A Review of the State of the Art of Data Quality in Healthcare

Caihua Liu, Guilin University of Electronic Technology, China
https://orcid.org/0000-0002-9641-4712

Amir Talaei-Khoei, University of Nevada, Reno, USA

Veda C. Storey, Georgia State University, USA

Guochao Peng, Sun Yat-sen University, China*
https://orcid.org/0000-0003-1206-5509

ABSTRACT

Effective implementation of strategic data-driven health analysis initiatives is heavily dependent on the quality of the electronic medical records that serve as the foundation from which to improve clinical decisions and, in turn, the quality of care. Although there is a large body of research on the quality of healthcare data, a systematical understanding of the methods used to address the issues of data quality is missing. This study analyzes research articles in health information systems/healthcare informatics on data quality to derive a set of dimensions for understanding data quality. Issues related to each dimension are identified and methods used to address them summarized. The issues and methods can inform healthcare professionals of how to improve data practices.

KEYWORDS

Data Quality, Healthcare, Literature Review

INTRODUCTION

Organizations and individuals increasingly rely on information systems (IS) for data-driven decision-making (ZareRavasan & Krčál, 2021), making it critical to ensure that the data being used is of sufficient quality. Poor quality data can cause the loss of revenue and even put lives at risk, especially in healthcare. For example, entry errors in a provider system resulted in an inappropriate treatment, with a patient developing seizures and requiring intubation (ECRI Institute, 2015). Duplicated drug orders entered into two separate prescribing systems used by a hospital resulted in nurses administering an excessive amount of insulin to a patient, resulting in death (Rowland, 2014). Furthermore, data not collected systematically in electronic medical records (EMR) or other operational systems can be of poorer quality, limiting its reuse (Kahn et al., 2016).

The poor delivery of evidence-based practice (James, 2013) and the high rate of adverse events (Landrigan et al., 2010) in hospital admissions and general practices motivate service providers to

DOI: 10.4018/JGIM.316236

*Corresponding Author

optimize clinical handovers between healthcare settings. This requires extensive efforts to improve the internal quality of data and advance the quality of information exchange programs. A substantial body of literature has identified the issues that may arise when undertaking these efforts, resulting in poor quality of medical or healthcare data (Michel-Verkerke, 2012; Vilic et al., 2016). Research on data quality for healthcare has developed several methods to resolve these issues (Prasser et al., 2018; Zięba, 2014). However, a systematic understanding of the methods is not available in the healthcare IS/informatics literature. It would be useful to link methods to data quality issues for two main reasons.

First, using EMR has created the potential to improve the quality of care and address cost-effectiveness. However, the scope of new risks (e.g., the quality of the data used in healthcare and research settings) for patients is still not fully understood (Rowland, 2014). Data quality issues are more prominent than in earlier healthcare technology initiatives but have not been taken seriously by practice. The widespread use of EMR technology is inevitable, so addressing data quality issues needs to be a priority.

Second, previous literature reviews of data quality for healthcare (Arts et al., 2002; Chen et al., 2014; Johnson et al., 2015; Kahn et al., 2016; Liaw et al., 2013; Thiru et al., 2003; Weiskopf & Weng, 2013) focus on describing and assessing data quality but do not provide a holistic picture that matches the methods to specific issues associated with different dimensions of data quality. As a result, there is a lack of studies to depict an overall understanding of methods used to address data quality issues in healthcare. Such a study could help healthcare and IT professionals ascertain what work has been done to address data quality issues in this field and identify possible gaps for further exploration.

In addition, a challenge arises in that the current data quality literature uses different terms to describe data quality dimensions, thus limiting our understanding of different or similar data quality dimensions discussed (Kahn et al., 2016; Liaw et al., 2013; Wang & Strong, 1996; Weiskopf & Weng, 2013). Furthermore, the lack of consistent use of terms that describe data quality dimensions makes it difficult to explain different or similar issues for a specific data quality dimension and establish relationships between methods and issues associated with different dimensions. Thus, standardizing the terms used in the literature for data quality dimensions, identifying the issues associated with the different dimensions, and understanding the methods used to address these issues could improve our understanding of this phenomenon, as well as the strengths and limitations of current methods, and future research needs.

This paper aims to identify definitions proposed for data quality in healthcare, and related data quality issues and methods for resolving them. Focusing on the healthcare IS/informatics field, this research identifies statistical and computational methods used to address data quality. We propose four research questions (RQs) to guide this study:

RQ1: What are the dimensions of data quality in healthcare?
RQ2: What are the relevant issues related to these dimensions of data quality?
RQ3: What are the methods used to address these issues?
RQ4: What are the strengths and limitations of these methods?

We reviewed the current research to identify relevant articles published from the start of 2012 to June 2022, from highly ranked journals within healthcare IS/informatics journals following the guidelines of Wolfswinkel et al. (2013). The review includes relevant articles cited by these papers and articles that cite them.

Our study differs from prior literature reviews from three perspectives. First, we adopt a manual search from the top healthcare IS/informatics journals based on a set of inclusion and exclusion criteria. The consistency of the literature in the highly prestigious outlets could offer a better chance to structure and discuss statistical and computational methods to improve healthcare quality enabled by IT. Second, we refer to a well-established taxonomy development approach (Nickerson et al., 2013) in IS to categorize 21 data quality issues under seven unique dimensions of data quality and six methods. Third, we establish relationships between the methods and specific issues under different dimensions of data quality that are not fully developed in the related literature reviews (Arts, De Keizer, & Scheffer, 2002; Chen et al., 2014; Johnson et al., 2015; Kahn et al., 2016; Liaw et al., 2013; Thiru et al., 2003; Weiskopf & Weng, 2013).

Academic contributions include the development of a systematic understanding of methods that can deal with data quality issues for achieving quality clinical decision-making and care. In addition, we relate the methods to specific issues under different dimensions of data quality. Practical contributions include specifying implications for healthcare professionals to improve data practices and helping assess healthcare organizations' data quality management process.

The rest of the paper is structured as follows. We next define the dimensions, issues, and methods used in this research and review related studies. Then, we explain the research methods. Next, we provide the method used for the data extraction and analysis results to address our research questions. The next section discusses our findings and contribution and suggests future research areas before the conclusion to the research is provided.

DEFINITIONS AND RELATED STUDIES

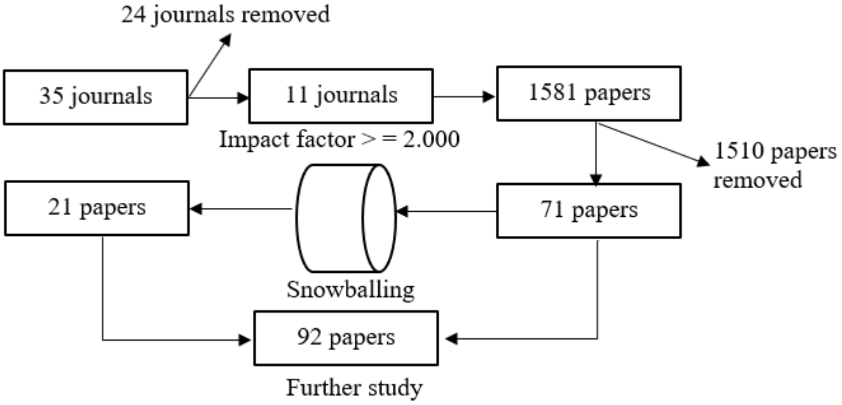
The most important concepts for this study are dimensions, issues, and methods, as they relate to data quality. Data quality refers to "fitness for use". A "data quality dimension" is a set of data quality attributes representing an individual aspect of data, such as completeness, accuracy, and timeliness (Storey et al., 2012; Wang et al., 1995). Issues of data quality are simply quality problems associated with data. These can be errors or anomalies, focusing on the manifestation of the issues under different data quality dimensions. Finally, we identified methods used to address data quality issues from strategies explicitly described in the literature.

We examine a body of literature that reviews articles regarding the quality of medical and healthcare records (Arts, De Keizer, & Scheffer, 2002; Chen et al., 2014; Johnson et al., 2015; Kahn et al., 2016; Liaw et al., 2013; Thiru et al., 2003; Weiskopf & Weng, 2013). Arts et al. (2002) summarized studies concerning the analysis of data quality and the procedures for data quality assurance in medical registries. Thiru et al. (2003) systematically reviewed the literature concentrating on the measurement of data quality in primary care and the reference standards used to assess data quality between 1980 and 2001. Weiskopf and Weng (2013) and Chen et al. (2014) developed systematic methods to assess data quality in electronic health records for research. Liaw et al. (2013) reviewed data quality assessed and managed in integrated chronic disease management. Johnson et al. (2015) mined terms used in the healthcare data quality literature to describe data quality and organized them into an ontology. Kahn et al. (2016) analyzed and organized existing published terms that describe data quality into a conceptual framework to support defining, assessing, and reporting data quality findings.

These studies take different approaches to analyzing the data quality literature to advance our understanding of healthcare. This literature, however, does not completely integrate the dimensions, issues, and methods of data quality in healthcare. For example, some reviews identified several methods used to measure specific issues of EMR data under two or three data quality dimensions (Chen et al., 2014; Kahn et al., 2016; Thiru et al., 2003; Weiskopf & Weng, 2013). Arts et al. (2002) disclosed that issues of completeness and correctness could be reduced using quality assurance and control procedures. In contrast, Liaw et al. (2013) proposed that ontology-based approaches that were used to solve semantic interoperability problems (concerning consistency) have the potential to assess data quality from multiple dimensions for medical data. Then, Johnson et al. (2015) developed a harmonized data quality assessment terminology using an ontology. However, identifying specific issues for each data quality dimension was not a focus of their study.

In the present study, we apply an iterative taxonomy development process, as Nickerson et al. (2013) suggested. Moreover, our review summarizes the strengths and limitations of the methods applied to address issues concerning the quality of medical or healthcare data that have not been

Figure 1. Research process in this review



highlighted in the previous reviews. Finally, our study considers methods used to define and assess data quality dimensions and explores methods adopted to reduce and prevent poor data based on empirical evidence.

The data quality domain has grown, and its reference disciplines expanded (Sadiq et al., 2011). An Electronic Data Methods Forum (Agency for Healthcare Research and Quality, 2017a) facilitated the use of EMR for research and quality improvement. The sponsored initiatives of quality measurement and reporting since 2012 could culminate in developing strategies to improve healthcare quality enabled by IT (Agency for Healthcare Research and Quality, 2017b). Addressing the current data quality status in healthcare IS/informatics would advance health IT to improve patient care and outcomes.

RESEARCH METHODS

We follow the guidelines of Wolfswinkel et al. (2013) to conduct the review, including (1) defining the review scope; (2) searching for the initial list of articles; (3) selecting relevant papers; (4) analyzing data from the included studies. Figure 1 shows the research process employed in this review.

Defining the Review Scope

The four activities in this step are establishing the inclusion and exclusion of an article in the data set, identifying appropriate fields of research, selecting probable corresponding outlets, and formulating search terms (Wolfswinkel et al., 2013).

The inclusion criteria are: (1) we limit the search to articles published in English; (2) We selected articles with a publication date between January 2012 and June 2022; and (3) the keyword(s) (e.g., “data quality,” “quality of data,” “information quality,” and “quality of information”) emerge three or more times in the text’s body (Sadiq et al., 2011; Wang et al., 1995). The removal of papers is based on the following exclusion criteria: (1) the theme of the paper does not pertain to data quality in the realm of healthcare; (2) the studies do not provide empirical findings themselves; (3) the articles are editorials or commentary; (4) the latest impact factor for the source of the paper is below 2.000 in the Journal Citation Reports (Thomson Reuters, 2022); (5) the papers are not accessible online; or (6) the studies do not answer any question from the RQs. We focused on a corpus of research outlets spanning the healthcare IS/informatics field to ensure the quality of the data set in a broad coverage of statistical and computational contributions.

To achieve a balanced data set of articles for the analysis (Sadiq et al., 2011; Wolfswinkel et al., 2013), we selected the outlets from a survey based on the top journal basket for healthcare IS/

informatics (GeorgiaState University, 2016). They conducted the survey to identify the highly regarded outlets from health IT experts, to serve as the initial resources in this review. In addition, the consistency of the literature reviews in highly prestigious outlets offers the opportunity to discuss the papers more deeply (Hanafizadeh & Zareravasan, 2020; Liu et al., 2017; Sadiq et al., 2011). Our search keywords contained "data quality," and alternative terms, including "quality of data," "information quality," and "quality of information" to screen the papers.

Searching for the Initial List of Articles

Twenty-three journals were selected from the top journal basket for healthcare IS/informatics as our initial resources because their impact factors are greater than or equal to 2.000 that meet our inclusion criteria. We visited the official websites of these journals and considered each volume and issue manually. We screened the candidates from the link of each publication. If "data quality", "quality of data", "information quality", or "quality of information" emerged three or more times in the text's body, we included the paper in our initial list of articles. This manual search resulted in the identification of 1581 publications.

Selecting Relevant Papers

We eliminated 1510 papers from the initial list of the articles (n = 1581) according to their abstract and full-text review based on the exclusion criteria including (1) the theme of the paper does not pertain to data quality in the realm of healthcare; (2) the studies do not provide empirical findings themselves; (3) the articles are published in editorials and commentary; (4) the latest impact factor for the source of the paper is below 2.000 in the Journal Citation Reports (Thomson Reuters, 2022); (5) the papers are not accessible online; or (6) the studies do not answer any question from the RQs. As a result, 71 papers were selected for further, detailed study. In addition, backward and forward snowballing approaches (Jalali & Wohlin, 2012) were adopted to exhaust the data (Wolfswinkel et al., 2013) and enrich our sample size. Regarding the inclusion and exclusion criteria, 21 papers remained after the abstract and full-text review in the snowballing process. Thus, 92 papers remained for further analysis. (The following link is available to see the included papers: https://www.dropbox.com/s/yvj8mi6ysmnx021/ Appendix%201% 20Included%20papers.docx?dl=0

ANALYZING DATA FROM THE INCLUDED STUDIES

To address the research questions, we first studied four main outcomes of the relevant papers including (1) definitions of data quality in healthcare which used in the articles; (2) issues that the papers were addressed; (3) methods used to respond to the issues; and (4) strengths and limitations of these methods. However, not every paper contained a complete set of definitions, issues, and methods. Second, we extracted and recorded direct responses from publications for the three main outcomes (definitions, issues, and methods) in a form. Third, after completing the form, we reviewed the definitions of data quality used in the papers and removed duplicated terms that describe data quality. Then, we categorized these definitions into distinct dimensions of data quality in healthcare using the taxonomy development approach (Nickerson et al., 2013). We also analyzed the lists of issues and methods of resolution similarly. Last, we mapped the issues related to the different data quality dimensions from the papers to those dimensions of data quality and their methods identified in the literature.

RESULTS

The first portion of the research focuses on the dimensions of data quality. The second portion of the studies in the review highlight the issues that healthcare organizations face to achieve better data quality. These issues are related and dependent on the dimensions, and therefore, we discuss them together. While there has been a consensus on the need for better adoption of healthcare IS (Li et al., 2013; Najaftorkaman et al., 2013; Najaftorkaman et al., 2015), because of the expensive process of fully customized IS (Saleem et al., 2015) and socio-technical issues that occurred because of human interventions (Sittig & Singh, 2010), data quality issues are unavoidable (Alshawi et al., 2003). So, making the best out of generated healthcare data is necessary to achieve quality in clinical decisions and care. For this to happen, the literature has adopted statistical and computational methods to improve the data quality depending on issues related to data quality dimensions. We also discussed these methods and their relationships with the issues they are resolving below.

Dimensions of Data Quality in Healthcare: Definitions and Related Issues

The overall starting point has been mentioned in the Introduction, and one of our objectives is to identify definitions proposed for data quality in healthcare and related data quality issues (see Table 1). We developed Table 1 through taxonomy development, which includes (1) having a flat list of definitions of data quality in healthcare and the related issues based on the literature; (2) separating the list into several relevant groups; and (3) giving each group an appropriate name according to their similarities. Using the taxonomy development approach, we identified and classified the dimensions that were used to construct data quality in healthcare, including: (1) Completeness, (2) Correctness, (3) Currency, (4) Consistency, (5) Usability, (6) Relevance, and (7) Duplication. Table 1 presents the keywords associated with, and used to identify, the seven dimensions. These keywords can search for similar research. We also present the definitions extracted from the relevant papers in Table 1. We detail these dimensions below:

- **Completeness:** Completeness is defined regarding relevance. Completeness can be interpreted as coverage of baseline features or data required for a particular disease [S17, S52, S90, S91]. It can also refer to the right amount of available data [S76, S78, S79, S80, S81, S82, S83, S86]. Although several attempts in the literature have addressed missing values, completeness has been repeatedly reported as a challenge in practicing healthcare data programs. Completeness has been related to Availability, Coverage, Presence, and Comprehensiveness.
- **Correctness:** Correctness in healthcare data has been defined concerning validity [S5, S36, S53, S58, S70] and accuracy [S49, S62, S68, S76, S77, S78, S79, S80]. The literature has identified several issues resulting from incorrect data in healthcare settings. First, human errors can significantly contribute to incorrect data in healthcare settings. The errors can occur when entering data into a computerized system [S25, S34, S38, S65], or writing paper-based records [S2, S6, S14, S34]. Second, values can be out of the defined range of measurement [S54]. Third, in some situations, data has been misused in terms of what they were supposed to mean [S68]. Other reasons for incorrect data include illogical patterns [S26], typographical problems [S9, S27], and media discontinuity [S2, S6]. Correctness has been related to Accuracy, Validity, Trustworthiness, and Plausibility.
- **Currency:** Currency is defined as whether the data in a system is associated with a recent measurement [S5, S22], and is known as timeliness [S5, S47, S50, S59, S68, S69, S70, S76, S78, S81], including temporal stability and spatial stability [S56]. Sáez et al. [S1] state that currency is not only related to time but can also be special. For instance, there are treatment and drug advocacies related to particular regions. Commonly, one treatment or drug can be acceptable in one medical authority while banned in another. Hence, they define currency as a measure that includes both current conditions and the place of patients [S1].

The papers addressing currency in the data argue that, in healthcare settings, the data may lack currency because of either unexpected and undesired changes in, for example, patients' conditions, or data entry delays. For instance, a patient's condition may change when it was expected to be stable,

Table 1. A summary of keywords and issues related to dimensions of data quality in this review

Dimension	Definitions	Related issues
Completeness (Availability) (Coverage) (Presence) (Comprehensiveness)	Presence of all data [S58, S87] or a value for a given data element [S2, S60] for a patient Items that should be recorded and be available for each patient [S25, S57] Availability of baseline features or data that is required for a particular disease [S17, S52, S90] Comprehensiveness of the content and whether enough data is provided for a specific task [S23]	CI1: Missing values [S2, S6, S7, S20, S21, S25, S26, S27, S30, S31, S32, S34, S35, S38, S43, S45, S58, S59, S60, S63, S64, S72, S73, S74, S75, S76, S77, S79, S80, S82, S83, S84, S85, S89, S91, S92] CI2: Inappropriate measurement of missing values (e.g., zeros) [S9, S35] CI3: No source documents [S54, S85, S86, S92]
Correctness (Accuracy) (Validity) (Trustworthiness) (Plausibility)	Valid responses [S5, S59] for requests A truth [S3, S12] for the details of patients Degree of accuracy and precision within a real-world situation [S56] A valid and appropriate record with correct measurements between acceptable ranges [S17, S87]	AI1: Random errors of data entry [S25, S34, S38, S65] AI2: Transcription errors or typographical problems [S2, S6, S9 S14, S27, S34, S75, S76, S77, S79, S84, S85] AI3: Errors that occur during media discontinuity [S2, S6, S72, S83, S85] AI4: Value lies outside the defined range of measurement [S54, S85, S88] AI4: Misuse in respect to the “meaning of data” [S68] AI5: Illogical patterns[S26]
Currency (Timeliness) (Temporal stability) (Spatial stability)	Degree to which measurement is current with a patient’s conditions [S1, S22] Degree to which measurement is current with a patient’s place [S1] An element in the electronic health record that is a representation of the patient state at a given point of time [S3, S12, S41, S56]	EI1: Unexpected or undesired changes in patients’ conditions through time [S1, S51] EI2: Late data entry [S5, S6, S76, S92]
Consistency (Comparability) (Concordance) (Reliability) (Conformance)	Adherence to other data sets and if the measurements were to be repeated the same results would be obtained [S2, S3, S12, S28, S41, S58, S62] Degree to which data meet specific constrains and rules [S56, S87] A unified data type, format and standard [S17, S47] Ease of access and understanding of database or repository [S47]	SI1: Different content between past and present [S25] SI2: Contradiction of standards in integrity [S28], terminology and coding rules [S34, S57, S68, S71, S75, S84, S85] SI3: Lack of relations between data sources or between diagnoses and procedures [S25, S54, S84, S89]
Usability (Contextualization) (Accessibility)	A data element that makes sense in using other knowledge about what that element is [S3, S12, S41]	UI1: Non-structured free text [S25] UI2: Non-documentation or documented in the wrong place [S15, S17] UI3: Complexity in representation (Sariyar et al., 2013) UI4: Inappropriate granularity (Sariyar et al., 2013)
Relevance (Predictive value)	Degree to which data meets current and potential needs from users [S47, S56]	RI1: Lack of planning or lack of knowledge for future analysis [S47, S56]
Duplication (Repetition) (Uniqueness)	Degree to which data contains repetitions that represent the same entity [S56]	DI1: Lack of integrated systems [S4] DI2: Lack of relationships between data sources [S35, S85, S88]

resulting in no scheduled examination [S1, S51]. The literature in this area has used Timeliness, Temporal and Spatial stability as relevant keywords for currency.

- **Consistency:** Three different definitions of data consistency have been identified in the literature on IS healthcare. These definitions refer to the adherence of the data to other data sets, with standards and constrains [S2, S3, S12, S28, S41, S58, S62, S81, S82]. Laberge and Shachak [S47] argue that, in the cycle of care for patients, the value of each piece of data should be comparable at each point of care with the other data sets and, therefore, the comparability of the data with its constraints and rules is also another perspective of consistency. Rahimi et al. [S17] defined consistency as meeting the data format and standards set in the data.

One of the main reasons for inconsistency in healthcare settings is related to the currency of data [S25]. In practice, data is inconsistent among different resources because some of the data sets have not been updated with the latest data. The lack of integrated systems may also create inconsistency in the data used and generated in healthcare settings. Dentler et al. [S25] provide an example of poor integration of systems that coordinate diagnoses and procedures. Consistency has been used in the literature related to Comparability, Concordance, Reliability, and Conformance.

- **Usability:** Data usability in the healthcare IS/informatics literature has been defined as the ease of understanding and accessibility of data [S47]. One of the main issues in the usability of healthcare data is non-structured free texts. Dentler et al. [S25] state that healthcare practitioners are used to free-text descriptive inputs, which makes the process of the data and, accordingly, the usability of it more difficult. Furthermore, because of a busy schedule, events in the healthcare settings or even the patients' data may lack documentation or be documented incorrectly or timely [S15, S17]. Medical professionals abstractly document diagnoses, resulting in poor data utilization. In contrast, listing too much detail can lead to complex data and a consequent lack of usability [44]. To improve data usability, data is appropriately annotated in the context where it was acquired [S56]. The papers in our data set concerning healthcare IS/informatics employed usability with Contextualization and Accessibility.
- **Relevance:** The relevance of data in healthcare settings is defined as whether the data represents the needs for current or potential analysis [S47]. The lack of awareness of the type of analysis needed when constructing a healthcare IS is one reason for poor relevance [S47]. García-de-León-Chocano et al. [S56] note that the lack of planning for future analysis may cause irrelevant data collection in healthcare. Predictive value has been used in association with the relevance of healthcare data.
- **Duplication:** Duplication is defined as the multiple existences of the same data entity (e.g., a patient in the data set) [S56]. Poor system integrations [S4] and the lack of relationships between data sets [S35] may cause data duplication. Duplication of healthcare data could concern repetition and uniqueness.

METHODS

In addition to the dimensions of data quality associated with the issues summarized in this study, we also identified the methods used to address these data quality issues in this review. These methods were generated by a similar process as the development of Table 1 and divided into six categories in addressing the issues of data quality: (1) Percentage Estimate; (2) Measure of Jensen-Shannon Divergence; (3) Measure of Inter-Rater Reliability; (4) Patient's Flow Model based on Event Logs; (5) Data Quality Ontology; and (6) Fellegi–Sunter Record Linkage Extension. Some of these methods were used in multiple dimensions of data quality. In addition, researchers have attempted to address the issues by using both qualitative and quantitative methods. Table 2 outlines these methods with

Table 2. A summary of methods used to address data quality issues in this review

Method	Description	Strengths and Limitations
Percentage estimate (PE)	A particular ratio is used as a benchmark for the users' acceptance of data quality and therefore, if data elements do not match with that ratio, they will not be considered in the analysis requested from users. The ratio could be estimated when using previously successful analyses.	*Strengths* quantitatively evaluate and compare discrepancies of data items easy and efficient to apply queries in medical data when the removed data items can be traced back for improvement [S5] *Limitations* only select the data that is most appropriate regarding its quality
Measure of Jensen-Shannon divergence (MJSD)	It is a method of measuring the similarity between two probability distributions.	*Strengths* convenient to compare the datasets by a convergent measure bounded between zero and one [S1] calculate the degree of similarity between the datasets and decide if the degree of the change over time or sources of data entry indicates currency or correctness[S1] *Limitations* the probabilistic space becomes sparser leading to ineffective comparison when the number of variables increase [S51]
Measure of Inter-Rater Reliability (MIRR)	A statistical method is used to assess the degree to which different data entry points are consistent of the same data item.	*Strengths* not limited to the changes of a particular split of data [S1, S51] independent on the rubric against which the data item is being measured for continuous data [S1, S51] *Limitations* unable to distinguish the data entry point that has entered the data incorrectly or inconsistent with other entries (Killen, 2005)
Patient's Flow Model based on Event Logs (PFM)	Ordering of activities with the intention is used to discover all possible paths. The activities are ordered chronologically using timestamp data extracted from EMR. This gains insights into the data quality of the various fields used in each phase.	*Strengths* identify the underlying phases of the patient's journey in which the medical records may miss values or provide duplication [S35] trackback to the source of the problem by knowing exactly the time stamp for which the issue occurred [S35] generalize a best practice process to program medical record software packages and adopt them in medical practices [S35] *Limitations* lack an automatic time stamp; instead different attempts of data entry have been used as time stamps [S35]
Data Quality Ontology (DQO)	An ontology provides a means for its users to consistently and accurately use uniform terminology about the same entities in some domain and enables automated auditing of data quality for requests.	*Strengths* an automated auditing of data quality for an analysis [S17, S41, S42] can be used as data quality assurance in the data entry points where the medical records that do not match the ontology automatically cannot get accepted by the system (Sahoo et al., 2013; Vandenbussche et al., 2013), and be served for data quality assessment [S46]. *Limitations* the dependency of the ontology model to the practitioners that should be interviewed and the data sample that should be tested [S17, S41, S42]
Fellegi–Sunter Record Linkage Extension (FSRLE)	A method provides more data for the same entity through combining independent data sources.	Strengths enable development of automated imputation for missing values [S43] has high sensitivity to detect most match pairs and generate no false positive matches [S43]. Limitations the repetition of pre-computing the imputation rule set when linking each pair of data sets [S43] the inability of precise imputation for the distance of missing values [S43]

their strengths and limitations. We discuss these methods by addressing the relevant data quality issues in the literature below:

- **Percentage estimate:** A percentage estimate is a ratio-based calculation showing the degree to which a data element meets the needs of data users. A specific ratio is used as a benchmark for data quality. Therefore, if data elements do not match that ratio, we will not consider them in the analysis requested by the users. Tosti et al. [S50] recommend using past EMR as a pool of historical data to estimate the ratio. Puttkammer et al. [S5] used the percentage estimate to filter data with missing values. Percentage estimate has, similarly, been used to address: errors of data entry (removing the data item entered by a particular entry node that has demonstrated the accuracy of less than the percentage estimate); lack of integrated relationships between data sources, complexity in representation; inappropriate granularity; late data entry (removing the data item entered by delay more than the percentage estimate); and contradiction of standards.

- **Measure of Jensen-Shannon divergence:** The Jensen–Shannon divergence is a method of measuring the similarity between two probability distributions. This method can compare two datasets. Sáez et al. [S1, S51] utilize this method to identify unexpected or undesired changes in EMR through time or places. Sáez et al. [S1, S51] identify the temporal changes in the patient's conditions and therefore notify the need to update the data for better currency. However, the application of the Jensen-Shannon divergence measure is not limited to the currency of medical data. Although it has not been found in the set of relevant papers, the method can also compare the data entered by different resources and identify the errors in data entry, helping to increase the correctness of medical data. The correctness and the currency of electronic medical data can, thus, be verified through comparison with datasets entered by other sources or through time and space [S74].
- **Measure of Inter-Rater Reliability:** The measure of inter-rater, also known as inter-observer, reliability is a statistical method used to assess the degree to which different data entry points are consistent with the same data item. Hirdes et al. [S31] used this method to identify random errors regarding the correctness of medical data. They calculate the correlation between data entry points on continuously monitored patient data, which provides the data's reliability measure. As the correlation between data entered at various points increases, the reliability of the data items increases as well. To have consistent data, Lambdin et al. [S52] use inter-rater reliability to address the lack of relationships between data sources used by physicians during the diagnosis and procedures in a hospital. Although Hirdes et al. [S31] used this method to measure the reliability of continuous data, Lambdin et al. [S52] measured categorical medical data reliability. Here, the percentage of agreement among data entry points in a category in the hospital calculates the consistency of the data item.
- **Patient's Flow Model based on Event Logs:** The Patient's Flow Model focuses on ordering activities intending to discover all paths. The activities are ordered chronologically using timestamp data extracted from EMR. The patient's flow model is suited for discovering abnormal flow because it relies on timestamp data. Perimal-Lewis et al. [S35] used Disco (Fluxicon, 2012) as a software tool to extract the phases based on the timestamps logged in EMR and constructed the patient's flow model, called the patient's journey, in an emergency department. They then used this model to gain insights into the data quality of the various fields used in each phase. Similarly, Benevento et al. [S88] modeled treatment processes of lung cancer in the healthcare environment using process mining techniques based on system logs, to identify unacceptable events or redundant activities.

This method can identify missing values of each patient's record by comparing them with those entered in the same time frame in other records. Then, a missing value in a particular timestamp can be likely assumed to have the value of the same variable in another time stamp if, in other records, the value has not changed between these two stamps. The method argues that if the value has not been entered and is missing, it probably was not highlighted to medical staff, despite its changes [S35, S88]. Duplicated records occur because of a lack of integrated systems, but can be identified and removed by comparing the medical records in time stamps [S35, S88].

- **Data Quality Ontology:** An ontology provides a vocabulary of terms, meanings, and relationships used in various application contexts (Borst, 1997). Ontologies presenting the relationships between entities in the application domain have been proposed to assure the quality of information by representing the relationships between the entities and the requirements of information quality in a domain (Tartir et al., 2005). For example, Rahimi et al. [S17] constructed an ontology model that has been implemented in EMR to automatically ensure data quality for querying Type 2 Diabetes Mellitus (T2DM) patients. The researchers incorporated completeness, correctness and consistency as dimensions of data quality. They generated their ontology model based on the terminology given in chronic disease management and interviews with practitioners. The ontology captures the terminology used in T2DM and the related information in EMR. This could prevent the contradictions of standards of integrity, terminology, and coding rules. Mapping the ontology to medical records resulted in identifying the records that do not satisfy these conditions and also helped address duplicate data [S88].

- **Fellegi–Sunter Record Linkage Extension:** Record linkage methods provide integrated information for the same patient by combining different data sources with one or more nonunique fields (also quasi-identifiers) [S27, S43], entailing deterministic (Durham et al., 2010) and probabilistic approaches (Herzog et al., 2007) to matching records. By extending the Fellegi–Sunter method of record linkage (Fellegi & Sunter, 1969), Ong et al. [S43] developed three methods to use the data available better and discard fewer data in record linkage, including Weight Redistribution, Distance Imputation, and Linkage Expansion. The researchers first removed fields with missing values from the set of quasi-identifiers and reconstructed the weight from the missing attribute according to relative proportions in the remaining available linked fields, and then represented the distance between the fields of missing values instead of the missing values. Finally, they added prior non-linkage fields to the linkage field to complement the corresponding responses for the missing values.

DISCUSSION

The literature on dimensions of healthcare data quality appears to be sparse and inconsistent (Kahn et al., 2016; Liaw et al., 2013; Wang & Strong, 1996; Weiskopf & Weng, 2013). Our findings agree with “fit to the purpose” definition of data quality. We identified relationships amongst the dimensions of data quality, the issues that challenge these dimensions in healthcare settings, and the methods that can be implemented to improve the already generated data. We can also use these findings to manage and improve healthcare organizations’ data quality.

Depending on the application of data in the healthcare setting, various dimensions of data quality suggest different issues. The methods attempt to resolve the issues of achieving quality data that could cause high-quality clinical decisions and quality of care. Improving healthcare data quality relates to pre- and post-data generation efforts. Before generating data along a healthcare or clinical process, use managerial practices to align information systems with daily practices and produce better data quality. In addition, after generating data, statistical methods and technical solutions can improve the quality of already generated data, although this may be limited depending on the context of the application domain.

RQ1: What are the dimensions of data quality in healthcare? Findings: Seven dimensions of data quality are identified: (1) Completeness, (2) Correctness, (3) Currency, (4) Consistency, (5) Usability, (6) Relevance, and (7) Duplication.

RQ2: What are the relevant issues related to these dimensions of data quality? Findings: Twenty-one issues are summarized in Table 1. More and more data is constantly being collected. However, the amount of valuable resources, with their potential to improve the clinical decisions and quality of healthcare settings, compounds the data quality issues specified in Table 1. Our findings contrast with prior notions blaming poor data quality on data entry (Grimes, 2010; Johansen et al., 2008). Challenges in healthcare data quality are much more complex, with the main issues summarized in Table 1.

RQ3: What are the methods used to address these issues? Findings: we categorize six methods from the included studies: (1) Percentage Estimate; (2) Measure of Jensen-Shannon divergence; (3) Measure of Inter-Rater Reliability; (4) Patient's Flow Model based on Event Logs; (5) Data Quality Ontology; and (6) Fellegi-Sunter Record Linkage Extension.

RQ4: What are the strengths and limitations of these methods? Findings: The strengths and limitations of each method used to address data quality issues are summarized in Table 2.

Table 3 presents a holistic picture of future research into appropriate methods to improve healthcare data quality.

Whereas statistical and computational methods have addressed various data quality, a general agreement on how these methods can guarantee data quality is missing. There have been several issues that have not been addressed. For instance, missing values causing concerns about the completeness of data in healthcare settings has brought great attention to these methods. Several methods have addressed errors in data entry, resulting in correctness issues in data quality. Several methods have attempted to resolve the lack of relationships between data sources. While the literature has been

Table 3. Data quality in healthcare: dimensions, issues and methods

Dimensions of Data Quality in Healthcare	Issues	Methods					
		PE	MJSD	MIRR	PFM	DQO	FSRLE
Completeness	CI1	×			×	×	×
	CI2						
	CI3						
Correctness	AI1	×	×	×			
	AI2						×
	AI3						
	AI4					×	
	AI5						
	AI6						
Currency	EI1		×				
	EI2	×					
Consistency	SI1						
	SI2					×	
	SI3	×		×			
Relevance	RI1	×					
Usability	UI1						
	UI2						
	UI3	×					
	UI4	×					
Duplication	DI1				×		
	DI2	×			×	×	

extensive in identifying the relevant dimensions of data quality to healthcare settings and their issues, the methods proposed to resolve them cover little of these issues. This lack of coverage motivates researchers for future developments.

CONTRIBUTIONS AND LIMITATIONS

The healthcare IS/informatics literature adopted inconsistent terminologies. Hence, our study first categorizes seven single dimensions in data quality. "Duplication" is identified in the current work that is not included in previous reviews. We then developed a taxonomy of data quality dimensions for the healthcare IS/informatics community that could facilitate using consistent terms to describe data quality. We also present the keywords under each dimension, contributing to the bibliographic searches' basket.

The second contribution is a unique study that describes specific issues for each data quality dimension in healthcare. The study reveals 21 specific data quality issues under different dimensions and shows the overlap among these dimensions. For example, we can consider the lack of relationships among data sources as an issue related to consistency, resulting in incomplete data, inaccurate data, or duplicates in data linkage. Thus, the data quality dimensions cannot be isolated directly and heavily relate to users' purposes.

Finally, this paper identified the methods adopted to resolve issues for achieving quality data, providing a clear picture of linking the methods to the issues under different data quality dimensions that were not delineated in previous literature reviews. The findings from Table 3 imply that no single method can address all the issues regarding data quality. Table 3 also discloses some issues that require further exploration. Table 4 summarizes the differences between this review and related studies and proposes potential areas for future research.

Table 4. A summary of the differences between this review and related studies and proposed potential areas for future research

Findings of data quality in healthcare	Prior studies	This review	Potential areas for future research
Dimensions	Multiple dimensions of data quality have been identified in prior reviews (Arts et al., 2002; Chen et al., 2014; Johnson et al., 2015; Kahn et al., 2016; Liaw et al., 2013; Thiru et al., 2003; Weiskopf & Weng, 2013).	Categorizes seven single dimensions of data quality, including duplication dimension that is not included in prior studies.	To develop an assessment indicator system with specific measures and indicator weights for each data quality dimension in healthcare, serving as the guidelines for data quality assessment in healthcare.
Issues	Identifying and conceptualizing data quality issues under each dimension has received limited attention from prior reviews.	Specifies relevant issues for each data quality dimension in healthcare.	To identify the relationships between data quality issues, contributing to better root cause analysis and problem solving.
Methods	Data quality methods used to address the issues under each data quality dimensions have not been completely integrated in prior reviews.	Identifies methods used to address data quality issues with their strengths and limitations. Links these methods to the identified issues under different data quality dimensions.	To specify strengths and limitations of methods that assist in better addressing the latter. To further explore the methods used to respond to the issues that have not been addressed as shown in Table 3.

Practically, the study informs healthcare professionals how to determine their data quality using multiple dimensions, which were categorized to help assess the data quality improvement process. In particular, machine learning is modern and widely used to discover patterns from healthcare data sources and provide strong capabilities to predict diseases. Incomplete, inaccurate, and inconsistent data can lead to drastic degradation in prediction for machine learning models (Gudivada et al., 2017). Defining data quality dimensions (e.g., completeness, correctness, and consistency) to assess the datasets and ensure their data quality helps develop better machine learning models. For 21 specific data quality issues identified, healthcare professionals could better understand how poor data occurs and how to improve data practices.

The study is limited in that the literature review was limited to selected journals and publications within ten years. Furthermore, relying on our classification, we did not contact original researchers for consultation.

FUTURE WORK

- **Dimensions:** The data quality dimensions are needed to use consistently. Additionally, "Duplication" can be used as one dimension to construct data quality for a given purpose.
- **Issues:** Identifying specific issues under different data quality dimensions with root causes could help decision-makers determine an appropriate solution and guild audiences' efforts on the same issues.
- **Methods:** The strengths and limitations of methods used to address data quality issues should be specified. Recognizing such strengths and limitations could help address the latter.

Researchers would benefit from the body of data quality research from healthcare IS/informatics and other areas and integrate the existing theories and methods into their research. Therefore, our literature review also proposes the following opening questions for further study: (1) For each of the seven dimensions of data quality, what are the characteristics of healthcare IS/informatics to avoid the associated issues? (2) How can statistical and computational methods be assessed and used to improve the quality of already generated healthcare data to achieve a better quality of clinical decisions and overall better quality of care? We present additional potential research areas in Table 4.

In practice, the findings of data quality dimensions identified in this review can assess how good is the data at hand and determine whether this data can support decision-making. Practitioners would also benefit from our findings to find a data quality issue and identify its root causes for resolution. In contrast, the data quality methods discussed in the study can serve as references (e.g., select an appropriate method or combine methods from Table 2 based on their strengths and limitations) to address data quality issues and improve healthcare data quality.

CONCLUSION

Through a systematic literature review, this paper has identified a set of dimensions that are important when dealing with data quality in healthcare IS. The extracted data was aggregated and analyzed to answer 4 RQs. Dimensions, issues, and methods to address the issues related to healthcare data quality were discussed, and we made proposals for future research. In addition, we presented the importance of improving data quality, patient care, and better clinical decisions as this research's motivation and potential impact.

This study identified seven dimensions including (1) Completeness, (2) Correctness, (3) Currency, (4) Consistency, (5) Usability, (6) Relevance, and (7) Duplication, with 21 specific issues under different dimensions. Furthermore, we retrieved six methods used to address these issues: (1) Percentage Estimate; (2) Measure of Jensen-Shannon divergence; (3) Measure of Inter-Rater

Reliability; (4) Patient's Flow Model based on Event Logs; (5) Data Quality Ontology; and (6) Fellegi–Sunter Record Linkage Extension. After mapping the methods to related data quality issues, we presented a few opening questions as potential future research topics. We hope these questions outlined in this paper will stimulate growing interest in improvements for addressing data quality issues by healthcare IT in teams. In addition, an appreciation of this topic can also be of practical use, as healthcare professionals increasingly use EMR and other operational systems in healthcare settings and improve their data practices.

ACKNOWLEDGMENT

This research was supported and funded by the Humanities and Social Sciences Youth Foundation, Ministry of Education of the People's Republic of China (Grant No.21YJC870009).

REFERENCES

Agency for Healthcare Research and Quality. (2017a). *Electronic Data Methods Forum (2010-2017).* Health IT. https://healthit.ahrq.gov/ahrq-funded-projects/past-health-it-initiatives/electronic-data-methods-edm-forum

Agency for Healthcare Research and Quality. (2017b). *Health IT-Enabled Quality Measurement (2012-2013).* Health IT. https://healthit.ahrq.gov/ahrq-funded-projects/past-health-it-initiatives/health-it-enabled-quality-measurement

Alshawi, S., Missi, F., & Eldabi, T. (2003). Healthcare information management: The integration of patients' data. *Logistics Information Management*, *16*(3/4), 286–295. doi:10.1108/09576050310483772

Arts, D. G., De Keizer, N. F., & Scheffer, G.-J. (2002). Defining and improving data quality in medical registries: A literature review, case study, and generic framework. *Journal of the American Medical Informatics Association: JAMIA*, *9*(6), 600–611. doi:10.1197/jamia.m1087 PMID:12386111

Borst, W. N. (1997). *Construction of engineering ontologies for knowledge sharing and reuse* [Doctoral thesis, Universiteit Twente, the Netherlands]. https://ris.utwente.nl/ws/portalfiles/portal/6036651/t0000004.pdf

Chen, H., Hailey, D., Wang, N., & Yu, P. (2014). A review of data quality assessment methods for public health information systems. *International Journal of Environmental Research and Public Health*, *11*(5), 5170–5207. doi:10.3390/ijerph110505170 PMID:24830450

Durham, E., Xue, Y., Kantarcioglu, M., & Malin, B. (2010). Private medical record linkage with approximate matching. *AMIA ... Annual Symposium Proceedings - AMIA Symposium. AMIA Symposium*, *2010*, 182–186. PMID:21346965

ECRI Institute. (2015). *Wrong-record, wrong-data errors with health IT systems*. ECRI. https://www.ecri.org/Resources/In_the_News/PSONavigator_Data_Errors_in_Health_IT_Systems.pdf

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210.

Fluxicon. (2012). *Products*. Fluxicon. https://fluxicon.com/disco/

Georgia State University. (2016). *SIG-Health Survey for "top journals" list*. GSU. https://gsu.qualtrics.com/jfe/form/SV_cwr6Gn4lBSsTZyJ

Grimes, D. A. (2010). Epidemiologic research using administrative databases: Garbage in, garbage out. *Obstetrics and Gynecology*, *116*(5), 1018–1019. doi:10.1097/AOG.0b013e3181f98300 PMID:20966682

Gudivada, V., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software*, *10*(1), 1–20.

Hanafizadeh, P., & Zareravasan, A. (2020). A systematic literature review on IT outsourcing decision and future research directions. *Journal of Global Information Management*, *28*(2), 160–201. doi:10.4018/JGIM.2020040108

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. Springer.

Jalali, S., & Wohlin, C. (2012). Systematic literature studies: Database searches vs. backward snowballing. *Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, (pp. 29–38). ACM, IEEE. doi:10.1145/2372251.2372257

James, J. T. (2013). A new, evidence-based estimate of patient harms associated with hospital care. *Journal of Patient Safety*, *9*(3), 122–128. doi:10.1097/PTS.0b013e3182948a69 PMID:23860193

Johansen, M. A., Scholl, J., Hasvold, P., Ellingsen, G., & Bellika, J. G. (2008). Garbage in, garbage out: Extracting disease surveillance data from EPR systems in primary care. *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, (pp. 525–534). ACM. doi:10.1145/1460563.1460646

Johnson, S. G., Speedie, S., Simon, G., Kumar, V., & Westra, B. L. (2015). A data quality ontology for the secondary use of EHR data. *AMIA Annual Symposium Proceedings*. ACM.

Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., Estiri, H., Goerg, C., Holve, E., Johnson, S. G., Liaw, S.-T., Hamilton-Lopez, M., Meeker, D., Ong, T. C., Ryan, P., Shang, N., Weiskopf, N. G., Weng, C., Zozus, M. N., & Schilling, L. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Washington, DC)*, *4*(1), 1244. doi:10.13063/2327-9214.1244 PMID:27713905

Killen, R. (2005). *Programming and assessment for quality teaching and learning*. Thomson Social Science Press.

Landrigan, C. P., Parry, G. J., Bones, C. B., Hackbarth, A. D., Goldmann, D. A., & Sharek, P. J. (2010). Temporal trends in rates of patient harm resulting from medical care. *The New England Journal of Medicine*, *363*(22), 2124–2134. doi:10.1056/NEJMsa1004404 PMID:21105794

Li, J., Talaei-Khoei, A., Seale, H., Ray, P., & MacIntyre, C. R. (2013). Health care provider adoption of eHealth: Systematic literature review. *Interactive Journal of Medical Research*, *2*(1). doi:10.2196/ijmr.2468 PMID:23608679

Liaw, S.-T., Rahimi, A., Ray, P., Taggart, J., Dennis, S., de Lusignan, S., Yeo, A. E. T., & Talaei-Khoei, A. (2013). Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. *International Journal of Medical Informatics*, *82*(1), 10–24. doi:10.1016/j.ijmedinf.2012.10.001 PMID:23122633

Liu, C., Talaei-Khoei, A., Zowghi, D., & Daniel, J. (2017). Data completeness in healthcare: A literature survey. *Pacific Asia Journal of the Association for Information Systems*, *9*(2), 75-100. https://aisel.aisnet.org/pajais/vol9/iss2/5

Michel-Verkerke, M. B. (2012). Information quality of a nursing information system depends on the nurses: A combined quantitative and qualitative evaluation. *International Journal of Medical Informatics*, *81*(10), 662–673. doi:10.1016/j.ijmedinf.2012.07.006 PMID:22898320

Najaftorkaman, M., Ghapanchi, A. H., Talaei-Khoei, A., & Ray, P. (2013). Recent research areas and grand challenges in electronic medical record: A literature survey approach. *The International Technology Management Review*, *3*(1), 12–21. doi:10.2991/itmr.2013.3.1.2

Najaftorkaman, M., Ghapanchi, A. H., Talaei-Khoei, A., & Ray, P. (2015). A taxonomy of antecedents to user adoption of health information systems: A synthesis of thirty years of research. *Journal of the Association for Information Science and Technology*, *66*(3), 576–598. doi:10.1002/asi.23181

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, *22*(3), 336–359. doi:10.1057/ejis.2012.26

Prasser, F., Kohlmayer, F., Spengler, H., & Kuhn, K. (2018). A scalable and pragmatic method for the safe sharing of high-quality health data. *IEEE Journal of Biomedical and Health Informatics*, *22*(2), 611–622. doi:10.1109/JBHI.2017.2676880 PMID:28358693

Rowland, C. (2014, July 20). Hazards tied to medical records rush - Subsidies given for computerizing, but no reporting required when errors cause harm. *Boston Globe*. https://www.bostonglobe.com/news/nation/2014/07/19/obama-pushed-electronic-health-records-with-huge-taxpayer-subsidies-but-has-rebuffed-calls-for-hazards-monitoring-despite-evidence-harm/OV4njlT6JgLN67Fp1pZ01I/story.html

Sadiq, S., Yeganeh, N. K., & Indulska, M. (2011). 20 years of data quality research: Themes, trends and synergies. *Proceedings of the Twenty-Second Australasian Database Conference (*Volume 115, pp. 153-162*)*. ACM.

Sahoo, S. S., Lhatoo, S. D., Gupta, D. K., Cui, L., Zhao, M., Jayapandian, C., Bozorgi, A., & Zhang, G.-Q. (2013). Epilepsy and seizure ontology: Towards an epilepsy informatics infrastructure for clinical research and patient care. *Journal of the American Medical Informatics Association: JAMIA*, *21*(1), 82–89. doi:10.1136/amiajnl-2013-001696 PMID:23686934

Saleem, J. J., Plew, W. R., Speir, R. C., Herout, J., Wilck, N. R., Ryan, D. M., Cullen, T. A., Scott, J. M., Beene, M. S., & Phillips, T. (2015). Understanding barriers and facilitators to the use of clinical information systems for intensive care units and anesthesia record keeping: A rapid ethnography. *International Journal of Medical Informatics*, *84*(7), 500–511. doi:10.1016/j.ijmedinf.2015.03.006 PMID:25843931

Sariyar, M., Borg, A., Heidinger, O., & Pommerening, K. (2013). A practical framework for data management processes and their evaluation in population-based medical registries. *Informatics for Health & Social Care*, *38*(2), 104–119. doi:10.3109/17538157.2012.735731 PMID:23323639

Sittig, D. F., & Singh, H. (2010). A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. *Quality & Safety in Health Care*, *19*(Suppl 3), i68–i74. doi:10.1136/qshc.2010.042085 PMID:20959322

Storey, V. C., Dewan, R. M., & Freimer, M. (2012). Data quality: Setting organizational policies. *Decision Support Systems*, *54*(1), 434–442. doi:10.1016/j.dss.2012.06.004

Tartir, S., Arpinar, I. B., Moore, M., Sheth, A. P., & Aleman-Meza, B. (2005). OntoQA: Metric-based ontology quality analysis. *Proceedings of IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources*. IEEE.

Thiru, K., Hassey, A., & Sullivan, F. (2003). Systematic review of scope and quality of electronic patient record data in primary care. *BMJ (Clinical Research Ed.)*, *326*(7398), 1070. doi:10.1136/bmj.326.7398.1070 PMID:12750210

Thomson Reuters. (2022). *Journal citation reports.* Thomson Reuters. https://jcr.incites.thomsonreuters.com/

Vandenbussche, P.-Y., Cormont, S., André, C., Daniel, C., Delahousse, J., Charlet, J., & Lepage, E. (2013). Implementation and management of a biomedical observation dictionary in a large healthcare information system. *Journal of the American Medical Informatics Association: JAMIA*, *20*(5), 940–946. doi:10.1136/amiajnl-2012-001410 PMID:23635601

Vilic, A., Hoppe, K., Petersen, J., Kjaer, T., & Sorensen, H. (2016). Simplifying EHR overview of critically ill patients through vital signs monitoring. *IEEE Journal of Biomedical and Health Informatics*, PP(99), . doi:1

Wang, R. Y., Storey, V. C., & Firth, C. P. (1995). A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering*, *7*(4), 623–640. doi:10.1109/69.404034

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33.

Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association: JAMIA*, *20*(1), 144–151. doi:10.1136/amiajnl-2011-000681 PMID:22733976

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. (2013). Using grounded theory as a method for rigorously reviewing literature. *European Journal of Information Systems*, *22*(1), 45–55. doi:10.1057/ejis.2011.51

ZareRavasan, A., & Krčál, M. (2021). A systematic literature review on 30 years of empirical research on information systems business value. *Journal of Global Information Management*, *29*(6), 1–37. doi:10.4018/JGIM.288894

Zięba, M. (2014). Service-oriented medical system for supporting decisions with missing and imbalanced data. *IEEE Journal of Biomedical and Health Informatics*, *18*(5), 1533–1540. doi:10.1109/JBHI.2014.2322281 PMID:24816614

Caihua Liu is an Associate Professor at the School of Artificial Intelligence, Guilin University of Electronic Technology. Dr. Liu received her master and doctoral degrees in Information Systems from Hong Kong Baptist University and University of Technology Sydney, respectively. She was a visiting PhD researcher at the Enterprise of Things (EoT) Lab of University of Koblenz-Landau, Germany, in 2018. She has worked as a Postdoctoral researcher and the Associate Research Fellow at the School of Information Management, Sun Yat-sen University, China for two years. Her current research interests include data and information quality governance in healthcare, Big Data technology and general artificial intelligence.

Amir Talaei-Khoei is an Associate Professor and Chair of the Department of Information Systems at the Ansari College of Business and Interim Associate Dean of Research at the School of Social Work in the University of Nevada, Reno (UNR). Prior to joining UNR, Amir spent almost five years in Australia as a faculty member. His research on healthcare analytics has been funded by international, federal and state agencies. His focus is on predictive analytics for health outcomes as well as abnormality analysis for public health surveillance and pandemic outbreaks. He has received his PhD in Information Systems from the University of New South Wales (UNSW), Australia and holds MSc of Information Technology from Royal Institute of Technology, Sweden. Prior to academia, Dr. Talaei-Khoei worked in software engineering industry in Europe.

Veda C. Storey is the Tull Professor of Computer Information Systems and Professor of Computer Science at the J. Mack Robinson College of Business, Georgia State University. Her research interests are in data management, conceptual modelling, and design science research. She is particularly interested in the assessment of the impact of new technologies on business and society from a data management perspective. Dr. Storey is a member of AIS College of Senior Scholars and the steering committee of the International Conference of Conceptual Modeling. She is a recipient of the Peter P. Chen Award, an ER Fellow, an AIS Fellow, and an INFORMS Fellow.

Guochao Peng is based at the School of Information Management at Sun Yat-sen University, China. He holds a BSc in Information Management (1st Class Honours) and a PhD in Information Systems (IS), both from the University of Sheffield. Prof Peng has over 90 publications in the IS field. He is the co-founder and co-chair of the IADIS International Conference on Information Systems Post-Implementation and Change Management (ISPCM) since 2012. He has also conducted peer review of submissions to more than 15 leading IS journals and international conferences.