

Assessing Data Quality in Heterogeneous Healthcare Integration: The AIDAVA Framework

Jens Declerck, Ömer Durukan Kiliç, Ensar Emir Erol, Shervin Mehryar, Dipak Kalra, Isabelle de Zegher, Remzi Celebi

Submitted to: JMIR Medical Informatics
on: March 31, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
---------------------------------	----------

Preprint
JMIR Publications

Assessing Data Quality in Heterogeneous Healthcare Integration: The AIDAVA Framework

Jens Declerck^{1,2} MSc; Ömer Durukan Kiliç³ MSc; Ensar Emir Erol⁴ MSc; Shervin Mehryar⁵ MSc; Dipak Kalra^{1,2} Prof Dr; Isabelle de Zegher⁶ Dr med; Remzi Celebi³ Dr.

¹ The European Institute for Innovation through Health Data Oosterzele BE

² Ghent University Unit of Medical Informatics and Statistics Department of Public Health and Primary care Ghent BE

³ Maastricht University Institute of Data Science Department of Advanced Computing Sciences Maastricht NL

⁴ Maastricht University Institute of Data Science Department of Advanced Computing Sciences Maastricht NL

⁵ Maastricht University Institute of Data Science Department of Advanced Computing Science Maastricht NL

⁶ B!lboa Tervuren BE

Corresponding Author:

Jens Declerck MSc

The European Institute for Innovation through Health Data
Merebaastraat 10
Oosterzele
BE

Abstract

Background: Integrated health data is foundational for secondary use, research, and policy making. However, data quality issues – such as missing values and inconsistencies – are common due to the heterogeneity of health data sources. Existing frameworks often apply static, one-time assessments, limiting their ability to address quality problems across evolving data pipelines.

Objective: This study evaluates the AIDAVA data quality framework, which introduces dynamic, lifecycle-based validation of health data using knowledge graph technologies and SHACL-based rules. The framework is assessed for its ability to detect and manage data quality issues – specifically, completeness and consistency – during integration.

Methods: Using the MIMIC-III dataset, we simulated real-world data quality challenges by introducing structured noise, including missing values and logical inconsistencies. The data was transformed into Source Knowledge Graphs (SKGs) and integrated into a unified Personal Health Knowledge Graph (PHKG). SHACL validation rules were applied iteratively during the integration process, and data quality was assessed under varying noise levels and integration orders.

Results: The AIDAVA framework effectively detected completeness and consistency issues across all scenarios. Completeness was shown to influence the interpretability of consistency scores, and domain-specific attributes (e.g., diagnoses, procedures) were more sensitive to integration order and data gaps.

Conclusions: AIDAVA supports dynamic, rule-based validation throughout the data lifecycle. By addressing both dimension-specific vulnerabilities and cross-dimensional effects, it lays the groundwork for scalable, high-quality health data integration. Future work should explore deployment in live clinical settings and expand to additional quality dimensions.

(JMIR Preprints 31/03/2025:75275)

DOI: <https://doi.org/10.2196/preprints.75275>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in [JMIR Publications](#), I will be able to make my accepted manuscript PDF available to anyone.

No. Please do not make my accepted manuscript PDF available to anyone.



Original Manuscript

Assessing Data Quality in Heterogeneous Healthcare Integration: The AIDAVA Framework

Abstract

Background: Integrated health data is foundational for secondary use, research, and policy making. However, data quality issues – such as missing values and inconsistencies – are common due to the heterogeneity of health data sources. Existing frameworks often apply static, one-time assessments, limiting their ability to address quality problems across evolving data pipelines.

Objective: This study evaluates the AIDAVA data quality framework, which introduces dynamic, lifecycle-based validation of health data using knowledge graph technologies and SHACL-based rules. The framework is assessed for its ability to detect and manage data quality issues – specifically, completeness and consistency – during integration.

Methods: Using the MIMIC-III dataset, we simulated real-world data quality challenges by introducing structured noise, including missing values and logical inconsistencies. The data was transformed into Source Knowledge Graphs (SKGs) and integrated into a unified Personal Health Knowledge Graph (PHKG). SHACL validation rules were applied iteratively during the integration process, and data quality was assessed under varying noise levels and integration orders.

Results: The AIDAVA framework effectively detected completeness and consistency issues across all scenarios. Completeness was shown to influence the interpretability of consistency scores, and domain-specific attributes (e.g., diagnoses, procedures) were more sensitive to integration order and data gaps.

Conclusion: AIDAVA supports dynamic, rule-based validation throughout the data lifecycle. By addressing both dimension-specific vulnerabilities and cross-dimensional effects, it lays the groundwork for scalable, high-quality health data integration. Future work should explore deployment in live clinical settings and expand to additional quality dimensions.

Introduction

The integration of high-quality, complete, and interoperable patient health records is essential to modern healthcare and medical research [1-3]. Accurate and well-structured data enhances

research reproducibility [4-6], which in turn drives more effective clinical decision-making and improved patient outcomes. However, as health data is collected across diverse and heterogeneous sources [7, 8], its quality can be compromised by fragmentation [7, 9, 10], variability [9, 11, 12], and incomplete information [13-15]. These challenges compromise data usability and hinder the development of unified, clinically meaningful datasets suitable for both primary and secondary uses [14, 16, 17].

Existing efforts on health data quality often focus on defining standardized quality dimensions and organizing these into structured frameworks [18-22]. However, many current approaches rely on static, one-time evaluations that do not reflect the dynamic and iterative nature of the entire data lifecycle [23, 24]. This limits their effectiveness in identifying evolving quality issues that emerge across stages of data transformation - such as extraction [25, 26], harmonization [27, 28], or final validation [20] – as they do not provide continuous, iterative assessment.

To address this gap, the AIDAVA project [29], launched in 2022 as part of a Horizon Europe initiative, proposes a dynamic data quality framework that enables continuous assessment throughout the data lifecycle. At its core is the Personal Health Knowledge Graph (PHKG), a patient-centered, interoperable data model built using Knowledge Graph technologies and validated with SHACL-based rules. This approach allows for the assessment of data quality constraints across multiple integration stages.

This paper evaluates the AIDAVA framework's effectiveness in detecting and improving data quality issues, with a particular focus on completeness and consistency. It also investigates how these dimensions evolve across the data transformation pipeline, highlighting their interdependencies and implications for integrated health data.

Related work

Ensuring the quality of integrated healthcare data, particularly for the secondary use of electronic health records (EHRs), has been the subject of extensive research [30]. Prior efforts generally focus on defining and measuring data quality along dimensions such as completeness, consistency, conformance, and plausibility [19-23, 31, 32]. These dimensions serve as the foundation for several

frameworks and tools developed to detect anomalies and enforce standards across clinical datasets [14, 20, 21, 31].

Framework-based approaches have sought to structure and formalize the assessment of health data quality. Notably, Kahn et al. [20] introduced a widely adopted framework that groups quality concerns into three core dimensions – completeness, conformance, and plausibility – further divided into verification and validation contexts. This framework laid the groundwork for systematic quality checks but remains limited to static evaluations that do not adapt to the changing nature of data pipelines.

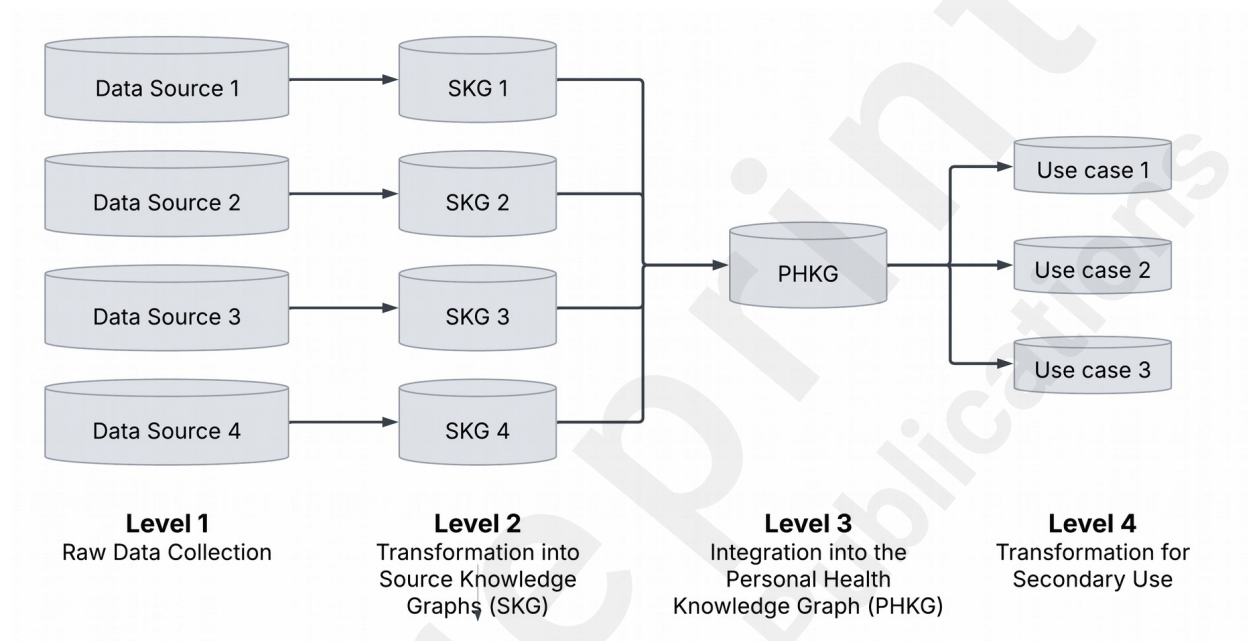
Dimension-specific studies have also provided deeper insights. Issa et al. [33] conducted a comprehensive review of completeness in knowledge graphs, identifying seven distinct subtypes, including three previously unclassified forms. Their findings emphasized the interplay between completeness and other quality dimensions such as consistency and correctness – further highlighting the need for more context-aware, dynamic quality assessments.

Tool-based solutions, such as ACHILLES Heel [34], offer practical mechanisms for identifying data quality issues. As part of the OHDSI ecosystem, ACHILLES Heel applies 70 pre-defined validation rules to detect anomalies in large-scale clinical datasets. While effective in identifying static inconsistencies, these tools typically evaluate data quality at a single point in time and cannot accommodate the iterative transformations that occur during integration workflows.

In contrast to these prior approaches, the AIDAVA framework introduces a lifecycle-based, dynamic model of data quality validation. By embedding SHACL-based validation rules directly into a knowledge graph pipeline, AIDAVA allows for continuous monitoring and enforcement of constraints during data ingestion, transformation, and integration. It extends existing methodologies by addressing data quality not as a one-time task, but as an ongoing process – ensuring that completeness and consistency are maintained across evolving data structures.

The AIDAVA Data Quality Framework

The AIDAVA data quality framework is designed to ensure semantic and structural standardisation across the entire health data lifecycle, addressing the complexities of integrating heterogeneous health data sources. The framework operates across four levels, each targeting specific points in the health data integration process. These four levels of the framework are illustrated in Figure 1.



Level 1: Raw data collection

At the initial level, the framework relies on data providers to provide data that meets baseline standards. Given the variability of data sources, direct transformations or validations at this stage are limited. Instead, the framework ensures compliance with transfer specifications, verifying that incoming data adheres to structural and format requirements before progressing to the next level.

Level 2: Transformation into Source Knowledge Graphs (SKG)

In this stage, raw data is transformed into Source Knowledge Graphs (SKG) by executing a data curation tool from the library of tools available in AIDAVA and delivering a data sources in a knowledge graph format aligned with the AIDAVA Reference Ontology. This process ensures semantic and structural standardisation of each data source, which is critical for interoperability. Standardizing the data at this level minimizes inconsistencies and enables proper integration with

multiple sources.

Level 3: Integration into the Personal Health Knowledge Graph (PHKG)

In this phase, multiple SKGs are integrated into a unified Personal Health Knowledge Graph (PHKG), creating a longitudinal representation of an individual's health history. This step consolidates fragmented patient data sources while maintaining consistency and logical coherence in the integrated patient record. The integration process ensures that patient data is well-structured, complete, and free of inconsistent and redundant information to support clinical decision-making and research applications.

Level 4: Transformation for secondary use

In the final stage, the PHKG is transformed into formats tailored for secondary use. This step adapts integrated data to align with specific output formats enabling accurate analysis (e.g., Breast Cancer registry in OMOP compliant format), improved patient care (e.g., patient IPS in FHIR format), and effective reuse across clinical and research settings.

This study focuses on Levels 2 and 3, as they are the earliest feasible and scalable stages for enforcing semantic standardisation and integrating fragmented, heterogeneous data sources. Performing this at Level 1 is not practical due to diverse formats and limited control over data provider systems. Level 4 is only partially addressed, as downstream transformation is performed, but full data quality assessment from the perspective of a specific use case lies outside the scope of this paper.

Data quality dimensions and categories

To evaluate data quality, the AIDAVA framework adopts a structured approach based on dimensions and categories. Dimensions provide a high-level perspective on data quality, grouping related categories to help identify systemic issues. Recent research underscores the increasing need to evaluate diverse data quality dimensions, especially when repurposing health data for secondary use [18]. While multiple dimensions contribute to ensuring high-quality health data, the AIDAVA framework currently focuses on two data quality dimensions: completeness and consistency.

Completeness ensures that all necessary data elements are present, preventing critical information gaps. Consistency verifies that data adheres to defined constraints and logical relationships, such as ensuring that diagnoses and procedures align with a patient's demographic details. Within these dimensions, categories provide a more granular level of assessment, as outlined in Table 1.

Dimension	Category	Description
Completeness	Essential variable completeness	Identifies when critical data elements are missing
	Conditional variable completeness	Detects cases where missing variables prevent consistency checks.
Consistency	Data type for property	Ensures values conform to expected data types.
	Time sequence consistency	Identifies implausible event sequences.
	Diagnosis for gender consistency	Detects diagnoses incompatible with a patient's gender.
	Diagnosis for age consistency	Flags diagnoses that are incompatible with a patient's age.
	Procedure for gender consistency	Detects gender-incompatible procedures.
	Procedure for age consistency	Flags procedures that are inappropriate for the patient's age.

Table 1: Overview of data quality dimensions and categories in the AIDAVA framework

The decision to focus on completeness and consistency is driven by their widespread use in data quality research (REF) and their suitability for automated assessment (REF). Unlike other dimensions, which often require subjective interpretation or manual validation, completeness and consistency can be systematically measured using predefined rules and automated validation techniques. By prioritizing completeness and consistency, the AIDAVA framework establishes a scalable approach to data quality assessment. These dimensions not only support automated validation but also provide a foundation for expanding the framework to address more complex data quality dimensions.

Data quality instruments

The AIDAVA framework relies on a robust instrument for scalable and automated data quality assessment: the Shapes Constraint Language (SHACL). As a W3C standard, SHACL defines and enforces semantic and structural constraints on RDF knowledge graphs, enabling rule-based validation to detect inconsistencies, missing elements, and structural misalignments.

Within the AIDAVA framework, SHACL validation rules are categorized into two types, as outlined in Table 2: ontology-based and domain-specific rules. Each category includes several validation rules, technically known as SHACL shapes – a SHACL term referring to rule templates that define how data should conform to expected structures or values. The number of rules differs between categories

due to the nature of the information being encoded. For instance, completeness checks such as ‘Essential variable’ rely on the fully connected structure of the *Admission node* to validate the presence of key attributes (e.g., admission date, discharge status) across sources. In contrast, rules like ‘Diagnosis for gender’ require a distinct SHACL shape for each valid gender related pairing (e.g., “prostate cancer – male”, “ovarian cancer – female”), increasing the number of shapes. Table 2 shows the number of SHACL shapes associated with each rule category, reflecting the complexity and granularity of the validations.

To enable meaningful evaluation of data quality across different rule categories, we use normalized data quality scores rather than raw violation counts. This is necessary because the number of rules (SHACL shapes) and corresponding checked nodes varies significantly by category (see Table 2). Violation counts alone would disproportionately reflect rule volume rather than actual quality trends. We calculate a category-specific quality score using the formula:

$$\text{Quality score} = 1 - e$$

Where e represents the error rate, calculated as the violation count divided by the total number of checked nodes. This provides a relative measure of how well the data conforms to the specified rules within each category.

Type	Category	SHACL Shape Count
Ontology based checks	Essential variable completeness	14
	Conditional variable completeness	1
	Data type for property	6
Medical and common-sense checks	Time sequence consistency	6
	Diagnosis for gender consistency	5208
	Diagnosis for age consistency	130
	Procedure for gender consistency	640
	Procedure for age consistency	79

Table 2: Categories and counts of SHACL validation rules in the AIDAVA framework

Ontology-based rules in the AIDAVA framework are derived from the AIDAVA reference ontology [35], which builds on established standards like HL7 FHIR, SNOMED, and LOINC to ensure interoperability and support automated curation. These rules ensure conformance with predefined semantic standards, such as verifying data types, relationships, and the presence of mandatory variables. Domain-specific rules are informed by healthcare-specific knowledge. These rules address

real-world data quality challenges, including validating gender-appropriate procedures, and ensuring consistency in clinical attributes. The definition and validation of domain-specific rules require expert consensus. However, data quality research has established a variety of validated domain-specific rules over time [36]. The AIDAVA framework leverages these existing, literature-validated rules, allowing it to build upon proven methodologies while avoiding redundancy, ensuring alignment with best practices in health data quality management.

These SHACL shapes are applied within SKGs (intra source consistency and completeness) and PHKGs (across data source consistency and completeness), ensuring semantic and structural standardisation at each stage. Figure 2 illustrates this validation process, applied at both the SHKG and PHKG levels.

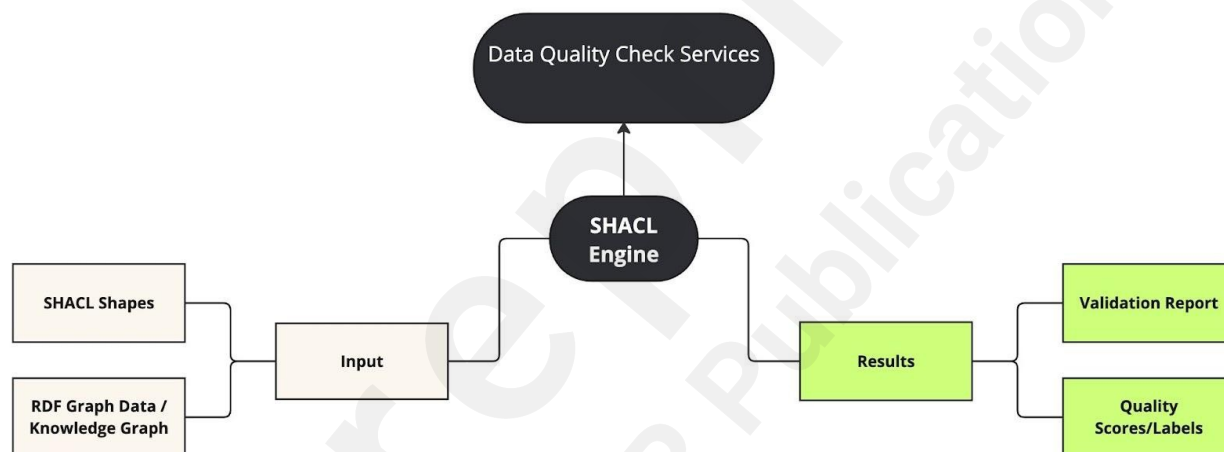


Figure 2: SHACL-based data quality check pipeline

The SHACL engine processes health data represented as an RDF graph – a structured format used to build Knowledge Graphs (KGs) – by applying predefined constraints (SHACL shapes). These rules define the expected structure and content of the data, including semantic relationships and required elements. As the engine executes these rules, it produces a SHACL validation report that highlights data quality violations, such as missing values, incorrect data types, or inconsistent relationships. This validation step is critical for identifying and addressing quality issues prior to integration into the PHKG.



Method

To evaluate the robustness and effectiveness of the AIDAVA framework in detecting and managing data quality issues, we conducted a controlled experiment using the MIMIC-III dataset, a publicly available and well-structured critical care database. While MIMIC-III offers a rich and diverse set of clinical variables, it does not contain the types of data quality issues typically encountered in real-world health information systems. As such, this study introduced artificial noise to simulate common completeness and consistency problems found in heterogeneous clinical data. This approach allows us to systematically assess how well the AIDAVA framework, particularly its SHACL-based validation rules, can identify quality issues under different conditions and at multiple stages of data integration. The following subsections describe the dataset, noise injection methodology, and the process of knowledge graph construction and validation.

Data source and data preparation

We used the MIMIC-III dataset due to its structured format and breadth of clinical variables, offering an optimal foundation for simulating real-world integration scenarios. The dataset includes de-identified health records from over 58,000 admissions. For this study, hospital admissions lacking relevant diagnosis or procedure codes (as required for SHACL validation) were excluded, resulting in a final cohort of 13,607 admissions. Four tables were used: PATIENTS, ADMISSIONS, DIAGNOSES_ICD and CPTEVENTS.

Data cleaning was performed using OpenRefine [37], with date fields standardized to ISO 8601 format (YYYY-MM-DDThh:mm:ss+zz:zz) and gender values mapped to SNOMED CT codes to support interoperability and non-binary. Classifications. These steps ensured alignment with the AIDAVA reference ontology and interoperability across sources.

Adding noise to the dataset

To simulate real-world data quality challenges [14, 38-40], artificial noise was introduced into the dataset based on two parameters: noise level and completeness ratio. Noise Level ($\in [0, 1]$) defines the proportion of knowledge graph statements impacted by errors. The completeness ratio specifies the share of this noise that results in missing values (as opposed to logically inconsistent entries). For example, a noise level of 0.50 and completeness ratio of 0.25 implies that 50% of selected statements are altered, with 25% of them made incomplete and the remainder made inconsistent.

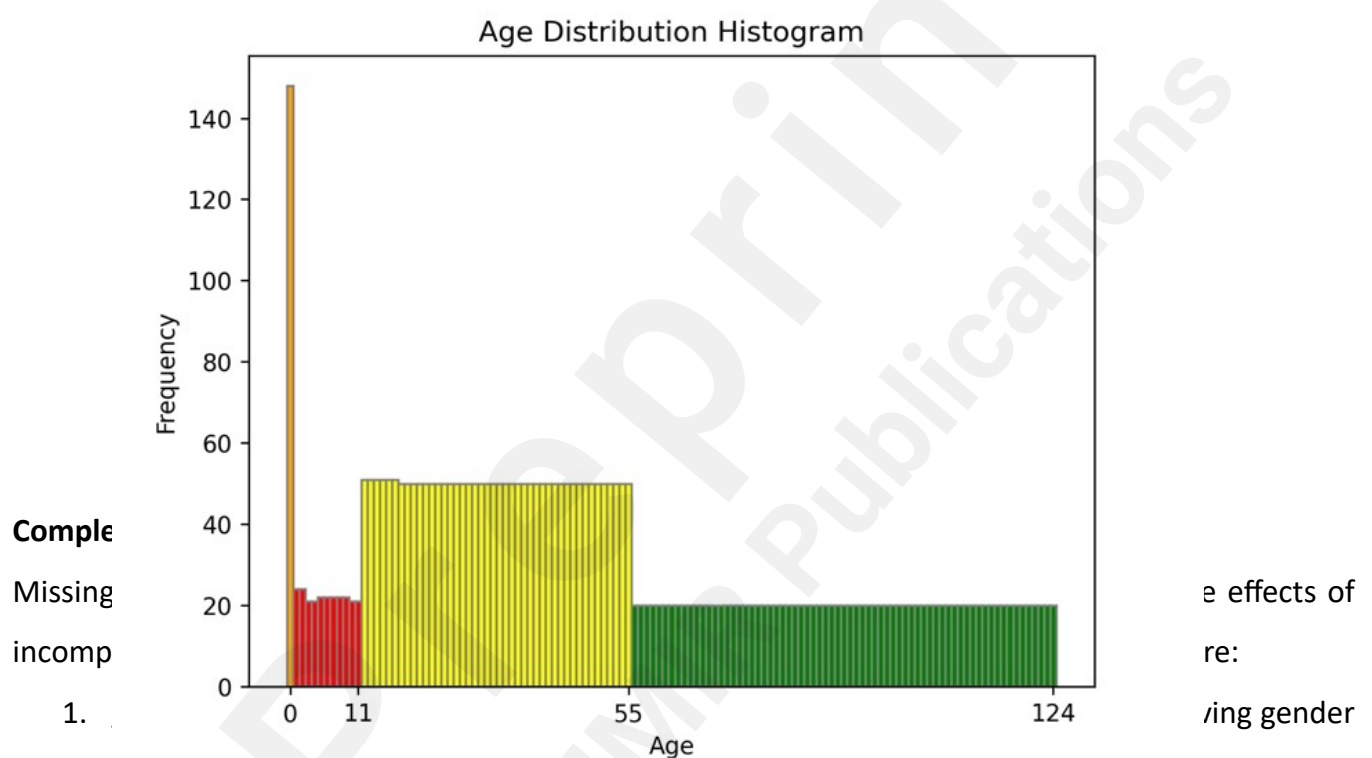
Noise was introduced across two categories: consistency noise reflecting logical contradictions and completeness noise simulating missing information. All injections were performed in a reproducible manner using a fixed randomization seed.

Consistency Noise

We targeted error types that frequently occur in clinical data entry or integration processes [14, 39-41]. These methods included:

1. Gender swapping: Male and female values in the PATIENTS table were randomly exchanged. This modification introduced inconsistencies in gender-specific diagnoses and procedures, affecting validation rules such as diagnosis for gender consistency and procedure for gender consistency. Gender values are sampled from a binomial distribution with $n=2$ trials and success probability, $p=0.5$, ensuring a balanced distribution of changes.
2. Age alteration: Instead of randomly changing a patient's age, we introduced interval-based errors by defining plausible age groups. These age groups were determined based on clustering patterns observed in diagnosis for age consistency and procedure for age consistency rules (Figure 3). In this implementation, age groups were defined as [0,1] for infants, [1,12] for children, [12,56] for teens/adults, and [56,124] for seniors. Birth years in the PATIENTS table were altered so that a new age was randomly selected from another age group, triggering violations in age-dependent diagnosis and procedure rules. To ensure realistic distribution, the selection of a new age group was sampled using a multinomial distribution with $n=4$ trials, where the event probabilities were weighted based on the frequency of each age group in the dataset.
3. Swapping admission and discharge dates: In the ADMISSIONS table, admission and discharge dates were swapped. This modification primarily disrupted time-sequence validation rules and triggered errors in age calculations that depended on the admission date.
4. Day-month swapping in dates: Day and month values in date fields across PATIENTS, ADMISSIONS, and CPTEVENTS were randomly swapped, leading to invalid date formats where month values exceeded 12, or sequences where the chronological order of events became disrupted. These errors mirrored common data entry mistakes in hospital settings, where clinicians or administrative staff may mistakenly invert date components.

5. Inconsistencies in medical coding: We randomly changed diagnosis and procedure codes in the DIAGNOSES_ICD and CPTEVENTS tables. ICD-9 diagnosis codes and CPT procedure codes were replaced with alternative codes randomly drawn from their respective rulesets. This change indirectly caused invalid gender and age violations in both diagnoses and procedures, since certain codes are only applicable to specific demographic groups. To implement this, each affected record had its original code excluded and replaced with another randomly sampled value from the remaining choices, ensuring a uniform distribution of errors across the dataset.



1. (GENDER) or date of birth (DOB) fields in the PATIENTS table with equal probabilities of being chosen. This noise is intended to represent missing or de-identified patient information.
2. Hospital admission records: Missing values were introduced by deleting admission or discharge timestamps (ADMITTIME, DISCHTIME) as well as admission or discharge locations (ADMISSION_LOCATION, DISCHARGE_LOCATION) from the ADMISSIONS table. As with demographic records, the field to be deleted is chosen with equal percentages of 25%. The selection of omitted records was performed randomly yet consistently across the experiment, preserving the dataset's structural integrity while reflecting real-world gaps in clinical administrative case documentation.

3. Procedure records: Missing values were introduced by removing procedure chart dates (CHARTDATE) or procedural codes (CPT_CD) in the CPTEVENTS table. Same with the others, equal probabilities of 0.5 decided which cell type should be deleted. The omission process was randomized across the experiment, replicating common record-keeping errors that lead to incomplete procedural documentation.
4. Diagnosis records: Missing values were introduced by deleting ICD-9 diagnosis codes (ICD9_CODE) from the DIAGNOSES_ICD table. As there is only one column type able to be deleted, it is chosen with the probability of 1, dissimilar to other categories of introducing completeness noise. This modification simulates errors observed by not typing ICD codes for billing purposes.

While these noise injection scenarios may overlap in their effects, the incompleteness was introduced independently at this stage. The combined effects of the above, as it relates to triage and patient data acquisition progressions, are further elaborated and investigated in the following section. To account for the randomness in choosing the cell types to be deleted, this process was carried out in a reproducible manner using a fixed randomisation seed.

Data integration and data quality assessment

After noise injection, the dataset was mapped into SKGs using RDFCraft [42], aligning data elements to the AIDAVA reference ontology. These SKGs were then integrated into a unified PHKG, following a typical data ingestion workflow. Beginning with demographic information from the PATIENTS table, which provided key patient attributes. Next, temporal data from the ADMISSIONS table was incorporated, establishing admission and discharge events. This was followed by procedural details from the CPTEVENTS table, and finally, diagnostic information from the DIAGNOSES_ICD table. This stepwise integration reflects a typical hospital workflow, where patient registration occurs first, followed by admissions, treatments, and recorded diagnoses. At each step, SHACL validation rules were applied to assess data quality in terms of completeness and consistency. This staged validation aligns with levels 2 and 3 of the AIDAVA framework and reflects both intra-source and cross-source quality checks.

To evaluate the effect of the order of data addition on data quality, alternative sequences were also tested by changing the ingestion order (e.g., loading procedures or diagnoses before

demographics). This allows us to observe how quality issues propagate or get masked depending on integration order.

Results

The integration process followed a sequential order, beginning with the PATIENTS table, followed by ADMISSIONS, CPTEVENTS, and DIAGNOSES_ICD tables. The framework assessed completeness and consistency at each stage, tracking how data quality changed throughout the process. To evaluate the robustness of the framework, we also conducted an alternative integration sequence, starting with CPTEVENTS, followed by DIAGNOSES_ICD, ADMISSIONS, and finally PATIENTS.

In the following section, we first present the baseline data quality assessment without noise, establishing a reference for comparison. We then provide the final data quality scores after full integration across varying noise levels for both integration orders, highlighting key trends. A detailed breakdown of progressive changes at specific noise levels is available in the appendix.

Data quality assessment without noise

The baseline analysis, presented in Table 3, was conducted without artificial noise to establish a reference for data quality. Results showed that the majority of dimensions and categories achieved nearly perfect quality scores, indicating that the integration process preserved data integrity and did not amplify errors. 'Essential variable' completeness remained at 100%, confirming that all mandatory data elements were present in the knowledge graph. The domain consistency checks for diagnosis and gender as well as diagnosis and age yielded perfect quality scores. Similarly, data type adherence at all integration stages is achieved with no errors.

Dimension	Category	PATIENTS	ADMISSIONS	CPTEVENTS	DIAGNOSES_ICD
Completeness	Essential variable	100%	100%	100%	100%
	Conditional variable	NA	NA	NA	100%
Consistency	Time sequence	100%	99.90%	99.90%	99.90%
	Diagnosis for gender	100%	100%	100%	100%
	Procedure for gender	100%	100%	100%	100%
	Diagnosis for age	100%	100%	100%	100%
	Procedure for age	100%	100%	99.89%	99.89%

	Data type for property	100%	100%	100%	100%
--	------------------------	------	------	------	------

Table 3: Baseline data quality analysis results

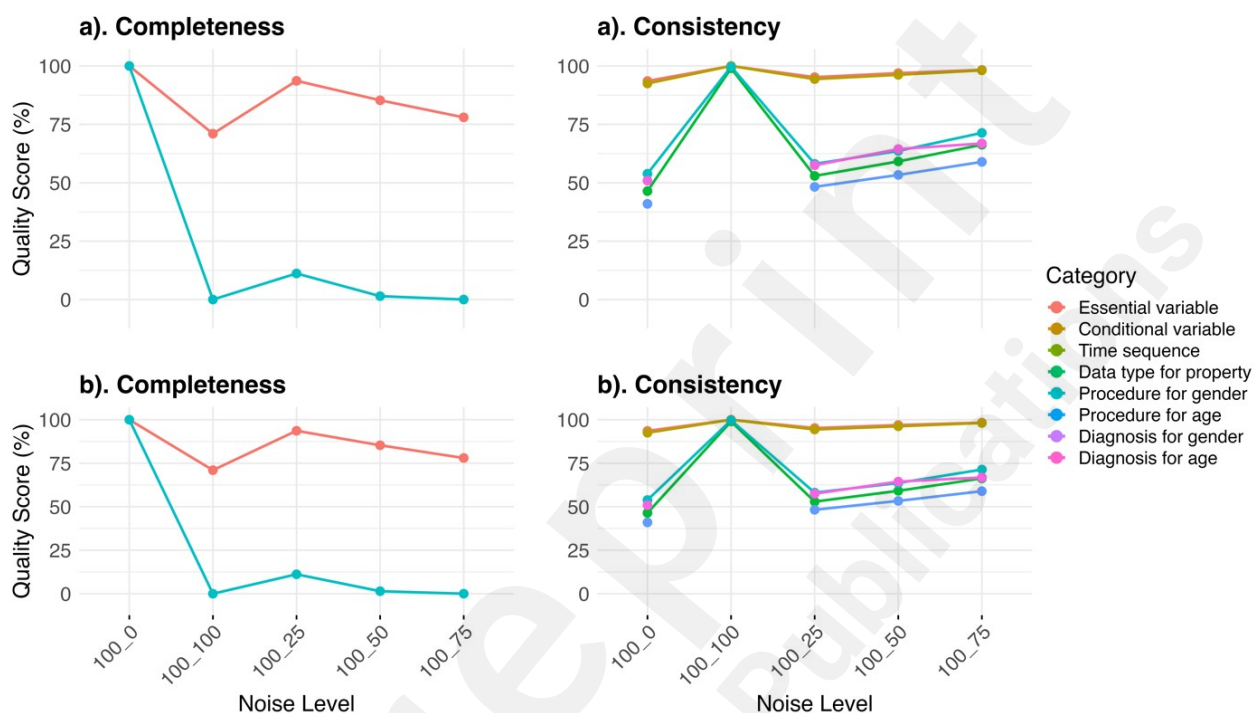
Although overall data quality is high, minor discrepancies were observed in specific categories. 'Conditional variable' completeness category assesses whether all necessary concepts from different sources are properly integrated and complete. Since SHACL rules evaluate conditional variables across every table, an incomplete graph would return 0% quality score until the last stage of integration. To reflect this dependency, the 'Conditional variable' completeness scores for PATIENTS, ADMISSIONS, AND CPTEVENTS were marked as "NA" since these values would not be meaningful to assess. Only at the DIAGNOSES_ICD stage could this category be properly evaluated where all variable nodes will be reachable.

The procedure for gender and age categories received near-perfect scores of 99.99% and 99.89%, respectively. These errors were not due to integration issues, but rather pre-existing errors in the MIMIC-III dataset. The errors in the 'procedure for age' category originated from the dataset's de-identification process, which assigned a few birth dates to years in the 1800s, resulting in implausible patient ages exceeding 300 years - a violation of age-related consistency rules. Similarly, errors in the 'procedure for gender' category were traced to a data entry mistake in MIMIC-III, where a Laparoscopic procedure on the Oviduct/Ovary (CPT 58660) was attributed to a male patient. Additionally, time sequence consistency scored 99.90%, with errors primarily linked to discharge times recorded earlier than admission times. This discrepancy is a known artifact of the MIMIC-III data collection process, likely caused by inconsistent timestamp recording practices [42].

Data quality assessment with noise

Overall trends across all noise levels

To simulate real-world data quality challenges, we introduced artificial noise and repeated the integration process under two different orders. Figure 5 presents the final data quality scores after full integration for both orders. The results confirm that data quality scores remained consistent regardless of integration sequence, indicating that the order of data source integration does not alter overall data quality scores. While intermediate values may vary slightly, the progression of completeness and consistency follows the same overall patterns.

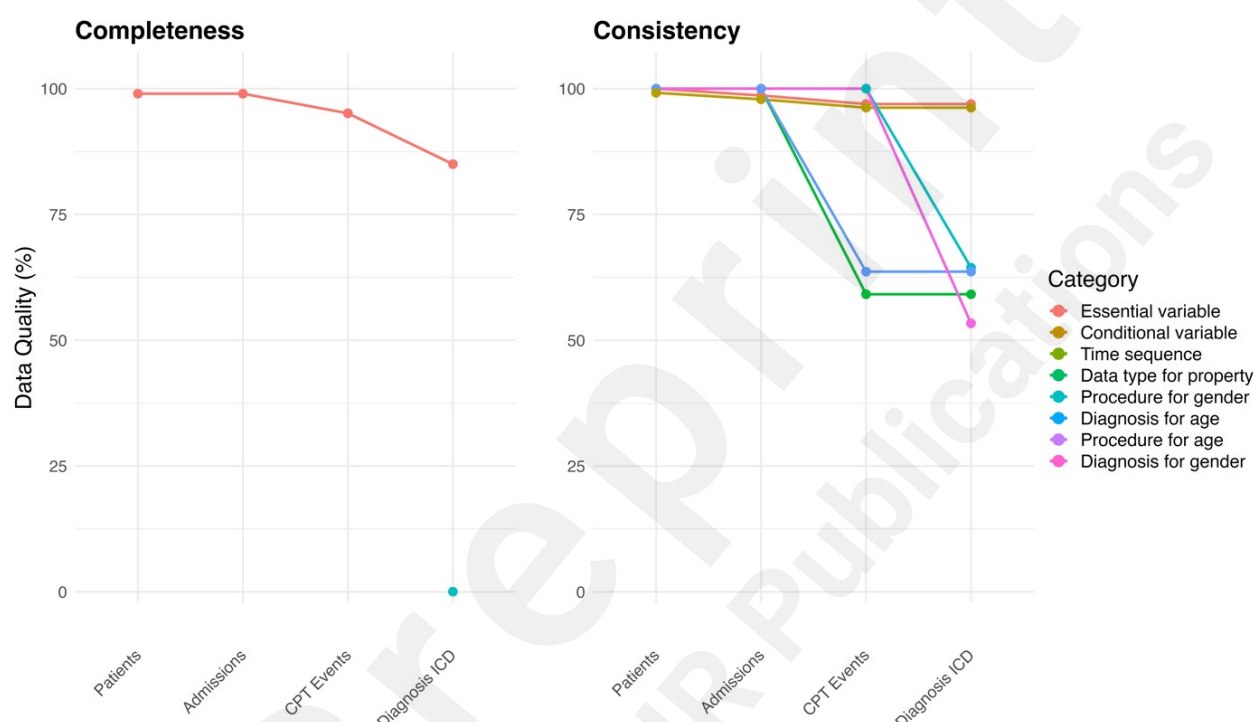


‘Essential variable’ completeness declined gradually from 100% to 71%, with the largest drop occurring between 100_0 (100%) and 100_25 (93.65%), continuing to decrease at higher noise levels. ‘Conditional variable’ completeness dropped sharply from 100% to 11.17% at 100_25, reaching near zero at higher noise levels, reflecting its strong dependence on cross-source relationships.

Time sequence and data type consistency improved throughout the integration, rising from 93.57% and 92.51% at 100_0 to 100% at 100_100. Procedure consistency for gender and age increased, starting at 46.43% and 53.86%, respectively, and reaching 99.01% and 99.68% at 100_100. Diagnosis consistency for gender and age followed a different pattern, increasing steadily from 40.98% and 50.91% at 100_0 to 58.94% and 66.82% at 100_75, but both became unmeasurable at 100_100.

Detailed analysis at 100% consistency and 50% completeness noise (100_50)

To gain deeper insights into how data quality changes during data integration under specific noise levels, this section provides a detailed analysis of the 100_50 noise level. Figure 5 presents the data quality scores at each step of the integration process, illustrating how completeness and consistency evolve as new data sources are incorporated.



At this noise level, where consistency noise is set to 100% and completeness noise to 50%, all relevant knowledge graph statements (100%) are modified, with 50% of them containing missing data. The introduction of ADMISSIONS caused time sequence consistency to decrease slightly to 98.64%. With CPTEVENTS added, this metric dropped to 96.94%, where it remained stable after DIAGNOSES_ICD. Procedure consistency for gender declined to 59.14%, while procedure consistency for age dropped to 63.63%.

As expected, missing data impacted completeness measures. 'Essential variable' completeness declined sharply, reaching 85% after DIAGNOSES_ICD. 'Conditional variable' completeness remained "NA" for PATIENTS, ADMISSIONS, and CPTEVENTS. Once DIAGNOSES_ICD was incorporated,

'Conditional variable' completeness increased slightly to 1.45%, reflecting a minimal recovery of required data.

An analysis of data quality trends across other noise levels is provided in the appendix, detailing variations in completeness and consistency scores under different conditions.



Discussion

Our findings demonstrate consistent patterns in data quality dynamics across all noise levels, characterized by a progressive degradation in consistency scores and a steady decline in completeness as noise levels increase. The category of 'Essential variable' completeness deteriorates gradually as noise accumulates, while 'Conditional variable' completeness declines more sharply due to its dependence on relationships across data sources. Meanwhile, 'Time sequence' and 'Data type consistency' remain relatively stable, whereas procedure and diagnosis consistency degrade significantly based on integration order, underscoring the importance of cross-source attribute alignment.

These trends confirm the adaptability of the AIDAVA data quality framework, which maintains stability in core structural checks (e.g., time and data type validation) even as data volume and disorder increase. However, the results also revealed that procedure and diagnosis consistency were more sensitive to integration order, emphasizing the need for targeted, domain-specific validation strategies during data merging.

To more explicitly frame these insights, the AIDAVA framework not only detects data quality issues with high granularity but also enables their continuous assessment throughout the integration pipeline. By embedding SHACL-based rules directly into the SKG and PHKG construction process, AIDAVA captures completeness and consistency violations both within and across data sources. This staged, rule-based validation empowers the framework to not only detect anomalies but also manage them by pinpointing their origin and timing – enabling early intervention and downstream reliability.

Completeness and conditional dependencies in data integration

When examining the completeness dimension, the results reveal distinct patterns of degradation across our two categories: 'Essential variable' and 'Conditional variable' completeness. As expected, both categories are increasingly affected as the completeness ratio level rises. However, their rates of decline differ, highlighting important nuances in their behaviour during the integration process. 'Essential variable' completeness steadily declines, suggesting that core patient attributes are progressively impacted by noise – particularly those embedded in the PATIENTS table. In contrast,

'Conditional variable' completeness cannot be meaningfully assessed until the final integration step, as it depends on the availability of interconnected variables across all datasets. This explains why intermediate scores are marked as "NA" and only become valuable after the final source (DIAGNOSES_ICD) is added.

These findings emphasize that 'Conditional variable' completeness is inherently linked to the integration process itself rather than noise alone. Unlike 'Essential variable' completeness, which is primarily affected by missing value within a single dataset, 'Conditional variable' completeness is more directly influenced by the presence or absence of cross-source relationships. As a result, its behavior differs from other completeness measures, demonstrating that missing values alone do not dictate 'Conditional variable' completeness trends - rather, it is the sequence and completeness of integrated sources that determine its final value.

Consistency challenges in stepwise data integration

The consistency results reveal that step-by-step integration of data sources introduces challenges related to compliance with data type and time sequence consistency. However, these declines are relatively minor, suggesting that the overall format of the ontology remains stable and that temporal relationships are preserved throughout the integration process. Notable, time sequence consistency exhibits only minor fluctuations, reinforcing its resilience to integration steps. In contrast, the categories related to procedure consistency and diagnosis consistency show more pronounced declines, highlighting the critical role of integration order in determining data quality outcomes. The results indicate that integrating the CPTEVENTS table first has a substantial impact on procedure consistency, causing notable drops in gender and age consistency. Similarly, the integration of DIAGNOSES_ICD exacerbates diagnosis-related inconsistencies, suggesting that diagnostic data is particularly vulnerable to integration-induced errors. This trend highlights that procedure and diagnosis consistency are more reliant on cross-source relationships and attribute alignment than other categories, such as data type and time sequence consistency.

The findings suggest that a case-specific data quality strategy is essential when integrating different types of health data. This underscores the need for tailored validation approaches that consider the vulnerabilities of various clinical data types during integration. This need for targeted strategies is

evident both within and across data quality dimensions. For instance, within consistency, procedure and diagnosis categories exhibit greater instability, highlighting that medical domain-specific content is more sensitive to inconsistencies than general data attributes like timestamps or data types.

Interdependencies between data quality dimensions

Beyond individual dimensions, these findings offer an in-depth perspective on the interrelationships between data quality dimensions, reinforcing the complex and dynamic nature of data quality in health data integration. While previous research has suggested that data quality dimensions are interrelated (REF), our results offer a unique demonstration of how these interdependencies manifest in real-world integration scenarios. A key finding from the results is that consistency measures alone cannot be fully trusted without ensuring adequate completeness. The results reveal that missing data directly impacts the interpretability of consistency scores, making it difficult to determine whether the observed consistency is genuine or merely an artifact of incomplete data. When key information is missing, certain contradictions - such as conflicts between diagnoses and procedures - may go undetected, creating a false impression of data reliability.

These insights highlight the critical need for health data integration strategies that not only address individual data quality dimensions but also account for their interdependencies. Overlooking these relationships can lead to misleading assessments, where high consistency scores mask underlying data gaps, or missing values distort the true extent of inconsistencies. To ensure reliable data quality assessment, integration of data must prioritise completeness validation before consistency assessment, ensuring that inconsistencies are accurately detected rather than artificially hidden. By adopting an approach that considers both dimension-specific vulnerabilities and their cross-dimensional effects, we can enhance the integrity of downstream analyses, clinical decision-making, and secondary research applications - ultimately improving the reliability of integrated health data.

Limitations

Several limitations must be acknowledged to contextualize the findings and guide future research directions. First, this study was conducted using the MIMIC-III dataset, a structured and de-identified critical care database. While this dataset provides a controlled environment for testing data integration, it does not fully capture the heterogeneity and complexity of live healthcare data environments. Real-world healthcare systems exhibit variations in data collection methods, coding practices, and patient populations, which may introduce additional challenges that were not observed in this controlled setting. Future research should evaluate the framework on diverse, non-deidentified hospital datasets that better reflect real-world conditions. Second, this study simulated real-world data quality issues by introducing structured artificial noise, allowing for a systematic evaluation of the framework under different levels of data inconsistencies. However, artificial noise does not fully replicate the unpredictability of errors found in operational health data. Health datasets often contain context-dependent inconsistencies, undocumented missingness patterns, and human-introduced biases that cannot be easily simulated. Future research should explore how the framework performs when applied in a real-world scenario. Third, the AIDAVA framework evaluated consistency and completeness as the core data quality dimensions. However, other important dimensions, such as timeliness and uniqueness, were not explicitly assessed in this study. Future work should extend the framework to incorporate a broader range of quality dimensions, ensuring a more complete evaluation of integrated health data. Last, this study demonstrated the effectiveness of the AIDAVA framework in a controlled dataset, its scalability for large-scale, high-velocity health data integration was not examined in depth. Future research should investigate the framework's performance in large-scale deployments.

Addressing these limitations will be essential to ensure the framework's adaptability, scalability, and utility across diverse healthcare applications. Future research should extend the evaluation to include live hospital datasets, explore additional stages of the data lifecycle, and assess the framework's efficiency at scale. By overcoming these challenges, the AIDAVA framework can further contribute to improving data quality and advancing the integration of heterogeneous health datasets.

Conclusion

This study evaluated the AIDAVA data quality framework for its effectiveness in detecting and managing data quality issues – specifically completeness and consistency – during the integration of heterogeneous health data. Using the MIMIC-III dataset, we simulated real-world challenges by introducing structured noise and systematically assessed how the framework performed across different integration sequences and noise levels. The staged SHACL-based validation enabled fine-grained analysis of data quality at multiple points along the integration pipeline. Our findings show that the AIDAVA framework successfully identifies both missing and inconsistent data elements and provides interpretable feedback at each stage of transformation. Together, these findings support the AIDAVA framework's suitability for dynamic, lifecycle-based data quality assessment. By enabling validation at each transformation step, the framework allows early detection, interpretable tracking, and strategic mitigation of data quality issues. Importantly, it encourages a holistic view of data quality – one that considers not only dimension-specific weaknesses but also how dimensions influence each other across the pipeline. As health systems increasingly rely on integrated datasets for clinical and research applications, frameworks like AIDAVA provide essential infrastructure for building trust in secondary health data use.

Future research

Future work should focus on improving the AIDAVA data quality framework by aligning SHACL shapes with the OMOP common data model [42] and comparing multiple integration orders to assess their impact on data quality. Expanding SHACL validation to lab results, prescriptions, and other hospital records will enhance its clinical applicability. Additionally, developing more realistic noise introduction methods will better simulate real-world inconsistencies, strengthening the framework's robustness.

Funding

The author(s) declare financial support was received for the research, authorship, and/or

publication of this article. This work was supported by the European Union's Horizon Europe research and innovation programme under grant agreements No. 101057062 (AIDAVA).

Supported by a licence waiver from SNOMED International from a period of 2 years (2024 – 2025), on condition that SNOMED CT is not used and/or deployed commercially.

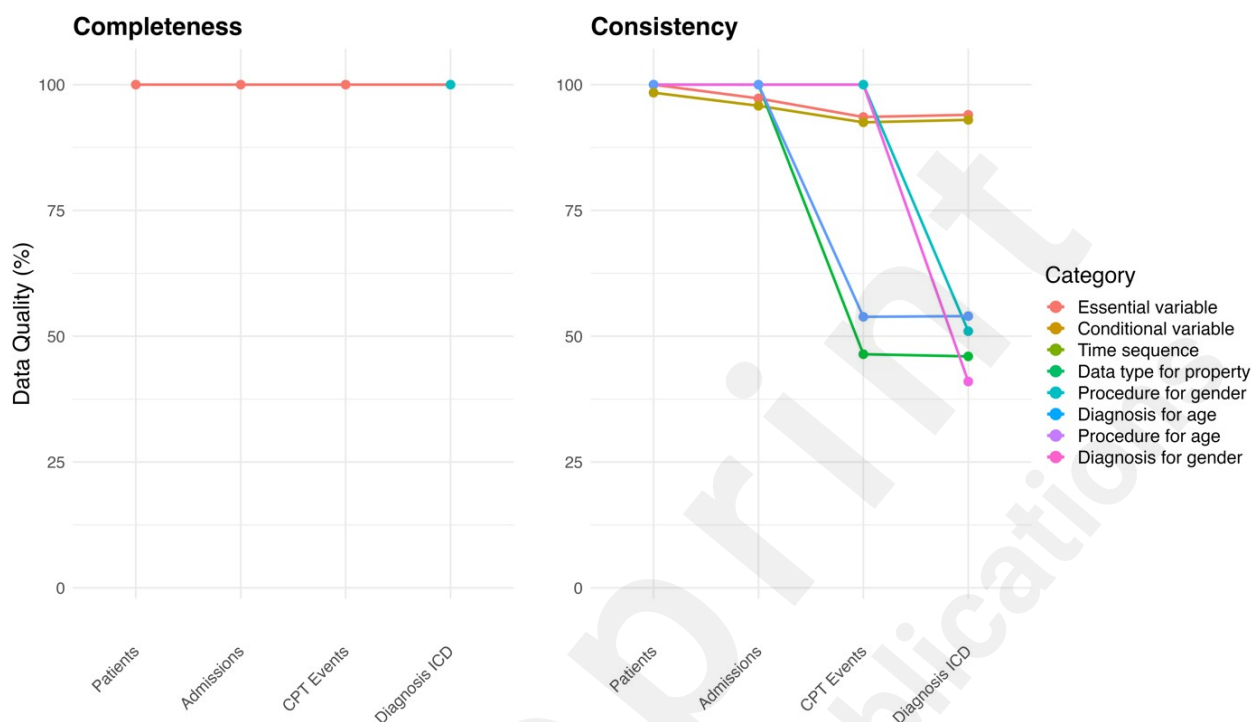
The work of the Swiss Partner (MIDATA) received funding by the Swiss State Secretariat for Education, Research and Innovation (SBFI), subvention contract 22.00093, REF-1131-52104.

Acknowledgments

The authors are grateful to all the members of the AIDAVA consortium who diligently support the project with their expertise. A special thanks goes to the team at the Department of Advanced Computing Sciences, Institute of Data Science at the Maastricht University for their invaluable contributions in setting up the experiment and developing the mechanism for noise addition, which was essential for this study.

APPENDIX

100% consistency and 0% completeness noise



As visualised in Figure 6, introducing 100% consistency and 0% completeness noise resulted in a progressive decline in consistency scores with each integrated table. However, completeness remained unchanged, as no data was missing. After integrating ADMISSIONS, time sequence consistency declined slightly to 97.26%. The addition of CPTEVENTS introduced inconsistencies in 'procedure for gender' and 'procedure for age', reducing their quality scores to 46.43% and 53.86%, respectively. These values remained stable following the integration of DIAGNOSES_ICD. However, this final step further impacted diagnostic consistency, with 'diagnosis for gender' dropping to 41% and 'diagnosis for age' consistency falling to 51%.

Interestingly, data type consistency declined to 92.51% after CPTEVENTS but showed a slight recovery to 93% after DIAGNOSES_ICD. This slight improvement in data type consistency after integrating DIAGNOSES_ICD suggests that some records may have aligned with expected formats at this stage.

100% consistency and 25% completeness noise

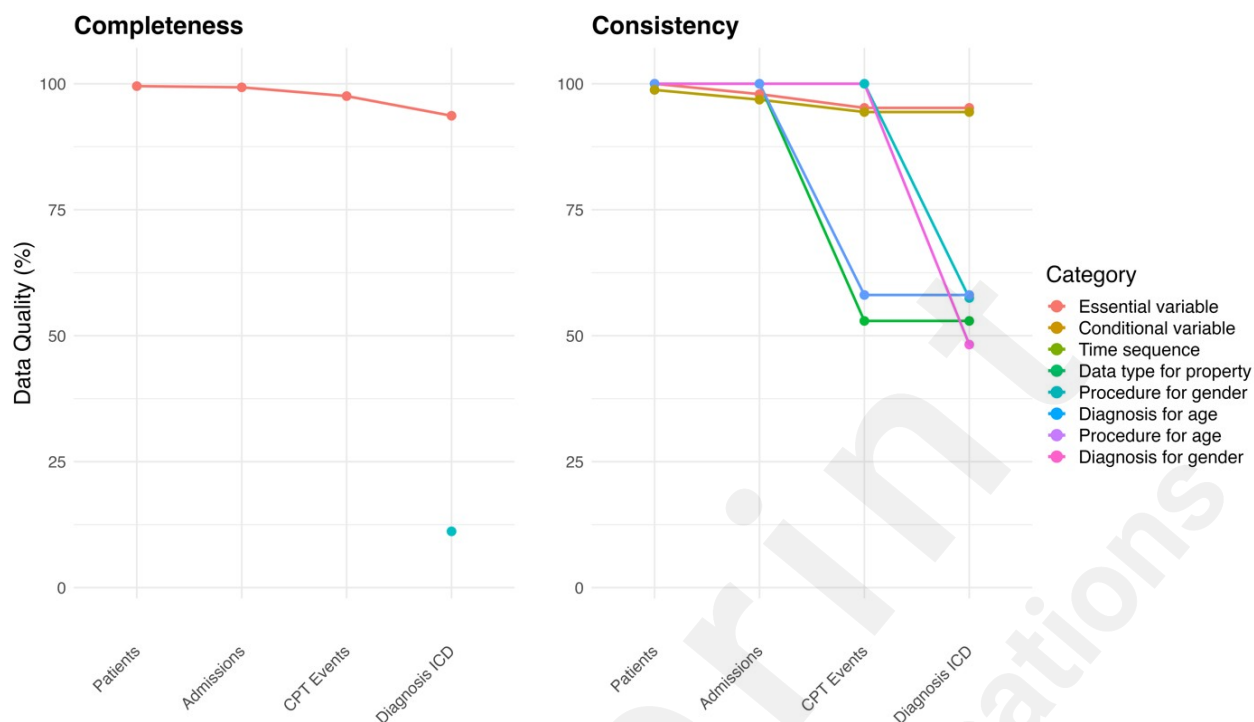


Figure 7: Data quality scores at 100% consistency noise and 25% completeness noise (100_25)

With 25% of the introduced noise consisting of missing, while 75% remained consistency errors, Figure 7 illustrates how the integration process was affected by both inconsistency and missing data. As ADMISSIONS was integrated, time sequence consistency declined slightly to 97.94%. When CPTEVENTS was introduced, procedure consistency for gender decreased to 52.93%, while procedure consistency for age declined to 58.05%. These values remained unchanged after integrating DIAGNOSES_ICD. The integration of DIAGNOSES_ICD further impacted diagnostic consistency, causing diagnosis-gender consistency to drop to 48.24% and diagnosis-age consistency to decline to 57.49%.

The presence of missing data had a visible impact on completeness. 'Essential variable' completeness declined at each step, reaching 93.65% after integrating DIAGNOSES_ICD. 'Conditional variable' completeness was marked as "NA" for PATIENTS, ADMISSIONS, and CPTEVENTS, as this dimension requires full dataset integration. However, once DIAGNOSES_ICD was integrated, 'Conditional variable' completeness reached 11.17%, indicating that some interrelated data elements became available in the final step.

100% consistency and 75% completeness noise

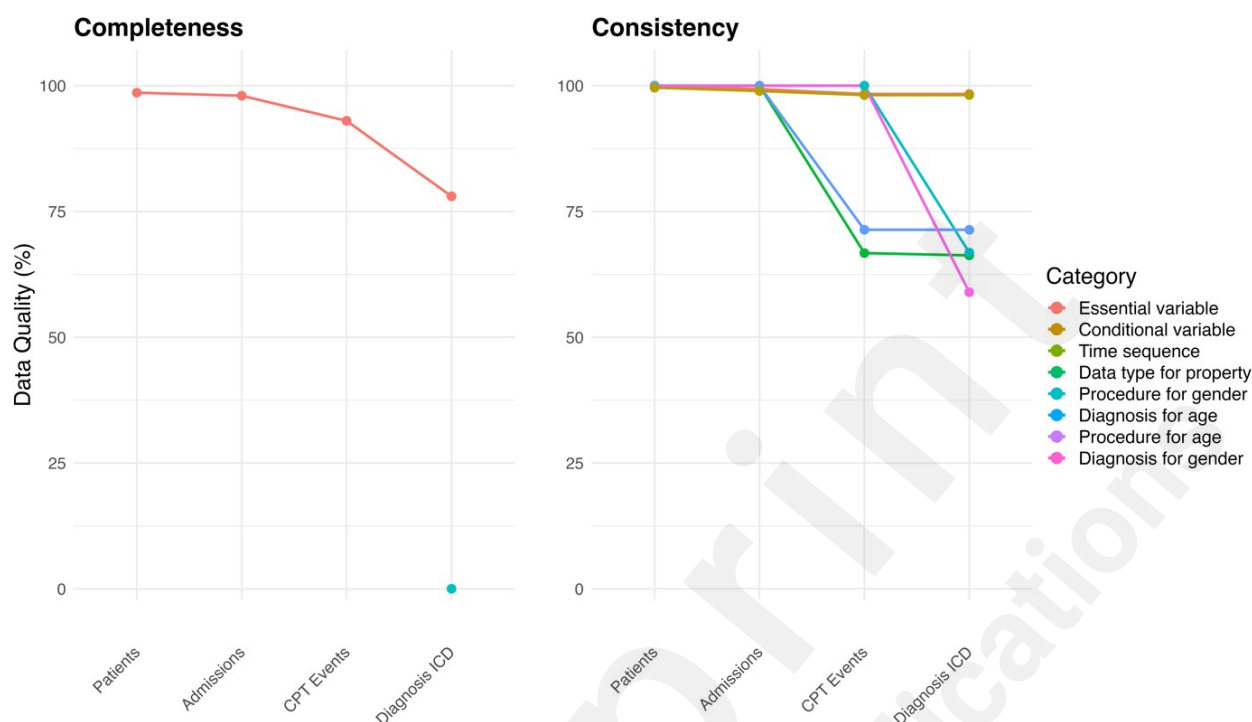


Figure 8: Data quality scores at 100% consistency noise and 75% completeness noise (100_75)

With 75% missing data, completeness degradation became the dominant effect (Figure 8). The integration of ADMISSIONS led to only a minor decline in time sequence consistency (99.32%), and this metric stabilized at 98.34% following CPTEVENTS. However, procedure consistency for gender dropped to 66.72%, and procedure consistency for age declined to 71.36%. 'Essential variable' completeness dropped to just 78% after DIAGNOSES_ICD, demonstrating the severe loss of data availability. After DIAGNOSES_ICD, 'Conditional variable' completeness remained close to 0%, showing that missing data severely impacted the availability of required interrelationships.

100% consistency and 100% completeness noise

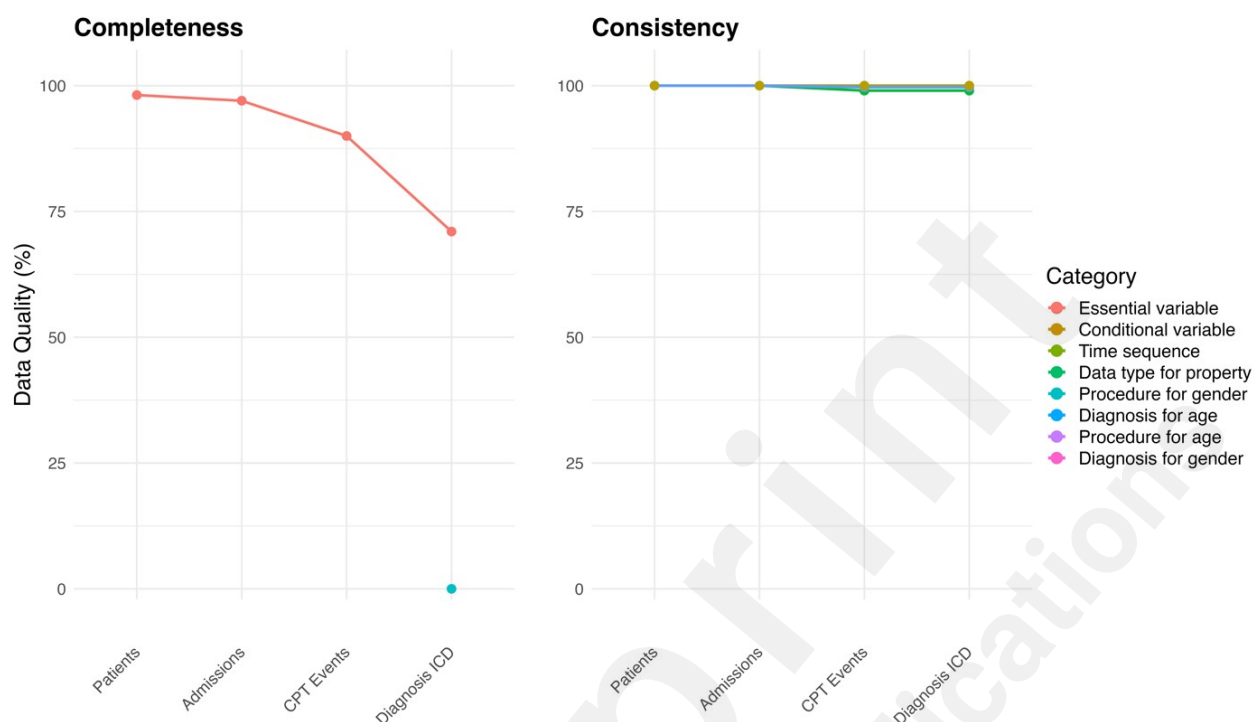


Figure 9: Data quality scores at 100% consistency noise and 100% completeness noise (100_100)

At 100% completeness noise, data loss was so severe that meaningful validation became impossible (Figure 9). Unlike previous noise levels, where inconsistencies could still be tracked, the overwhelming loss of data prevented effective validation. As CPTEVENTS was integrated, procedure consistency for gender and age could no longer be measured, as cross-source relationships had broken down. Time sequence consistency remained at 100% throughout, but this was misleading - the absence of data meant inconsistencies could not be detected rather than being resolved. Similarly, data type consistency remained at 100%, despite validation no longer being possible. The most severe impact was observed in 'Essential variable' completeness, which declined to 71% after DIAGNOSES_ICD - the lowest recorded across all noise levels. 'Conditional variable' completeness was entirely unassessable after integrating DIAGNOSES_ICD, as there was insufficient data to evaluate interdependencies between sources.

1. Beresniak A, Schmidt A, Proeve J, Bolanos E, Patel N, Ammour N, et al. Cost-benefit assessment of using electronic health records data for clinical research versus current practices: Contribution of the Electronic Health Records for Clinical Research (EHR4CR) European Project. *Contemp Clin Trials*. 2016 Jan;46:85-91. PMID: 26600286. doi: 10.1016/j.cct.2015.11.011.
2. De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, et al. Using electronic health records for clinical research: the case of the EHR4CR project. *J Biomed Inform*. 2015 Feb;53:162-73. PMID: 25463966. doi: 10.1016/j.jbi.2014.10.006.
3. Dupont D, Beresniak A, Sundgren M, Schmidt A, Ainsworth J, Coorevits P, et al. Business analysis for a sustainable, multi-stakeholder ecosystem for leveraging the Electronic Health Records for Clinical Research (EHR4CR) platform in Europe. *Int J Med Inform*. 2017 Jan;97:341-52. PMID: 27919393. doi: 10.1016/j.ijmedinf.2016.11.003.
4. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol*. 2019 May;37(5):555-60. PMID: 30858580. doi: 10.1038/s41587-019-0054-x.
5. Alonzo TA. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating: By Ewout W. Steyerberg. *American Journal of Epidemiology*. 2009;170(4):528-. doi: 10.1093/aje/kwp129.
6. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*. 2018 2018/05/08;1(1):18. doi: 10.1038/s41746-018-0029-1.
7. Bernardi FA, Alves D, Crepaldi N, Yamada DB, Lima VC, Rijo R. Data Quality in Health Research: Integrative Literature Review. *J Med Internet Res*. 2023 Oct 31;25:e41446. PMID: 37906223. doi: 10.2196/41446.
8. Shara N, Anderson KM, Falah N, Ahmad MF, Tavazoei D, Hughes JM, et al. Early Identification of Maternal Cardiovascular Risk Through Sourcing and Preparing Electronic Health Record Data: Machine Learning Study. *JMIR Med Inform*. 2022 2022/2/10;10(2):e34932. doi: 10.2196/34932.
9. Wei WQ, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, Chai HS, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J Am Med Inform Assoc*. 2012 Mar-Apr;19(2):219-24. PMID: 22249968. doi: 10.1136/amiajnl-2011-000597.
10. Turbow SD, Ali MK, Culler SD, Rask KJ, Perkins MM, Clevenger CK, et al. Association of Fragmented Readmissions and Electronic Information Sharing With Discharge Destination Among Older Adults. *JAMA Network Open*. 2023;6(5):e2313592-e. doi: 10.1001/jamanetworkopen.2023.13592.
11. Razzaghi H, Greenberg J, Bailey LC. Developing a systematic approach to assessing data quality in secondary use of clinical data based on intended use. *Learning Health Systems*. 2022;6(1). doi: 10.1002/lrh2.10264.
12. Huang Y, Guo J, Chen Z, Xu J, Donahoo WT, Carasquillo O, et al., editors. The impact of electronic health records (EHR) data continuity on prediction model fairness and racial-ethnic disparities. 2023.
13. Li Y, Sperrin M, Martin GP, Ashcroft DM, van Staa TP. Examining the impact of data quality and completeness of electronic health records on predictions of patients' risks of cardiovascular disease. *International Journal of Medical Informatics*. 2020 2020/01/01;133:104033. doi: <https://doi.org/10.1016/j.ijmedinf.2019.104033>.
14. Declerck J, Vandenberk B, Deschepper M, Colpaert K, Cool L, Goemaere J, et al. Building a Foundation for High-Quality Health Data: Multihospital Case Study in Belgium. *JMIR Med Inform*. 2024 Dec 20;12:e60244. PMID: 39727158. doi: 10.2196/60244.
15. Deng Y, Jiang X, Long Q. Privacy-Preserving Methods for Vertically Partitioned Incomplete Data. *AMIA Annu Symp Proc*. 2020;2020:348-57. PMID: 33936407.

16. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. *J Med Internet Res*. 2018 2018/05/29;20(5):e185. doi: 10.2196/jmir.9134.
17. Lucero RJ, Kearney J, Cortes Y, Arcia A, Appelbaum P, Fernández RL, et al. Benefits and Risks in Secondary Use of Digitized Clinical Data: Views of Community Members Living in a Predominantly Ethnic Minority Urban Neighborhood. *AJOB Empir Bioeth*. 2015;6(2):12-22. PMID: 26101782. doi: 10.1080/23294515.2014.949906.
18. Declerck J KD, Vander Stichele R, Coorevits P. Frameworks, dimensions, definitions of aspects and assessment methods for the appraisal of quality of health data for secondary use: a review of reviews. *JMIR Medical Informatics*. 2024.
19. Bian J, Lyu T, Loiacono A, Viramontes TM, Lipori G, Guo Y, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. *J Am Med Inform Assoc*. 2020 Dec 9;27(12):1999-2010. PMID: 33166397. doi: 10.1093/jamia/ocaa245.
20. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Wash DC)*. 2016;4(1):1244. PMID: 27713905. doi: 10.13063/2327-9214.1244.
21. Liaw ST, Guo JGN, Ansari S, Jonnagaddala J, Godinho MA, Borelli AJ, et al. Quality assessment of real-world data repositories across the data life cycle: A literature review. *J Am Med Inform Assoc*. 2021 Jul 14;28(7):1591-9. PMID: 33496785. doi: 10.1093/jamia/ocaa340.
22. Ozonze O, Scott PJ, Hopgood AA. Automating Electronic Health Record Data Quality Assessment. *Journal of Medical Systems*. 2023 2023-02-13;47(1). doi: 10.1007/s10916-022-01892-2.
23. Weiskopf NG, Bakken S, Hripcsak G, Weng C. A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. *EGEMS (Wash DC)*. 2017 Sep 4;5(1):14. PMID: 29881734. doi: 10.5334/egems.218.
24. Sendak M, Sirdeshmukh G, Ochoa T, Premo H, Tang L, Niederhoffer K, et al. Development and Validation of ML-DQA – a Machine Learning Data Quality Assurance Framework for Healthcare. In: Zachary L, Rajesh R, Mark S, Michael S, Serena Y, editors. *Proceedings of the 7th Machine Learning for Healthcare Conference; Proceedings of Machine Learning Research*: PMLR; 2022. p. 741--59.
25. Seitzl F, Kovářík T, Mirshahi S, Krystufek J, Dujava R, Ondreicka M, et al. Assessing the quality of information extraction. *ArXiv*. 2024;abs/2404.04068.
26. Azeroual O, Saake G, Abuosba M, Schöpfel J. Text data mining and data quality management for research information systems in the context of open data and open science. *ArXiv*. 2018;abs/1812.04298.
27. Palojoki S, Lehtonen L, Vuokko R. Semantic Interoperability of Electronic Health Records: Systematic Review of Alternative Approaches for Enhancing Patient Information Availability. *JMIR Med Inform*. 2024 2024/4/25;12:e53535. doi: 10.2196/53535.
28. Hendler J. Data Integration for Heterogeneous Datasets. *Big Data*. 2014 Dec 1;2(4):205-15. PMID: 25553272. doi: 10.1089/big.2014.0068.
29. AIDAVA (2023) Project website.; Available from: www.aidava.eu.
30. Juran JM, Gryna FM, Bingham RS. *Quality control handbook*: McGraw-Hill New York; 1974.
31. Liaw ST, Rahimi A, Ray P, Taggart J, Dennis S, de Lusignan S, et al. Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. *Int J Med Inform*. 2013 Jan;82(1):10-24. PMID: 23122633. doi: 10.1016/j.ijmedinf.2012.10.001.
32. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc*. 2013 Jan 1;20(1):144-51.

PMID: 22733976. doi: 10.1136/amiajnl-2011-000681.

33. Issa S, Adekunle O, Hamdi F, Cherfi SS-S, Dumontier M, Zaveri A. Knowledge Graph Completeness: A Systematic Literature Review. *IEEE Access*. 2021;9:31322-39.

34. Huser V, DeFalco FJ, Schuemie M, Ryan PB, Shang N, Velez M, et al. Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Data Sets. *EGEMS (Wash DC)*. 2016;4(1):1239. PMID: 28154833. doi: 10.13063/2327-9214.1239.

35. de Zegher I, Norak K, Steiger D, Müller H, Kalra D, Scheenstra B, et al. Artificial intelligence based data curation: enabling a patient-centric European health data space. *Front Med (Lausanne)*. 2024;11:1365501. PMID: 38813389. doi: 10.3389/fmed.2024.1365501.

36. Mohamed Y, Song X, McMahon TM, Sahil S, Zozus M, Wang Z, et al. Electronic health record data quality variability across a multistate clinical research network. *J Clin Transl Sci*. 2023;7(1):e130. PMID: 37396818. doi: 10.1017/cts.2023.548.

37. Ham K. OpenRefine (version 2.5). <http://openrefine.org>. Free, open-source tool for cleaning and transforming data. *J Med Libr Assoc*. 2013 Jul;101(3):233-4. doi: 10.3163/1536-5050.101.3.020.

38. Lewis AE, Weiskopf N, Abrams ZB, Foraker R, Lai AM, Payne PRO, et al. Electronic health record data quality assessment and tools: a systematic review. *J Am Med Inform Assoc*. 2023 Sep 25;30(10):1730-40. PMID: 37390812. doi: 10.1093/jamia/ocad120.

39. Syed R, Eden R, Makasi T, Chukwudi I, Mamudu A, Kamalpour M, et al. Digital Health Data Quality Issues: Systematic Review. *J Med Internet Res*. 2023 Mar 31;25:e42615. PMID: 37000497. doi: 10.2196/42615.

40. Fraser HSF, Mugisha M, Bacher I, Ngenzi JL, Seebregts C, Umubyeyi A, et al. Factors Influencing Data Quality in Electronic Health Record Systems in 50 Health Facilities in Rwanda and the Role of Clinical Alerts: Cross-Sectional Observational Study. *JMIR Public Health Surveill*. 2024 2024/7/3;10:e49127. doi: 10.2196/49127.

41. Lewis AE, Weiskopf N, Abrams ZB, Foraker R, Lai AM, Payne PRO, et al. Electronic health record data quality assessment and tools: a systematic review. *Journal of the American Medical Informatics Association*. 2023;30(10):1730-40. doi: 10.1093/jamia/ocad120.

42. ; Available from: <https://ohdsi.github.io/CommonDataModel/>.