Assessing and Measuring Quality of Health Data -A Review

Albert Wong[1], Brian Nyakundi[2], Andres Viloria Garcia[2], Kent Liu[2], Keegan Pereira[2], and Yew-Wei Lim[3]

[1]Statistics and Data Analytics, Langara College Vancouver
[2]Statistics and Data Analytics Langara College Vancouver
[3]Sidus Insights Inc. Kelowna

January 02, 2026

Abstract

Data-driven decision-making is paramount across various industries. The quality of the data used in the decision making process is critical. This is especially true in health care. High-quality data is not just a matter of efficiency or accuracy. It directly impacts patient safety, treatment results, and the effectiveness of public health initiatives. This paper reviewed various recent research studies on the understanding and definition of data quality. Using the "fit to use" definition, ideas were extracted from recent literature on the different approaches used in a data quality assessment project that is typically a one-time event. Ideas on the various dimensions of data quality and ways to measure them to ensure data quality on an ongoing basis were also extracted. The importance of quality in healthcare data and the issues arising from its increased use were discussed in several research projects. Finally, research on the challenges of ensuring quality in health care data was highlighted. The results of this research can be used as a roadmap for the development of a data quality measurement subsystem in a health data collection system.

Assessing and Measuring Quality of Health Data - A Review

Albert Wong
Statistics and Data Analytics
Langara College
Vancouver, Canada
0000-0002-0669-4352

Brian Nyakundi
Statistics and Data Analytics
Langara College
Vancouver, Canada
0009-0007-1099-6349

Andres Viloria Garcia
Statistics and Data Analytics
Langara College
Vancouver, Canada
0009-0005-6667-8900

Kent Liu
Statistics and Data Analytics
Langara College
Vancouver, Canada
0009-0000-3627-049X

Keegan Pereira
Sidus Insights Inc.
Kelowna, Canada
0000-0002-2893-3406

Yew-Wei Lim
Statistics and Data Analytics
Langara College
Vancouver, Canada
0009-0008-8999-4633

***Abstract*—Data-driven decision-making is paramount across various industries. The quality of the data used in the decision-making process is critical. This is especially true in health care. High-quality data is not just a matter of efficiency or accuracy. It directly impacts patient safety, treatment results, and the effectiveness of public health initiatives. This paper reviewed various recent research studies on the understanding and definition of data quality. Using the "fit to use" definition, ideas were extracted from recent literature on the different approaches used in a data quality assessment project that is typically a one-time event. Ideas on the various dimensions of data quality and ways to measure them to ensure data quality on an ongoing basis were also extracted. The importance of quality in healthcare data and the issues arising from its increased use were discussed in several research projects. Finally, research on the challenges of ensuring quality in health care data was highlighted. The results of this research can be used as a roadmap for the development of a data quality measurement subsystem in a health data collection system.**

***Index Terms*—Data Quality , Data Quality Score, Health Data**

I. Introduction

The increasing reliance on data-driven decision making across industries, particularly in healthcare, has highlighted the critical importance of data quality. In healthcare, high-quality data is not just a matter of efficiency or accuracy; it directly impacts patient safety, treatment outcomes, and the effectiveness of public health initiatives. As stated in Ehsani-Moghaddam et al. [1], poor data quality can lead to misdiagnoses, ineffective treatments, and flawed research findings, all of which have significant consequences.

This paper extracts ideas from the recent literature on key concepts and issues surrounding data quality, particularly in the health data domain. We will explore the data quality dimensions, examining how each contributes to the overall "fitness for use" of healthcare data. We will delve into traditional and modern approaches to data quality assessment, extracting ideas about the tools and technologies used in this area. In addition, we will examine the frameworks and metrics that were studied to measure data quality. Research on the quality of health data will also be identified and discussed.

The main contribution of this paper is to provide a summary of ideas on concepts and issues on data quality, especially in the field of health data. The paper will also cover ideas for data quality assessment, as well as tools and technologies that can be used to measure data quality.

We organize the paper as follows. Data Quality is defined and discussed in Section 2. The work on dimensions of data quality and data quality measurement is identified and presented in Section 3. Section 4 focuses on ideas generated from research and writing that deal with considerations and issues related to health data. Section 5 addresses selected ideas on other data quality and measurement topics, including case studies that we find helpful for quality management and measurement projects in health data. We discuss possible future work and directions in Section 6. The paper concludes with a summary of what has been presented and discussed.

II. Defining Data quality

To understand data quality and how it should be measured, one needs to define this term carefully in context. Several authors have provided excellent and workable definitions that could be useful.

According to Wang and Strong [2], data quality is defined as "fitness for use," indicating that the data must meet the requirements of users in terms of these attributes. This concept is also supported by Karr et al. [3], who define data quality as the ability of data that can be used effectively, economically, and quickly to inform and evaluate decisions.

Data quality is related to "the processes and technologies for identifying, understanding, and correcting data flaws that support effective information governance in operational business processes and decision making" in Chien et al. [4].

This perspective helps to develop data quality management strategies and processes within an organization.

Other research emphasizes the planning, implementation, and control of quality management techniques to ensure the data are suitable for its intended purpose. See, for example, [5]–[8].

The above points of view are not as valid for organizations that want to share their data or users of the shared data. In this case, the issue is not about governance and correcting flaws but about whether the shared data is fit to use. In this regard, the definition given by Wang and Strong [2] is relevant.

Although the general principles of data quality apply to all domains, the specific requirements for "fitness for use" can vary significantly depending on the context. Ehsani-Moghaddam et al. [1] emphasize that the quality of health data is not just a theoretical ideal, but a practical necessity with tangible consequences for patient care, research validity, and public health surveillance. Inaccurate or incomplete data can lead to misdiagnoses, ineffective treatments, and flawed research findings. Therefore, health data must be "as clean and free of errors as possible" [1]. The only question is, How do we know?

As evidenced by Liu et al. [9], data quality is not static. It evolves with changing user requirements or in the different stages of the so-called Data Evolution Life Cycle [9]. Therefore, the quality of health data must be continuously measured to ensure that it remains fit for its intended purpose.

Based on the above, we will adopt the "fitness for use" perspective in this review.

III. Existing Work

This section presents a comprehensive review of the literature on measuring data quality, emphasizing the work completed on the measurement approach for health data. The various dimensions of measurement data quality are examined and discussed. Different approaches to assessing data quality, a closely related concept to data quality measurement, are also reviewed and presented. The important work on measuring data quality is then presented to close out this review.

Bernardi et al. [10] give a recent review of the literature on data quality in health research. The objective of the review is "to identify and evaluate digital health technology interventions designed to support the conducting of health research based on data quality." The research addresses "the need for standardized practices and collaborative efforts to enhance data quality." However, it does not directly deal with the measurement of data quality. The Canadian Institute of Health Information (CIHI) developed a data quality strategy. In addition to learning about the framework to "systematically analyze, evaluate, document, and improve the data quality of CIHI databases", one could also learn from the evaluation instrument and scoring algorithm developed to measure data quality.

Other reviews include the work by Stausberg [11]. This review covers 39 research papers on data quality between 2005 and 2013 and focuses on "quality indicators". We believe that it is relevant from a measurement point of view.

A. Dimensions of Data Quality

Data quality is a multifaceted concept and its measurement should involve various dimensions that collectively determine its general fitness for use. Laranjeiro et al. [12] and Erlinger et al. [13] both give an overview of the various possible dimensions and references for research in this area.

One can imagine that no consensus could ever be reached on a complete list of dimensions for data quality. With this in mind, the following is a list of data quality dimensions that we believe could be used to develop one or more data quality metrics. See also [1], [5], [13]–[17]:

- Relevancy
- Completeness
- Accuracy
- Validity
- Consistency
- Uniqueness
- Heterogeneity
- Timeliness
- Currency

As highlighted by Mohammed et al. [18], relevancy refers to the extent to which a dataset contains the necessary information to realize an underlying task from the user's perspective. In other words, data are considered relevant if they directly address the needs of the data consumer and effectively support their specific objectives. This concept aligns with the perspective of Black [17], who defines relevancy as the degree to which the composition of the data sets meets the needs of the data consumer.

Completeness refers to the extent to which all required data are present. Ensuring that the data sets are comprehensive and can support accurate analysis is essential. As defined by Weiskopf et al. [19], completeness means recording all the necessary and relevant information about a patient. This is vital for both clinical and research purposes, as incomplete data can lead to incorrect diagnoses, ineffective treatments, and unreliable research results. They highlight that completeness is a fundamental dimension of data quality, directly impacting healthcare data's accuracy and reliability. Loshin [14] explains that completeness is critical, as incomplete data could lead to flawed analyses and biased decisions.

Accuracy refers to the degree to which the data reflect the correct or proper values [5]. In essence, accuracy measures whether the observed data value represents the actual value of the entity or phenomenon in the real world that it is intended to capture [20]. Accurate data are essential for making informed decisions, drawing valid conclusions, and ensuring the effectiveness of interventions. In healthcare, inaccurate data can lead to misdiagnosis, incorrect treatment plans, and potentially harmful consequences for patients. Accuracy is often measured by comparing data against a known gold standard or using statistical techniques to identify outliers.

Validity is the degree to which the data conform to defined business rules or constraints [5]. It ensures that the data are logically and statistically sound. Using the same description, validity ensures that the data values make sense and are logically coherent. Validity checks are essential to assess whether electronic health records (EHR) data are plausible and consistent with other known information. This involves using various techniques to determine whether the values recorded in the EHR are within expected ranges and logically consistent with other data elements. Furthermore, Ehsani-Moghaddam et al. [1] state that without validity, data could lead to incorrect diagnoses, ineffective treatments, and unreliable research results.

Consistency encompasses uniform presentation and compatibility with previous data [16]. Data should be captured and formatted consistently and adhere to established semantic rules [21]. For example, one should consistently record patient weight in the same unit (e.g., kilograms or pounds) across all records. Ehsani-Moghaddam et al. [1] highlight that consistency is often verified through statistical procedures that evaluate the data for uniformity and logical coherence.

The timeliness is a crucial dimension of data quality. It refers to the availability of data within the required time frame. Timely data are critical for making informed decisions in real-time scenarios, such as monitoring disease outbreaks or managing patient care. While Sidi et al. [16] define timeliness narrowly as the delay between a real-world change and its reflection in the information system, Batini and Scannapieco [22] offer a broader perspective. They consider timeliness as the extent to which the age of the data is appropriate for the task at hand, encompassing concepts such as currency and volatility.

Another dimension of time is currency. Black and Nederpelt [17] describe currency as the measurement of the degree to which the data values are up-to-date. As mentioned by Batini and Scannapieco [22], the concept of currency is closely related to timeliness and encapsulates the degree to which data reflect the most current available information. This dimension is vital for applications that require real-time data to make informed decisions.

Mohammed et al. [18] state that uniqueness is whether each entity in the real world is represented by one entity without duplicates in the data set. This perspective aligns with del Pilar Angeles and García-Ugalde [23], who define uniqueness as the extent to which an entity in the real world is represented only once. In healthcare, uniqueness is essential to avoid issues such as duplicate patient records, which can lead to overtreatment, inaccurate analyses, and administrative burdens [1].

Finally, heterogeneity refers to the variation in data formats and structures when data are sourced from multiple sources that health data providers must use as input. Standardizing these diverse data formats is crucial to ensure seamless integration and analysis. According to Ehrlinger and Wöß [13], addressing heterogeneity involves harmonizing data standards to facilitate consistent data usage across different systems.

B. Approaches to Data Quality Assessment

There are many papers on the assessment of data quality. Refer to Chen et al. [24] for an overview. Mohammed et al. [18] also provide a more recent review of data quality assessment (DQA).

Although the term assessment is often used as a synonym for measurement, there is a clear distinction between them. Assessment is a critical process for ensuring the reliability and usability of the data [19]. Therefore, assessment is most likely a one-time event as it serves as an "important starting point for any data quality project to detect critical data that do not meet expectations and to define improvement goals for data cleansing activities." [25]. DQA is also described as the detection and initial estimation of data quality and the impact analysis of DQ problems [26].

Data Quality Assessment (DQA) is a multifaceted process that goes beyond simple measurement. According to Ehrlinger et al. [13], it involves evaluating and interpreting the measured results to conclude the overall quality of a dataset relative to its intended use. Although measurement focuses on quantifying specific characteristics of data quality, assessment extends this process by identifying potential issues and understanding their impact on data utility. It includes detecting and analyzing data quality problems [15], incorporating subjective user perceptions and objective metrics [27], and tailoring the assessment to specific use cases [18]. DQA is a comprehensive approach that combines quantitative and qualitative methods to assess the fitness for the use of data within a specific context.

Given this understanding of DQA, several approaches have been developed to evaluate and enhance the fitness for the use of data. Data auditing is a foundational technique in traditional DQA. As described by Batini et al. [15], data auditing involves systematically examining information bases, schemas, and metadata to identify inconsistencies, errors, or missing values. Chen et al. [24] believe that the major quantitative assessment method (audits) involves a human reviewer examining the data records to identify errors, inconsistencies, and missing values. Although thorough, manual auditing is time-consuming, labor-intensive, and prone to human error. It also notes that, while this is usually acceptable for one-time data quality assessments (DQA), manual auditing can be valuable for small datasets or targeted investigations requiring expert judgment.

Data profiling is often used in conjunction with data auditing. It can be described as the process of analyzing a dataset by collecting data (metadata) on it [28], [29]. In addition to being used in the assessment process, the profiling task is an essential step toward any measurement or monitoring activity. The metadata collected in the profiling step includes the number of distinct or missing (i.e., null) values in a variable (column), the data types of attributes or patterns, and their frequency [30].

Other traditional tools are surveys and questionnaires, which provide insight into subjective perceptions and experiences of data users and stakeholders. As highlighted by Pipino et al. [27], subjective data quality assessments can vary significantly depending on the roles and perspectives of different stakeholders. For example, data custodians (often IT professionals) may prioritize technical aspects such as timeliness, while data consumers (business users) may focus on usability and relevance to their specific tasks. Surveys and questionnaires can help capture these diverse viewpoints and identify potential areas of conflict or misalignment.

Statistical techniques could be used in a DQA study to analyze data distributions, identify outliers, and detect anomalies. As discussed by J. Alipour et al. [29], statistical profiling provides a quantitative overview of data quality, revealing patterns and trends that manual inspection may not reveal. However, specialized knowledge may be required to interpret the results effectively.

As emphasized by Mohammed et al. [18], metadata is crucial for incorporating external knowledge into the DQA process, enabling the validation and contextualization of data. It also plays a vital role in ensuring the scalability of assessment methods, as metadata can help manage and organize large volumes of data efficiently. The importance of metadata is further highlighted by Batini et al. [22], who stress its relevance at all stages of the assessment process. Metadata provides valuable information on data schemas, architectural rules, and management processes, facilitating a comprehensive understanding of the data landscape. However, as noted by Mohammed et al. [18], the availability and quality of metadata can pose challenges in DQA. Data catalogs can provide support, but may need to be extended to assess metadata quality. Al-Salim [5] echoes this concern, identifying gaps in metadata management within the UN Sustainable Development Goals report, including inconsistencies and lack of completeness.

In the healthcare context, Weiskopf and Weng [31] used semi-structured interviews to explore the views of clinical researchers on EHR data quality and reuse. This approach helped uncover the practical challenges and concerns of researchers when working with healthcare data. Chen et al. [24] further emphasize the importance of surveys and interviews in public health DQA, as they can be used to assess data use and the data collection process, two critical dimensions often overlooked in quantitative assessments. Although surveys and interviews offer valuable qualitative insights, Mohammed et al. [18] acknowledge that these methods are inherently subjective and may not be objectively quantifiable. However, they stress the importance of incorporating user feedback in DQA, particularly to assess dimensions such as ease of manipulation and understanding, which are best evaluated through subjective experiences. They propose using automated tools to generate and assess questionnaires based on sound user survey design principles while allowing domain experts to design surveys manually.

As noted by Kahn et al. [32], health data are often aggregated from multiple sources, and therefore DQA is particularly important. They proposed a "fit-for-use" conceptual model for DQA and a process model to plan and conduct one. From their perspective, a DQA should include a standardized approach, a targeted and prioritized list of variables and data quality dimensions and domains vulnerable to data quality problems, and an "iterative cycle of assessments within and between data collection sites". It is also essential to document the rationale and results of DQA.

Weiskopf et al. [31] introduce the 3x3 Data Quality Assessment (3x3 DQA), a set of guidelines for DQA and its associated reporting for health data in clinical research. Developed through a triangulation of the results of three different studies and reviewed by a panel of data quality experts, this tool includes the three key constructs of data quality: completeness, correctness, and currency, which are "operationalized according to the three primary dimensions of EHR data: patients, variables, and time."

Although traditional approaches such as auditing, profiling, and surveys remain valuable for in-depth and contextual understanding, contemporary approaches to DQA leverage automation and advanced technologies to address the challenges of increasing data volume and complexity. These approaches include the use of machine learning for anomaly detection discussed by Mohammed et al. [18], predictive modeling discussed by Bayram et al. [20], as well as automated tools for data profiling, cleaning and monitoring which Ehrlinger and Wöß [13] discuss.

A tool developed to automate data quality verification and allow assessment and measurement is described in Schelter et al. [33]. This tool provides for a "declarative API" that supports, through machine learning, quality assessment and validation incrementally as the dataset grows. A similar tool, called QualIe (Quality Assessment for Integrated Information Environments), was developed and described by Ehringer et al. [25]

C. Measuring Data Quality

The increasing importance of data in decision making and business operations has highlighted the need for effective measurement and, therefore, the management of its quality "to ensure that the data continue to comply with requirements and to detect unexpected changes in the data" [25]. According to Kaiser et al., [34], measuring data quality is essential to compare data sets, track changes, and evaluate improvement initiatives. After all, what gets measured gets managed. This section explores the various approaches and metrics used to measure data quality.

Stausberg et al. [11] present a comprehensive survey of articles on data quality (DQ) management in registries and cohort studies. This review highlights the importance of quality indicators, feedback, and source data verification in measurement. Similarly, Ehrlinger and Wöß [13] emphasize the need for systematic reviews of the literature to bridge the gap between theoretical research and practical implementation. Both studies highlight the value of detailed reviews in identifying critical metrics and best practices to ensure high data quality, thus facilitating better data-driven decision making.

Gray and Weng [19] argue that effective data quality measurement goes beyond simply applying metrics. A holistic approach requires a deep understanding of the broader context in which the data exist. This includes understanding the data's origins, storage, movement within the organization, and intended use. The author stressed the concept of "fitness for use", emphasizing that data quality is relative to how well it serves its specific purpose, with different use cases demanding varying quality requirements. Consequently, the research focuses on five key dimensions: completeness, timeliness, validity, consistency, and integrity. This user-centric perspective aligns with Helfert's [35] notion that data quality assessment should be tailored to the specific task and context. However, Helfert proposes a framework grounded in semiotics and quality's intrinsic and contextual aspects. Despite these differences, both frameworks advocate for a comprehensive and adaptable approach, using statistical methods, data mining, and qualitative research tools such as questionnaires to quantify data quality characteristics. In addition, both authors champion the idea of continuous measurement to track trends, identify emerging issues, and continuously improve data quality over time.

Statistical approaches offer robust quantitative methods to measure data quality. Chug et al. [36] present an empirical study to formulate an automated data quality platform to evaluate the quality of a data set and generate a quality label using principal component analysis (PCA). The concepts are demonstrated using data from healthdata.gov, open-data.nhs, and the Demographics and Health Surveys (DHS) Program. Similarly, Farzi and Dastjerdi [37] propose a novel approach utilizing data mining algorithms to extract association rules, which are then used to assess the quality of the input data. Their three-step algorithm simplifies the extraction process, focusing only on relevant rules, thus reducing the time and space complexity. In contrast, Islam et al. [38] concentrate on the impact of data perturbations on the predictive accuracy of decision trees and neural networks, using datasets from the UCI Machine Learning Repository. Their findings reveal discrepancies between predictive accuracy and decision tree similarity as data quality measures. The authors suggested that the choice of statistical metric should be tailored to the specific application and desired outcomes.

Modern approaches to DQ scoring aim to provide a comprehensive and efficient assessment of "fitness for use". Bayram et al. [20] introduce the Data Quality Scoring Operations (DQSOps) framework, which leverages machine learning (ML) to predict DQ scores based on accuracy, completeness, consistency, timeliness, and skewness. This framework combines the efficiency of ML predictions with periodic ground-truth calculations to ensure reliability. Although DQSOps offers a streamlined approach, Vaziri et al. [39] emphasize the importance of considering the varying significance of data elements within an organization. Their proposed "weighted metrics" assign weights to data based on their relevance to business goals, providing a more nuanced and context-aware measure of DQ. By combining these perspectives, organizations can develop a DQ scoring system that provides accurate and timely assessments, aligns with the "fitness for use" principle, tailors the metrics to the specific context and purpose of the data, and therefore improves data-driven decision-making.

Specialized approaches have emerged to address the unique challenges posed by different data types and applications in DQ measurement. Mishra et al. [40] introduce the Data Quality Index (DQI) framework, designed specifically to measure the quality of Natural Language Processing (NLP) datasets. The DQI utilizes multiple granularities (e.g., sentences, words) and various metrics (e.g., consistency, coverage) to assess NLP data quality comprehensively. Similarly, Swazinna et al. [41] focus on the specific challenges of DQ measurement in offline reinforcement learning (RL). They introduce two indicators, Estimated Relative Return Improvement and Estimated Action Stochasticity, designed to evaluate the potential value of datasets for training RL models. Both approaches highlight the importance of tailoring data quality measurement methodologies to the specific characteristics and requirements of different types and applications of data, emphasizing the need for specialized metrics beyond traditional general-purpose data quality dimensions.

As emphasized by Chien and Jain [4], DQ measurement tools have evolved to incorporate automation, machine learning, and cloud-based deployment models, addressing the complexities of modern data environments and the need for real-time analytics. They offer a comprehensive suite of functionalities, including profiling, cleansing, standardization, matching, and monitoring, to ensure data quality across diverse business applications.

Specialized tools such as ACHILLES Heel offer predefined data quality rules and customization options tailored to patient-level clinical datasets [42]. Although valuable, such specialized tools are often used in conjunction with other methodologies. For example, Huser et al. [42] find that organizations frequently employ a combination of SQL queries, custom scripts, and additional tools such as WhiteRabbit and Rabbit-In-a-Hat for comprehensive DQA. Using common data models further enhances the ability to apply uniform quality checks across organizations, fostering collaboration and data sharing. The market for these tools, as described by Chien and Jain [4], comprises established vendors offering comprehensive solutions with advanced analytics and smaller niche players focusing on specific industries or data domains. The optimal choice of tools and technologies ultimately depends on the organization's particular needs, data types, available resources, and desired level of automation.

IV. Quality of Health Data

In addition to the work on data quality presented above, ideas from the work completed on the quality of health data are presented below.

A. Importance of Quality in Health Data

The volume of health data has increased dramatically in recent years, driven by the proliferation of digital health records, wearable devices, and other data-generating technologies. This exponential growth has created significant challenges in managing and processing the vast amounts of information generated in the healthcare sector. As highlighted by Al-Salim et al. [5], advanced methodologies are crucial to ensure the quality of these data and prevent erroneous decisions. Furthermore, increasing digitization of medical records contributes to data deluge, providing vast amounts of information for analysis (Dash et al., 2019) [43]. Alberto et al. [44] further emphasize the role of cloud data storage, distributed computing, and machine learning in facilitating this expansion. Integrating new data sources, such as big data, remote sensing, and satellite imagery, into health data systems presents opportunities and challenges. As Al-Salim et al. [5] note, while these innovative data sources can enhance data quality and fill existing gaps, they also require the development of robust methodologies to ensure their effective integration and utilization.

High-quality health data are crucial for informed decision-making, effective research, and better patient care. Navaz et al. (2023) and Daneshkohan et al. [45] note the critical role of high-quality data in primary care, clinical decision-making, and effective health policies. Schmidt et al. [46] stress the need for consistent data quality assessments and proper metadata use to ensure data sharing and reuse across studies, ultimately benefiting healthcare practices. Studies show that data errors can result in adverse outcomes, including medication errors and misdiagnoses [31].

Furthermore, the exponential growth in health data, as noted by Lewis et al. [47], has expanded its use in biomedical research, making robust data quality assessments even more critical to ensure the reliability of these datasets. Alan F. Karr et al. [3] underscore the need for strategic data management in the face of increasing data volume and complexity, emphasizing the crucial role of data quality frameworks in facilitating effective decision-making and positive health outcomes.

High-quality health data is essential for accurate clinical decision-making and effective treatment outcomes, as it directly impacts the credibility and reliability of healthcare systems [48]. Juddoo et al. [49] reinforce this perspective by highlighting the importance of maintaining high data quality within the health industry, particularly in terms of accuracy, completeness, and consistency. They argue that robust data governance frameworks are critical in managing the increasing volume of health data, ensuring that the data remain reliable and fit for use. This is crucial for the integrity of data-driven decisions that directly affect patient outcomes.

Operational efficiency also depends on accurate data for resource management and cost reduction [1]. Regulatory compliance requires accurate data to meet standards and avoid penalties [13]. On the other hand, strategic decision making benefits from high-quality data, enabling administrators to make informed decisions about resource allocation and policy development [1], [50]. Public health surveillance relies on accurate and timely data for an effective response to health emergencies [42].

The quality of health data is crucial to maximize the benefits of Big Data in healthcare. High-quality data supports these advances, from improving personalized medicine and reducing patient readmission rates to detecting fraud. Juddoo et al. [49] emphasize that without ensuring key dimensions like accuracy, completeness, and timeliness, the effectiveness of Big Data applications in healthcare can be significantly compromised, potentially leading to inaccurate insights and poor decision-making.

With the rapid advancement in the use of artificial intelligence (AI) models in health care [51], the quality of the data used in the development process is significantly elevated due to "its heightened downstream impact, impacting predictions like cancer detection." [52]. Unfortunately, data can be the most undervalued aspect of an AI project.

B. Data Quality Issues due to Increased Use and Sharing

The exponential increase in data generation across various sectors, driven by technological advancements and the utilization of new data forms, has significantly expanded data use and sharing. However, this surge in data usage raises several challenges that must be carefully managed to maintain data quality.

One of the primary challenges is the presence of inaccuracies and inconsistencies in the data, which can undermine the reliability of insights derived from big data. Lindström et al. [53] emphasize the critical need to address these issues to ensure data remains trustworthy and valuable. Adding to this issue is maintaining consistent data quality between different systems and stakeholders. Al-Salim et al. [5] highlight how varying quality assessment frameworks between countries can lead to difficulty integrating data from multiple sources, often resulting in incomplete or inaccurate data sets. Thus, harmonizing these frameworks is crucial for ensuring consistent data quality across platforms.

In addition to consistency, the timeliness and completeness of the data are essential for effective decision making. Al-Salim et al. [5] further note that outdated and incomplete data can severely hinder decision-making processes, particularly in critical fields like healthcare, where timely data are vital for responding to emerging health issues. Ehsani-Moghaddam et al. [1] stress the importance of continuously updating health databases to maintain their relevance and utility. Achieving this in a shared data environment requires robust mechanisms for regular updates and comprehensive data collection processes.

Another significant challenge arises from the increasing prevalence of unstructured, schema-less data, often collected from multiple sources. Taleb et al. [54] point out that managing and ensuring the quality of such diverse data types requires a robust integrated framework that spans the entire data lifecycle, from collection to processing and analysis.

As data sharing becomes more widespread, these challenges are further complicated by the heterogeneity of the data sources, making traditional data quality assessment methods less effective. Byabazaire et al. [55] argue that each Internet of Things (IoT) domain presents unique data quality challenges, requiring customized approaches to ensure data integrity. Therefore, the need for comprehensive data quality frameworks is more pressing than ever, as these frameworks must manage the complexities introduced by large volumes and varied data types. Cai and Zhu [56] emphasize that structured approaches are essential to maintain data integrity and support effective decision-making.

C. Challenges in Ensuring Data Quality in Healthcare

Ensuring data quality in healthcare is particularly challenging due to the fragmented nature of patient care and the complexity of electronic health records (EHRs). D'Amore et al. [57] discuss how EHRs often contain incomplete and poorly integrated data, leading to significant inaccuracies in quality measurements. They emphasize that the lack of integration between different healthcare systems causes gaps in patient records, which, in turn, can result in clinical decision-making based on partial or outdated information. Syed et al. [58] build on this by highlighting the critical importance of consistency, completeness, and accuracy in digital health data, noting that errors arising from inconsistent or incomplete records not only affect quality assessments, but also pose serious risks to patient safety. This issue is further illustrated by Moloko and Ramukumba [59] and Gass et al. [60], who highlight how organizational factors and complex data collection processes in both the Tshwane District, South Africa, and the BetterBirth Trial in Uttar Pradesh, India, undermine data accuracy and reliability.

Furthermore, Cho et al. [61] provide examples of how the lack of standardization in data from wearable devices, such as varying measurement units and proprietary algorithms, complicates the integration of data across different devices, leading to challenges in conducting accurate and consistent biomedical research. Syed et al. [58] further highlights that even within more traditional healthcare settings, the lack of standardized terminology, diagnostic codes, and workflows contributes to inconsistencies in data entry, affecting the reliability of health records. They note that despite efforts to establish standards, differences in health information systems between settings and varying adherence to these standards by staff continue to hinder data concordance. Addressing these challenges through improved training, supportive supervision, simplified data tools, and robust standardization efforts is essential to improve healthcare data quality and ensure better patient outcomes.

V. Discussions

As documented above, before putting health data to good use, one has to recognize the importance of data quality and deal effectively with its challenges. Understanding data quality and its various dimensions is an important first step.

To effectively manage the quality of health data, one must measure it consistently using the various ideas and tools available. Implementing a data quality measurement subsystem as part of the health data collection system is critical. The research reviewed here should be sufficient as a starting point for researchers and developers embarking on a challenging project in this area.

VI. Conclusions

This paper reviews ideas in the recent literature on key concepts and issues surrounding data quality, particularly in the health data domain. We explore the dimensions of data quality, examining how each contributes to the overall "fitness for use" of healthcare data. We dive into traditional and modern approaches to assessing data quality [24], highlighting the tools and technologies used in this area. In addition, we examine frameworks and metrics that were studied to measure data quality. Research on the quality of health data is also reviewed and discussed.

By addressing these multifaceted aspects of data quality, this review aims to equip healthcare professionals, researchers, and data managers with the knowledge and tools needed to ensure that the data they rely on is truly fit for its intended purpose, ultimately leading to improved patient care and better health.

Acknowledgements

The authors acknowledge financial support through the National Science and Engineering Research Council (NSERC) and administrative support from the Applied Research Center of Langara College.

References

[1] B. Ehsani-Moghaddam, K. Martin, and J. A. Queenan, "Data quality in healthcare: A report of practical experience with the Canadian Primary Care Sentinel Surveillance Network data," *Health Information Management Journal*, vol. 50, no. 1-2, pp. 88–92, 1 2021.

[2] R. Y. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996.

[3] A. F. Karr, A. P. Sanil, and D. L. Banks, "Data quality: A statistical perspective," *Statistical Methodology*, vol. 3, no. 2, pp. 137–173, 2006.

[4] M. Chien and A. Jain, "Magic Quadrant for Data Quality Tools," Gartner, Tech. Rep., 2019.

[5] W. Al-Salim, A. S. K. Darwish, and P. Farrell, "Analysing data quality frameworks and evaluating the statistical output of United Nations Sustainable Development Goals' reports," *Renewable Energy and Environmental Sustainability*, vol. 7, p. 17, 2022.

[6] C. Cichy and S. Rass, "An overview of data quality frameworks," *IEEE Access*, vol. 7, pp. 24 634–24 648, 2019.

[7] J. Merino, I. Caballero, B. Rivas, M. Serrano, and M. Piattini, "A data quality in use model for big data," *Future Generation Computer Systems*, vol. 63, pp. 123–130, 2016.

[8] S. Dungey, N. Beloff, S. Puri, R. Boggon, T. Williams, and A. R. Tate, "A pragmatic approach for measuring data quality in primary care databases," in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2014, pp. 797–800.

[9] L. Liu and L. Chi, "Evolutional Data Quality: A Theory-Specific View." in *ICIQ*, 2002, pp. 292–304.

[10] 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, 2023.

[11] J. Stausberg, D. Nasseh, and M. Nonnemacher, “Measuring data quality: a review of the literature between 2005 and 2013,” *Digital Healthcare Empowering Europeans*, pp. 712–716, 2015.

[12] N. Laranjeiro, S. N. Soydemir, and J. Bernardino, “A Survey on Data Quality: Classifying Poor Data,” in *IEEE 21st Pacific rim international symposium on dependable computing (PRDC)*, 2015, pp. 179–188.

[13] L. Ehrlinger and W. Wöß, “A survey of data quality measurement and monitoring tools,” *Frontiers in big data*, vol. 5, p. 850611, 2022.

[14] D. Loshin, *The practitioner’s guide to data quality improvement*. Elsevier, 2010.

[15] C. Batini and M. Scannapieco, “Data and Information Quality,” *Cham, Switzerland: Springer International Publishing, 63.*, 2016. [Online]. Available: http://www.springer.com/series/5258

[16] F. Sidi, P. H. S. Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, “Data quality: A survey of data quality dimensions,” in *2012 International Conference on Information Retrieval & Knowledge Management*. IEEE, 2012, pp. 300–304. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6204995

[17] A. Black and P. van Nederpelt, “Dimensions of data quality (DDQ): research paper,” *DAMA NL Foundation*, pp. 1–113, 2020. [Online]. Available: https://www.dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper-version-1.2-d.d.-3-Sept-2020.pdf

[18] S. Mohammed, H. Harmouch, F. Naumann, and D. Srivastava, “Data Quality Assessment: Challenges and Opportunities,” *arXiv preprint arXiv:2403.00526.*, 3 2024. [Online]. Available: http://arxiv.org/abs/2403.00526

[19] N. G. Weiskopf and C. Weng, “Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research,” *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 144–151, 2013.

[20] F. Bayram, B. S. Ahmed, E. Hallin, and A. Engman, “DQSOps: Data Quality Scoring Operations Framework for Data-Driven Applications,” in *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, 3 2023, pp. 32–41. [Online]. Available: http://arxiv.org/abs/2303.15068

[21] A. Bronselaer, R. De Mol, and G. De Tré, “A measure-theoretic foundation for data quality,” *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 2, pp. 627–639, 2017.

[22] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, “Methodologies for data quality assessment and improvement,” *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–52, 2009.

[23] M. del Pilar Angele and F. García-Ugalde, “A Data Quality Practical Approach,” *International Journal on Advances in Software, 1(2&3)*, 2009. [Online]. Available: http://www.iariajournals.org/software/www.iaria.org

[24] H. Chen, D. Hailey, N. Wang, and P. Yu, “A review of data quality assessment methods for public health information systems,” pp. 5170–5207, 5 2014.

[25] L. Ehrlinger, B. Werth, and W. Wöß, “Automated continuous data quality measurement with QualIe,” *International Journal on Advances in Software*, vol. 11, no. 3, pp. 400–417, 2018.

[26] L. P. English, *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. John Wiley & Sons, Inc., 1999.

[27] L. L. Pipino, Y. W. Lee, R. Y. Wang, and R. Y. Yang, “Data Quality Assessment,” *Communications of the ACM*, vol. 45, no. 4, pp. 211–218, 2002.

[28] F. Naumann, “Data profiling revisited,” *ACM SIGMOD Record*, vol. 42, no. 4, pp. 40–49, 2014.

[29] J. Alipour and M. Ahmadi, “Dimensions and assessment methods of data quality in health information systems,” *Acta Medica Mediterranea*, vol. 33, no. 2, pp. 313–320, 2017.

[30] Z. Abedjan, L. Golab, and F. Naumann, “Profiling relational data: a survey,” *The VLDB Journal*, vol. 24, pp. 557–581, 2015.

[31] N. G. Weiskopf, S. Bakken, G. Hripcsak, and C. Weng, “A Data Quality Assessment Guideline for Electronic Health Record Data Reuse,” *Egems, 5(1).*, 2017.

[32] M. G. Kahn, M. A. Raebel, J. M. Glanz, K. Riedlinger, and J. F. Steiner, “A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research,” *Medical care*, vol. 50, 2012.

[33] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Grafberger, “Automating large-scale data quality verification,” *Proceedings of the VLDB Endowment*, vol. 11, no. 12, pp. 1781–1794, 2018.

[34] M. Kaiser, M. Klier, and B. Heinrich, “How to measure data quality?-a metric-based approach,” in *ICIS 2007 Proceedings*, 2007, p. 108.

[35] M. Helfert, “Managing and measuring data quality in data warehousing,” in *Proceedings of the world multiconference on systemics, cybernetics and informatics*, vol. 38. Citeseer, 2001.

[36] S. Chug, P. Kaushal, P. Kumaraguru, and T. Sethi, “Statistical learning to operationalize a domain agnostic data quality scoring,” *arXiv preprint arXiv:2108.08905*, 2021.

[37] S. Farzi and A. B. Dastjerdi, “Data quality measurement using data mining,” *International Journal of Computer Theory and Engineering*, vol. 2, no. 1, p. 115, 2010.

[38] M. Z. Islam, P. M. Barnaghi, and L. Brankovic, “Measuring data quality: Predictive accuracy vs. similarity of decision trees,” in *6th International Conference on Computer & Information Technology*, vol. 2. Citeseer, 2003, pp. 457–462.

[39] R. Vaziri, M. Mohsenzadeh, and J. Habibi, “Measuring data quality with weighted metrics,” *Total Quality Management & Business Excellence*, vol. 30, no. 5-6, pp. 708–720, 2019.

[40] S. Mishra, A. Arunkumar, B. Sachdeva, C. Bryan, and C. Baral, “Dqi: Measuring data quality in nlp,” *arXiv preprint arXiv:2005.00816*, 2020.

[41] P. Swazinna, S. Udluft, and T. Runkler, “Measuring data quality for dataset selection in offline reinforcement learning,” in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2021, pp. 1–8.

[42] V. Huser, F. J. DeFalco, M. Schuemie, P. B. Ryan, N. Shang, M. Velez, R. W. Park, R. D. Boyce, J. Duke, R. Khare, L. Utidjian, and C. Bailey, “Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Datasets,” *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, vol. 4, no. 1, p. 24, 11 2016.

[43] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, “Big data in healthcare: management, analysis and future prospects,” *Journal of Big Data*, vol. 6, no. 1, 12 2019.

[44] I. R. I. Alberto, N. R. I. Alberto, A. K. Ghosh, B. Jain, S. Jayakumar, N. Martinez-Martin, N. McCague, D. Moukheiber, L. Moukheiber, M. Moukheiber, S. Moukheiber, A. Yaghy, A. Zhang, and L. A. Celi, “The impact of commercial health datasets on medical research and health-care algorithms,” pp. e288–e294, 5 2023.

[45] A. Daneshkohan, M. Alimoradi, M. Ahmadi, and J. Alipour, “Data quality and data use in primary health care: A case study from Iran,” *Informatics in Medicine Unlocked*, vol. 28, 1 2022.

[46] C. O. Schmidt, S. Struckmann, C. Enzenbach, A. Reineke, J. Stausberg, S. Damerow, M. Huebner, B. Schmidt, W. Sauerbrei, and A. Richter, “Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R,” *BMC Medical Research Methodology*, vol. 21, no. 1, 12 2021.

[47] A. E. Lewis, N. Weiskopf, Z. B. Abrams, R. Foraker, A. M. Lai, P. R. Payne, and A. Gupta, “Electronic health record data quality assessment and tools: a systematic review,” *Journal of the American Medical Informatics Association : JAMIA*, vol. 30, no. 10, pp. 1730–1740, 9 2023.

[48] M. N. Zozus, W. E. Hammond, B. B. Green, M. G. Kahn, R. L. Richesson, S. A. Rusincovitch, G. E. Simon, and M. M. Smerek, “Assessing Data Quality,” NIH Health Care Systems Research Collaboratory., Tech. Rep. 2, 2014.

[49] S. Juddoo, C. George, P. Duquenoy, and D. Windridge, “Data governance in the health industry: Investigating data quality dimensions within a big data context,” *Applied System Innovation*, vol. 1, no. 4, pp. 1–16, 12 2018.

[50] T. Williamson, M. E. Green, R. Birtwhistle, S. Khan, S. Garies, S. T. Wong, N. Natarajan, D. Manca, and N. Drummond, “Validating the 8 CPCSSN case definitions for chronic disease surveillance in a primary care database of electronic health records,” *Annals of Family Medicine*, vol. 12, no. 4, pp. 367–372, 2014.

[51] R. Manne and S. C. Kantheti, “Application of artificial intelligence in healthcare: chances and challenges,” *Current Journal of Applied Science and Technology*, vol. 40, no. 6, pp. 78–89, 2021.

[52] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, ““Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI,” in *CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–15.

[53] V. Lindström, F. Persson, A. P. C. Viswanathan, and M. Rajendran, "Data quality issues in production planning and control – Linkages to smart PPC," *Computers in Industry*, vol. 147, 5 2023.

[54] I. Taleb, M. A. Serhani, C. Bouhaddioui, and R. Dssouli, "Big data quality framework: a holistic approach to continuous quality management," *Journal of Big Data*, vol. 8, no. 1, 12 2021.

[55] J. Byabazaire, G. O'hare, and D. Delaney, "Data quality and trust: Review of challenges and opportunities for data sharing in IoT," *Electronics (Switzerland)*, vol. 9, no. 12, pp. 1–22, 12 2020.

[56] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," in *Data Science Journal*, vol. 14. Committee on Data for Science and Technology, 2015.

[57] J. D. D'amore, L. K. Mccrary, J. Denson, C. Li, C. J. Vitale, P. Tokachichu, D. F. Sittig, A. B. Mccoy, and A. Wright, "Clinical data sharing improves quality measurement and patient safety," *Journal of the American Medical Informatics Association*, vol. 28, no. 7, pp. 1534–1542, 7 2021.

[58] R. Syed, R. Eden, T. Makasi, I. Chukwudi, A. Mamudu, M. Kamalpour, D. K. Geeganage, S. Sadeghianasl, S. J. Leemans, K. Goel, R. Andrews, M. T. Wynn, A. ter Hofstede, and T. Myers, "Digital Health Data Quality Issues: Systematic Review," 2023.

[59] S. M. Moloko and M. M. Ramukumba, "Healthcare providers' views of factors influencing family planning data quality in Tshwane District, South Africa," *African Journal of Primary Health Care and Family Medicine*, vol. 14, no. 1, 2022.

[60] J. D. Gass, A. Misra, M. N. S. Yadav, F. Sana, C. Singh, A. Mankar, B. J. Neal, J. Fisher-Bowman, J. Maisonneuve, and M. M. Delaney, "Implementation and results of an integrated data quality assurance protocol in a randomized controlled trial in Uttar Pradesh, India," *Trials*, vol. 18, pp. 1–9, 2017.

[61] S. Cho, C. Weng, M. G. Kahn, and K. Natarajan, "Identifying Data Quality Dimensions for Person-Generated Wearable Device Data: Multi-Method Study," *JMIR mHealth and uHealth*, vol. 9, no. 12, 12 2021.