Digital Health Data Quality Issues: Systematic Review

Abstract

Background: The promise of digital health is principally dependent on the ability to electronically capture data which can be analysed to improve decision making. Yet, the ability to effectively harness data has proven elusive, which has largely been due to the quality of data captured. Despite the importance of data quality (DQ), an agreed upon DQ taxonomy evades literature. When consolidated frameworks are developed, the dimensions are often fragmented, without consideration of the interrelationships between the dimensions or their resultant impact.
Objective: The aim of this study was to develop a consolidated digital health DQ dimensions and outcomes framework, which provided insights into the three research questions: 1) What are the dimensions of digital health DQ? 2) How are the dimensions of digital health DQ related? And 3) What are the impacts of digital health DQ?
Methods: Following PRISMA guidelines, a developmental systematic literature review was conducted of peer-reviewed literature focussing on digital health DQ in predominately hospital settings. A total of 227 relevant articles were retrieved that were inductively analysed to identify digital health DQ dimensions and outcomes. The articles were inductively analysed, using open coding, constant comparison, and card-sorting with subject matter experts to identify the digital health DQ dimensions and digital health DQ outcomes. Subsequently, computer-assisted analysis was performed and verified by DQ experts to identify: the interrelationships between the DQ dimensions; and, relationships between DQ dimensions and outcomes. The analysis resulted in the development of the DQ dimensions and outcomes (DQ-DO) framework.
Results: The digital health DQ-DO framework consists of 1) six dimensions of DQ: accessibility, accuracy, completeness, consistency, contextual validity, and currency; 2) interrelationships amongst the dimensions of digital health DQ, with consistency being the most influential dimensions impacting all other digital health DQ dimensions; 3) Five digital health DQ outcomes: clinical, clinician, research-related, business processes, and organizational outcomes; and 4) relationships between the digital health DQ dimensions and DQ outcomes; with the consistency and accessibility dimensions impacting all DQ outcomes.
Conclusions: The DQ-DO framework developed in this study demonstrates the complexity of digital health data quality and the necessity for reducing digital health data quality issues. The framework further provides healthcare executives with holistic insights into DQ issues and resultant outcomes, which can help them prioritise which DQ-related problems to tackle.

Keywords: Data quality; digital health; electronic health record; eHealth; systematic reviews.

Introduction

Background

The healthcare landscape is changing globally owing to substantial investments in health information systems which seek to improve healthcare outcomes [1]. Despite the rapid adoption of health information systems [2] and the perception of digital health as

a panacea [3] for improving healthcare quality, the outcomes have been mixed [4, 5]. As Reisman [6] notes, despite substantial investment, effort, and widespread application of digital health, many of the promised benefits have yet to be realized.

The promise of digital health is principally dependent on the ability to electronically capture data which can be analysed to improve decision making at local, national [6], and global levels [7]. However, the ability to harness data effectively and meaningfully has proven difficult and elusive, which has largely been due to the quality of data captured. Darko-Yawson and Ellingsen [8] highlight that digital health has resulted in more bad data rather than improving the quality of data. It is widely accepted that the data from digital health are plagued by accuracy and completeness concerns [9-12]. Poor data quality (DQ) can be detrimental to continuity of care [13], patient safety [14], clinician productivity [15], and research [16].

To assess DQ, scholars have developed numerous DQ taxonomies, which evaluate the extent to which the data contained within digital health systems adhere to multiple dimensions (i.e., measurable components of DQ). Weiskopf and Weng [17] identified five dimensions of DQ spanning completeness, correctness, concordance, plausibility, and currency. Subsequently, Weiskopf et al. [18] refined the typology to consist of only three dimensions: completeness, correctness, and currency. Similarly, Puttkammer et al. [13] focused on completeness, accuracy, and timeliness, whereas Kahn et al. [19] examined conformance, completeness, and plausibility. Others identified 'fitness of use' [20] and the validity of data to a specific context [21] as key DQ dimensions. Overall, there are wide ranging definitions of DQ, with an agreed upon taxonomy evading the literature. In this paper, through synthesising literature, we define data quality as the extent to which digital health data is accessible, accurate, complete, consistent, contextually valid, and current. When consolidated frameworks are developed, the dimensions are often treated in a fragmented way, with little attempt to understand the relationships between the dimensions, and the resultant outcomes. This is substantiated by Bettencourt-Silva et al. [22] who indicated that DQ is not systematically or consistently assessed.

Research Aims and Questions

Failure of health organisations to leverage high quality data will compromise the sustainability of an already strained healthcare system [23]. Therefore, we undertook a systematic literature review to answer the following research questions: 1) What are the dimensions of digital health DQ? 2) How are the dimensions of digital health DQ related? And 3) What are the impacts of digital health DQ? The aim of this research is to develop, from synthesizing literature, a consolidated digital health DQ dimensions and outcomes framework, which demonstrates the DQ dimensions and their interrelationships as well as their impact on core healthcare outcomes. The consolidated data quality dimensions and outcomes framework will be beneficial to both research and practice. For researchers our review consolidates digital health DQ literature and provides core areas for future research to rigorously evaluate and improve digital health DQ. For practice, this study provides healthcare executives and strategic decision makers with insights into both the criticality of digital health DQ through exemplifying the impacts and the complexity of digital health DQ through demonstrating the interrelationships between the dimensions.

This paper is structured as follows: first, we provide details of the systematic literature review method; second, in line with the research questions, we present our three key findings: 1) DQ dimensions; 2) DQ interrelationships; and 3) DQ outcomes; third, we compare the output of our findings to previous literature and discuss the implications of this work.

Method

We followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and Webster and Watson's [24] guidelines for systematic literature reviews. Specifically, consistent with Templier and Paré [25], this systematic literature review was developmental in nature with the goal of developing a consolidated digital health DQ framework.

Literature Search and Selection

To ensure the completeness of the review [24] and consistent with interdisciplinary reviews, the literature search spanned multiple fields and databases (i.e., PubMed, Public Health, Cochrane, SpringerLink, EBSCOhost (Medline and PsycINFO), ABI/Inform, AISel, Emerald Insight, IEEE Xplore digital library, Scopus, and ACM Digital Library). The search was conducted in October, 2021 and was not constrained by year of publication because the concept of data quality has a long-standing academic history. The search terms were reflective of our research topic and research questions. To ensure comprehensiveness, the search terms were broadened by searching their synonyms. For example, we used search terms, such as 'electronic health record', 'digital health record', 'e-health', 'electronic medical record', 'EHR', 'EMR', 'data quality, 'data reduction', 'data cleaning', 'data pre-processing', 'information quality', 'data cleansing', 'data preparation', 'intelligence quality', 'data wrangling', and 'data transformation'. Keywords and search queries were reviewed by the reference librarian and subject matter experts in digital health (appendix 1).

The papers returned from the search were narrowed down in a four-step process (Figure 1Error: Reference source not found). In the identification step 5177 studies were identified through multiple database searches with 3856 duplicates removed resulting in 1321 articles. The 1321 articles were randomly divided into six batches, which were assigned to separate researchers who applied the inclusion and exclusion criteria (Table 1). As a result of abstract screening 896 articles were excluded, resulting in 425 articles remaining. Following a similar approach to the abstract screening, the 425 articles were again randomly divided into six batches and assigned to one of six researchers to read and assess the relevance of the article in line with the selection criteria. The assessment of each of the 425 articles was then verified by the research team resulting in the final set of 227 relevant articles. During this screening phase (i.e., abstract and full-text), daily meetings were held with the research team with any uncertainties raised and discussed until consensus was reached by the team as to whether the article should be included or excluded from the search. In line with Templier and Paré [25], as this systematic literature review was developmental in nature rather than an aggregative meta-analysis, quality appraisals were not performed on the individual articles.

Table 1. Inclusion and exclusion criteria

Inclusion	Exclusion Criteria
Specifically focuses on data quality in digital health.	Development of algorithms for advanced analytics techniques (e.g., machine learning, artificial intelligence) without application within hospital settings.
Empirical papers or review articles where conceptual frameworks were either developed or assessed.	Descriptive papers without a conceptual framework or empirical analysis.
Considers digital health within hospital settings.	Focused only on primary care (e.g., general practice)
Published in peer reviewed outlets within any timeframe	Pre go-live considerations (e.g., software development)
Published in English	Theses and non-peer reviewed (e.g., white papers, editorials).

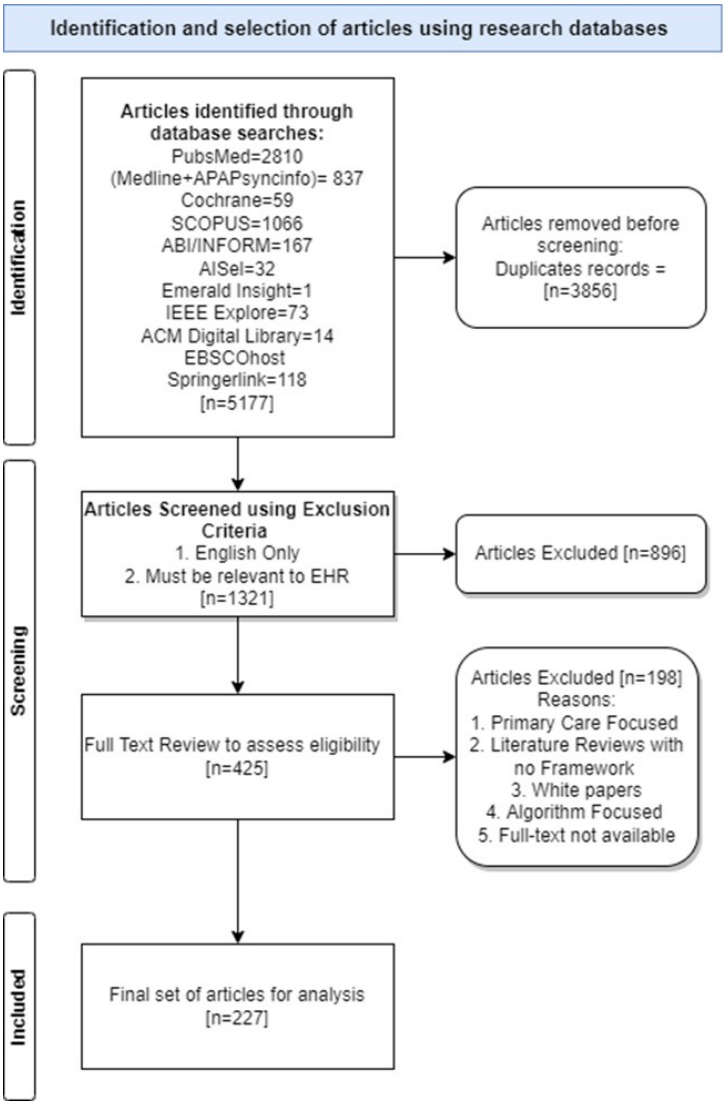


Figure 1. PRISMA Inclusion Process

Literature Analysis

The relevant articles were imported in NVivo (v.12) where analysis was iteratively performed. To ensure reliability and consistency in coding, a coding rule book [26] was developed and progressively updated to guide the coding process. The analysis process involved six steps (Figure 2).

In the first step of the analysis, the research team performed open coding [26] where relevant statements from each article were extracted using verbatim codes while allowing initial grouping of similar concepts [27]. The first round of coding resulted in 1298 open codes. Second, the open codes were segmented into two high level themes, the first group contained 1044 open codes pertaining directly to DQ dimensions (e.g., data accuracy); the second group contained 254 open codes related to DQ outcomes (e.g., financial outcomes).

In the third step, through constant comparison [28] the 1044 raw DQ codes were combined into 29 DQ sub-themes based on commonalities (e.g., contextual DQ, fitness for use, granularity, relevancy, accessibility, availability). In the fourth step, again through performing iterative and multiple rounds of constant comparison, the 254 open codes related to DQ outcomes were used to construct 22 initial DQ outcome sub-themes (e.g., patient safety, clinician-patient relationship, continuity of care). The DQ outcomes sub-themes were further compared to each other resulting in 5 DQ outcome dimensions (e.g., clinical, business process, research-related, clinician, and organisational). For the DQ sub-themes, constant comparison was performed facilitated by the card sorting method [29] where an expert panel of 8 DQ researchers formed into four groups assessed the sub-themes for commonalities and differences. The expert groups presented their categorisation to each other until a consensus was reached. This resulted in a consolidated set of six DQ dimensions (accuracy, consistency, completeness, contextual validity, accessibility, and currency). Appendix 2 provides an example of how the open codes, were reflected in sub-themes, and themes.

After identifying the DQ dimensions and outcomes, the next stage of coding progressed to identifying the interrelationships (Step 5) between the DQ dimensions as well as the relationships (Step 6) between the DQ dimensions and DQ outcomes. To do so, the matrix coding query function using relevant Boolean operators (AND, Near) in NVivo was performed. The outcomes of the matrix queries were reviewed and verified by an expert researcher in the health domain.

Throughout the analysis, steps were performed to provide credibility into our findings. Firstly, prior to commencing the analysis, the research team members who were extracting the verbatim codes initially independently reviewed three common articles, then convened to review any variations in coding. In addition, they reconvened multiple times a week to discuss their coding and update the codebook to ensure a consistent approach was followed. Coder corroboration was performed throughout the analysis with two experienced researchers independently verifying all verbatim nodes until consensus was reached [26]. Subsequent coder corroboration was performed by two experienced researchers to ensure the open codes accurately mapped to the themes and the dimensions. This served to provide internal reliability. Steps were also performed to improve external reliability [107]. Namely, the card-sorting method provided an expert

appraisal. In addition, the findings were presented to and confirmed by three digital healthcare professionals.

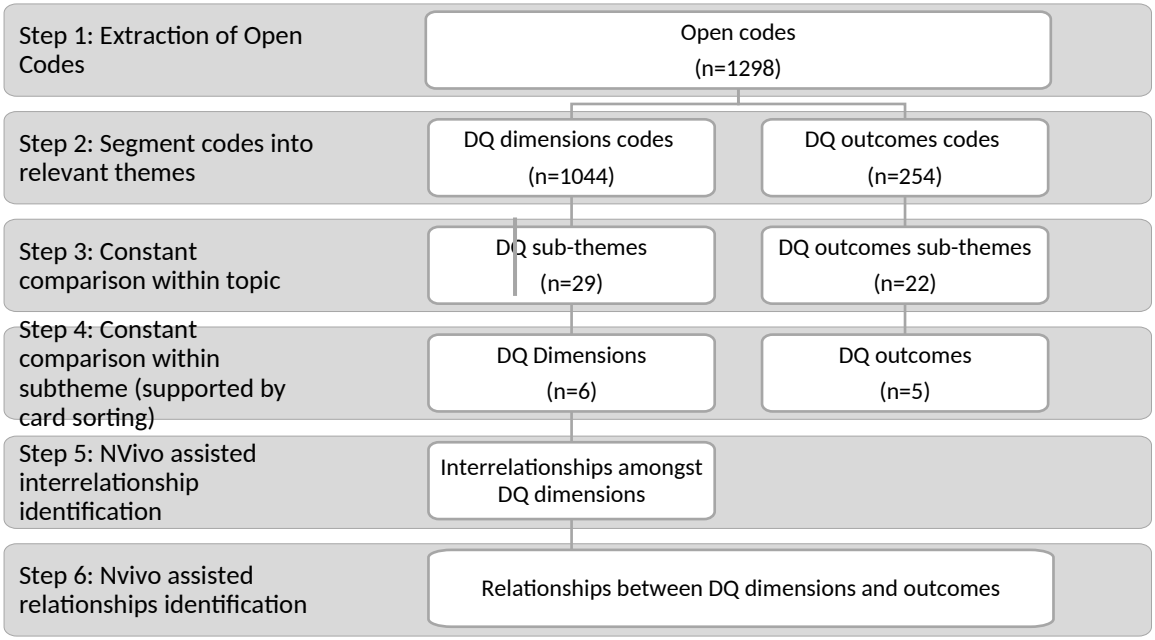


Figure 2. Analysis Process

Results

The vast majority of relevant articles were published in journal outlets (n=169), followed by conference proceedings (n=42), and book sections (n=16). The 169 journal articles were published in 107 journals, with 12% of the journals publishing more than one study (illustrated in Figure 3). The complete breakdown of how many articles have been published within each outlet is detailed in Appendix 3.

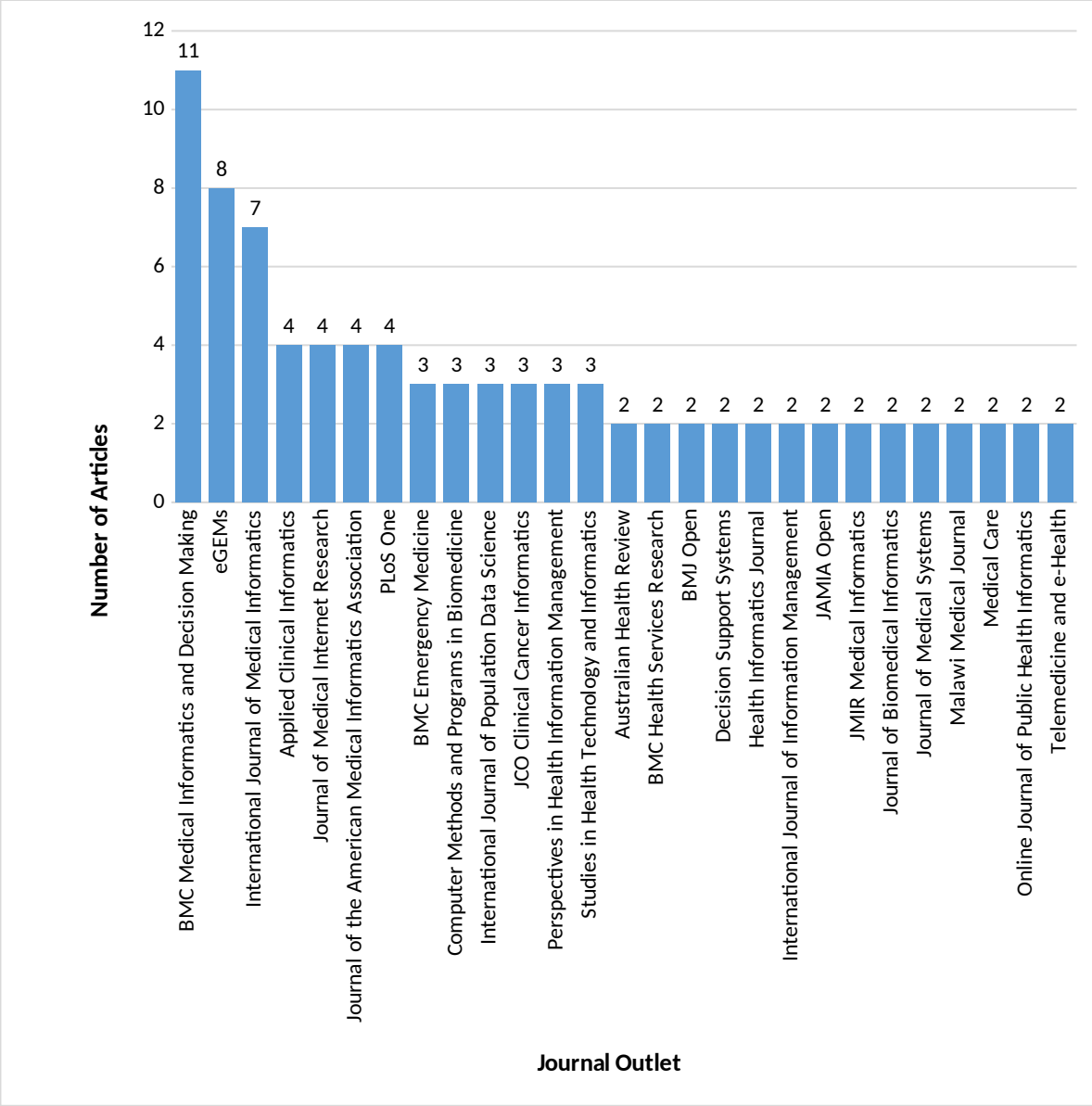


Figure 3: List of top 12% journals

Overall, as illustrated in Figure 4, the interest in digital health data quality has been increasing over time, with sporadic interest prior to 2006.

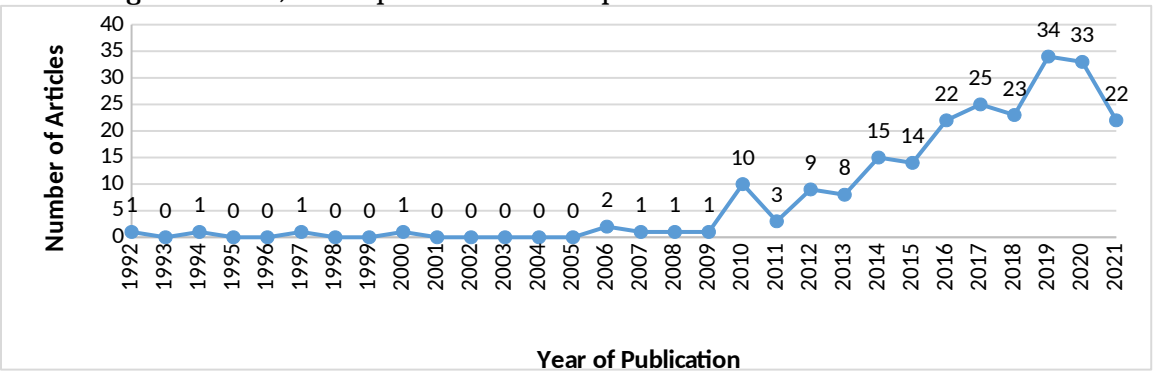


Figure 4: Publications by year

Below, we provide an overview of the DQ definitions, DQ dimensions, their interrelationships, and outcomes to develop a consolidated digital health DQ framework.

Data Quality Definitions

Multiple definitions of DQ are discussed in the literature (Appendix 4). There is no consensus on a single definition of DQ, however analysis of the definitions reveals two perspectives, which we label as the 1) context-agnostic perspective, and 2) context-aware perspective. The context-agnostic perspective defines DQ based on a set of dimensions regardless of the context within which the data is used. For instance, as [30] notes "documentation and contents of data within an electronic medical record (EMR) must be accurate, complete, concise, consistent and universally understood by users of the data, and must support the legal business record of the organization by maintaining the required parameters such as consistency, completeness and accuracy." Conversely the context-aware perspective evaluates the dimensions of DQ with recognition of the context within which the data is being used. For instance, as [31, 32] notes DQ is "the degree to which data satisfy the requirements defined by the product-owner organization", and can be reflected through its dimensions such as completeness and accuracy.

Data Quality Dimensions

In total, 30 sub-themes were identified, which were grouped into six DQ dimensions: accuracy, consistency, completeness, contextual validity, accessibility, and currency (Table 1, Appendix 5). Consistency (n=164), completeness (n=137), and accuracy (n=123) are the main DQ dimensions. Comparatively, less attention has been paid to accessibility (n=28), currency (n=18), and contextual validity (n=26).

Table 2. Description of the DQ dimensions

Dimension	Description	Sub-Themes
Accuracy	"The degree to which data reveal the truth about the event being described". [33]	Validity, correctness, integrity, conformance, plausibility, veracity, accurate diagnostic data
Consistency	"Absence of differences between data items representing the same objects based on specific information requirements. Consistent data contain the same data values when compared between different databases". [33]	Inconsistent data capturing, standardisation, concordance, uniqueness, data variability, temporal variability, system differences, semantic consistency, structuredness, representational consistency
Completeness	"The absence of data at a single moment over time or when measured at multiple moments over time". [34]	Missing data, level of completeness, representativeness, fragmentation, breadth of documentation
Contextual Validity	Assessment of DQ is "dependent on the task at hand" [18].	Contextual DQ, fitness for use, granularity, relevancy
Accessibility	The extent to which it is "feasible it is for users to extract the data	Accessible DQ, availability

	of interest" [18]	
Currency	"The degree to which data represent reality from the required point in time" [35]	Timeliness

Data Quality Dimension: Accessibility

The accessibility dimension (n=28, 12.3%) is composed of both the accessibility (n=15) and availability sub-themes reflecting the feasibility for users to extract data of interest [18]. Scholars regularly view the *accessibility* sub-theme favourably with the increased adoption of electronic health record systems (EHRs) overcoming physical and chronological boundaries associated with paper records by allowing access to information from multiple locations at any time [36, 37]. Top et al. [36] notes that EHR made it possible for nurses to access patient data, resulting in improved decision making. Rosenlund et al. [38] further notes that EHRs benefit healthcare professionals through providing increased opportunities for searching and utilising information. The *availability* sub-theme is an extension of the accessibility sub-theme and examines whether the data exists and whether the data is in a format that is readily usable [39]. For instance, Dentler et al. [39] notes that pathology reports although accessible are recorded in a non-structured, free-text format making it challenging to readily use the data. While structured data may make data more available, Yoo et al. [40] highlights that structured data entry in the form of drop-down lists and check boxes tend to reduce the narrative description of patients' medical conditions. While not explicitly investigating accessibility, Makeleni and Cilliers [33] also note the challenges associated with structured data entry.

Data Quality Dimension: Accuracy

The accuracy dimension (n=123, 54%) is composed of seven sub-themes, correctness (n=42), validity (n=23), integrity (n=19), plausibility (n=17), accurate diagnostic data (n=13), conformance (n=7), and veracity (n=2). Accuracy refers to the extent to which data reveals the truth about the event being described [33] and conforms to its actual value [41].

Studies often referred to accuracy as the '*correctness*' of data, which is the degree to which data correctly communicates the parameter being represented [35]. Conversely, others focus on *plausibility*, the extent to which data points are believable [42]. While accuracy concerns were present for all forms of digital health data, some studies focused specifically on *inaccuracies with diagnostic data,* where "the accurate and precise assignment of structured [diagnostic] data within EHRs is crucial" [43], which is "key to supporting secondary clinical data" [44].

To assess accuracy, the literature regularly asserts that data needs to be *validated* against metadata constraints, system assumptions, and local knowledge [19] and *conform* to structural and syntactical rules. According to Kahn et al. [19], Sirgo et al. [45], conformance focusses on compliance of data with internal or external formatting, relational, or computational definitions. Accurate, verified, and validated data, as well as data conforming to standards contributes to *integrity* of data. Integrity requires that the data stored in health information systems is accurate and consistent, where the "improper use of [health information systems] can jeopardise the integrity of a patient's information" [33]. An emerging sub-theme of accuracy was the veracity of data, which represents uncertainty in the data due to inconsistency, ambiguity, latency, deception,

and model approximations [21]. It is particularly important in the context of the secondary use of big data, where "data veracity issues can arise from attempts to preserve privacy, …and is a function of how many sources contributed to the data." [46]

Data Quality Dimension: Completeness

The completeness dimension (n=114, 50%) is composed of six sub-themes: missing data (n=66), level of completeness (n=25), representativeness (n=13), fragmentation (n=8), and breadth of documentation (n=2). A well-accepted definition of data completeness considers four perspectives: documentation (the presence of observations regarding a patient in data), breadth (the presence of all desired forms of data), density (the presence of a desired frequency of data values over time), and predictive (the presence of sufficient data to predict an outcome) [47]. Our analysis revealed that these four perspectives, while accepted, are rarely systematically examined in extant literature, rather papers tend to discuss completeness or the lack thereof as a whole.

Missing data is a prominent sub-theme and represents a common problem in EHR data. For instance, Gloyd et al. [48] argue that incomplete, missing and implausible data "was by far the most common challenge encountered". Scholars regularly identified that data fragmentation contributed to incompleteness, with a patient's medical record deemed incomplete due to data being required from multiple systems and EHRs [18, 49-55]. "Data were also considered hidden within portals, outside systems, or multiple EHRs, frustrating efforts to assemble a complete clinical picture of the patient" [50]. More positive perspectives pertaining to data completeness focus on the *level of completeness*, with studies reporting relatively high completeness rates in health datasets [37, 39, 56-59]. For data to be considered complete it needs to be captured at sufficient breadth and depth over time [12, 18].
Some studies have proposed techniques to improve completeness, which include: developing fit-for-purpose user interfaces [60-62], standardizing documentation practices, [63, 64], automating documentation [65], and performing quality control [64].

In some instances, the *level of completeness* and *extent of missing data* differed depending on the nature of the patient [15, 16, 18, 20, 46, 51, 59, 66-72], which we classified into the sub-theme of *representativeness*. It has been found that there is "a statistically significant relationship between EHR completeness and patient health status" [70] with more data recorded for sick patients compared to less acute patients. This aligns strongly with the sub-theme of contextual validity.

Data Quality Dimension: Consistency

The consistency dimension (n=157, 69%) is composed of ten sub-themes: inconsistent data capturing (n=33), standardisation (n=28), concordance (n=22), uniqueness (n=14), data variability (n=14), temporal variability *(n=13),* system differences (n=12), semantic consistency (n=10), structuredness (n=7), and representational consistency (n=4).

Inconsistent data capturing is a prevalent sub-theme caused by the manual nature of data entry in healthcare settings [46], especially when data involves multiple times, teams, and goals [73]. Inconsistent data capturing results in *data variability* and *temporal variability*. *Data variability* refers to inconsistency in the data captured within and between health information systems, whereas *temporal variability* reflects inconsistencies that occur over time and may be due to changes to policies or medical

guidelines [20, 48, 74-79]. *Semantic inconsistency (i.e.,* data with logical contradictions) and *representational inconsistency (i.e*., data variations due to multiple formats*)* can also result from inconsistent data capturing [80].

Standardization in terms of terminology, diagnostic codes, and workflows [64] are proffered to minimise inconsistency in data entry, yet in practice there is a "lack of standardized data and terminology" [9] and "even with a set standard in place not all staff accept and follow the routine" [64]. The lack of standardisation is further manifested due to health information *system differences* across settings [81]. As a result of the differences between systems, *concordance* - the extent of "agreement between elements in the EHR, or between the EHR and another data source" is hampered [82].

Inconsistent data entry can be further caused by redundancy within the system due to structured versus unstructured data [83], which we label as the sub-theme '*structuredness'* and duplication across systems [66, 78, 84-87], which we label as the sub-theme '*uniqueness'.* While structured data entry, "facilitates information retrieval" [36] and is "in a format that enables reliable extraction" [18], the presence of unstructured fields leads to data duplication efforts, hampering uniqueness as data is recorded in multiple places with varying degrees of granularity and level of detail.

Data Quality Dimension: Contextual Validity

The contextual validity dimension (n=26, 11%) is composed of four sub-themes: fitness for use (n=11), contextual DQ (n=9), granularity (n=4), and relevancy (n=2). Contextual validity requires a deep understanding of the context which gives rise to data [46], including technical, organisational, behavioural, and environmental factors [88].

Contextual DQ is often described as '*fitness of use*' [20] for which understanding the context in which data is collected is deemed important [18, 51]. Another factor that contributes to data being fit for use is *granularity* of data. Adequate *granularity* of timestamps [89], patient information [16], and data present in EHR (e.g., diagnostic code [16]) was considered important to make data fit for use. Finally, for data to be fit for use it needs to be *relevant*. As indicated by Schneeweiss and Glynn [69], for data to be meaningful healthcare databases need to contain relevant information of sufficient quality, which can help answer specific questions. The literature clearly demonstrates the need to take context into consideration when analysing data and the need to adapt technologies to the healthcare context so that appropriate data is collected for reliable analysis to be performed.

Data Quality Dimension: Currency

The currency dimension (n=18, 8%) was formed by the single sub-theme of *timeliness*. Currency or timeliness, is defined in Afshar et al. [35] and Makeleni and Cilliers [33] as the degree to which data represents reality from the required point in time. From an EHR perspective, the data should be up to date, available, and reflect the profile of the patient at the time the data is accessed [35, 90]. Lee et al. [42] extends this to include the recording of an event at the time it occurs such that a value is deemed current if it is representative of the clinically relevant time of the event. Frequently mentioned causes for lack of currency of data include: (i) recording of events (long) after the event actually occurred [52, 64, 91, 92], (ii) incomplete recording of patient characteristics over time [16], (iii) system/interface design not matching workflow and impeding timely recording of data [64], (iv) mixed mode recording – paper and electronic [64],

and (v) lack of timestamp metadata meaning the temporal sequence of events is not reflected in the recorded data [16].

Interrelationships between the Data Quality Dimensions

As illustrated in Figure 5 and evidenced in Appendix 6, interrelationships were found between the digital health DQ dimensions.

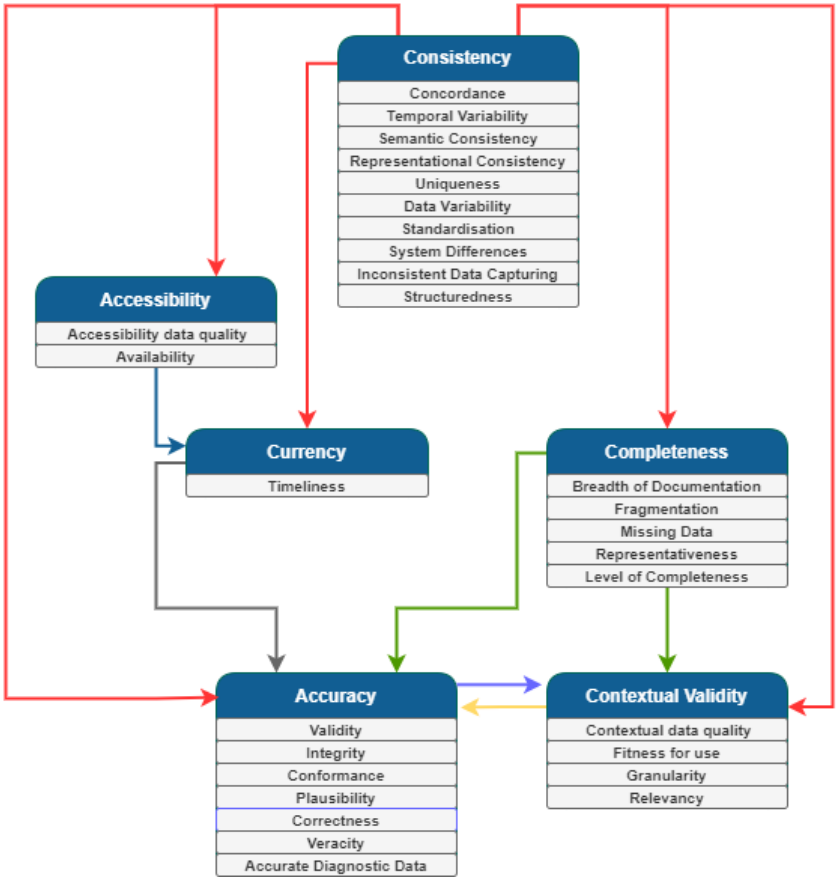


Figure 5. Interrelationships between DQ Dimensions

Consistency influenced all DQ dimensions. Commonly these relationships were expressed in terms of the presence of structured and consistent data entry prompting complete and accurate data to be entered into the health information system, which provides more readily accessible and current data for healthcare professionals when treating patients. As Roukema et al. [37] notes "structured data entry applications can prompt for completeness, provide greater accuracy and better ordering for searching and retrieval, and permit validity checks for DQ monitoring, research, and especially decision support". When data is entered inconsistently it impedes the accuracy of the medical record and the contextual validity for secondary uses of data [67].

Accessibility of data was found to influence the currency dimension of DQ. When data is not readily accessible it seldom satisfies the timeliness of information for healthcare or research purposes [39]. Currency also influenced the accuracy of data. In a study investigating where DQ issues in EHR arise , it was found that "false negatives and false positives in the problem list sometimes arose when the problem list ... [was] out-of-date, either because a resolved problem was not removed or because an active problem was not added" [51].

Completeness further influenced the accuracy of data as [33] notes "data should be complete to ensure it is accurate". The presence of inaccurate data was regularly linked to information fragmentation [49], incomplete data entry [86], and omissions [42]. Completeness also influenced contextual validity as it is necessary to have all the data available to complete specific tasks [32]. When it comes to the secondary use of EHR data, evaluation of "completeness becomes extrinsic, and is dependent upon whether or not there are sufficient types and quantities of data to perform a research task of interest" [70].

Accuracy and contextual validity exhibited a bidirectional relationship with each other. The literature suggests that accuracy influences contextual validity, however data cannot simply be extracted from structured form fields, free text fields will also need to be consulted. For instance, Kim and Kim [93] identifies "it is sometimes thought that structured data are more completely optimized for clinical research. However, this is not always the case, particularly given that extracted EMR data can still be unstable and contain serious errors." Conversely, other literature suggests that when only a segment of information regarding a specific clinical event (i.e., contextual validity) is captured inaccuracy can result [16].

Outcomes of Digital Health Data Quality

The analysis of literature identified five types of digital health DQ outcomes: 1) clinical, 2) business process, 3) clinician, 4) research related, and 5) organisational outcomes (Appendix 7). Through utilising NVivo's built-in crosstab query coupled with subject matter expert analysis, it was identified that different DQ dimensions were related to DQ outcomes in different ways (Table 3). Currency was the only dimension that did not have a direct effect on DQ outcomes. However, as discussed later (Figure 6), it is plausible that currency affects DQ outcomes through impacting other DQ dimensions. Below, we discuss each DQ dimension and their respective outcomes.

Table 3. The Relationships between DQ Dimensions and Data Outcomes

DQ Dimension	Outcomes*				
	Research	Organisational	Business Process	Clinical	Clinician
Accessibility	X	X	X	X	X
Accuracy	X			X	
Completeness	X	X	X	X	
Consistency	X	X	X	X	X
Contextual Validity					
Currency					

*Note: X denotes relationship between DQ dimension and outcome is reported in literature. Blank cells denote that there is no evidence to support the relationship.

We identified that the accessibility DQ dimension influenced clinical, clinician, business process, research-related, and organisational outcomes. In terms of *clinical outcomes,* Roukema et al. [37] indicates that EHRs through improving accessibility and legibility of healthcare data significantly enhances the quality of patient care. The increased accessibility of medical records during the delivery of patient care is further proffered

to benefit *clinicians* through reducing data entry burden [36]. Conversely, inconsistency in the availability of data across health settings increases clinician workload, as Wiebe et al. [15] notes "given the predominantly electronic form of communication between hospitals and general practitioners in Alberta, the inconsistency in availability of documentation in one single location can delay processes for practitioners searching for important health information". When data is accessible and available it can improve *business processes* (e.g., quality assurance) and *research-related* (e.g., outcomes-oriented research) *outcomes* and is able to support *organisational outcomes* with improved billing and financial management [94].

The literature demonstrates that data accuracy influences *clinical outcomes* [14, 66, 95] and *research-related outcomes* [14, 96], as Wang et al. [14] describes, "errors in healthcare data are numerous and impact secondary data use and potentially patient care and safety". Downey et al. [66] observe the negative impact on quality of care (i.e., *clinical outcomes)* resulting from incorrect data and state "manual data entry remains a primary mechanism for acquiring data in EHRs, and if the data is incorrect then the impact to patients and patient care could be significant" [66]. Poor data accuracy also diminishes the quality of *research outcomes*. Precise data is beneficial in producing high quality research outcomes as Gibby [96] explains, "computerized clinical information systems have considerable advantages over paper recording of data, which should increase the likelihood of their use in outcomes research. Manual records are often inaccurate, biased, incomplete, and illegible". Closely related to accuracy, contextual validity is an important DQ dimension which considers the fitness for *research* as stated by Weiskopf et al. [70] "[w]hen repurposed for secondary use, however, the concept of "fitness for use" can be applied".

The consistency DQ dimension was related to all DQ outcomes. It was commonly reported that inconsistency in data negatively impacts the *reusability* of EHR data for research purposes hindering *research-related outcomes* and negatively impacting *business processes* and *organizational outcomes*. For example, Kim et al. [97] acknowledge that inconsistent data labelling in EHR systems may hinder accurate research results noting, "a system may use local terminology that allows unmanaged synonyms and abbreviations. … If local data are not mapped to terminologies, … performing multicentre research would require extensive labour". Alternatively, von Lucadou et al. [16] indicates the impact of inconsistency on *clinical outcomes* reporting that the existence of inconsistencies in captured data "could explain the varying number of diagnoses throughout the encounter history of some subjects". Whereas, Diaz-Garelli et al. [43] demonstrate the negative impact that inconsistency has on *clinicians* in terms of increased workload.

Incomplete EMR data was found to impact *clinical outcomes* (e.g., reduced quality of care), *business process outcomes* (e.g., interprofessional communication), *research-related* (e.g., research facilitation), and *organizational outcomes* (e.g., key performance indicators related to readmissions) and *research related outcomes*[15]. For example, while reviewing the charts of 3011 non-obstetric inpatients, Wiebe et al. [15] found that missing discharge summary within an EHR "can present several issues for healthcare processes, including hindered communication between hospitals and general practitioners, heightened risk of readmissions, and poor usability of coded health data", among other widespread implications. Liu et al. [98] further reports that "having

incomplete data on patients' records has posed the greatest threat to patient care". Due to the heterogenous nature (with multiple data points) of EHR data, Richesson et al. [20] emphasise that access to large, complete data will allow clinical investigators "to detect smaller clinical effects, identify and study rare disorders, and produce robust, generalisable results".

Discussion

The following sections describe the three main findings of this research: 1) identification of the dimensions of data quality, 2) the interrelationships between the dimensions of data quality, and 3) the outcomes of data quality. As described in the 'Summary of Key Findings' section, these three findings led to the development of the DQ Dimensions and Outcomes (DQ-DO) framework. Subsequently, we compare the DQ-DO framework with related work. This leads to the generation of implications for future research. The discussion concludes with a reflection of the limitations of this study.

Summary of key findings

In summary, we unearthed three core findings. Firstly, we identified six dimensions of DQ within the digital health domain: consistency, accessibility, completeness, accuracy, contextual validity, and currency. These dimensions were synthesised from 30 sub-themes described in the literature. We found that consistency, completeness, accuracy are the predominant dimensions of DQ. Comparatively speaking, limited attention has been paid to the dimensions of accessibility, currency, and contextual validity. Secondly, we identified interrelationships between these six dimensions of digital health DQ (Table 2). The literature indicates that data inconsistencies can influence all other DQ dimensions. The accessibility of data was found to influence the currency of data. Completeness impacts accuracy and contextual validity, with these dimensions serving as dependent variables and exhibiting a bidirectional relationship with each other. Thirdly, we identified five types of data outcomes (Table 2, Appendix 7): research-related, organisational, business process, clinical, and clinician. Consistency was found to be a very influential dimension impacting all types of DQ outcomes. Contextual validity on the other hand, was shown to be particularly important for data reuse (e.g. performance measurement, outcome-oriented research etc.). Whilst currency does not directly impact any outcomes, it impacts the accuracy of data, which impacts clinical and research-related outcomes. Therefore, if currency is not resolved, accuracy issues would still prevail. If the objective is to improve organisational outcomes, consistency, accessibility, and completeness were shown to be important considerations. Through consolidating our three core findings, we developed a consolidated DQ Dimensions and Outcomes framework, DQ-DO (Figure 6).

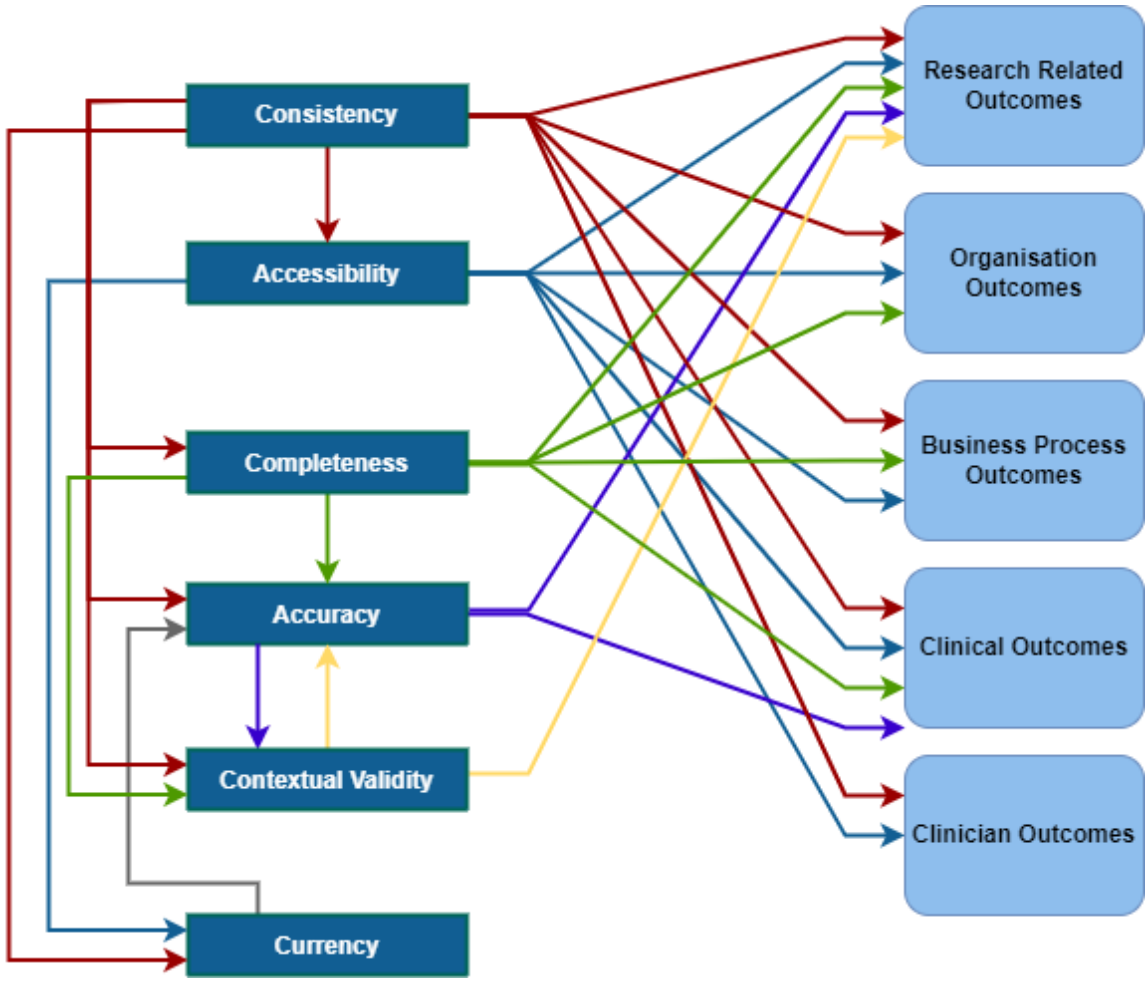


Figure 6. Consolidated Digital Health Data Quality Dimensions and Outcomes (DQ-DO) Framework

Comparison to literature

Our findings extend previous studies on digital health DQ in three ways. Firstly, through our rigorous approach, we identified a comprehensive set of DQ dimensions, which both confirms and extends existing literature. For instance of Weiskopf and Weng [17] identified five DQ dimensions including completeness, correctness, concordance, plausibility, and currency, all of which are present within our DQ framework, although in some instances, we use slightly different terms (referring to correctness as accuracy and concordance as consistency). Extending the framework of Weiskopf and Weng [17], we view plausibility as a sub-theme of accuracy, disentangle accessibility from completeness, and we also stress the importance of contextual validity per Richesson et al. [20]. Others have commonly had a narrower perspective of DQ focusing on completeness, correctness, and currency [18], or on completeness, timeliness, and accuracy [13]. In other domains of digital health, such as physician-rating systems, Wang and Strong's [99] data quality dimensions of intrinsic, contextual, representational, and accessibility have been adopted. Such approaches to assessing data quality are appropriate although it removes a level of granularity that is necessary to understanding relationships and outcomes. This is particularly necessary given the salience of consistency in our data set and the important role it plays in generating outcomes.

Secondly, unlike previous studies on DQ dimensions, we also demonstrate how these dimensions are all related to each other. By analysing the interrelationships between these DQ dimensions, we can determine how a particular dimension influences another and in which direction this relationship is unfolding. This is an important implication for digital health practitioners as whilst several papers have examined how to validate [57] and resolve data quality issues [16], to resolve issues with a specific DQ dimension requires awareness of the interrelated DQ dimensions. For instance, to improve accuracy, one also needs to consider improving consistency and completeness.

Thirdly, although previous studies describe how DQ can impact a particular outcome (e.g., [18, 100, 101]), they largely focus broadly on data quality, or a specific dimension of data quality, or on a specific outcome. For instance, Sung et al. [102] notes that poor quality data were a prominent barrier hindering adoption of digital health systems. Conversely, Kohane et al. [103] focus on research-related outcomes in terms of publication potential and identified that incompleteness and inconsistency can serve as core impediments. To summarise, the DQ-DO framework (Figure 6) developed through this review provides not only the dimensions and the outcomes but also the interrelationships between these dimensions and how they influence outcomes.

Implications for Future Work

Implication 1: Equal Consideration across Data Quality Dimensions

This study highlights the importance of each of the six DQ dimensions: consistency, accessibility, completeness, accuracy, contextual validity, and currency. These dimensions have received varying attention in the literature. Although we observe that some DQ dimensions such as accessibility, contextual validity, and currency are discussed less frequently than others, it does not mean that these dimensions are not important for assessment. This is evident in Figure 6, which identifies that all DQ dimensions except for currency directly influence DQ outcomes. Whilst we did not identify a direct relationship between the currency of data and the six types of data outcomes it is likely that the currency of data influences the accuracy of data, which subsequently influences the research-related and clinical outcomes. Future research, including consultation with a range of stakeholders, needs to further delve into understanding the under-researched DQ dimensions. For instance, both currency and accessibility of data are less frequently discussed dimensions in the literature yet, with the advances in digital health technologies, both have become highly relevant for real-time clinical decisions [21, 104].

Implication 2: Empirical Investigations of the Impact of the Data Quality dimensions

The DQ-DO framework identified in this study has been developed through a rigorous systematic literature review process, which synthesised literature related to digital health DQ. To extend this study, we advocate for empirical mixed-methods case studies to validate the framework, including an examination of the interrelationships between DQ dimensions and DQ outcomes, based on real-life data and consultation with a variety of stakeholders. To identify the presence of issues with DQ dimensions within digital health system logs existing approaches could be used [57, 105]. The DQ-outcomes could be assessed by extracting pre-recorded key performance indicators from case hospitals and be triangulated with interview data to capture patients, clinicians, and hospitals

perspectives of impacts of DQ. This could then be incorporated into a longitudinal study, where data collection is performed prior to and after a DQ improvement intervention being performed, which would provide efficacy to the digital health DQ intervention.

Implication 3: Understanding the Root Causes of Data Quality Challenges

Although this study provides a first step towards a more comprehensive understanding of DQ dimensions for digital health data and their influences on outcomes, it does not explore potential causes of such DQ challenges. Without understanding the reasons behind these DQ issues, the true potential of evidence-based healthcare decision-making remains unfulfilled. Future research should examine the root causes of DQ challenges in healthcare data with a view to prevent such errors from occurring in the first place. One framework that may prove useful to illuminating the root-causes of DQ is the Odigos framework, which indicates that DQ issues emanates from the social world (i.e., macro and situational structures, roles, and norms), material world (e.g., quality of the EHR system and technological infrastructure), and the personal world (e.g., characteristics and behaviours of healthcare professionals) [105]. These insights could then be incorporated into a data governance roadmap for digital hospitals.

Implication 4: Systematic assessment and remedy of Data Quality Issues

Though prevention remains better than the cure (see previous limitation), not all DQ errors can be prevented or mitigated. It is common for many healthcare organisations to dedicate resources to data cleaning in order to obtain high quality data in a timely manner and this will remain necessary (though hopefully to a lesser degree). Some studies (e.g., [18]) advocate evidence-based guidelines and frameworks for a detailed assessment of the quality of digital health data. However, there is little work focusing on a systematic and automated way of assessing and remedying common DQ issues. Future research should also focus on evidence-based guidelines, best practices, and automated means to assess and remedy digital health data.

Limitations

This review is scoped to studying digital health data generated within a hospital setting and not to other healthcare settings. This is necessary because of the vast differences between acute health care settings and primary care. Future research should seek to investigate the digital health data of primary care settings to identify the DQ dimensions and outcomes relevant to these settings. In addition, this literature review has been scoped to peer-reviewed outlets, with “grey” literature excluded, which could have led to publication bias. Although this scoping may have missed some articles, it was necessary to ensure quality behind the development of the digital health DQ framework. An additional limitation that may be raised by our method is that due to the sheer amount of articles returned by our search, we did not perform double coding (where independent researchers analyse the same article). To mitigate this limitation, steps were taken to minimise bias through conducting coder corroboration sessions and group validation as mentioned in the Methods section with the objective of improving internal and external reliability [107]. To further improve internal reliability two experienced researchers verified the entirety of the analysis in NVivo and for external reliability card sorting assessments were performed with data quality experts and the findings were presented and confirmed by three digital healthcare professionals.

Furthermore, empirical validation of the framework is required, both in terms of real-life data and input from a range of experts.

Conclusions

The multidisciplinary systematic literature review conducted in this study resulted in the development of a consolidated digital health DQ framework comprised of six DQ dimensions, the interrelationships between these dimensions, six DQ outcomes, and relationships between these dimensions and outcomes. We identified four core implications to motivate future research: specifically researchers should: 1) pay equal consideration to all dimensions of data quality as the dimensions can both directly and/or indirectly influence DQ outcomes; 2) seek to empirically assess the DQ-DO framework using a mixed-methods case study design; 3) identify the root causes of the digital health DQ issues; 4) develop interventions to mitigate and prevent DQ issues from arising . The DQ-DO framework provides healthcare executives (e.g., chief information officers, chief clinical informatics officers) with insights into DQ issues, and which digital health-related outcomes they have an impact on - this can help them prioritise tackling DQ-related problems.

Acknowledgements

We acknowledge support provided by the Centre of Data Science, Queensland University of Technology.

Conflicts of Interest

None declared

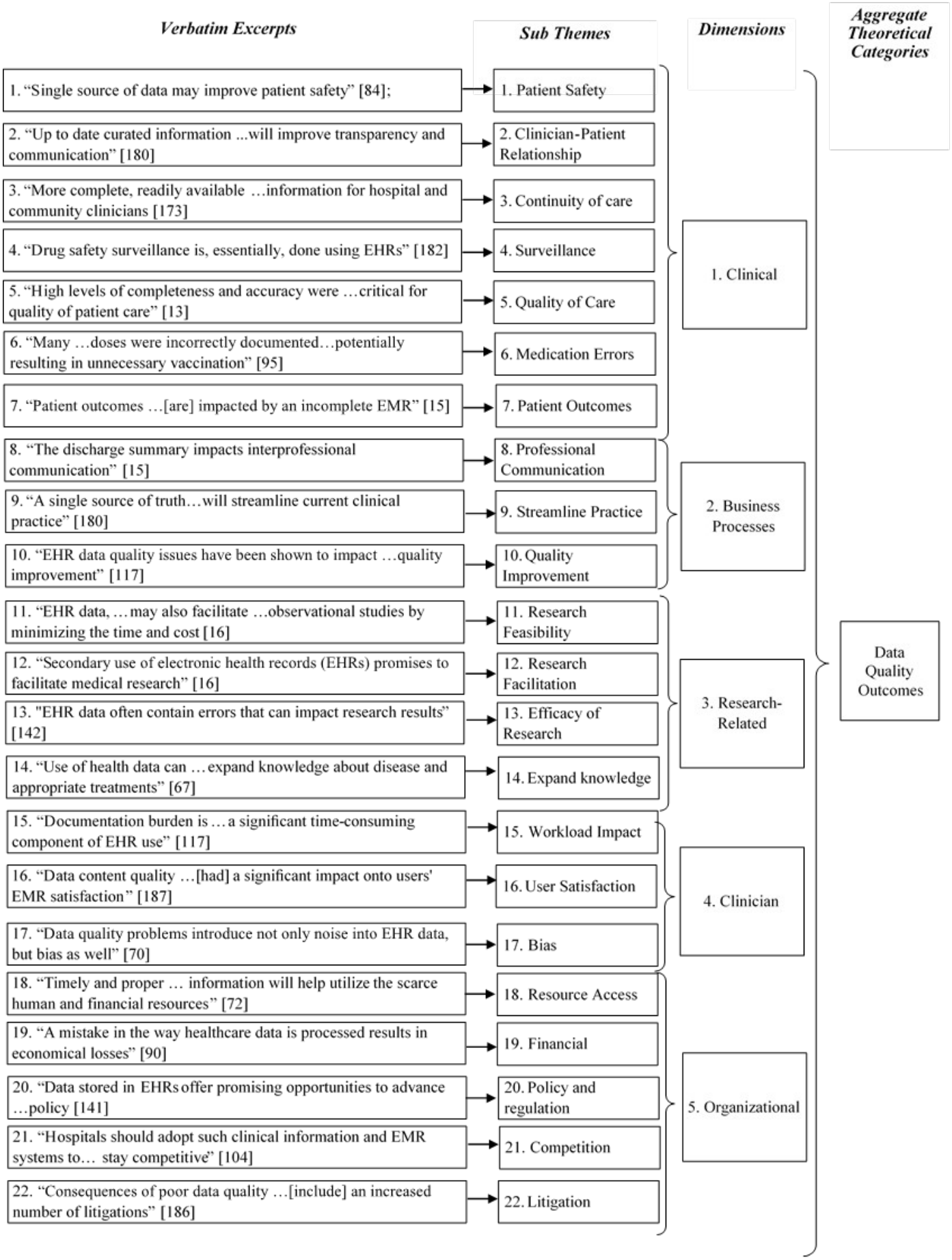
Abbreviations

Acronym	Explication	Description
DQ	Data Quality	The extent to which digital health data is accessible, accurate, complete, consistent, contextually valid, and current.
DQ Dimensions	Data Quality Dimensions	The components used to evaluate data quality (i.e., accessibility, accuracy, completeness, consistency, contextual validity, currency)
DQ-DO Framework	Data Quality – Data Outcomes Framework	The consolidated framework developed in this study demonstrating the interrelationships between data quality dimensions and their relationships with data quality outcomes.
EHR	Electronic Health Records	A longitudinal and electronic collection of patients' clinical information available across case settings [108].
EMR	Electronic Medical Records	Synonymous to EHR.

Appendix 1: Verification of Search Strategy

Area	Researchers (CIs)	Subject Expert 1	Reference Librarian	Subject Expert 2	Co-Researchers
Research Questions	☒	R	C	.	V	C
Keywords	☒	R	V	C	V	C
Subject Area /Domain	☒	R	C	.	V	C
Search Databases	☒	R	V	V	V	C
Journals	☒	R	C	V	V	C
Conferences	☒	R	C	V	V	C
Search Engine	☒	R	.	V	.	R
Relevance of Selected Seminal Articles	☒	R	V	.	V	V
R= Responsible, V= Verifier, C = Contributor						

Appendix 2: Data Coding Structures


Verbatim Excerpts
Sub Themes
Dimensions
Aggregate Theoretical Categories
1. "Single source of data may improve patient safety" [84];
1. Patient Safety
2. "Up to date curated information ...will improve transparency and communication" [180]
2. Clinician-Patient Relationship
3. "More complete, readily available …information for hospital and community clinicians [173]
3. Continuity of care
4. "Drug safety surveillance is, essentially, done using EHRs" [182]
4. Surveillance
5. "High levels of completeness and accuracy were …critical for quality of patient care" [13]
5. Quality of Care
6. "Many …doses were incorrectly documented…potentially resulting in unnecessary vaccination" [95]
6. Medication Errors
7. "Patient outcomes …[are] impacted by an incomplete EMR" [15]
7. Patient Outcomes
1. Clinical
8. "The discharge summary impacts interprofessional communication" [15]
8. Professional Communication
9. "A single source of truth…will streamline current clinical practice" [180]
9. Streamline Practice
10. "EHR data quality issues have been shown to impact …quality improvement" [117]
10. Quality Improvement
2. Business Processes
11. "EHR data, … may also facilitate …observational studies by minimizing the time and cost [16]
11. Research Feasibility
12. "Secondary use of electronic health records (EHRs) promises to facilitate medical research" [16]
12. Research Facilitation
13. "EHR data often contain errors that can impact research results" [142]
13. Efficacy of Research
14. "Use of health data can … expand knowledge about disease and appropriate treatments" [67]
14. Expand knowledge
3. Research-Related
15. "Documentation burden is … a significant time-consuming component of EHR use" [117]
15. Workload Impact
16. "Data content quality …[had] a significant impact onto users' EMR satisfaction" [187]
16. User Satisfaction
17. "Data quality problems introduce not only noise into EHR data, but bias as well" [70]
17. Bias
4. Clinician
18. "Timely and proper … information will help utilize the scarce human and financial resources" [72]
18. Resource Access
19. "A mistake in the way healthcare data is processed results in economical losses" [90]
19. Financial
20. "Data stored in EHRs offer promising opportunities to advance …policy [141]
20. Policy and regulation
21. "Hospitals should adopt such clinical information and EMR systems to… stay competitive" [104]
21. Competition
22. "Consequences of poor data quality …[include] an increased number of litigations" [186]
22. Litigation
5. Organizational
Data Quality Outcomes

Appendix 3: Publication outlets

Outlet	N*
Abdominal Radiology	1
American Journal of Emergency Medicine	1
American Journal of Law and Medicine	1
America's Conference on Information Systems 2017	1
AMIA Annual Symposium	16
AMIA Joint Summits on Translational Science Proceedings	2
Anesthesia and analgesia	1
Anesthesiology Clinics	1
Annals of Internal Medicine	1
Applied Clinical Informatics	4
Applied Network Science	1
Asian Bioethics Review	1
Asia-Pacific Conference on Business Process Management	1
Australasian Computer Science Week 2016	1
Australasian Conference on Information Systems	2
Australian Health Review	2
BioMedicine	1
BMC Emergency Medicine	3
BMC Health Services Research	2
BMC Infectious Diseases	1
BMC Medical Informatics and Decision Making	11
BMC Medical Research Methodology	1
BMC Medicine	1
BMC Pediatrics	1
BMJ	1
BMJ Open	2
Building Capacity for Health Informatics in the Future	1
Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth	1
Business & Information Systems Engineering	1
Canadian Journal of Diabetes	1
Clinical Epidemiology	1
Computer Methods and Programs in Biomedicine	3
Computers in Biology and Medicine	1
Decision Support Systems	2
Deeble Institute for Health Policy Research	1
Digital Personalized Health and Medicine	2
eGEMs	8
e-Health – For Continuity of Care	1
Electronic Journal of Health Informatics	1
Emergency Medicine Australasia	1
Endocrinol Metabolism	1
European Journal of Cardio-thoracic Surgery	1

Frontiers in Medicine	1
German Medical Data Sciences: Bringing Data to Life	1
GigaScience	1
Government Information Quarterly	1
Hawaii International Conference on System Sciences	4
Health Informatics Journal	2
Health Information Management Journal	1
Health Policy and Technology	1
Health Research Policy and Systems	1
Health Services Research	1
Healthcare	1
Healthcare Executive	1
Healthcare Quarterly	1
Healthcare Technology Letters	1
Hong Kong Law Journal	1
IEEE EMBS International Conference on Biomedical and Health Informatics 2016	1
IEEE International Conference on Healthcare Informatics 2018	1
IEEE International Symposium on Computer-Based Medical Systems 2008	1
Industrial and Systems Engineering Research Conference 2018	1
Informatics for Health and Social Care	1
Information and Software Technology	1
Information Systems International Conference	1
Information Technology and Communications in Health Conference	1
Injury Prevention	1
International Conference On Computational And Bio Engineering	1
International Conference on Computer and Information Science 2022	1
International Conference on Computer Modeling, Simulation and Algorithm 2020	1
International Conference on e-Health Networking, Applications and Services 2016	1
International Conference on Emerging Ubiquitous Systems and Pervasive Networks 2016	1
International Conference on Information Quality 2010	1
International Conference on Information Society (i-Society 2013)	1
International Conference on Information Systems	1
International Congress of the European Federation for Medical Informatics 2006	1
International Joint Conference on Biomedical Engineering Systems and Technologies 2019	1
International Journal of E-Health and Medical Communications	1
International Journal of Health Care Quality Assurance	1
International Journal of Healthcare Information Systems and Informatics	1
International Journal of Healthcare Management	1
International Journal of Information Management	2
International Journal of Medical Informatics	7
International Journal of Pediatric Obesity	1

International Journal of Population Data Science	3
International Journal of Social Research Methodology	1
IST-Africa Conference 2011	1
JCO Clinical Cancer Informatics	3
Joint Conference on Knowledge-Based Software Engineering	1
Journal of Biomedical Informatics	2
Journal of Cardiothoracic and Vascular Anesthesia	1
Journal of Clinical Epidemiology	1
Journal of General Internal Medicine	1
Journal of Healthcare Engineering	1
Journal of Healthcare Informatics Research	1
Journal of Korean Medical Science	1
Journal of Medical Internet Research	4
Journal of Medical Internet Research Medical Informatics	2
Journal of Medical Systems	2
Journal of Medicine & Public Health	1
Journal of Nursing Care Quality	1
Journal of Oncology Practice	1
Journal of Public Health Management and Practice	1
Journal of the American College of Surgeons	1
Journal of the American Medical Informatics Association	4
Journal of the American Medical Informatics Association Open	2
Journal of the International AIDS Society	1
Malawi Medical Journal	2
Medical Care	2
MEDINFO 2010	7
Neurology	1
Obstetrics & Gynecology	1
Online Journal of Public Health Informatics	2
Open Access Journal of Clinical Trials	1
Orphanet Journal of Rare Diseases	1
Pacific Asia Journal of the Association for Information Systems	1
Pediatric Critical Care Medicine	1
Pediatrics	1
Perspectives in Health Information Management	3
Pharmacy and Therapeutics	1
PLoS One	4
Policy, Politics, & Nursing Practice	1
Public Health Management Practice	1
Public Health Reports	1
Respir Care	1
SA Journal of Information Management	1
Saudi Pharmaceutical Journal	1
Scientific Reports	1
Statistical Methods in Medical Research	1
Studies in Health Technology and Informatics	3

Summit on Translational Bioinformatics	1
Systemic Practice and Action Research	1
Telemedicine and e-Health	2
The Annals of Family Medicine	1
The Conversation	1
The Lancet Digital Health	1
Topics in Health Information Management	1
Vaccine	1
Wireless Personal Communications	1
Yearbook of Medical Informatics	1

Appendix 4: Data Quality Definitions

DQ Definition	Reference
DQ: Context Aware Perspective	
The totality of features & characteristics of an entity that bears on its ability to satisfy stated and implied needs	[109]
Data's "fitness for use" and can be described by a set of dimensions (e.g., accuracy and completeness)	[110]
The ability of the data to fulfil the purpose for which they were collected or fit for use. The concept of 'fitness for use' emphasises the importance of taking the end user's perspective of quality into account because it is the end users who will decide whether a product is fit for use or is conforming to specific requirements	[33]
Data which is accurate, reliable, "fit for use" and relevant	[111]
Data fit for use, where fitness for use produces accurate, complete, and timely data accessible to stakeholders and relevant to their tasks	[112]
DQ is "fit-for-use" in that its determinants are dependent on the data consumer's expectations, in the context of a specific purpose for data use.	[113]
DQ is most commonly defined as 'fitness for use'	[20]
DQ: Context Agnostic Perspective	
Documentation and contents of data within an electronic medical record (EMR) must be accurate, complete, concise, consistent and universally understood by users of the data, and must support the legal business record of the organization by maintaining the required parameters such as consistency, completeness and accuracy.	[30]
EHR DQ dimensions: completeness, correctness, concordance, plausibility, and currency.	[22]
Relevant, necessary, accurate, complete, and updated data	[114]
Data that are accurate, relevant, valid, reliable, legible, complete, and available when it is needed by decision-makers for healthcare delivery and planning purposes; DQ consists of six primary dimensions, which includes completeness, consistency, conformity, accuracy, integrity and timeliness	[33]
Variations in expected data versus collected data (e.g., timeliness, accuracy) are collectively referred to as DQ	[115]
Three DQ categories: conformance, completeness, and plausibility	[116]

EMR DQ dimensions: correctness (i.e., accuracy), completeness, concordance (i.e., accessibility), currency (i.e., timeliness), and plausibility (i.e., relevancy).	[17]
The core framework includes three constructs of DQ: complete, correct, and current data… EHR data completeness can be defined in multiple ways, depending upon intended use, and that, in turn, efforts to calculate rates of records completeness would vary based upon these different definitions and uses	[18]
"accuracy, believability, reputation, objectivity, factuality, consistency, freedom from bias, correctness, and unambiguousness."	[21]
Three categories: currency, completeness, and correctness. To estimate correctness, two further categories—plausibility and concordance—were used	[117]
A proper assessment of DQ will examine the data from several perspectives or dimensions including validity, accuracy, completeness, relevance, timeliness, availability, comparability, consistency, duplication, integrity and conformity	[118]
Accuracy, availability, usability, integrity, consistency, standardisation and timeliness are some characteristics of high-quality data	[62]

Appendix 5: Evidence of the Sub-Theme for Each DQ Dimension

Dimension	Sub-Theme	Reference
Accuracy	Validity	[19, 20, 37, 45, 51, 57, 58, 69, 93, 97, 100, 119-129]
	Correctness	[9, 11, 14, 16, 18, 21, 30, 35, 39, 45, 46, 49-51, 54, 57, 58, 60, 63, 64, 66, 69, 70, 76, 77, 80-82, 93, 95, 96, 117, 119, 122, 124, 126, 128, 130-134]
	Integrity	[8, 10, 33, 39, 46, 49, 53, 58, 86, 93, 95, 97, 101, 112, 122, 135-139]
	Conformance	[19, 33, 42, 45, 58, 116, 140]
	Plausibility	[14, 16, 18, 19, 35, 42, 45, 57, 58, 68, 82, 93, 101, 117, 140-142]
	Veracity	[21, 46]
	Accurate Diagnostic Data	[16, 39, 43, 44, 46, 51, 52, 93, 123, 126, 131, 143, 144]
Consistency	Inconsistent data capturing	[16, 20, 33, 34, 43, 44, 46, 48, 49, 54, 56, 64, 67, 73, 77, 81, 83, 96, 97, 100, 123, 124, 127, 128, 133, 135, 143, 145-149]
	Standardisation	[9, 11, 16, 39, 42, 43, 54, 57, 64, 65, 67, 74, 76, 80, 81, 97, 100, 120, 122, 127, 130, 133, 141, 145, 150-153]
	Concordance	[12, 14-16, 18, 20, 30, 37, 51, 54, 57, 61, 82, 97, 111, 127, 137, 142, 148, 154, 155]
	Uniqueness	[39, 44, 48, 53, 66, 78, 81, 84-87, 97, 131, 143]
	Data variability	[11, 39, 63, 68, 79, 93, 131, 133, 145, 152,

		156-159]
	Temporal variability	[51, 54, 57, 67, 72, 74, 75, 79, 124, 127, 130, 153, 160]
	System differences	[34, 39, 43, 44, 49, 66, 73, 76, 84, 143, 145, 146]
	Semantic consistency	[16, 20, 39, 54, 80, 93, 97, 100, 124, 128]
	Structuredness	[9, 18, 20, 36, 37, 40, 93]
	Representational consistency	[15, 20, 67, 80]
Completeness	Missing data	[10-12, 14-16, 22, 30, 33-35, 37, 39, 42, 45, 46, 48-51, 53, 54, 56-58, 61, 63, 64, 66, 69, 70, 73-75, 78, 80, 86, 93, 100, 110, 111, 125, 126, 128, 133, 135, 136, 141, 145, 147, 149, 161-172]
	Level of Completeness	[9, 16, 20, 35, 37, 39, 46, 56-58, 60-62, 64, 65, 84, 104, 110, 128, 141, 146, 167, 173, 174]
	Representativeness	[15, 16, 18, 20, 46, 51, 66, 68-72, 175]
	Fragmentation	[18, 49-55]
	Breadth of documentation	[12, 18]
Contextual Validity	Contextual DQ	[8, 11, 18, 32, 46, 69, 88, 93, 135]
	Fitness for use	[20, 46, 57, 70, 117, 119, 142, 176-178]
	Granularity	[16, 18, 67, 89]
	Relevancy	[69, 128]
Accessibility	Accessibility DQ	[18, 36-38, 40, 66, 104, 130, 143, 147]
	Availability	[15, 35, 36, 39, 66, 96, 147, 165]
Currency	Timeliness	[16, 18, 22, 33, 35, 41, 51, 63-65, 78, 82, 90-92, 95, 117, 179]

Appendix 6: Evidence for the interrelationships between the dimensions of DQ

Relationship	Evidence
Availability -> Currency	“Given the predominantly electronic form of communication between hospitals and general practitioners in Alberta, the inconsistency in availability of documentation in one single location can delay processes for practitioners searching for important health information.” [39]
Accuracy <- -> Contextual Validity (bidirectional)	“Counting complications would require interpretations of plausible temporal and causal relationships, which we were not always able to infer from observable codes. When a subject had received more than one intervention during an encounter, for example, it was difficult to determine which of the corresponding clinical events happened first and caused each other.” [16] “We believe a lack of granularity provokes incorrectness as only part of the true clinical course of a subject can be portrayed” [16]
Completeness ->	“Some providers questioned the integrity of EHR data and the

Accuracy	potential perpetuation of errors through incomplete or repeated data entry." [86]
Completeness -> Contextual validity	"In secondary use settings, EHR data completeness becomes extrinsic, and is dependent upon whether or not there are sufficient types and quantities of data to perform a research task of interest." [70]
Consistency -> Accessibility	"Structured data entry (SDE) applications can prompt for completeness, provide greater accuracy and better ordering for searching and retrieval, and permit validity checks for DQ monitoring, research, and especially decision support" [37]
Consistency -> Contextual validity	"Information inaccuracy was also frequently observed. It was reflected as poor granularity of the diagnosis terms or disease classification codes and inadequate or non-standardized documentation of disease status or treatment details. Consequently, such information could not satisfy the information needs of a survival analysis study." [67]
Consistency -> Accuracy	"We found two factors related to EHR documentation practices. False negatives and false positives in the problem list sometimes arose when the problem list was not consistently maintained and was therefore out-of-date, either because a resolved problem was not removed or because an active problem was not added (or was added after the measurement period concluded)." [51]
Consistency -> Completeness	"The actually corresponding procedure codes for the described operation techniques in the original study were not frequently used in our EHR, which instead employed different procedure codes; this suggests that documentation habits may have affected frequency estimates. We were unable to clearly ascertain which procedure codes represented treatment of conditions that had been documented via simultaneous diagnostic codes."[32]
Consistency -> Currency	"Documentation factors: We found two factors related to EHR documentation practices. False negatives and false positives in the problem list sometimes arose when the problem list was not consistently maintained and was therefore out-of-date, either because a resolved problem was not removed or because an active problem was not added (or was added after the measurement period concluded)." [51]
Currency -> Accuracy	"Data is entered at different times. Some data are entered into the electronic system in real-time during admissions but other data are recorded on paper and only entered into EHR at the end of patient's admission to the hospital. This can result in some of the data not entered into system or data recorded with errors." [52]

Appendix 7: Evidence for the outcomes of Data Quality

Outcome	Description	Evidence
Clinical	The extent to which digital health DQ	• "Healthcare professional access to complete lifelong patient information will facilitate more

	impacts healthcare consumers.	effective, personalised delivery of care and increased patient safety"[180] • "When there is a gap or incomplete data from what is expected can lead to poor or delayed patient care that can lead to death, e.g., wrong results to wrong patient" [33]
Business process	The extent to which digital health DQ impacts the efficiency and effectiveness of healthcare-related business processes.	• The "timely and efficient access to all relevant information" [66] streamlines clinical practice and minimises unnecessary tasks [15, 180, 181] • The absence of a discharge summary can hinder communication between hospitals and general practitioners [15]
Clinician	The extent to which digital health DQ impacts frontline healthcare professionals.	• Nurses identified that EHR data will eliminate paperwork, improve ability to monitor patients, and decrease their workflow [36] • Poor data quality increases workload due to the documentation burden associated with inconsistent diagnosis codes [43] and inconsistency between data recorded across health settings [15]
Research-related	The extent to which the reusability of digital health DQ impacts clinical research outcomes.	• Well managed, high-quality digital health data facilitates data analytics [79], data retrieval [100], supporting the reusability of data [15, 16, 51, 66, 128, 165] and can be applied in medical research related to clinical trials [20, 67, 70, 95, 96, 100, 128, 141, 165, 177, 182-185] • The efficacy and quality of the research depend on the quality of the healthcare records [15, 20, 43, 67, 70, 81, 97, 100, 142, 185]
Organisational	The extent to which digital health DQ impacts institutional finances, policy, and regulation compliance.	• "High DQ in medical records is fundamental to good clinical practice, program management and ultimately to policy decisions" [30] and further supports auditing and monitoring [30, 67, 81, 183] • DQ issues can negatively impact institutional finances and regulatory compliance. [117]

References

1 Duncan R, Eden R, Woods L, et al. Synthesizing Dimensions of Digital Maturity in Hospitals: Systematic Review. Journal of medical Internet research. 2022 Mar 30;24(3):e32994.

2 Eden R, Burton-Jones A, Scott I, et al. Effects of eHealth on hospital practice: synthesis of the current literature. Australian Health Review. 2018;42(5):568-78.

3 Rahimi K. Digital health and the elusive quest for cost savings. The Lancet Digital Health. 2019;1(3):e108-e9.

4 Eden R, Burton-Jones A, Scott I, et al. The impacts of eHealth upon practice; Synthesis of the current literature. Deeble Institute for Health Policy Research. 2017(16):1-7.

5 Zheng K, Abraham J, Novak LL, et al. A survey of the literature on unintended consequences associated with health information technology: 2014-2015. Yearbook of Medical Informatics. 2016;25(1):13-29.

6 Reisman M. EHRs: The challenge of making electronic data usable and interoperable. Pharmacy and Therapeutics. 2017;42(9):572-5.

7 World Health Organization. Global Strategy of Digital Health 2020-2025. 2021 [cited 2021 15/10/2021]; Available from: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf.

8 Darko-Yawson S, Ellingsen G. Assessing and Improving EHRs Data Quality through a Socio-technical Approach. 7th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2016)/The 6th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2016)/Affiliated Workshops; 2016; London, United Kingdom: 98; 2016. p. 243-50.

9 Afzal M, Hussain M, Ali Khan W, et al. Comprehensible knowledge model creation for cancer treatment decision making. Computers in Biology and Medicine. 2017;82:119-29.

10 Assale M, Dui LG, Cina A, et al. The revival of the notes field: Leveraging the unstructured content in electronic health records. Frontiers in medicine. 2019;6:66.

11 Savitz ST, Savitz LA, Fleming NS, et al. How much can we trust electronic health record data? Healthcare. 2020;8(3):1-4.

12 Beauchemin M, Weng C, Sung L, et al. Data quality of chemotherapy-induced nausea and vomiting documentation. Applied Clinical Informatics. 2021;12(2):320-8.

13 Puttkammer N, Baseman JG, Devine EB, et al. An assessment of data quality in a multi-site electronic medical record system in Haiti. International Journal of Medical Informatics. 2016;86(February 2016):104-16.

14 Wang Z, Penning M, Zozus M. Analysis of anesthesia screens for rule-based data quality assessment opportunities. Studies in health technology and informatics. 2019;257:473-8.

15 Wiebe N, Xu Y, Shaheen AA, et al. Indicators of missing electronic medical record (EMR) discharge summaries: A retrospective study on Canadian data. International Journal of Population Data Science. 2020;5(1):1-10.

16 von Lucadou M, Ganslandt T, Prokosch HU, et al. Feasibility analysis of conducting observational studies with the electronic health record. BMC Medical Informatics and Decision Making. 2019;19(1):1-14.

17 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. Journal of the American Medical Informatics Association. 2013;20(1):144-51.
18 Weiskopf NG, Bakken S, Hripcsak G, et al. A data quality assessment guideline for electronic health record data reuse. eGEMs. 2017;5(1):1-19.
19 Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. eGEMs. 2016;4(1):1-21.
20 Richesson RL, Hammond WE, Nahm M, et al. Electronic health records based phenotyping in next-generation clinical trials: A perspective from the NIH Health care systems collaboratory. Journal of the American Medical Informatics Association. 2013;20(e2):e226-e31.
21 Reimer AP, Madigan EA. Veracity in big data: How good is good enough. Health Informatics Journal. 2019;25(4):1290-8.
22 Bettencourt-Silva JH, Clark J, Cooper CS, et al. Building data-driven pathways from routinely collected hospital data: A case study on prostate cancer. JMIR Medical Informatics. 2015;3(3):e26.
23 Britt H, Pollock A, Wong C, et al. Can Medicare sustain the health of our ageing populate. The Conversation. 2015.
24 Webster J, Watson RT. Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Quarterly. 2002;26(2):xiii-xxiii.
25 Templier M, Paré G. A framework for guiding and evaluating literature reviews. Communications of the Association for Information Systems. 2015;37(1):6.
26 Saldaña J. The coding manual for qualitative researchers: Sage; 2015.
27 Dubois A, Gadde L-E. "Systematic combining"-A decade later. Journal of Business Research. 2014;67(6):1277.
28 Gioia DA, Corley KG, Hamilton AL. Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. Organ Res Methods. 2013;16(1):15-31.
29 Best P, Badham J, McConnell T, et al. Participatory theme elicitation: open card sorting for user led qualitative data analysis. International Journal of Social Research Methodology. 2021:1-19.
30 Abiy R, Gashu K, Asemaw T, et al. A comparison of electronic medical record data to paper records in antiretroviral therapy clinic in Ethiopia: What is affecting the quality of the data? Online Journal of Public Health Informatics. 2018;10(2):1-13.
31 ISO. ISO/IEC 25012. ISO 25000 STANDARDS: ISO25000.com; 2019.
32 Liu C, Zowghi D, Talaei-Khoei A, et al. Empirical study of Data Completeness in Electronic Health Records in China. Pacific Asia Journal of the Association for Information Systems. 2020;12(2):1-26.
33 Makeleni N, Cilliers L. Critical success factors to improve data quality of electronic medical records in public healthcare institutions. SA Journal of Information Management. 2021;23(1):1-8.
34 Estiri H, Stephens KA, Klann JG, et al. Exploring completeness in clinical data research networks with DQe-c. Journal of the American Medical Informatics Association. 2018;25(1):17-24.
35 Afshar AS, Li Y, Chen Z, et al. An exploratory data quality analysis of time series physiologic signals using a large-scale intensive care unit database. JAMIA Open. 2021;4(3):1-6.

36 Top M, Yilmaz A, Gider Ö. Electronic medical records (EMR) and nurses in Turkish hospitals. Systemic Practice and Action Research. 2013;26(3):281-97.
37 Roukema J, Los RK, Bleeker SE, et al. Paper versus computer: Feasibility of an electronic medical record in general pediatrics. Pediatrics. 2006;117(1):15-21.
38 Rosenlund M, Kivekas E, Mikkonen S, et al. Health professionals' perceptions of information quality in the health village portal. Health Informatics Vision: From Data via Information to Knowledge. 2019 ed. IOS Press Ebooks; 2019. p. 300-3.
39 Dentler K, Cornet R, ten Teije A, et al. Influence of data quality on computed Dutch hospital quality indicators: A case study in colorectal cancer surgery. BMC Medical Informatics and Decision Making. 2014;14(1):1-10.
40 Yoo CW, Huang CD, Goo J, et al. Explaining Task Support Satisfaction on Electronic Patient Care Report (ePCR) in Emergency Medical Services (EMS): An Elaboration Likelihood Model Lens. ICIS 2017: Transforming Society with Digital Innovation; 2018; Seoul, South Korea; 2018.
41 Almutiry O, Wills G, Alwabel A, et al. Toward a framework for data quality in cloud-based health information system. International Conference on Information Society (i-Society 2013); 2013; Toronto, Canada; 2013. p. 153-7.
42 Lee K, Weiskopf N, Pathak J. A framework for data quality assessment in clinical research datasets. 2017 AMIA Annual Symposium; 2017; Washington, DC, USA; 2017. p. 1080-9.
43 Diaz-Garelli JF, Strowd R, Ahmed T, et al. A tale of three subspecialties: Diagnosis recording patterns are internally consistent but specialty-dependent. JAMIA Open. 2019;2(3):369-77.
44 Diaz-Garelli F, Strowd R, Ahmed T, et al. What Oncologists Want: Identifying Challenges and Preferences on Diagnosis Data Entry to Reduce EHR-Induced Burden and Improve Clinical Data Quality. JCO Clinical Cancer Informatics. 2021;5(2021):527-40.
45 Sirgo G, Esteban F, Gomez J, et al. Validation of the ICU-DaMa tool for automatically extracting variables for minimum dataset and quality indicators: The importance of data quality assessment. International Journal of Medical Informatics. 2018;112(2018):166-72.
46 Sukumar SR, Natarajan R, Ferrell RK. Quality of big data in health care. International Journal of Health Care Quality Assurance. 2015;28(6):621-34.
47 Weiskopf NG, Hripcsak G, Swaminathan S, et al. Defining and measuring completeness of electronic health records for secondary use. Journal of biomedical informatics. 2013;46(5):830-6.
48 Gloyd S, Wagenaar BH, Woelk GB, et al. Opportunities and challenges in conducting secondary analysis of HIV programmes using data from routine health information systems and personal health information. Journal of the International AIDS Society. 2016;19(5 Suppl 4):1-6.
49 Fox F, Aggarwal VR, Whelton H, et al. A data quality framework for process mining of electronic health record data. 2018 IEEE International Conference on Healthcare Informatics (ICHI); 2018; New York, USA: IEEE; 2018. p. 12-21.
50 McCormack JL, Ash JS. Clinician perspectives on the quality of patient data used for clinical decision support: A qualitative study. 2012 AMIA Annual Symposium; 2012; Chicago, USA: American Medical Informatics Association; 2012. p. 1302-9.

51 Weiskopf NG, Cohen AM, Hannan J, et al. Towards augmenting structured EHR data: A comparison of manual chart review and patient self-report. 2019 AMIA Annual Symposium 2019; Washington, USA; 2019. p. 903-12.
52 Wilk M, Marsh DWR, De Freitas S, et al. Predicting length of stay in hospital using electronic records available at the time of admission. Studies in health technology and informatics. 2020;270:377-81.
53 Kindermann A, Tute E, Benda S, et al. Preliminary analysis of structured reporting in the HiGHmed use case cardiology: Challenges and measures. German Medical Data Sciences: Bringing Data to Life. 2021 ed; 2021. p. 187-94.
54 Lanzola G, Parimbelli E, Micieli G, et al. Data quality and completeness in a web stroke registry as the basis for data and process mining. Journal of Healthcare Engineering. 2014;5(2):163-84.
55 Rule A, Rick S, Chiu M, et al. Validating free-text order entry for a note-centric EHR. 2015 AMIA Annual Symposium; 2015; San Francisco, USA; 2015. p. 1103-10.
56 Alwhaibi M, Balkhi B, Alshammari TM, et al. Measuring the quality and completeness of medication-related information derived from hospital electronic health records database. Saudi Pharmaceutical Journal. 2019;27(4):502-6.
57 van Hoeven LR, Bruijne MC, Kemper PF, et al. Validation of multisource electronic health record data: An application to blood transfusion data. BMC Medical Informatics and Decision Making. 2017;17(1):1-10.
58 Corey KM, Helmkamp J, Simons M, et al. Assessing quality of surgical real-world data from an automated electronic health Record pipeline. Journal of the American College of Surgeons. 2020;230(3):295-305.
59 Rutzner S, Ganslandt T, Fietkau R, et al. Noncurated data lead to misinterpretation of treatment outcomes in patients with prostate cancer after salvage or palliative radiotherapy. JCO Clinical Cancer Informatics. 2019;3(2019):1-11.
60 Hosseini N, Mostafavi SM, Zendehdel K, et al. Factors affecting clinicians' adherence to principles of diagnosis documentation: A concept mapping approach for improved decision-making. Health Information Management Journal. 2021:1-10.
61 Bae CJ, Griffith S, Fan Y, et al. The challenges of data quality evaluation in a joint data warehouse. eGEMs. 2015;3(1):1125.
62 Perimal-Lewis L, Teubner D, Hakendorf P, et al. Application of process mining to assess the data quality of routinely collected time-based performance data sourced from electronic health records by validating process conformance. Health Informatics Journal. 2016;22(4):1017-29.
63 Skyttberg N, Chen R, Blomqvist H, et al. Exploring vital sign data quality in electronic health records with focus on emergency care warning scores. Applied Clinical Informatics. 2017;8(3):880-92.
64 Skyttberg N, Vicente J, Chen R, et al. How to improve vital sign data quality for use in clinical decision support systems? A qualitative study in nine Swedish emergency departments. BMC Medical Informatics and Decision Making. 2016;16(1):1-12.
65 Skyttberg N, Chen R, Koch S. Man vs machine in emergency medicine - a study on the effects of manual and automatic vital sign documentation on data quality

and perceived workload, using observational paired sample data and questionnaires. BMC Emergency Medicine. 2018;18(1):1-9.
66 Downey S, Indulska M, Sadiq S. Perceptions and challenges of EHR clinical data quality. Australasian Conference on Information Systems 2019; 2019; Perth, Australia; 2019. p. 233-43.
67 Botsis T, Hartvigsen G, Chen F, et al. Secondary use of EHR: Data quality issues and informatics opportunities. Summit on Translational Bioinformatics. 2010;2010(2010):1-5.
68 Polubriaginof F, Salmasian H, Albert DA, et al. Challenges with collecting smoking status in electronic health records. 2017 AMIA Annual Symposium; 2017; Washington, USA: American Medical Informatics Association; 2017. p. 1392-400.
69 Schneeweiss S, Glynn RJ. Real-world data analytics fit for regulatory decision-making. American Journal of Law and Medicine. 2018;44(2-3):197-217.
70 Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: The non-random completeness of electronic health records. 2013 AMIA Annual Symposium; 2013; Washington, USA; 2013. p. 1472-7.
71 Li S, Zhang L, Liu S, et al. Surveillance of noncommunicable disease epidemic through the integrated noncommunicable disease collaborative management system: Feasibility pilot study conducted in the city of ningbo, China. Journal of Medical Internet Research. 2020;22(7):1-11.
72 Lingren T, Sadhasivam S, Zhang X, et al. Electronic medical records as a replacement for prospective research data collection in postoperative pain and opioid response studies. International Journal of Medical Informatics. 2018;111(March 2018):45-50.
73 Garg N, Kuperman G, Onyile A, et al. Validating health information exchange (HIE) data for quality measurement across four hospitals. 2014 AMIA Annual Symposium; 2014; Washington DC, USA: American Medical Informatics Association; 2014. p. 573-9.
74 Baker K, Dunwoodie E, Jones RG, et al. Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy. International Journal of Medical Informatics. 2017;103:32-41.
75 Callahan A, Shah NH, Chen JH. Research and reporting considerations for observational studies using electronic health record data. Annals of Internal Medicine. 2020;172(11 Supplement):S79-S84.
76 Deng X, Lin WH, E-Shyong T, et al. From descriptive to diagnostic analytics for assessing data quality: An application to temporal data elements in electronic health records. 3rd IEEE EMBS International Conference on Biomedical and Health Informatics (BHI 2016); 2016; Las Vegas, USA: IEEE; 2016. p. 236-9.
77 Duan R, Cao M, Wu Y, et al. An empirical study for impacts of measurement errors on EHR based association studies. AMIA Annual Symposium Proceedings. 2016 ed: American Medical Informatics Association; 2016. p. 1764-73.
78 Gumede-Moyo S, Todd J, Bond V, et al. A qualitative inquiry into implementing an electronic health record system (SmartCare) for prevention of mother-to-child transmission data in Zambia: A retrospective study. BMJ open. 2019;9(9):1-9.
79 Saez C, Gutierrez-Sacristan A, Kohane I, et al. EHRtemporalVariability: Delineating temporal data-set shifts in electronic health records. GigaScience. 2020;9(8):1-7.

80 Jetley G, Zhang H. Electronic health records in IS research: Quality issues, essential thresholds and remedial actions. Decision Support Systems. 2019;126(November 2019):1.
81 Deakyne Davies SJ, Grundmeier RW, Campos DA, et al. The pediatric emergency care applied research network registry: A multicenter electronic health record registry of pediatric emergency care. Applied Clinical Informatics. 2018;9(2):366-76.
82 Reimer AP, Milinovich A, Madigan EA. Data quality assessment framework to assess electronic medical record data for use in research. International Journal of Medical Informatics. 2016;90(June 2016):40-7.
83 Chen ES, Carter EW, Sarkar IN, et al. Examining the use, contents, and quality of free-text tobacco use documentation in the electronic health record. AMIA Annual Symposium; 2014; Washington DC, USA; 2014. p. 366-74.
84 El Fadly A, Lucas N, Rance B, et al. The REUSE project: EHR as single datasource for biomedical research. MEDINFO 2010. 2010 ed; 2010. p. 1324-8.
85 Garcia-de-Leon-Chocano R, Munoz-Soler V, Saez C, et al. Construction of quality-assured infant feeding process of care data repositories: Construction of the perinatal repository (Part 2). Computers in Biology and Medicine. 2016;71(April 2016):214-22.
86 Hamamura FD, Withy K, Hughes K. Identifying barriers in the use of electronic health records in Hawai'i. Hawaii J Med Public Health. 2017;76(3 Suppl 1):28-35.
87 Just BH, Marc D, Munns M, et al. Why patient matching is a challenge: Research on master patient index (MPI) data discrepancies in key identifying fields. Perspectives in health information management. 2016;13(Spring):1-20.
88 Daniel C, Serre P, Orlova N, et al. Initializing a hospital-wide data quality program. The AP-HP experience. Computer methods and programs in biomedicine. 2019;181:1-8.
89 Suriadi S, Mans RS, Wynn MT, et al. Measuring patient flow variations: A cross-organisational process mining approach. Asia-Pacific Conference on Business Process Management; 2014; Brisbane, Australia: Springer; 2014. p. 43-58.
90 Chiasera A, Toai TJ, Bogoni LP, et al. Federated EHR: How to improve data quality maintaining privacy. 2011 IST-Africa Conference; 2011; Gaborone, Botswana; 2011. p. 1-8.
91 Boyle DIR, Cunningham SG. Resolving fundamental quality issues in linked datasets for clinical care. Health Informatics Journal. 2016;8(2):73-7.
92 Tilahun B, Fritz F. Modeling antecedents of electronic medical record system implementation success in low-resource setting hospitals. BMC Medical Informatics and Decision Making. 2015;15(1):1-9.
93 Kim HS, Kim JH. Proceed with caution when using real world data and real world evidence. Journal of Korean Medical Science. 2019;34(4):1-5.
94 Yoo CW, Huang CD, Goo J. Task support of electronic patient care report (ePCR) systems in emergency medical services: An elaboration likelihood model lens. Information & Management. 2020;57(6):1-12.
95 Koepke R, Petit AB, Ayele RA, et al. Completeness and accuracy of the wisconsin immunization registry: An evaluation coinciding with the beginning of meaningful use. Journal of Public Health Management and Practice. 2015;21(3):273-81.

96 Gibby GL. Anesthesia information-management systems: Their role in risk-versus cost assessment and outcomes research. Journal of Cardiothoracic and Vascular Anesthesia. 1997;11(2):2-5.
97 Kim M, Shin SY, Kang M, et al. Developing a standardization algorithm for categorical laboratory tests for clinical big data research: Retrospective study. JMIR Medical Informatics. 2019;7(3):1-13.
98 Liu C, Zowghi D, Talaei-Khoei A. Empirical Evaluation of the Influence of EMR Alignment to Care Processes on Data Completeness. 2020.
99 Wang RY, Strong DM. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems. 1996;12(4):5-33.
100 Johnson SG, Pruinelli L, Hoff A, et al. A framework for visualizing data quality for predictive models and clinical quality measures. AMIA Joint Summits on Translational Science Proceedings. 2019 ed. USA: AMIA; 2019. p. 630-8.
101 Dantanarayana G, Sahama T. Metrics for eHealth services improvement. 18th International Conference on e-Health Networking, Applications and Services (Healthcom 2016); 2016; Munich, Germany; 2016. p. 1-7.
102 Sung M, He J, Zhou Q, et al. Using an Integrated Framework to Investigate the Facilitators and Barriers of Health Information Technology Implementation in Noncommunicable Disease Management: Systematic Review. J Med Internet Res. 2022 Jul 20;24(7):e37338.
103 Kohane IS, Aronow BJ, Avillach P, et al. What every reader should know about studies using electronic health record data but may be afraid to ask. Journal of medical Internet research. 2021;23(3):e22219.
104 Top M, Yilmaz A, Karabulut E, et al. Validation of a nurses' views on electronic medical record systems (EMR) questionnaire in Turkish health system. Journal of Medical Systems. 2015;39(6):1-7.
105 Andrews R, Emamjome F, ter Hofstede AH, et al. Root-cause analysis of process-data quality problems. Journal of Business Analytics. 2022;5(1):51-75.
106 Poppe E, Pika A, Wynn MT, et al. Extracting Best-Practice Using Mixed-Methods. Business & Information Systems Engineering. 2021 2021/12/01;63(6):637-51.
107 Yin RK. Qualitative research from start to finish: Guilford publications; 2015.
108 Black AD, Car J, Pagliari C, et al. The impact of eHealth on the quality and safety of health care: a systematic overview. PLoS medicine. 2011;8(1):e1000387.
109 Liaw ST, Taggart J, Yu H. EHR-based disease registries to support integrated care in a health neighbourhood: An ontology-based methodology. e-Health – For Continuity of Care. 2014 ed. IOS Press Ebooks; 2014. p. 171-5.
110 Liu C, Zowghi D, Talaei-Khoei A. Empirical evaluation of the influence of EMR alignment to care processes on data completeness. 53rd Hawaii International Conference on System Sciences; 2020; Grand Wailea, Hawaii; 2020. p. 3519-28.
111 Muthee V, Bochner AF, Osterman A, et al. The impact of routine data quality assessments on electronic medical record data quality in Kenya. PLoS One. 2018;13(4):1-10.
112 Ryan J, Doster B, Daily S, et al. Data quality assurance via perioperative EMR reconciliation. 24th Americas Conference on Information Systems 2018; 2018; New Orleans, USA: Association for Information Systems; 2018.
113 Landis-Lewis Z, Manjomo R, Gadabu OJ, et al. Barriers to using eHealth data for clinical performance feedback in Malawi: A case study. International Journal of Medical Informatics. 2015;84(10):868-75.

114 Chunyan D. Patient privacy protection in China in the age of electronic health records. Hong Kong Law Journal. 2013;43:245-78.
115 Penning ML, Blach C, Walden A, et al. Near real time EHR data utilization in a clinical study. Digital Personalized Health and Medicine. 2020 ed. IOS Press Ebooks; 2020. p. 337-41.
116 Rajamani S, Kayser A, Emerson E, et al. Evaluation of data exchange process for interoperability and impact on electronic laboratory reporting quality to a state public health agency. Online Journal of Public Health Informatics. 2018;10(2):1-13.
117 Wang Z, Talburt JR, Wu N, et al. A rule-based data quality assessment aystem for electronic health record data. Applied Clinical Informatics. 2020;11(4):622-34.
118 Hart R, Kuo MH. Better data quality for better healthcare research results - A case study. Building Capacity for Health Informatics in the Future. 2017 ed. Netherlands: IOS Press Ebooks; 2017. p. 161-6.
119 Garcia-de-Leon-Chocano R, Saez C, Munoz-Soler V, et al. Robust estimation of infant feeding indicators by data quality assessment of longitudinal electronic health records from birth up to 18 months of life. Computer methods and programs in biomedicine. 2021;207(August 2021):1-10.
120 Liaw ST, Chen HY, Maneze D, et al. The quality of routinely collected data: Using the "principal diagnosis" in emergency department databases as an example. Electronic Journal of Health Informatics. 2012;7(1):1-8.
121 Baernholdt M, Dunton N, Hughes RG, et al. Quality measures: A stakeholder analysis. Journal of Nursing Care Quality. 2018;33(2):149-56.
122 Curtis MD, Griffith SD, Tucker M, et al. Development and validation of a high-quality composite real-world mortality endpoint. Health Services Research. 2018;53(6):4460-76.
123 Gupta S, Liu L, Patterson OV, et al. A framework for leveraging "big data" to advance epidemiology and improve quality: Design of the VA colonoscopy collaborative. eGEMs. 2018;6(1):1-9.
124 Knake LA, Ahuja M, McDonald EL, et al. Quality of EHR data extractions for studies of preterm birth in a tertiary care center: Guidelines for obtaining reliable data. BMC Pediatrics. 2016;16(1):1-8.
125 Rees S, Akbari A, Collins H, et al. Developing a standardised approach to the aggregation of inpatient episodes into person-based spells in all specialties and psychiatric specialties. BMC Medical Informatics and Decision Making. 2019;19(1):1-12.
126 Salomon RM, Blackford JU, Rosenbloom ST, et al. Openness of patients' reporting with use of electronic records: Psychiatric clinicians' views. Journal of the American Medical Informatics Association. 2010;17(1):54-60.
127 Tantoso E, Wong WC, Tay WH, et al. Hypocrisy around medical patient data: Issues of access for biomedical research, data quality, usefulness for the purpose and omics data as game changer. Asian bioethics review. 2019;11(2):189-207.
128 Toftdahl AKS, Pape-Haugaard LB, Palsson TS, et al. Collect once - Use many times: The research potential of low back pain patients' municipal electronic healthcare records. Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth. IOS Press Ebooks; 2018. p. 211-5.
129 Zaghir J, Rodrigues-Jr JF, Goeuriot L, et al. Real-world patient trajectory prediction from clinical notes using artificial neural networks and UMLS-based

extraction of concepts. Journal of Healthcare Informatics Research. 2021;5(4):474-96.
130 Cohen B, Vawdrey DK, Liu J, et al. Challenges associated with using large data sets for quality assessment and research in clinical settings. Policy Polit Nurs Pract. 2015;16(3-4):117-24.
131 Diaz-Garelli F, Strowd R, Lawson VL, et al. Workflow differences affect data accuracy in oncologic EHRs: A first step toward detangling the diagnosis data babel. JCO Clinical Cancer Informatics. 2020;4(2020):529-38.
132 Smith N, Coleman KJ, Lawrence JM, et al. Body weight and height data in electronic medical records of children. International Journal of Pediatric Obesity. 2010;5(3):237-42.
133 Welch G, von Recklinghausen F, Taenzer A, et al. Data cleaning in the evaluation of a multi-site intervention project. eGEMs. 2017;5(3):1-7.
134 Wennberg S, Karlsen LA, Stalfors J, et al. Providing quality data in health care - almost perfect inter-rater agreement in the Norwegian tonsil surgery register. BMC Medical Research Methodology. 2019;19(1):1-9.
135 Aalsma MC, Schwartz K, Haight KA, et al. Applying an electronic health records data quality framework across service sectors: A case study of juvenile justice system data. eGEMs. 2019;7(1):1-10.
136 Alves DS, Maranhão PA, Pereira AM, et al. Can openEHR represent the clinical concepts of an obstetric-specific EHR - Obscare software? MEDINFO 2019: Health and Wellbeing e-Networks for All. IOS Press; 2019. p. 773-7.
137 Kheterpal S. Clinical research using an information system: The multicenter perioperative outcomes group. Anesthesiology Clinics. 2011;29(3):377-88.
138 O'Shea MP, Kennedy C, Relihan E, et al. Assessment of an electronic patient record system on discharge prescribing errors in a Tertiary University Hospital. BMC Medical Informatics and Decision Making. 2021;21(1):1-11.
139 Ward MJ, Froehle CM, Hart KW, et al. Operational data integrity during electronic health record implementation in the ED. American Journal of Emergency Medicine. 2013;31(7):1029-33.
140 Kamdje-Wabo G, Gradinger T, Lobe M, et al. Towards structured data quality assessment in the German medical informatics initiative: initial approach in the MII demonstrator study. MEDINFO 2019: Health and Wellbeing e-Networks for All. 2019 ed. IOS Press Ebooks; 2019. p. 1508-9.
141 Estiri H, Klann JG, Murphy SN. A clustering approach for detecting implausible observation values in electronic health records data. BMC Medical Informatics and Decision Making. 2019;19(1):1-16.
142 Johnson SG, Speedie S, Simon G, et al. A data quality ontology for the secondary use of EHR data. 2015 AMIA Annual Symposium; 2015; San Francisco, USA; 2015. p. 1937-46.
143 Diaz-Garelli JF, Strowd R, Wells BJ, et al. Lost in translation: Diagnosis records show more inaccuracies after biopsy in oncology care EHRs. AMIA Joint Summits Translational Science Proceedings. 2019 ed. United States: AMIA; 2019. p. 325-34.
144 DeShazo JP, Hoffman MA. A comparison of a multistate inpatient EHR database to the HCUP nationwide inpatient sample. BMC Health Services Research. 2015;15(1):1-8.

145 Fu S, Leung LY, Raulli AO, et al. Assessment of the impact of EHR heterogeneity for clinical research through a case study of silent brain infarction. BMC Medical Informatics and Decision Making. 2020;20(1):1-12.
146 Funkner AA, Egorov MP, Fokin SA, et al. Citywide quality of health information system through text mining of electronic health records. Applied Network Science. 2021;6(1):1-21.
147 Potter LE, Purdie C, Nielsen S. The view from the trenches: Satisfaction with eHealth systems by a group of health professionals. 23rd Australasian Conference on Information Systems; 2012; Geelong, Australia; 2012.
148 Spuhl J, Doing-Harris K, Nelson S, et al. Concordance of electronic health record (EHR) data describing delirium at a VA hospital. 2014 AMIA Annual Symposium 2014; Washington DC, USA; 2014. p. 1066-71.
149 Zakim D, Brandberg H, El Amrani S, et al. Computerized history-taking improves data quality for clinical decision-making-Comparison of EHR and computer-acquired history data in patients with chest pain. PLoS One. 2021;16(9):1-13.
150 Groenhof TKJ, Koers LR, Blasse E, et al. Data mining information from electronic health records produced high yield and accuracy for current smoking status. Journal of clinical epidemiology. 2020;118(February 2020):100-6.
151 Sachdeva S, Batra S, Bhalla S. Evolving large scale healthcare applications using open standards. Health Policy and Technology. 2017;6(4):410-25.
152 Li Y, Sperrin M, Martin GP, et al. Examining the impact of data quality and completeness of electronic health records on predictions of patients' risks of cardiovascular disease. International Journal of Medical Informatics. 2020;133(January 2020):1-9.
153 Rockenschaub P, Nguyen V, Aldridge RW, et al. Data-driven discovery of changes in clinical code usage over time: a case-study on changes in cardiovascular disease recording in two English electronic health records databases (2001-2015). BMJ open. 2020;10(2):1-9.
154 Ansari S, Jain D, Harikumar H, et al. Identification of predictors and model for predicting prolonged length of stay in dengue patients. Health Care Management Science. 2021;24(4):786–98.
155 Flint A, Chaudhry NA, Riverso M, et al. Effective communication of cross-sectional imaging findings in Crohn's disease: comparing conventional EMR reporting to a published scoring system. Abdominal Radiology. 2018;43(7):1798-806.
156 Zozus MN, Young LW, Simon AE, et al. Training as an intervention to decrease medical record abstraction errors multicenter studies. Studies in health technology and informatics. 2019;257:526-39.
157 Kahn MG, Raebel MA, Glanz JM, et al. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Medical Care. 2012;50 1-16.
158 Reeves RM, FitzHenry F, Brown SH, et al. Who said it? Establishing professional attribution among authors of veterans' electronic health records. 2012 AMIA Annual Symposium; 2012; Chicago, USA; 2012. p. 753-62.
159 Spil TAM, Katsma CP, Stegwee RA, et al. Value, participation and quality of electronic health records in the Netherlands. 43rd Hawaii International Conference on System Sciences; 2010; Honolulu, Hawaii: IEEE; 2010.

160 Erlirianto LM, Ali AHN, Herdiyanti A. The implementation of the human, organization, and technology-fit (HOT-Fit) framework to evaluate the electronic medical record (EMR) system in a hospital. 3rd Information Systems International Conference 2015; 2015; Surabaya, Indonesia; 2015. p. 580-7.
161 Carr LL, Zelarney P, Meadows S, et al. Development of a cancer care summary through the electronic health record. Journal of Oncology Practice. 2016;12(2):e231-e40.
162 Fort D, Wilcox AB, Weng C. Could patient self-reported health data complement EHR for phenotyping? 2014 AMIA Annual Symposium; 2014; Washington DC, USA; 2014. p. 1738-47.
163 Hawley S, Yu J, Bogetic N, et al. Digitization of measurement-based care pathways in mental health through REDCap and electronic health record integration: Development and usability study. Journal of Medical Internet Research. 2021;23(5):1-32.
164 Miettinen M, Korhonen M. Information quality in healthcare: Coherence of data compared between organization's electronic patient records. 21st IEEE International Symposium on Computer-Based Medical Systems; 2008; Jyvaskyla, Finland; 2008. p. 488-93.
165 Yamamoto, Matsumoto, Yanagihara K, et al. A data-capture system for post-marketing surveillance of drugs that integrates with hospital electronic health records. Open Access Journal of Clinical Trials. 2011;3(2011):21-6.
166 Liaw ST, Chen HY, Maneze D, et al. Health reform: is routinely collected electronic information fit for purpose? Emerg Med Australas. 2012;24(1):57-63.
167 Liu C, Zowghi D, Talaei-Khoei A. An empirical study of the antecedents of data completeness in electronic medical records. International Journal of Information Management. 2020;50:155-70.
168 Piri S. Missing care: A framework to address the issue of frequent missing values;The case of a clinical decision support system for Parkinson's disease. Decision Support Systems. 2020;136:1.
169 Qiu H, Yu HY, Wang LY, et al. Electronic health record driven prediction for gestational diabetes mellitus in early pregnancy. Scientific Reports. 2017;7(1):1-13.
170 Razavi AR, Gill H, Åhlfeldt H, et al. A data pre-processing method to increase efficiency and accuracy in data mining. 2005 Conference on Artificial Intelligence in Medicine in Europe; 2005; Aberdeen, United Kingdom: Springer; 2005. p. 434-43.
171 Salati M, Pompili C, Refai M, et al. Real-time database drawn from an electronic health record for a thoracic surgery unit: High-quality clinical data saving time and human resources. European Journal of Cardio-thoracic Surgery. 2014;45(6):1017-9.
172 Weller GB, Lovely J, Larson DW, et al. Leveraging electronic health records for predictive modeling of post-surgical complications. Statistical Methods in Medical Research. 2018;27(11):3271-85.
173 Hawley G, Jackson C, Hepworth J, et al. Sharing of clinical data in a maternity setting: How do paper hand-held records and electronic health records compare for completeness? BMC Health Services Research. 2014;14(1):1-9.
174 Ni K, Chu H, Zeng L, et al. Barriers and facilitators to data quality of electronic health records used for clinical research in China: A qualitative study. BMJ open. 2019;9(7):1-7.

175 Polubriaginof FCG, Ryan P, Salmasian H, et al. Challenges with quality of race and ethnicity data in observational databases. Journal of the American Medical Informatics Association. 2019;26(8-9):730-6.
176 Griffiths R, Schluter DK, Akbari A, et al. Identifying children with cystic fibrosis in population-scale routinely collected data in Wales: A retrospective review. International Journal of Population Data Science. 2020;5(1):1-9.
177 Callahan T, Barnard J, Helmkamp L, et al. Reporting data quality assessment results: Identifying individual and organizational barriers and solutions. eGEMs. 2017;5(1):1-25.
178 Hausvik GI, Thapa D, Munkvold BE. Information quality life cycle in secondary use of EHR data. International Journal of Information Management. 2021;56(February 2021):1-14.
179 Colquhoun DA, Shanks AM, Kapeles SR, et al. Considerations for integration of perioperative electronic health records across institutions for research and quality improvement: The approach taken by the multicenter perioperative outcomes group. Anesthesia and analgesia. 2020;130(5):1133-46.
180 Black AS, Sahama T. Chronicling the patient journey: Co-creating value with digital health ecosystems. Australasian Computer Science Week (ACSW 2016); 2016; Canberra, Australia; 2016.
181 Mkalira Msiska KE, Kumitawa A, Kumwenda B. Factors affecting the utilisation of electronic medical records system in Malawian central hospitals. Malawi Medical Journal. 2017;29(3):247-53.
182 Almeida JR, Fajarda O, Pereira A, et al. Strategies to access patient clinical data from distributed databases. 12th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2019; 2019; Prague, Czech Republic; 2019. p. 466-73.
183 Laird-Maddox M, Mitchell SB, Hoffman M. Integrating research data capture into the electronic health record workflow: real-world experience to advance innovation. Perspectives in health information management. 2014;11(Fall):1-10.
184 Mathieu A, Sauthier M, Jouvet P, et al. Validation process of a high-resolution database in a paediatric intensive care unit-Describing the perpetual patient's validation. Journal of evaluation in clinical practice. 2021;27(2):316-24.
185 Kim HS, Kim H, Jeong YJ, et al. Development of clinical data mart of HMG-CoA reductase inhibitor for varied clinical research. Endocrinol Metab. 2017;32(1):90-8.
186 Makeleni, N. and L. Cilliers. Critical success factors to improve data quality of electronic medical records in public healthcare institutions. SA Journal of Information Management, 2021;23(1): p. 1-8.
187 Chang, I-C. Li, Y-C. Wu, T-Y. Yen, D. C. Electronic medical record quality and its impact on user satisfaction – Healthcare providers' point of view. Government Information Quaterly, 2012;29(2): p235-242.