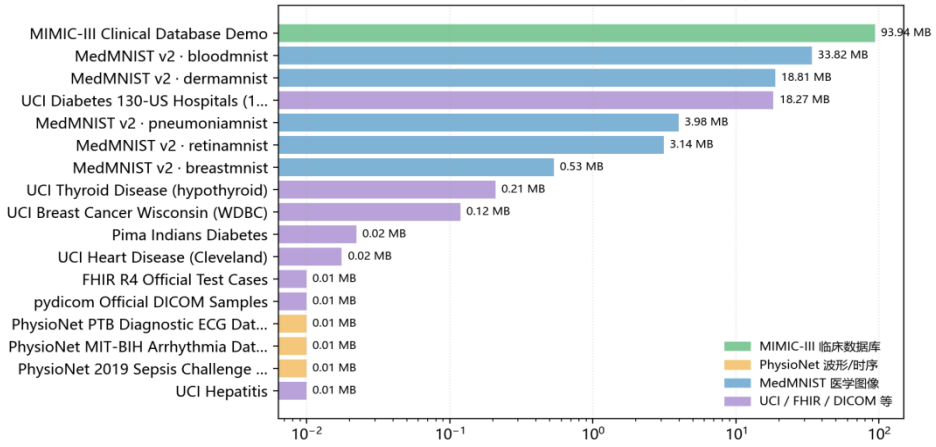
© 2026 The Author(s). Authors retain copyright in their articles.

Table 1: Representative datasets used in the experiment

Dataset	Scale	Modality / Format	Usage
MIMIC-III demo (CHARTEVENTS)	758,355 rows × 15 cols	Structured CSV	Large-scale end-to-end evaluation
MIMIC-III demo (LABEVENTS)	76,074 rows × 9 cols	Structured CSV	Real-data quality assessment
UCI Diabetes 130-US hospitals	101,766 rows × 50 cols	Structured CSV	Natural defect profiling, cross-tool comparison
PhysioNet 2019 Sepsis (training A)	24 hourly cases	PSV time-series	Time-series missing profiling
PhysioNet MIT-BIH arrhythmia	15 records	WFDB ECG	Waveform ingestion and quality check
MedMNIST v2 (5 subsets)	Max 33.8 MB	NPZ image	Image modality ingestion and statistics
FHIR R4 official samples / pydicom samples	8 files	JSON / DICOM	Standard format ingestion

The experimental dataset includes a total of 17 public datasets with a total size of 183.3 MB. Representative datasets are shown in Table 1.

Figure 2: Scale distribution of included real public datasets



4. Experimental Results and Analysis

4.1 Detection Capability under Controlled Defect Injection

The test benchmark is a clean clinical structured dataset with 2000 rows and 24 columns. Seven types of common medical defects are artificially injected, totaling 500 defect instances. The detection results under controlled injection are shown in Table 2.

Table 2: Per-type detection results under controlled defect injection (2000 rows × 24 cols)

Defect Type	Injected Count	Detected Count	Detection Rate
Missing values	380	370	97.37%
Numerical out-of-range	20	20	100%
Illegal coding	20	20	100%
Temporal anomaly	10	10	100%
Unit conflict	40	40	100%
Cross-field contradiction	10	10	100%
Duplicate records	20	20	100%
Total	500	490	98.00%

The overall recall is 98.00% and precision is 91.91%. All 10 undetected instances belong to sparse random missing: their single-column missing rate is below the 5% alarm threshold, and the field-by-field threshold

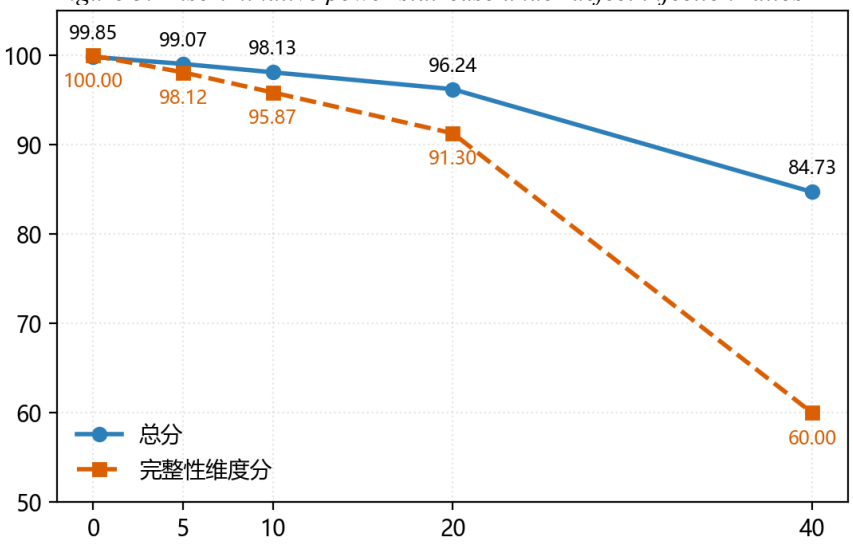
© 2026 The Author(s). Authors retain copyright in their articles.

approach is not sensitive to such dispersed missing values. It should be emphasized that the above table reflects the detection capability for injected defects, not the clinical error detection accuracy.

4.2 Discriminative Power Staircase

To verify the sensitivity of the score to quality degradation, missing defects are injected into the same clean dataset at proportions of 0%, 5%, 10%, 20% and 40%, and the changes in total score and completeness dimension score are observed (Figure 3). The total score decreases monotonically with injection proportion (99.85, 99.07, 98.13, 96.24, 84.73), and the completeness dimension score drops from 100.00 to 60.00. This shows that the score neither over-alarms on minor defects nor fails to respond to severe quality degradation, demonstrating discriminative power.

Figure 3: Discriminative power staircase under defect injection ratios



4.3 Natural Defect Profiling on Real Data

Injected defects are all outlier-type, which can be challenged as self-proving. To address this, profiling is directly performed on real public data that inherently has defects. The field missing rate of UCI Diabetes 130 can be independently compared with literature reported values (Table 3).

Table 3: Natural defect profiling of real public datasets

Dataset	Scale	Total Score / Rating	Most Severe Issue (Top-1)
UCI Diabetes 130	101,766 × 50	83.80 / Good	weight missing 96.9%
UCI Hepatitis	155 × 20	99.22 / Excellent	protime missing 43.2%
UCI Heart Disease	303 × 14	96.63 / Excellent	major_vessels missing 1.3%
PhysioNet Sepsis (single case)	54 × 41	84.11 / Good	DBP fully missing (100%)

Independent comparison with literature results: measured weight missing rate is 96.9% (literature 96.9%, deviation 0.04%), max_glu_serum 94.8% (literature 94.7%, deviation 0.05%), A1Cresult 83.3% (literature 83.3%, deviation 0.02%), medical_specialty 49.1% (deviation 0.02%), payer_code 39.6% (deviation 0.04%). The statistical results of multiple fields are highly consistent, proving that the system's missing rate statistical is consistent with public research.

4.4 Cross-tool Comparison with General Data Quality Tools

To answer the question of whether general tools can cover medical data quality, this system is compared with Great Expectations, DeepChecks and ydata-profiling under a unified protocol on the same injected defect dataset. For fair comparison, an oracle baseline derived from clean data is provided to general tools, while this system only uses factory-built-in medical knowledge and does not access the clean dataset. The comparison results are shown in Table 4 and Table 5.

© 2026 The Author(s). Authors retain copyright in their articles.

Table 4: Cross-tool per-defect-type detection comparison

Tool	Version	Manual Config	Missing	Out-of-range	Illegal Coding	Temporal	Unit Conflict	Duplicate
This system	V1.0	0 lines	Detected	Detected	N/A*	Detected	N/A*	Detected
Great Expectations	1.23.2	44 expectations	Detected	Detected	Detected	Not detected	Not expressible	Detected
DeepChecks	0.19.1	3 lines	Not detected	Not detected	Not detected	N/A	N/A	Not detected
ydata-profiling	4.18.4	1 line	Detected	Informational only	Not detected	N/A	N/A	Not detected

The two items marked with require special explanation: the unit consistency and coding compliance checks of this system are not attributed by column name in terms of evidence by design (unit consistency is grouped by “indicator”, coding compliance is aggregated by dictionary). Therefore, under the column-level, they should be recorded as “not column-attributable”, while their instance-level detections are 40/40 and 20/20 respectively.

On the comparable unit-level, Great Expectations declares 467 problem units and hits 467, with recall of 93.40% and precision of 100%; this system has recall of 98.00% and precision of 91.91% under the instance-level. The granularity differs between the two, and the denominators are noted in Table 4 and its explanation, so direct comparison is not appropriate.

Table 5: Cross-tool per-defect-type detection

Defect Type	MedQ	DeepChecks	Great Expectations	ydata-profiling
Duplicate Records	Detected 20/20 instances	Not detected	Detected	Detected
Cross-Field Contradiction	Detected 10/10 instances	Not detected	Detected	Not Expressable
Unit Conflict	Not Detected 40/40 instances	Not Expressable	Not Expressable	Not Expressable
Time Anomaly	Detected 10/10 instances	Not Expressable	Not detected	Not Expressable
Illegal Encoding	Not Detected 20/20 instances	Not detected	Detected	Not detected
Value Out-of-Bounds	Detected 20/20 instances	Not detected	Detected	Informational Only
Missing Value	Detected 370/380 instances	Not detected	Detected	Detected

4.5 Four-dimension Ablation

To verify whether all four dimensions have independent contributions, all checks under a certain dimension are removed one by one, and the detection capability is re-measured (Table 6).

Table 6: Four-dimension ablation results (300 injected defects)

Scenario	Number of Checks	Recall	Precision	Recall Change
Full set (four dimensions)	12	98.00%	89.47%	—
Remove completeness	9	26.00%	89.26%	−72.00%
Remove accuracy	8	86.00%	92.41%	−12.00%
Remove consistency	9	97.00%	88.06%	−1.00%
Remove timeliness	10	98.00%	88.81%	0.00%

Completeness and accuracy are the main sources of detection, corresponding respectively to missing-type defects with the highest injection proportion, and to accuracy defects such as out-of-range, coding, unit and outliers. Removing timeliness does not cause a decrease in total recall. Upon investigation, the reason is that the injected temporal anomalies can simultaneously be captured by checks from other dimensions. This does not mean that the timeliness dimension is useless, but rather indicates that injection-based evaluation has a structural bias in characterizing dimension contributions.

4.6 Dual-engine Consistency and Rule Transferability

The same data (5000 rows, with injected defects) is evaluated separately using the local engine and the Spark engine, and whether the conclusions are consistent is compared. The results show: the ratings on both

© 2026 The Author(s). Authors retain copyright in their articles.

paths are completely consistent (both Excellent), and the passed/failed check items are completely consistent (38/49). The total score has a difference of 0.31 points (98.86 vs 99.17). After item-by-check, the difference uniquely comes from the model performance check item.

In terms of rule transferability, the field applicability of the same set of 22 factory rules on four external real datasets is: UCI Diabetes 130 improved from 9% (2/22) to 23% (5/22), PhysioNet Sepsis improved from 9% to 41% (9/22), and UCI Hepatitis and Heart Disease are 23% and 27% respectively. Field naming differences are the biggest obstacle to cross-institutional reuse.

5. Discussion

5.1 Experimental Implications

Three practical implications are supported by the experiments in this paper. First, medical semantics is a shortcoming of general tools rather than a supplementary item: under a unified protocol, the unit consistency dimension cannot be natively expressed in any of the three general tools, while unit mixing silently pollutes cross-institutional statistics, which is a check that must be undertaken by domain knowledge. Second, zero configuration and computing adaptation are necessary conditions for primary healthcare deployment: the baseline requires 44 oracle expectations and relies on clean data, while this system is zero-config, which is more suitable for grassroots environments that often do not have a clean baseline. Third, engine consistency takes priority over speedup: the vast majority of check results on the Spark and pandas paths remain consistent, with only a few check items still to be aligned.

5.2 Limitations of the Evaluation Framework

The limitations and boundaries of this paper need to be truthfully listed. (1) Injected defects are synthetic defects, covering only "programmatically injectable" defect forms. Sparse cross-field missing values, semantic errors (such as diagnosis-medication mismatch), etc. are not covered. Therefore, the 98.00% recall rate only indicates the detection capability for injected defects, does not constitute clinical accuracy, and does not represent performance on any real dataset. (2) Four-dimension ablation shows that the timeliness dimension shows no contribution on this injection set, because the injection set does not include "data staleness" samples. (3) The model_performance check has not been fully aligned between the dual engines. (4) The field applicability of factory rules across institutions is at most 41%, relying on field mapping configuration. (5) The medical knowledge base of the system-under-test comes from public literature and has not undergone clinical expert blind annotation; the Spark cluster has only been verified on a single machine with multiple JVMs.

5.3 Comparison with Related Work

Some studies focus on algorithmic comparison of unsupervised anomaly detection methods. This paper focuses on end-to-end evaluation of a complete system, integrating defect injection, real-data profiling, cross-tool comparison and ablation into a reproducible evaluation workflow, which can subsequently be used as a benchmark for other medical quality tools.

6. Conclusion

This paper completes a full set of empirical evaluations based on the multimodal medical data quality assurance system developed in reference. Under controlled defect injection conditions, the system achieves a recall rate of 98.00% and precision of 91.91%. The quality score monotonically decreases with defect proportion. Profiling on real public datasets is highly consistent with published literature statistics. Cross-tool comparison confirms that general-purpose open-source tools lack medical semantic checks such as unit consistency and rely on clean baselines to generate large numbers of validation configurations. Ablation experiments show that completeness and accuracy are the two core dimensions for defect detection. This study also fully discloses the boundaries and limitations of the evaluation framework.

© 2026 The Author(s). Authors retain copyright in their articles.

Future work directions include: expanding the high-level semantic defect injection sample library, researching automatic field mapping algorithms, aligning all check logic between the dual engines item by item, and collecting real defect datasets annotated by clinical experts for further validation.

References

[1] Clemente F, Ribeiro G M, Quemy A, et al. ydata-profiling: accelerating data-centric AI with high-quality data. Neurocomputing. 2023; 554: 126585.

[2] Weiskopf N G, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association. 2013; 20(1): 144-151.

[3] Breunig M M, Kriegel H P, Ng R T, et al. LOF: identifying density-based local outliers. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. New York: ACM; 2000: 93-104.

[4] Liu F T, Ting K M, Zhou Z H. Isolation forest. 2008 Eighth IEEE International Conference on Data Mining (ICDM). Piscataway: IEEE; 2008: 413-422.

[5] Goldstein M, Uchida S. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLoS ONE. 2016; 11(4): e0152173.

[6] Johnson A E W, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data. 2023; 10: 1.

[7] Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. ACM Computing Surveys. 2009; 41(3): 1-58.

[8] Abhyankar S, Demner-Fushman D, Mcdonald C J. Standardizing clinical laboratory data for secondary use. Journal of Biomedical Informatics. 2012; 45(4): 642-650.

[9] Ta C N, Weng C. Detecting systemic data quality issues in electronic health records. Studies in Health Technology and Informatics. 2019; 264: 857-861.

[10] Batini C, Cappiello C, Francalanci C, et al. Methodologies for data quality assessment and improvement. ACM Computing Surveys. 2009; 41(3): 1-52.

[11] Schelter S, Lange D, Schmidt P, et al. Automating large-scale data quality verification. Proceedings of the VLDB Endowment. 2018; 11(12): 1781-1794.

[12] Yang J, Shi R, Wei D, et al. MedMNIST v2—a large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data. 2023; 10: 41.

[13] DAMA International. DAMA-DMBOK: data management body of knowledge. 2nd ed. Basking Ridge, NJ: Technics Publications; 2017.

[14] Sweeney L. k-anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002; 10(5): 557-570.

Funding

This work was supported by the College Students Innovation and Entrepreneurship Training Program of Xinjiang University of Science and Technology (No. X202613561140).

Data and Code Availability

All 17 public medical datasets used in this study are obtained from open-access academic repositories, with source URLs, license agreements and SHA-256 checksum provenance lists provided. Scripts for reproducing experiments include scripts/verify_paper_experiments.py, scripts/verify_baseline_compare.py and scripts/verify_spark_gate.py. Full source code of the system can be obtained from the corresponding author

© 2026 The Author(s). Authors retain copyright in their articles.

upon reasonable request. Public datasets shall be downloaded in compliance with each repository license terms.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights

Copyright for this article is retained by the author (s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

© 2026 The Author(s). Authors retain copyright in their articles.