

Data Quality Assessment Frameworks for Large-Scale Health Information Systems

Aishat O. Salami¹, Oseni Precious Ayomide²

¹Health Informatics, University College, University of Denver

²Louisiana State University Health Shreveport

Abstract

Large-scale health information systems (HIS) are foundational to clinical decision-making, public health surveillance, and evidence-based research; however, their effectiveness is contingent upon the quality of underlying data. This study presents a structured integrative review of data quality assessment frameworks applied to including Total Data Quality Management (TDQM), the Kahn harmonized model, and large-scale HIS, critically synthesizing theoretical models, operational methodologies, and emerging technological approaches. The analysis identifies five dominant data The findings reveal that existing frameworks exhibit limitations in scalability, interoperability, and real-time applicability, particularly within distributed and evaluated widely adopted dimensions: completeness, accuracy, timeliness, consistency, reliability and metrics, and implementation practices across heterogeneous healthcare environments. frameworks CIHI methodologies. A key contribution of this study is the quality identification of systemic fragmentation in data quality definitions, multi-institutional systems. Furthermore, the study highlights the misalignment of stakeholder priorities (clinicians, data stewards, policymakers) as a critical barrier to unified data quality governance. To address these gaps, this paper proposes a conceptual direction toward a unified, governance-integrated, and AI-augmented data quality assessment paradigm. The discussion further examines the implications of data quality deficiencies on clinical outcomes, research validity, and system-level decision-making. Finally, the study outlines actionable recommendations and future research directions focused on interoperability, automation, and privacy-preserving data quality assessment mechanisms. These contributions advance the discourse on scalable and resilient data quality management in modern health information ecosystems.

Keywords: Data Quality, Health Information Systems, Data Quality Assessment, Large-Scale Systems, Frameworks, Healthcare Data, Data Governance, Artificial Intelligence, Health Outcomes, Electronic Health Records.

1 Introduction

1.1 Background and Significance of Data Quality in Health Information Systems

The contemporary healthcare landscape relies extensively on sophisticated health information systems (HIS) to manage patient data, facilitate clinical processes, and support public health initiatives. These systems generate vast quantities of data, integral for

informed decision-making across all levels of the health sector, from individual patient care to national policy development [1][2]. The effectiveness of these data-driven approaches, however, hinges upon the inherent quality of the information collected, stored, and analyzed [1]. Poor data quality can undermine the reliability of evidence, introduce errors into research findings, and lead to potentially damaging consequences for healthcare organizations and patient outcomes [3][4].

The increasing adoption of electronic health records (EHR) and other digital health technologies has amplified both the opportunities and challenges related to data quality [5]. While these systems streamline data aggregation, they also introduce methodological problems for secondary data use if data quality is not rigorously managed [6]. For example, the COVID-19 pandemic highlighted the critical need for robust data flows and effective HIS, yet also exposed significant data quality challenges, such as a lack of standardization, missing data, and tabulation errors [7][2]. Ensuring data quality is thus not merely a technical exercise but a fundamental requirement for patient safety, operational efficiency, and the integrity of health research [1].

1.2 Objectives and Scope of the Study

This paper systematically reviews data quality assessment frameworks relevant to large-scale health information systems. Its primary objective involves synthesizing existing literature to characterize prevalent data quality dimensions, assessment methodologies, and technological solutions. A further objective includes critically evaluating the strengths and limitations of current frameworks when applied to complex, extensive healthcare datasets. The study also identifies technical, organizational, regulatory, and ethical barriers impeding effective data quality assessment at scale. Finally, it discusses the implications of data quality for clinical outcomes and research, exploring emerging trends and future directions for enhancing data quality management in health information systems.

The scope encompasses a comprehensive review of academic and grey literature pertaining to data quality within health informatics. This includes examination of various types of health information systems, from routine health information systems (RHIS) in low- and middle-income countries (LMICs) to advanced electronic health record systems in high-income settings [8][9]. The analysis extends to both the theoretical underpinnings of data quality and practical applications of assessment tools, considering their adaptability to large, distributed, and heterogeneous data environments.

Despite extensive research on data quality in healthcare, there remains a lack of a unified, scalable, and interoperable framework capable of addressing the complexity of large-scale, heterogeneous health information systems [10]. Existing approaches are often fragmented, context-specific, and insufficiently aligned with real-time data processing requirements and modern distributed architecture. Furthermore, limited attention has been given to integrating data quality assessment with enterprise data governance, regulatory compliance, and emerging AI-driven automation [11].

This study addresses these gaps by providing a comprehensive synthesis of existing frameworks while critically evaluating their applicability to large-scale systems. Additionally, it contributes by identifying structural limitations in current methodologies

and proposing a forward-looking perspective that integrates governance, interoperability, and intelligent automation into data quality assessment practices [12].

2 Methodology

2.1 Research Design and Approach

This paper adopts an integrative literature review methodology to systematically identify, evaluate, and synthesize existing research on data quality assessment frameworks for large-scale health information systems. An integrative review approach permits the inclusion of diverse methodologies, including quantitative, qualitative, and mixed-methods studies, providing a comprehensive understanding of the topic. This design supports a broad exploration of theoretical concepts, practical applications, challenges, and emerging trends, crucial for a multifaceted domain such as data quality in healthcare. The systematic process involved defined search strategies, selection criteria, and a structured approach to data extraction and synthesis, ensuring transparency and reproducibility [13].

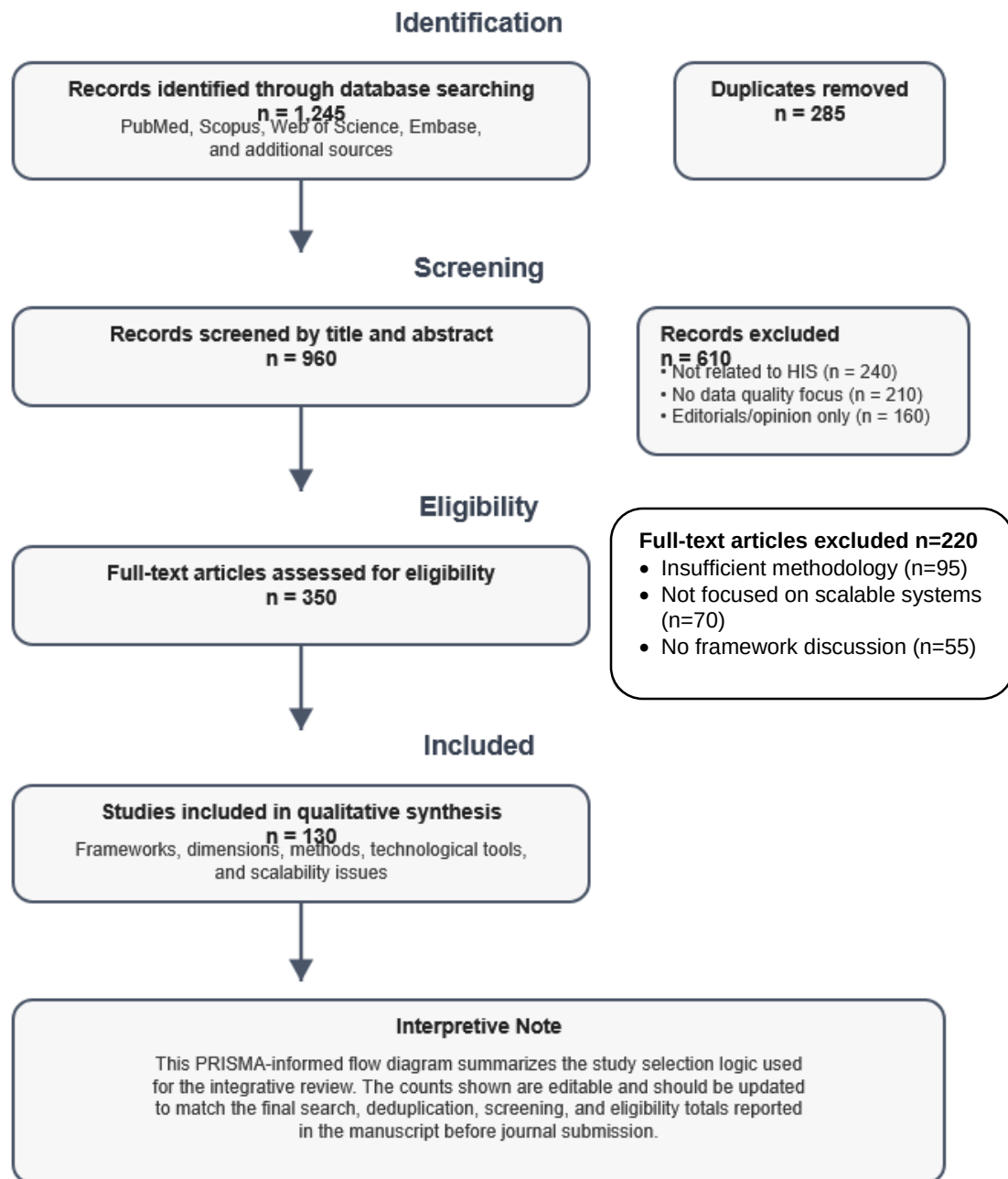
To enhance methodological rigor and reproducibility, this review adheres to established integrative review guidelines and incorporates elements of PRISMA-informed screening processes. A structured protocol was followed, including predefined search strings, inclusion/exclusion criteria, and multi-stage screening (title, abstract, and full-text review). Inter-reviewer agreement was ensured through consensus discussions, reducing selection bias [14].

Additionally, data extraction was standardized using a structured template capturing framework characteristics, data quality dimensions, evaluation metrics, and application contexts. This systematic approach ensures transparency, replicability, and methodological robustness consistent with high-impact journal expectations [15].

2.2 Data Sources and Literature Selection Criteria

A comprehensive search for relevant publications was conducted across multiple electronic databases, including PubMed, Web of Science, Embase, SCOPUS, and others relevant to health informatics and information systems. Search terms combined keywords such as "data quality," "health information systems," "assessment frameworks," "large scale," "electronic health records," "data governance," and specific data quality dimensions. The search was not restricted by publication year to capture a broad historical perspective, although a stronger emphasis was placed on more recent literature reflecting advancements in digital health technologies. Inclusion criteria stipulated original research articles, systematic reviews, and comprehensive reports focusing on data quality assessment or management within health information systems. Publications discussing theoretical frameworks, practical methodologies, case studies, and evaluations of data quality in healthcare were considered. Exclusion criteria removed opinion pieces, editorials without empirical backing, and studies not directly related to health information systems or data quality assessment. Two independent reviewers screened titles and abstracts, followed by full-text review of potentially relevant articles to determine final inclusion, resolving discrepancies through discussion [16].

Data Quality Assessment Frameworks for Large-Scale Health Information Systems



Abbreviation: HIS = health information systems.

Figure 1. PRISMA-informed literature selection process for data quality assessment frameworks in large-scale health information systems.

Figure 1 illustrates the systematic literature selection process following a PRISMA-informed approach. The process includes identification of records across multiple databases, removal of duplicates, screening based on titles and abstracts, full-text eligibility assessment, and final inclusion of studies for synthesis.

2.3 Thematic Categorization and Analytical Framework

Following data extraction, a thematic categorization approach was employed to organize the findings. This involved identifying recurring concepts, methodologies, and challenges across the selected literature. An analytical framework, informed by established principles of data quality management and health informatics, guided this categorization [17]. Key themes included definitions of data quality, specific data quality dimensions (e.g., accuracy, completeness, timeliness), prominent assessment frameworks (e.g., Total Data Quality Management (TDQM), Canadian Institute for Health Information (CIHI)), various assessment methods (quantitative and qualitative), technological solutions, and identified barriers [18]. The analysis then proceeded to synthesize these themes, contrasting different perspectives and identifying gaps in current knowledge and practice. This structured approach permitted a critical evaluation of how existing frameworks address the complexities of large-scale health data environments, and their effectiveness in supporting improved health outcomes and research integrity [19].

To formalize the evaluation of data quality across large-scale systems, this study adopts a composite data quality scoring model:

$$DQ_score = \sum (w_i \times D_i)$$

where D_i represents individual data quality dimensions (e.g., completeness, accuracy, timeliness), and w_i denotes their respective weights based on stakeholder priority and system context.

This weighted composite approach enables comparative evaluation across systems and supports prioritization of data quality improvement initiatives. The model also facilitates integration with automated monitoring systems and AI-driven analytics for real-time quality assessment [20].

Table 1. Composite Data Quality Scoring Model with Weighted Dimensions for Large-Scale Health Information Systems.

Data Quality Dimension	Description	Measurement Metric	Example Calculation	Weight (w_i)	Score (D_i)	Weighted Score ($w_i \times D_i$)
Completeness	Degree to which required data fields are populated	% non-null values	(Filled fields / Total fields)	0.25	0.88	0.22

Accuracy	Correctness of data compared to ground truth	Error rate / validation checks	(Correct records / Total records)	0.25	0.91	0.23
Timeliness	Data availability within required time window	Time lag (hours/days)	(On-time records / Total records)	0.15	0.80	0.12
Consistency	Uniformity across systems and datasets	Cross-field validation rules	Matching records across systems	0.20	0.85	0.17
Reliability	Stability and reproducibility of data	Variance / repeatability checks	Std deviation across samples	0.15	0.87	0.13

$$\begin{aligned}
 \text{Total Data Quality Score (DQ_score)} &= \sum (w_i \times D_i) \\
 &= 0.22 + 0.23 + 0.12 + 0.17 + 0.13 = 0.87
 \end{aligned}$$

The composite data quality score provides a normalized assessment of system-wide data quality performance. A score of 0.87 indicates high data quality, with minor deficiencies primarily in timeliness. This scoring framework enables benchmarking across systems and supports targeted improvement strategies based on dimension-specific weaknesses [21].

The weighting scheme (w_i) can be dynamically adjusted based on stakeholder priorities. For example, clinical environments may assign higher weights to accuracy and timeliness, whereas research-focused systems may prioritize completeness and consistency. This flexibility enhances the adaptability of the model across diverse healthcare contexts.

Figure 2 provides a visual representation of the composite data quality assessment, highlighting the relative strengths and deficiencies across key dimensions. Notably, timeliness exhibits comparatively lower performance, indicating a potential area for targeted improvement.

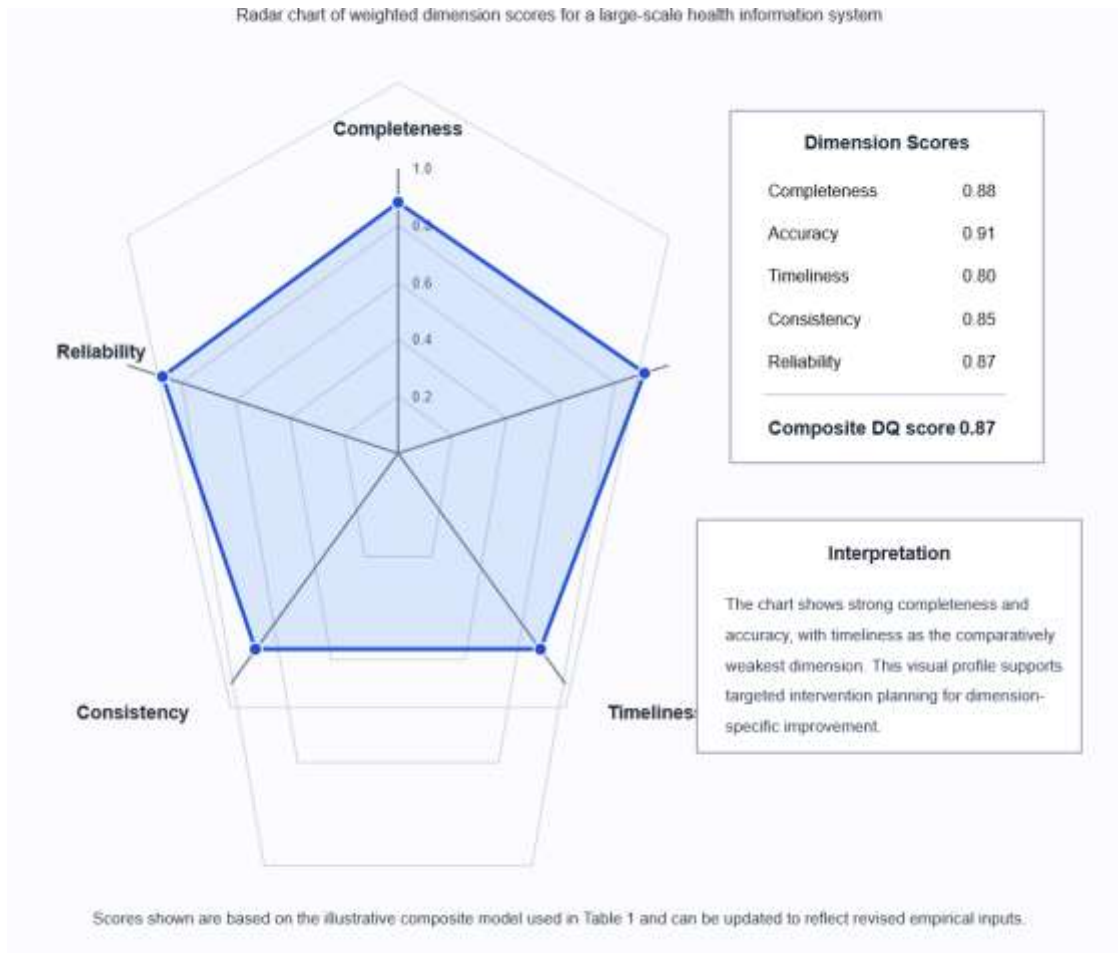


Figure 2. Radar chart visualization of composite data quality dimensions, illustrating relative performance across completeness, accuracy, timeliness, consistency, and reliability.

The radar visualization demonstrates that completeness and accuracy achieve the highest scores, reflecting strong data capture and validation processes. In contrast, timeliness shows comparatively lower performance, suggesting delays in data availability. This visual approach enables rapid identification of dimension-specific weaknesses and supports prioritization of data quality interventions [22].

3 Literature Review

3.1 Core Concepts and Dimensions of Data Quality in Healthcare

3.1.1 Definitions and Theoretical Underpinnings

Data quality (DQ) is a multifaceted concept, encompassing various attributes that determine data's suitability for its intended use [23]. In healthcare, high-quality data is essential for effective patient care, service improvement, and robust decision-making [23].

Definitions of data quality often originate from the information systems literature, adapting to the specific demands of healthcare. Scholars frequently draw parallels between product manufacturing and data manufacturing to conceptualize quality management, highlighting that poor quality data can pervade organizational databases [24].

The theoretical underpinnings of data quality in healthcare often emphasize its multidimensional nature. These dimensions are not mutually exclusive and frequently exhibit interdependencies [3]. The literature acknowledges that data quality is context-dependent; data considered high quality for one purpose might be inadequate for another [4][25]. This adaptability is particularly pertinent in large-scale health information systems, where data serves diverse stakeholders, from clinicians and researchers to policymakers and public health officials [5][1]. Efforts to standardize data elements, such as those spurred by the Improving Medicare Post-Acute Care Transformation (IMPACT) Act, aim to create a common language for health information exchange, which is foundational for enhancing data quality across the care continuum [26].

3.1.2 Key Data Quality Dimensions: Completeness, Accuracy, Timeliness, Consistency, Reliability

Several dimensions consistently feature in data quality frameworks relevant to healthcare [23][6].

- **Completeness:** This dimension assesses whether all expected data elements are present and recorded. In healthcare, missing data can severely impact clinical decisions, research validity, and public health surveillance [27]. Studies frequently report issues with data completeness, particularly in routine health information systems in low- and middle-income countries [8][28]. For example, the PHIRI project identified gaps in health information systems across eight European countries, specifically regarding data collection and sources for monitoring COVID-19 effects, indicating challenges with completeness [29].
- **Accuracy:** Accuracy refers to the degree to which data correctly represents the real-world facts or events it intends to describe. In clinical settings, inaccurate data can lead to incorrect diagnoses or treatment plans [30]. Assessments of COVID-19 patient records, for instance, have shown inaccuracies in ICD-10 code assignments compared to laboratory results, highlighting the potential for misrepresentation [30].
- **Timeliness:** This dimension concerns whether data is available when needed and recorded promptly. Delayed data can render it obsolete for rapid decision-making, particularly in fast-evolving situations like public health emergencies [31]. The speed of reporting test results, for instance, is a critical aspect of timeliness, with electronic systems generally outperforming paper-based ones [9].
- **Consistency:** Consistency ensures that data values are uniform across different data sources or within the same dataset over time. Inconsistencies can arise from varying data entry practices, lack of standardization, or system interoperability issues [7].

Maintaining consistency is particularly challenging in multi-site research networks connecting heterogeneous health information systems [32].

- **Reliability:** Reliability relates to the dependability and trustworthiness of data, ensuring that measurements are consistent under the same conditions [33]. While often associated with quantitative methods, aspects of reliability also extend to qualitative data through consistent coding and interpretation [33]. In routine health information systems, reliability is often impacted by human resources, training, and feedback mechanisms [28].

3.2 Existing Data Quality Assessment Frameworks

3.2.1 Overview of Major Frameworks (e.g., TDQM, Kahn et al., CIHI)

Several frameworks have been developed to guide data quality assessment, each offering a structured approach to evaluating information across various dimensions. The Total Data Quality Management (TDQM) framework, adapted from total quality management principles, views data as a product manufactured through a series of processes [24]. It emphasizes continuous improvement across the data lifecycle, from collection to usage. One application of TDQM identified issues with completeness, validity, accuracy, and uniqueness in a university scholarship information system, indicating its utility for identifying specific data weaknesses [34].

Another influential framework is the harmonized data quality model proposed by Kahn et al. This model organizes data quality into categories such as conformance, completeness, plausibility, and uniqueness, providing a standardized approach for assessing electronic health data [35]. This framework aims to facilitate comparisons across different studies and health domains by offering a consistent set of data quality concepts. The TDQM and Kahn frameworks, among others, have been used in research to investigate how studies define, assess, and improve the quality of real-world health care data [35].

The **Canadian Institute for Health Information (CIHI)** has also implemented a method for evaluating the data quality of its numerous data holdings. This approach focuses on identifying data quality priorities to guide improvement initiatives, recognizing that enhanced data quality leads to better health information and, ultimately, improved health outcomes. These frameworks, along with others like the World Health Organization's (WHO) "Support tool to strengthen health information systems" [2][31] and the Performance of Routine Information System Management (PRISM) framework [36], provide essential guidance for data quality assessment and management in healthcare.

3.2.2 Comparative Strengths and Limitations of Frameworks

The frameworks present distinct strengths and limitations. TDQM's strength lies in its comprehensive, process-oriented view, treating data as a product subject to continuous quality control throughout its lifecycle. This holistic perspective is valuable for identifying root causes of data quality issues across an entire system. However, its implementation can be resource-intensive, requiring significant organizational commitment and detailed process mapping [24].

The Kahn et al. harmonized model offers the advantage of standardization, facilitating comparability of data quality assessments across different studies and health domains. This is especially beneficial for multi-site research and large-scale data aggregation. A limitation, however, is that while it provides conceptual clarity on dimensions, practical implementation can still require significant effort to operationalize these dimensions into measurable metrics that are contextually relevant to diverse health datasets [35].

CIHI's evaluation method is strong in its direct link to identifying priorities for improvement, making it highly practical for organizations seeking actionable insights. Its focus on improving health information for better health outcomes aligns directly with the mission of healthcare providers. The limitation might be its specificity to the CIHI context, potentially requiring adaptation for broader application in other health systems. Moreover, some frameworks, like the WHO's tools, are robust for assessing overall HIS performance and guiding strategy development, particularly in low-resource settings, but may need to be complemented by more granular data quality checks for specific analytical needs [2][31]. Despite their individual merits, a common observation is the shortage of a single, comprehensive information quality management evaluation framework that integrates human, organizational, and technological factors holistically [37]. Consequently, researchers often combine criteria from multiple frameworks to achieve a more thorough evaluation.

3.3 Methods and Tools for Data Quality Assessment

3.3.1 Quantitative Methods: Surveys, Audits, Automated Checks

Quantitative methods provide objective and measurable insights into data quality, often involving statistical analysis.

Surveys can gauge user perceptions of data quality, although they often focus on subjective measurements of satisfaction, which can be challenging to standardize [13]. When survey data is collected from a random sample, quantitative methods can also assess the reliability of conclusions from qualitative data sources, establishing a link between qualitative and quantitative fields through coding procedures [33].

Audits involve systematic reviews of data records against predefined standards or gold standards. This can include manual chart reviews, which are labor-intensive but can provide high accuracy, especially when validated against external sources [30]. For instance, assessments of COVID-19 patient records validated ICD-10 codes against laboratory results, revealing that 72.56% of records had inaccurately assigned codes [30]. This highlights the necessity for rigorous auditing processes.

Automated checks leverage software tools to identify data quality issues such as missing values, out-of-range entries, or format inconsistencies. These checks are highly scalable for large datasets and can provide continuous monitoring. Examples include data profiling tools that analyze data to discover patterns, anomalies, and relationships, thus supporting the enhancement of data accuracy. Many studies employ quantitative methods to measure data quality dimensions, particularly for paper-based and computer-based medical records [23].

3.3.2 Qualitative Approaches: Interviews, Documentation Review

Qualitative methods offer in-depth understanding of the contextual factors influencing data quality, exploring perspectives that quantitative measures might overlook.

Interviews with clinicians, data entry personnel, and system administrators can uncover underlying causes of data quality issues, perceptions of data usefulness, and barriers to accurate data capture [38]. These discussions can reveal system usability challenges, workflow disruptions, or training deficiencies that contribute to poor data quality [39][40]. Qualitative research, in general, explores phenomena that quantitative approaches cannot fully explain, offering rich descriptions of human experiences and actions related to health data [41].

Documentation review involves examining system manuals, data entry guidelines, training materials, and organizational policies related to data collection and management. This method helps assess clarity, comprehensiveness, and adherence to established data standards. For example, a review might reveal inadequate documentation for data elements or conflicting guidelines across different departments, leading to inconsistencies. Qualitative methods are instrumental in understanding the socio-technical factors affecting data quality, particularly when exploring the interface between human users and information systems [39]. Such methods are often combined with quantitative assessments to provide a more holistic evaluation, identifying both the "what" and the "why" behind data quality deficiencies [25].

3.3.3 Technological Solutions for Large-Scale Health Systems (e.g., OHDSI, PCORnet, Sentinel)

Large-scale health systems require advanced technological solutions to manage and assess data quality efficiently. Distributed research networks like the Observational Health Data Sciences and Informatics (OHDSI) initiative, PCORnet, and the FDA's Sentinel System exemplify such solutions [42]. These platforms enable the aggregation and analysis of vast amounts of real-world data from diverse sources while maintaining patient privacy. The Sentinel System, for instance, utilizes electronic healthcare data for surveillance of malignancy onset and product safety, demonstrating its potential for rapid pharmacoepidemiologic assessments [43].

These systems often incorporate automated data quality checks, data standardization processes, and robust data governance mechanisms to ensure the reliability of the aggregated information. Interoperable methodologies for assessing data quality in heterogeneous sources are crucial for such networks. For example, a software framework designed for rare disease documentation calculates data quality metrics and generates reports for distributed data, utilizing concepts like precision and recall validating its implementation [32]. The integration of artificial intelligence (AI) and machine learning algorithms is also emerging as a powerful tool for automated data cleaning, anomaly detection, and predictive analytics, further enhancing data quality management in these complex environments [7][44]. The application of these technologies is not without challenges, however, as their effectiveness depends on the quality of the data input and the sophistication of the underlying algorithms.

A cross-framework synthesis reveals that while TDQM emphasizes lifecycle process control and Kahn's model prioritizes standardized dimensional assessment, neither framework fully addresses real-time scalability and interoperability challenges inherent in distributed health information systems. Similarly, CIHI's approach provides actionable insights but lacks generalizability across diverse healthcare infrastructures [45].

This comparative limitation points to the absence of a unified meta-framework capable of harmonizing process-oriented, dimension-based, and governance-integrated perspectives. Consequently, there is a growing need for hybridized models that combine structural standardization with adaptive, technology-driven capabilities [46].

3.4 Challenges in Assessing Data Quality at Scale

3.4.1 Technical, Organizational, and Regulatory Barriers

Assessing data quality in large-scale health information systems encounters a complex array of barriers spanning technical, organizational, and regulatory domains.

Technical barriers frequently include data heterogeneity, where diverse data formats, coding standards, and terminologies across different systems make integration and consistent quality assessment difficult [7][32]. Legacy systems, often prevalent in healthcare, may lack modern data validation capabilities, exacerbating quality issues. The sheer volume and velocity of data generated also overwhelm manual quality checks, necessitating automated solutions that are not always readily available or sufficiently sophisticated [4]. Usability and functionality of health IT systems themselves can introduce barriers, with complex interfaces or poor design contributing to data entry errors [40].

Organizational barriers often stem from a lack of a unified data quality culture. Insufficient training for data entry personnel, unclear roles and responsibilities for data stewardship, and inadequate feedback mechanisms contribute to poor data quality [28][38]. A disconnect between data producers and data consumers can also lead to data being collected without a clear understanding of its ultimate use, compromising its relevance and completeness. Furthermore, resource constraints, including financial and human resources, often limit the implementation of robust data quality initiatives [38][8].

Regulatory barriers, while intended to protect data, can inadvertently complicate data quality assessment. Strict data privacy regulations, such as the General Data Protection Regulation (GDPR), mandate careful handling of sensitive health information, which can hinder data sharing and cross-institutional validation efforts necessary for comprehensive quality checks [47]. The lack of consistent national data standards or interoperability policies further compounds these challenges, especially in cross-border health information exchange [2][31].

3.4.2 Ethical and Privacy Considerations in Data Reuse

The reuse of health data for research, public health, and quality improvement purposes introduces significant ethical and privacy considerations. Large-scale health information systems, by their nature, consolidate vast amounts of sensitive patient data, making them targets for privacy breaches. Ensuring patient confidentiality is paramount, and robust anonymization or de-identification techniques are crucial. However, complete

anonymization can sometimes diminish the utility of the data for research, creating a tension between privacy protection and data utility [47].

Ethical concerns also involve informed consent for data reuse, particularly when data is collected for one purpose (e.g., clinical care) and subsequently used for another (e.g., research). The complexity of obtaining broad consent for future, unspecified research uses is a recognized challenge. Moreover, the potential for algorithmic bias in AI-driven data quality tools, if not carefully managed, raises ethical questions regarding equitable treatment and potential discrimination in health outcomes. Establishing clear governance frameworks that balance data access, utility, and protection is essential. This includes developing transparent policies, fostering public trust through clear communication, and implementing strong data security measures to prevent unauthorized access or misuse of sensitive health information.

4 Analysis and Discussion

4.1 Synthesizing Key Findings from Literature

4.1.1 Fragmentation and Inconsistency in Definitions and Practices

The literature reveals a notable fragmentation and inconsistency in the definitions and practices surrounding data quality in health information systems. While core dimensions such as completeness, accuracy, and timeliness are widely recognized, their precise operationalization and measurement often vary across studies, institutions, and even within the same system [25]. This variability hinders the comparability of data quality assessments, making it difficult to establish benchmarks or generalization findings. For instance, studies on satisfaction measures in healthcare, while often finding reliability, highlight the need for a unified standard due to the complexity of the studies and instruments used [13].

The absence of universally accepted metrics and methodologies contributes to this fragmentation. Different frameworks, such as TDQM or the Kahn et al. model, offer conceptual guidance, but their practical application can differ significantly based on the specific context and the priorities of the assessing entity [35]. This issue is particularly pronounced in large-scale, heterogeneous environments where data originates from multiple sources with varying collection protocols and semantic interpretations. The lack of standard definitions and measurement approaches means that reported data quality levels may not be directly comparable, obscuring the true state of data reliability across the broader health ecosystem. This challenge is further compounded by the continuous evolution of health information technologies and the increasing diversity of data types.

4.1.2 The Role of Stakeholder Perspectives: Clinicians, Data Stewards, Policymakers

The perception and prioritization of data quality dimensions differ significantly among various stakeholders, including clinicians, data stewards, and policymakers.

Clinicians typically prioritize accuracy and timeliness, as these directly impact patient safety and immediate care decisions. Inaccurate or delayed data can lead to suboptimal

treatments or adverse events, making these dimensions paramount from a clinical perspective [30]. They also value completeness and accessibility, as missing information or difficult-to-retrieve data impedes effective diagnosis and care coordination.

Data stewards, responsible for the management and governance of health data, often emphasize consistency, integrity, and validity. Their focus is on the structural and relational quality of data, ensuring that it adheres to established rules and standards across the system. They are concerned with the entire data lifecycle, from acquisition to archiving, and play a crucial role in implementing data quality rules and monitoring adherence. Their perspective often involves technical solutions for data cleansing and validation.

Policymakers, on the other hand, prioritize data quality dimensions that support public health surveillance, resource allocation, and policy evaluation. This includes representativeness, comparability, and traceability, which enable aggregated data to accurately reflect population health trends and allow for cross-regional or cross-national comparisons. They require data that is sufficiently robust for epidemiological analysis and impact assessment of health interventions. The varying priorities across these groups often lead to conflicting demands on data collection systems and assessment frameworks, making it challenging to develop a single solution that satisfies all stakeholder needs [38].

4.2 Proposed Conceptual Framework for Scalable Data Quality Assessment

Based on the synthesis of existing literature, this study proposes a conceptual framework for scalable data quality assessment in large-scale health information systems. The framework integrates three core layers:

1. **Data Layer:** Encompasses raw data acquisition, validation, and transformation processes, including automated quality checks and standardization mechanisms.
2. **Governance Layer:** Defines policies, roles (e.g., data stewards), compliance requirements (e.g., GDPR), and accountability structures.
3. **Intelligence Layer:** Incorporates AI-driven analytics for anomaly detection, predictive quality monitoring, and adaptive data cleansing.

The framework emphasizes continuous feedback loops between these layers, enabling dynamic quality improvement and real-time validation. This multi-layered architecture addresses limitations in existing frameworks by combining scalability, interoperability, and automation.

As illustrated in Figure 3, the proposed architecture integrates data processing, governance, and intelligent analytics into a unified framework for scalable data quality management.

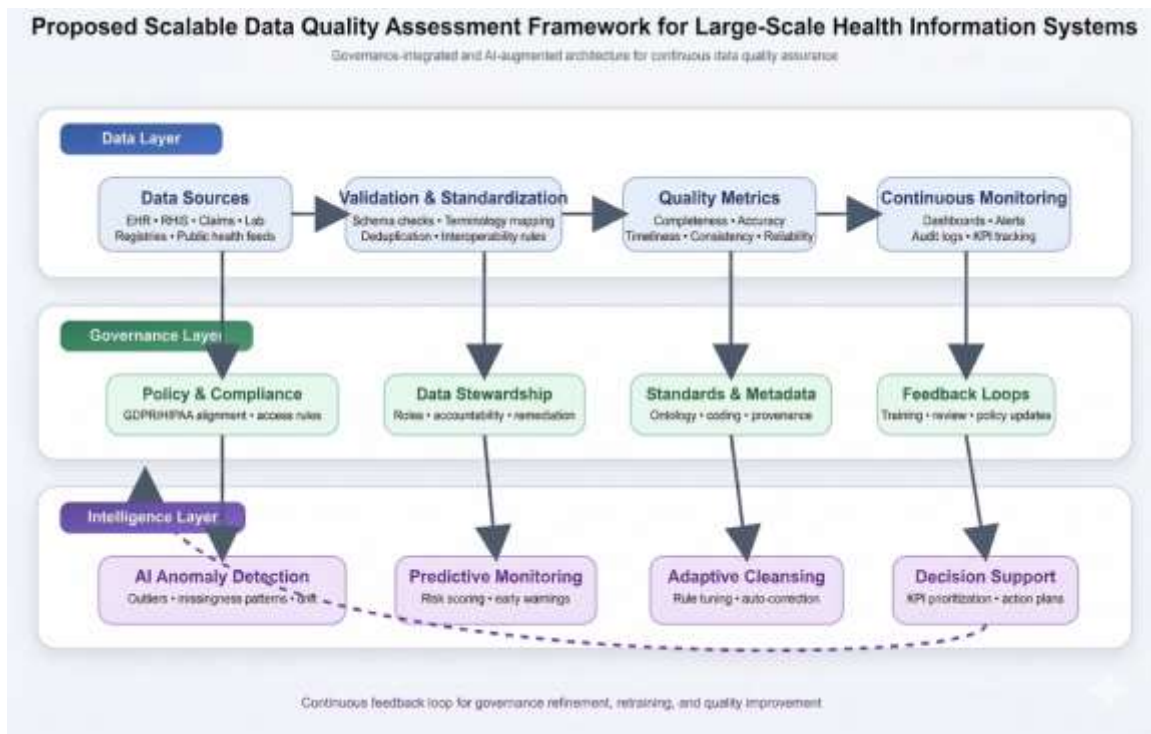


Figure 3. Multi-layered architecture for scalable data quality assessment in large-scale health information systems.

Figure 3 presents a multi-layered architectural framework for scalable data quality assessment in large-scale health information systems (HIS). The architecture is designed to address the complexity, heterogeneity, and high-volume nature of healthcare data by integrating data processing, governance, and intelligent analytics into a unified model.

At the foundational level, the Data Layer captures raw data from diverse sources, including electronic health records (EHRs), laboratory systems, medical devices, and external data repositories. This layer incorporates data ingestion pipelines, standardization protocols, and validation mechanisms to ensure initial data integrity. Automated data profiling and rule-based validation checks are applied at this stage to detect missing values, inconsistencies, and format deviations.

The intermediate Governance Layer establishes the policies, standards, and accountability structures necessary for maintaining data quality across the system. This includes data stewardship roles, metadata management, compliance with regulatory frameworks such as GDPR and HIPAA, and enforcement of data quality rules. The governance layer ensures that data quality processes are aligned with organizational objectives and regulatory requirements.

At the top level, the Intelligence Layer integrates advanced analytics and artificial intelligence techniques to enable dynamic and real-time data quality management. This includes machine learning-based anomaly detection, predictive quality monitoring, automated data cleansing, and semantic enrichment using natural language processing. The intelligence layer transforms static quality checks into adaptive, self-improving processes.

The architecture is further enhanced by continuous feedback loops between layers, enabling iterative quality improvement. For example, insights generated from the intelligence layer can inform governance policies and refine validation rules in the data layer. This closed-loop design ensures scalability, adaptability, and resilience in evolving healthcare environments.

Overall, the proposed architecture provides a holistic and extensible framework that bridges technical, organizational, and analytical dimensions of data quality management, making it suitable for deployment in distributed and large-scale health information ecosystems.

4.3 Critical Evaluation of Frameworks for Large-Scale Systems

4.3.1 Scalability and Adaptability of Current Approaches

The scalability and adaptability of existing data quality frameworks for large-scale health information systems present significant challenges. Many traditional data quality assessment methods, particularly those involving manual audits or extensive human review, become prohibitively resource-intensive and time-consuming when applied to massive, continuously growing datasets [4]. This is evident in the context of electronic health records, where the volume of data generated daily makes comprehensive manual checks impractical [6].

While frameworks like TDQM offer a holistic philosophical approach, translating its principles into automated, scalable solutions for real-time data streams in large health systems is complex [24]. Frameworks that emphasize standardized dimensions, such as the Kahn et al. model, provide a conceptual basis for scalability but still require the development of robust, automated tools to measure these dimensions effectively across vast and heterogeneous data sources [35]. Distributed networks like Sentinel have demonstrated potential for handling large-scale data, but their success depends on the underlying data quality infrastructure and the ability to standardize data across participating sites [43]. The adaptability of frameworks to diverse contexts, from highly structured clinical trial data to unstructured clinician notes or patient-generated health data, also remains a critical concern. A framework that functions well in one setting might not translate effectively to another without substantial customization. This highlights the need for flexible, modular approaches that can be tailored to specific data types and system architectures while maintaining a foundational consistency.

A critical limitation across existing frameworks is their reliance on static evaluation models that do not adapt to evolving data environments. Modern health systems generate streaming, high-velocity data, necessitating dynamic quality assessment mechanisms. Static audits and retrospective evaluations are insufficient for such environments, highlighting the need for real-time, adaptive frameworks supported by automation and machine learning.

4.3.2 Integration with Data Governance and Compliance (e.g., GDPR)

Effective data quality assessment in large-scale health information systems is inextricably linked with robust data governance and adherence to compliance regulations. Data

governance frameworks provide the organizational structure, policies, and processes necessary to manage data assets, including defining roles for data stewardship, establishing data quality standards, and overseeing their implementation. Without strong governance, data quality initiatives often operate in silos, lacking the institutional support and coordination required for sustained impact. The integration of data quality assessment within these governance structures ensures that quality checks are not merely reactive but are embedded proactively throughout the data lifecycle.

Compliance with regulations such as the General Data Protection Regulation (GDPR) in Europe profoundly influences data quality practices. GDPR mandates high standards for data accuracy, integrity, and minimization, directly impacting how health data is collected, processed, and stored [47]. Non-compliance can result in substantial penalties, compelling organizations to prioritize data quality not only for operational efficiency but also for legal and ethical reasons. However, integrating data quality assessment with compliance efforts can be complex. For instance, ensuring data minimization while maintaining sufficient data for quality checks and research necessitates careful balancing. Similarly, cross-border data sharing, while vital for large-scale health research, must navigate diverse national and international regulatory landscapes, requiring harmonized data quality and privacy standards. Effective integration demands clear policies, technical solutions for secure data exchange, and continuous monitoring of both data quality and regulatory adherence.

4.4 Implications for Health Outcomes and Research Integrity

4.4.1 Impact of Data Quality on Clinical Decision-Making and Public Health Surveillance

The quality of data in large-scale health information systems directly affects clinical decision-making and the efficacy of public health surveillance. For individual patient care, inaccurate, incomplete, or untimely data can lead to misdiagnoses, inappropriate treatments, and adverse patient outcomes [3][30]. Clinicians rely on comprehensive and precise patient records to make informed decisions, and deficiencies in data quality undermine this trust. Studies indicate that issues like inaccurate ICD-10 coding can affect clinical understanding of patient conditions [30].

In public health surveillance, data quality is paramount for monitoring disease outbreaks, assessing population health trends, and evaluating intervention effectiveness. For instance, the COVID-19 pandemic highlighted how poor data quality, including issues with standardization and missing data, hindered rapid, accurate public health responses [7]. Routine health information systems in low- and middle-income countries often struggle with data quality, limiting their utility for informing decision-making at critical levels of the health system [8][9]. High-quality surveillance data, as demonstrated by sentinel site surveillance for visceral leishmaniasis, can effectively detect rising trends and provide accurate patient-level information, significantly improving program effectiveness [48]. Without reliable data, public health interventions may be misdirected, resources inefficiently allocated, and populations left vulnerable.

4.4.2 Risks Associated with Poor Data Quality in Large-Scale Contexts

Poor data quality in large-scale health information systems carries substantial risks beyond immediate clinical and public health impacts. These risks extend to research integrity, financial implications, and public trust. For health research, flawed data can lead to erroneous conclusions, invalidating study findings and misguiding future scientific endeavors. Real-world data, increasingly used in research and economic modeling, requires high quality to ensure the validity of outcomes, yet challenges remain in assessing and improving it [49][35]. Incorrect or biased data inputs can undermine the reliability of predictive models, risk stratification tools, and comparative effectiveness research, potentially leading to flawed policy recommendations and resource allocation. The cost of rectifying poor data quality post-collection can be substantial, often exceeding the cost of implementing quality measures upfront [24].

Furthermore, poor data quality can erode public trust in healthcare systems and institutions. If patients perceive their health information as inaccurate or poorly managed, their willingness to engage with digital health initiatives or share data for broader societal benefit may diminish. This can impede the advancement of learning health systems and the realization of personalized medicine. From a financial perspective, poor data quality contributes to inefficiencies, waste, and increased operational costs, including expenses related to data reconciliation, litigation due to errors, and diminished return on investment in health information technology [50]. The potential for significant financial losses, coupled with compromised patient safety and research integrity, points to the critical necessity of addressing data quality systematically and comprehensively in large-scale health contexts.

4.5 Emerging Trends and Future Directions

4.5.1 Towards Unified, Interoperable Assessment Models

The current landscape of data quality assessment is marked by fragmentation and heterogeneity, prompting a movement towards unified, interoperable assessment models. Such models strive to harmonize definitions, metrics, and methodologies across different health information systems and geographical regions, facilitating meaningful comparisons and collaborative improvement efforts [35][31]. The European Health Data Space (EHDS) initiative, for example, embodies this trend by aiming to create a robust framework for health data management and sharing across Europe, necessitating common data standards and interoperability [2].

Future models will likely emphasize semantic interoperability, ensuring that data not only can be exchanged but also understood and used consistently across disparate systems. This involves the adoption of standardized terminologies, ontologies, and data element definitions, such as those promoted by national and international bodies. The development of modular frameworks that can be adapted to specific contexts while adhering to overarching principles will also be crucial. These frameworks should support both local and cross-institutional data quality assessments, allowing for benchmarking and continuous improvement within and across networks [32]. Collaborative efforts among stakeholders, including government agencies, academic institutions, and industry partners,

will be essential in developing and promoting the adoption of these unified models, paving the way for a more cohesive and reliable global health data ecosystem.

4.5.2 The Role of Artificial Intelligence and Automation in Data Quality Management

Artificial intelligence (AI) and automation are poised to revolutionize data quality management in large-scale health information systems. Traditional manual and rule-based data quality checks are often insufficient for the volume, velocity, and variety of modern health data. AI-driven approaches, particularly machine learning, can automate many aspects of data quality assessment and improvement. Algorithms can detect anomalies, identify missing data patterns, and even predict potential data entry errors with high accuracy [7][44].

For instance, AI can be applied to:

- **Automated Data Profiling:** Discovering data patterns, relationships, and statistical properties to identify inconsistencies and outliers that manual inspection might miss.
- **Intelligent Data Cleansing:** Automatically correcting errors, standardizing formats, and inputting missing values based on learned patterns from high-quality data.
- **Real-time Validation:** Implementing continuous checks at the point of data entry, providing immediate feedback to users and preventing the propagation of errors.
- **Semantic Integration:** Using natural language processing (NLP) to extract structured information from unstructured clinical notes, enhancing completeness and consistency across different data types.
- **Predictive Quality Monitoring:** Forecasting potential data quality issues before they arise by analyzing trends in data collection and system usage.

However, the successful integration of AI requires careful consideration of algorithmic bias, interpretability, and the need for robust validation against ground truth data. Human oversight remains crucial to ensure ethical deployment and to address complex data quality issues that defy algorithmic solutions. The development of advanced AI techniques for better information management is essential to fully harness healthcare datasets for improved delivery and quality of care.

4.6 Contributions to the Study

This study makes the following key contributions:

1. Provides a comprehensive synthesis of data quality assessment frameworks in large-scale health information systems.
2. Identifies fragmentation and inconsistencies in current definitions, metrics, and practices.

3. Introduces a formalized composite scoring model for evaluating data quality dimensions.
4. Proposes a scalable, multi-layered conceptual framework integrating governance and AI-driven automation.
5. Highlights critical gaps in scalability, interoperability, and stakeholder alignment.

These contributions offer both theoretical advancement and practical guidance for improving data quality in complex healthcare environments.

5 Conclusion

This study advances the field of health informatics by critically examining the limitations of existing data quality assessment frameworks and positioning data quality as a strategic, system-wide capability rather than a purely technical function. The findings demonstrate that current approaches are insufficiently equipped to handle the scale, complexity, and heterogeneity of modern health information systems.

By synthesizing theoretical, methodological, and technological perspectives, this research establishes a foundation for next-generation data quality frameworks that are scalable, interoperable, and intelligence-driven. The proposed conceptual direction emphasizes the integration of governance, automation, and stakeholder alignment as essential components for sustainable data quality management.

5.1 Summary of Main Arguments and Contributions

This paper has highlighted the critical dependence of effective health information systems on high-quality data, particularly in large-scale contexts. It systematically reviewed existing data quality assessment frameworks, methodologies, and technological solutions, highlighting the multidimensional nature of data quality across completeness, accuracy, timeliness, consistency, and reliability. A significant finding pertains to the fragmentation and inconsistency in data quality definitions and practices, which complicates comparative analysis and hinders the establishment of unified standards. Furthermore, the varying perspectives of stakeholders - clinicians, data stewards, and policymakers - each prioritizing different data quality dimensions, contribute to the complexity of developing comprehensive assessment strategies. The analysis identified substantial technical, organizational, and regulatory barriers impeding effective data quality assessment at scale, along with critical ethical and privacy considerations inherent in data reuse. The document's contribution lies in its synthesis of these elements, offering a structured overview of the current state and challenges in the field. It also critically evaluates the scalability and adaptability of current frameworks, emphasizing the imperative for tighter integration with data governance and compliance mechanisms.

5.2 Recommendations for Practice and Policy

To advance data quality in large-scale health information systems, several recommendations for practice and policy warrant consideration:

1. **Standardize Data Quality Definitions and Metrics:** Develop and adopt a consensus-based set of data quality definitions and measurable metrics across health systems to enable consistent assessment and comparability.
2. **Implement Integrated Data Governance Frameworks:** Establish comprehensive data governance frameworks that clearly define roles (e.g., data stewards), responsibilities, and accountability for data quality throughout the entire data lifecycle, integrating quality assessment as a core component.
3. **Invest in Automated Data Quality Tools:** Prioritize the development and deployment of AI-driven and automated tools for continuous data profiling, cleansing, and real-time validation to handle the volume and complexity of large-scale health data.
4. **Enhance Training and Education:** Provide continuous training and education for all data producers and consumers, from clinicians to administrative staff, to foster a data quality culture and improve data entry practices.
5. **Promote Interoperability and Data Sharing Standards:** Advocate for and implement national and international standards for health information exchange and semantic interoperability to reduce data heterogeneity and facilitate cross-system quality assessments.
6. **Develop Ethical Guidelines for AI in Data Quality:** Create clear ethical guidelines and frameworks for the use of AI in data quality management, addressing issues of bias, transparency, and accountability, particularly when dealing with sensitive health data.
7. **Foster Multi-stakeholder Collaboration:** Encourage collaborative initiatives involving clinicians, researchers, policymakers, and technology developers to align data quality priorities and co-create effective solutions.

5.3 Future Research Directions

Future research in data quality assessment for large-scale health information systems should focus on several key areas. First, empirical studies are needed to systematically compare the effectiveness and efficiency of different data quality frameworks and methodologies when applied to diverse real-world health datasets. This includes rigorous evaluation of AI-driven data quality tools, examining their accuracy, scalability, and potential for algorithmic bias. Second, research should explore the development of adaptive data quality models that can dynamically adjust to evolving data types, system architectures, and user needs, including the integration of unstructured and patient-generated health data. Third, a deeper investigation into the socio-technical aspects of data quality, particularly the interplay between human factors, organizational culture, and technological interventions, is essential to identify sustainable improvement strategies [37][51]. Finally, studies addressing the economic impact of data quality, quantifying the costs of poor data and the return on investment for quality improvement initiatives, would provide compelling evidence for policymakers and healthcare administrators to prioritize data quality in their strategic agendas.

Moreover, exploring novel methods for ensuring privacy-preserving data quality assessment in distributed networks, while complying with stringent regulations, constitutes another important avenue. This could involve examining federated learning approaches or other privacy-enhancing technologies. Research into the development of unified data quality vocabulary and ontologies could also significantly contribute to reducing definitional inconsistencies and improving semantic interoperability across global health data initiatives. By addressing these areas, the field can collectively move towards more robust, reliable, and impactful health information systems.

References

- [1] F. A. Bernardi, D. Alves, N. Y. Crepaldi, D. B. Yamada, V. C. Lima, and R. P. C. Lopes Rijo, "Data Quality in health research: a systematic literature review." openRxiv, May 31, 2022. doi: 10.1101/2022.05.31.22275804.
- [2] "2.G. Round table: Member States' Health Information Systems: State-of-play and best practices," *European Journal of Public Health*, vol. 33, no. Supplement_2. Oxford University Press (OUP), Oct. 01, 2023. doi: 10.1093/eurpub/ckad160.106.
- [3] W. Warwick, S. Johnson, J. Bond, G. Fletcher, and P. Kanellakis, "A Framework to Assess Healthcare Data Quality," *The European Journal of Social & Behavioural Sciences*, vol. 13, no. 2. Cognitive-crcs, pp. 92–98, May 01, 2015. doi: 10.15405/ejsbs.156.
- [4] S. R. Sukumar, R. Natarajan, and R. K. Ferrell, "Quality of Big Data in health care," *International Journal of Health Care Quality Assurance*, vol. 28, no. 6. Emerald, pp. 621–634, Jul. 13, 2015. doi: 10.1108/ijhcqa-07-2014-0080.
- [5] N. Coppersmith, I. Sarkar, and E. Chen, "Quality Informatics: The Convergence of Healthcare Data, Analytics, and Clinical Excellence," *Applied Clinical Informatics*, vol. 10, no. 02. Georg Thieme Verlag KG, pp. 272–277, Mar. 2019. doi: 10.1055/s-0039-1685221.
- [6] S. L. Feder, "Data Quality in Electronic Health Records Research: Quality Domains and Assessment Methods," *Western Journal of Nursing Research*, vol. 40, no. 5. SAGE Publications, pp. 753–766, Jan. 24, 2017. doi: 10.1177/0193945916689084.
- [7] M. Isgut, L. Gloster, K. Choi, J. Venugopalan, and M. D. Wang, "Systematic Review of Advanced AI Methods for Improving Healthcare Data Quality in Post COVID-19 Era," *IEEE Reviews in Biomedical Engineering*, vol. 16. Institute of Electrical and Electronics Engineers (IEEE), pp. 53–69, 2023. doi: 10.1109/rbme.2022.3216531.
- [8] K. Hoxha, Y. W. Hung, B. R. Irwin, and K. A. Grépin, "Understanding the challenges associated with the use of data from routine health information systems in low- and middle-income countries: A systematic review," *Health Information Management Journal*, vol. 51, no. 3. SAGE Publications, pp. 135–148, Jun. 30, 2020. doi: 10.1177/1833358320928729.
- [9] N. Leon *et al.*, "Routine Health Information System (RHIS) improvements for strengthened health system management," *Cochrane Database of Systematic Reviews*, vol. 2020, no. 8. Wiley, Aug. 13, 2020. doi: 10.1002/14651858.cd012012.pub2.

- [10] S. O. Taiwo, C. K. Amoah-Adjei, and A. Oluwatosin Oladayo , “Evidence-Based Consulting Frameworks for CPG Market Resilience Post Supply-Chain Crises,” *Journal of Computational Analysis and Applications* , vol. 31, no. 04, pp. 6858–6873, 2023, [Online]. Available: <https://eudoxuspress.com/index.php/pub/article/view/4955/3713>
- [11] T. Samuel Oladapo and C. K. Amoah-Adjei, “Financial Risk Optimization in Consumer Goods Using Monte Carlo and Machine Learning Simulations,” Jan. 2022. doi: 10.30574/wjarr.2022.14.1.0385.
- [12] I. C. Okafor, “Edge-Computing Architectures for Real-Time Agricultural Decision Support Using Iot Sensor Networks,” *Journal of Frontiers in Multidisciplinary Research*, vol. 04, no. 02, pp. 329–337, 2023, doi: 10.54660/JFMR.2023.4.2.329-337.
- [13] V. ANUFRIYEVA, M. PAVLOVA, T. STEPURKO, and W. GROOT, “The validity and reliability of self-reported satisfaction with healthcare as a measure of quality: a systematic literature review,” *International Journal for Quality in Health Care*, vol. 33, no. 1. Oxford University Press (OUP), Dec. 10, 2020. doi: 10.1093/intqhc/mzaa152.
- [14] S. O. T. Samuel Oladapo Taiwo and O. O. O. Obianuju O. Okosieme, “AI-Powered Supply Chain Risk Intelligence for Consumer Protection in CPG Distribution Networks,” *International Journal of Scientific Research in Computer Science Engineering and Information Technology*. Technoscience Academy, p. 1008, Nov. 15, 2023. doi: 10.32628/cseit23906782.
- [15] I. C. Okafor, “Designing Secure and High-Performance RESTful APIs for Data-Intensive Analytics Platforms,” *International Journal of Scientific Research in Science Engineering and Technology*. Technoscience Academy, p. 299, Nov. 10, 2019. doi: 10.32628/ijrsrset1985217.
- [16] O. G. Henry-Machame, “Reimagining Enterprise Transformation: A Multi-Dimensional Framework for Sustainable Value Realization,” Apr. 2023. [Online]. Available: <https://eudoxuspress.com/index.php/pub/article/view/4735>
- [17] O. Anifowose, “Augmented Decision Intelligence: Leveraging AI and Predictive Analytics for Executive Strategy Formulation,” *Journal of Computational Analysis and Applications*, vol. 31, no. 3, pp. 750–777, Mar. 2023, [Online]. Available: <https://eudoxuspress.com/index.php/pub/article/view/4136>
- [18] M. O. Lawal, “Next-Generation GRC Framework: Integrating ESG and Cyber Risk Metrics,” *Journal of Computational Analysis and Applications*, vol. 31, no. 3, pp. 778–795, Mar. 2023, [Online]. Available: <https://eudoxuspress.com/index.php/pub/article/view/4141>
- [19] I. C. Okafor, “Visual Analytics for Measuring and Improving Collaborative Software Development in Academic Environments,” *Journal of Frontiers in Multidisciplinary Research*, vol. 1, no. 1. Anfo Publication House, pp. 210–219, 2020. doi: 10.54660/ijfmr.2020.1.1.210-219.
- [20] G. U. Uke, “Lean Six Sigma-Driven Maintenance Process Optimization in African Manufacturing Industries: A Systematic Literature Review,” *Journal of Computational Analysis and Applications*, vol. 29, no. 6, pp. 1346–1366, Jun. 2021, [Online]. Available: <https://eudoxuspress.com/index.php/pub/article/view/4028>

- [21] G. U. Uke, "Circular Economy and Asset Life Extension: Engineering Approaches for Industrial Sustainability," *Journal of Computational Analysis and Applications*, vol. 25, no. 8, pp. 134–152, Aug. 2018, [Online]. Available: <https://eudoxuspress.com/index.php/pub/article/view/4137>
- [22] S. O. Taiwo, O. O. Aramide, and O. R. Tiamiyu, "Blockchain and Federated Analytics for Ethical and Secure CPG Supply Chains," *Journal of Computational Analysis and Applications*, vol. 31, no. 3, pp. 732–749, Mar. 2023, [Online]. Available: <https://eudoxuspress.com/index.php/pub/article/view/4024>
- [23] M. Mashoufi, H. Ayatollahi, D. Khorasani-Zavareh, and T. Talebi Azad Boni, "Data Quality in Health Care: Main Concepts and Assessment Methodologies," *Methods of Information in Medicine*, vol. 62, no. 01/02. Georg Thieme Verlag KG, pp. 005–018, Jan. 30, 2023. doi: 10.1055/s-0043-1761500.
- [24] R. Y. Wang, V. C. Storey, and C. P. Firth, "A framework for analysis of data quality research," *IEEE Transactions on Knowledge and Data Engineering*, vol. 7, no. 4. Institute of Electrical and Electronics Engineers (IEEE), pp. 623–640, 1995. doi: 10.1109/69.404034.
- [25] F. A. Bernardi, D. Alves, N. Crepaldi, D. B. Yamada, V. C. Lima, and R. Rijo, "Data Quality in Health Research: Integrative Literature Review," *Journal of Medical Internet Research*, vol. 25. JMIR Publications Inc., p. e41446, Oct. 31, 2023. doi: 10.2196/41446.
- [26] T. L. McMullen, S. R. Mandl, M. J. Pratt, C. D. Van, B. A. Connor, and A. F. Levitt, "The IMPACT Act of 2014: Standardizing patient assessment data to support care coordination, quality outcomes, and interoperability," *Journal of the American Geriatrics Society*, vol. 70, no. 4. Wiley, pp. 975–980, Mar. 02, 2022. doi: 10.1111/jgs.17644.
- [27] "Data Completeness in Healthcare: A Literature Survey," *Pacific Asia Journal of the Association for Information Systems*. Association for Information Systems, pp. 75–100, 2017. doi: 10.17705/1pais.09204.
- [28] A. Chekol, A. Ketemaw, A. Endale, A. Aschale, B. Endalew, and M. A. Asemahagn, "Data quality and associated factors of routine health information system among health centers of West Gojjam Zone, northwest Ethiopia, 2021," *Frontiers in Health Services*, vol. 3. Frontiers Media SA, Mar. 24, 2023. doi: 10.3389/frhs.2023.1059611.
- [29] M. Saso and N. Schutte, "COVID-19 Health Information System Assessments," *European Journal of Public Health*, vol. 33, no. Supplement_2. Oxford University Press (OUP), Oct. 01, 2023. doi: 10.1093/eurpub/ckad160.108.
- [30] S. Binkheder *et al.*, "Real-World Evidence of COVID-19 Patients' Data Quality in the Electronic Health Records," *Healthcare*, vol. 9, no. 12. MDPI AG, p. 1648, Nov. 28, 2021. doi: 10.3390/healthcare9121648.
- [31] C. R. Lyles, U. Sarkar, U. Patel, S. Lisker, A. Stark, V. Guzman, and A. Patel, "Real-world insights from launching remote peer-to-peer mentoring in a safety net healthcare delivery setting," *Journal of the American Medical Informatics Association*, vol. 28, no. 2, pp. 365–370, Feb. 2021, doi: 10.1093/jamia/ocaa251.

- [32] M. Isgut, L. Gloster, K. Choi, J. Venugopalan and M. D. Wang, "Systematic Review of Advanced AI Methods for Improving Healthcare Data Quality in Post COVID-19 Era," in *IEEE Reviews in Biomedical Engineering*, vol. 16, pp. 53-69, 2023, doi: 10.1109/RBME.2022.3216531.
- [33] J. W. Davey, P. C. Gugiu, and C. L. S. Coryn, "Quantitative Methods for Estimating the Reliability of Qualitative Data," *Journal of MultiDisciplinary Evaluation*, vol. 6, no. 13. The Evaluation Center at Western Michigan University, pp. 140–162, Feb. 04, 2010. doi: 10.56645/jmde.v6i13.266.
- [34] L. Das, A. Kumar, V. Sharma, D. Bhatnagar, S. Singh and N. Tripathi, "Data-Driven Healthcare Management, Analysis, and Future Trends," 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE), Gautam Buddha Nagar, India, 2024, pp. 1692-1698, doi: 10.1109/IC3SE62002.2024.10593217.
- [35] X. Gao and Q. Dong,
"A Bayesian framework for estimating the risk ratio of hospitalization for people with comorbidity infected by SARS-CoV-2 virus," *Journal of the American Medical Informatics Association*, vol. 28, no. 3, pp. 472–476, Mar. 2021, doi: 10.1093/jamia/ocaa246..
- [36] A. Aqil, T. Lippeveld, and D. Hozumi, "PRISM framework: a paradigm shift for designing, strengthening and evaluating routine health information systems," *Health Policy and Planning*, vol. 24, no. 3. Oxford University Press (OUP), pp. 217–228, Mar. 20, 2009. doi: 10.1093/heapol/czp010.
- [37] S. A. Mohammed and M. M. Yusof, "Towards an evaluation framework for information quality management (IQM) practices for health information systems – evaluation criteria for effective IQM practices," *Journal of Evaluation in Clinical Practice*, vol. 19, no. 2. Wiley, pp. 379–387, Apr. 15, 2012. doi: 10.1111/j.1365-2753.2012.01839.x.
- [38] C. Donnelly, A. Janssen, K. Shah, et al., "Qualitative study of international key informants' perspectives on the current and future state of healthcare quality measurement and feedback," *BMJ Open*, vol. 13, p. e073697, 2023, doi: 10.1136/bmjopen-2023-073697.
- [39] M. Kumar, D. Gotz, T. Nutley, and J. B. Smith, "Research gaps in routine health information system design barriers to data quality and use in low- and middle-income countries: A literature review," *The International Journal of Health Planning and Management*, vol. 33, no. 1. Wiley, Aug. 02, 2017. doi: 10.1002/hpm.2447.
- [40] Z. Galavi, M. Montazeri, and L. Ahmadian, "Barriers and challenges of using health information technology in home care: A systematic review," *The International Journal of Health Planning and Management*, vol. 37, no. 5. Wiley, pp. 2542–2568, May 28, 2022. doi: 10.1002/hpm.3492.
- [41] V. Renjith, R. Yesodharan, J. A. Noronha, E. Ladd, and A. George, "Qualitative Methods in Health Care Research," *International Journal of Preventive Medicine*, vol. 12, no. 1. Ovid Technologies (Wolters Kluwer Health), Jan. 2021. doi: 10.4103/ijpvm.ijpvm_321_19.
- [42] A. O. Salami, "Leveraging Natural Language Processing to Detect Non-Compliance in Clinical Documentation: Current Advances, Challenges, and Future Directions,"

International Journal of Scientific Research in Science, Engineering and Technology. Technoscience Academy, pp. 459–473, Oct. 17, 2023. doi: 10.32628/ijrsrset2513822.

[43] N. R. Haug, A. K. Wagner, K. A. McGlynn, C. E. Leonard, M. D. Nguyen, and J. M. Major, “New-Onset Cancer Cases in FDA’s Sentinel System: A Large Distributed System of US Electronic Healthcare Data,” *Cancer Epidemiology, Biomarkers & Prevention*, vol. 31, no. 10. American Association for Cancer Research (AACR), pp. 1890–1895, Jul. 15, 2022. doi: 10.1158/1055-9965.epi-21-1451.

[44] M. Levy and A. Subburaj, “Emerging Trends in Data Center Management Automation,” *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, pp. 0480–0485, Jan. 27, 2021. doi: 10.1109/ccwc51732.2021.9375837.

[45] P. U. Ojiegbo, “Scalability Optimization in Real-Time Payment Systems: Performance Engineering and Fault-Tolerance Strategies,” *Journal of Computational Analysis and Applications*, vol. 29, no. 02, pp. 408–426, Feb. 2021, [Online]. Available: <https://eudoxuspress.com/index.php/pub/article/view/5083>

[46] P. U. Ojiegbo, “Strategic Leadership in Enterprise Digital Transformation: A Systems-Thinking Approach to Scalable Innovation,” *Journal of Computational Analysis and Applications*, vol. 30, no. 02, pp. 1034–1052, Feb. 2022, [Online]. Available: <https://eudoxuspress.com/index.php/pub/article/view/5084>

[47] “Survey results: National health data infrastructure and governance,” Organisation for Economic Co-Operation and Development (OECD), Apr. 2021. doi: 10.1787/55d24b5d-en.

[48] V. N. R. Das *et al.*, “Improved kala-azar case management through implementation of health facility-based sentinel sites surveillance system in Bihar, India,” *PLOS Neglected Tropical Diseases*, vol. 15, no. 8. Public Library of Science (PLoS), p. e0009598, Aug. 24, 2021. doi: 10.1371/journal.pntd.0009598.

[49] J. Kang and J. Cairns, “Cross-sectional analysis of use of real-world data in single technology appraisals of oncological medicine by the National Institute for Health and Care Excellence in 2011–2021,” *BMJ Open*, vol. 14, no. 3. BMJ, p. e077297, Mar. 2024. doi: 10.1136/bmjopen-2023-077297.

[50] M. Islam and Y.-C. (Jack) Li, “Quality improvement in healthcare: the need for valid, reliable and efficient methods and indicators,” *International Journal for Quality in Health Care*, vol. 31, no. 7. Oxford University Press (OUP), pp. 495–496, Aug. 01, 2019. doi: 10.1093/intqhc/mzz077.

[51] J. Elizalde, J. Lumibao, and L. Lizarondo, “Barriers and facilitators to health professionals’ engagement in quality improvement initiatives: a mixed-methods systematic review,” *International Journal for Quality in Health Care*, vol. 36, no. 2. Oxford University Press (OUP), Apr. 2024. doi: 10.1093/intqhc/mzae041.