Data Completeness in Healthcare: A Literature Survey

Caihua Liu
University of Technology Sydney
Australia
Caihua.Liu@student.uts.edu.au

Amir Talaei-Khoei
University of Nevada, Reno
USA
University of Technology, Sydney
Australia
atalaeikhoei@unr.edu

Didar Zowghi
University of Technology Sydney
Australia
Didar.Zowghi@uts.edu.au

Jay Daniel
University of Technology Sydney
Australia
Jay.Daniel@uts.edu.au

Abstract

As the adoption of eHealth has made it easier to access and aggregate healthcare data, there has been growing application for clinical decisions, health services planning, and public health monitoring with daily collected data in clinical care. Reliable data quality is a precursor of the aforementioned tasks. There is a body of research on data quality in healthcare, however, a clear picture of data completeness in this field is missing. This research aims to identify and classify current research themes related to data completeness in healthcare. In addition, the paper presents problems with data completeness in the reviewed literature and identifies methods that have been adopted to address those problems. This study has reviewed 24 papers (January 2011–April 2016) published in information and computing sciences, biomedical engineering, and medicine and health sciences journals. The paper uncovers three main research themes, including design and development, evaluation, and determinants. In conclusion, this paper improves our understanding of the current state of the art of data completeness in healthcare records and indicates future research directions.

Keywords: Data Completeness, Data Quality, Healthcare, Literature Survey

Citation: Liu, C., Talaei-Khoei, A., Zowghi, D. and Daniel, J. (2017). "Data Completeness in Healthcare: A Literature Survey," *Pacific Asia Journal of the Association for Information Systems, 9*(2), pp. 63-88.

Introduction

Data quality is a multidimensional construct and is defined as fitness for purpose (Chapman, 2005; Gamble and Goble, 2011; Shaw and Norton, 2008). High quality data facilitates operation, decision making and planning in most industries. While if data stakeholders assess the quality of data as poor, this assessment will sway their behaviour. In healthcare industry, poor data quality could lead to increase in mortality and loss of revenue. In order to improve quality of care and address cost-effectiveness, some governments make commitment to electronic health record (EHR) systems and supporting technology (Menachemi and Collum, 2011). For example, the Health Information Technology for Economic and Clinical Health (HITECH) Act provides incentives to those healthcare organizations which can achieve "meaningful use" of their data with health information technology in the United States (Stark, 2010). The meaningful-use regulations strike a balance between investment and improvement in healthcare (Blumenthal and Tavenner, 2010). This creates an impetus for healthcare organizations to enhance clinical decision-making in using integrated healthcare data across departmental systems and repositories. Data quality plays an essential role in evaluating the safety and quality of care (Liaw et al., 2013) and therefore, data quality related issues have received extensive attention in healthcare.

There are at least three dominant literature reviews of data quality in healthcare, with a number of dimensions reviewed such as accuracy, completeness, and consistency (Liaw et al., 2013; Thiru et. al., 2003; Weiskopf and Weng, 2013). Particularly, data completeness is one of the most frequently assessed dimensions for data quality in the existing healthcare literature (Weiskopf and Weng, 2013), and is considered as the major impediment to the availability of data for secondary use (Nobles et al., 2015), because data incompleteness could lead to significant uncertainty in health indicators such as tuberculosis incidence, prevalence and mortality rates (WHO, 2016).

In healthcare, data incompleteness could trigger medical errors during the course of care and hinder further analysis from monitoring and research purposes. First, when incomplete data emerges at the point of care, this could impact accurate diagnosis of a patient's condition. It is reported that a clinical diagnostic support system could make inappropriate recommendations about the risk of gastrointestinal bleeding in 77% of patient encounters resting on missing data (Ray et al., 2005). This undoubtedly leads to serious patient harm. Second, data incompleteness is one of the biggest barriers for secondary use to understand real-world status of patients (Weiskopf et al., 2013), which then could impact strategic usage of patient data such as planning of health services and facilities. In addition, this requires more efforts on dealing with missing data and explanation of processed data so that interrupts public health and clinical workflows (Dixon et al., 2011).

We can see that data completeness related issues could result in severe consequences in the domain of healthcare. An understanding about concept of data completeness can be viewed as a starting point to address issues related to data completeness. Prior review studies about data quality in healthcare have only partially studied completeness dimension, while a comprehensive literature survey of data completeness in healthcare is not available. This has motived us to conduct a research literature review to investigate recent research progress about data completeness in healthcare and to identify gaps in the existing literature.

The objectives of our work are to investigate fundamental concepts of data completeness in healthcare records, give a summary of recent research themes in this field, and explore potential challenges for further

exploration. While data completeness has been the subject of much research in the last three decades, achieving data completeness for patient safety and quality of care remains as a significant challenge for the healthcare industry. An investigation about which forms of healthcare records have concerned and addressed the issues of data completeness could give us a clue to identify possible difficulties in achieving data completeness. Therefore, our study investigates three categories of healthcare records: paper-based records, electronic records and hybrid records. Furthermore, according to total data quality management (TDQM) methodology (Wang, 1998), defining, measuring, analysing, and improving data quality are considered as essential processes to ensure high-quality information products. The organizations must first define what they mean by data quality and then establish a usable metrics linked to their goals and objectives for measuring how good is the data (Kovac et al., 1997). In this case, a good understanding of definition and measure of data completeness could serve as a foundation for the clear alignment between intention of data creation and its usage within an assessment of data completeness. We could generate appropriate definitions and measures of data completeness for conducting an assessment based on a given task. In addition, a clear picture of study themes about data completeness in healthcare could help us to obtain what are data completeness that have been studied and what are other potential challenges related to data completeness that may be investigated. Accordingly, three research questions (RQ) guided our study:

RQ1: What are the forms of healthcare records investigated in the literature to address data completeness?

RQ2: How has data completeness in healthcare records been defined and measured in the literature?

RQ3: What are the themes of data completeness studied in healthcare records?

In this work, we conducted a literature survey to review the existing studies that have addressed data completeness in healthcare between January 2011 and April 2016. The result from data extraction and analysis based on the included studies in our survey could be benefit academics in terms of improved understanding of the field, defining research themes and ascertaining research gaps. In addition, practitioners may solve their practical problems related to data completeness in healthcare by using solutions summarised in this study.

The rest of this paper is organized as follows. Section 2 elaborates the methods to identify the relevant publications. Section 3 gives answers to our research questions. Section 4 outlines the main findings of this review and limitations of our study. Section 5 presents conclusion of this work.

Methods

To conduct this literature survey, we followed the guidelines used by previous literature (Najaftorkaman et al., 2013), including three main discrete activities: (i) searching the initial list of papers; (ii) appraising relevant papers; (iii) extracting and analysing data.

Initial Search

We identified initial search resources informed through prior work (Najaftorkaman et al., 2013). We considered that the relevant papers might distribute in the areas of research in Information and Computing Sciences, Biomedical Engineering, and Medicine and Health Sciences. Hence, we selected a set of journals ranked in CORE[1] (CORE, 2016b) referring to these three domains, and we only screened journals ranked A*, A, or B in CORE in this survey. Accordingly, the studies from these journals

[1] CORE is an association of university departments of computer science in Australia and New Zealand and works as a good reference for quality evaluation of research (CORE, 2016a).

are more reliable, and the research quality is assessed through CORE ranking and impact factors.

Firstly, we found 8 journals from the journal list of Information and Computing Sciences (FoR code[2] 08) overlapping bioinformatics, medical informatics, medicine and health sciences (UNSW, 2013). Secondly, we screened 7 journals from the journal list of Biomedical Engineering (FoR code 0903) covering information and computer sciences, and 3 of them were duplicate with those found in the first round. Thirdly, in the journal list concerning Medicine and Health Sciences (FoR code 11), we only collected 1 journal in the list involving information systems, but this journal was already included. Eventually, 12 journals remained as our initial search resources. Furthermore, we measured the latest impact factor of each journal from Journal Citation Reports (Thomson Reuters, 2016), and most impact factors are greater than 1.000. In addition, we checked the Google citation of each selected journal (Google Scholar, 2017). These journals are indexed in different databases such as IEEE Xplore, ScienceDirect and Scopus (see Table 1).

Table 1 - List of journals selected in this review

No.	Title	Ranking in CORE	Impact factor	h5-index	Database	Identified papers	Relevant papers
1	IEEE Transactions on Information Technology in Biomedicine	A*	2.493	46	IEEE Xplore (IEEE, 2016)	2	1
2	Computer Methods and Programs in Biomedicine	A*	1.862	39	ScienceDirect (Elsevier, 2016a)	15	0
3	International Journal of Medical Informatics	A	2.363	43	ScienceDirect (Elsevier, 2016a)	43	12
4	BMC Bioinformatics	A	2.435	66	Scopus (Elsevier, 2016b)	0	0
5	Artificial Intelligence in Medicine	A	2.142	28	ScienceDirect (Elsevier, 2016a)	2	0
6	Computer Methods in Biomechanics and Biomedical Engineering	A	1.850	27	Scopus (Elsevier, 2016b)	0	0
7	Medical and Biological Engineering and Computing	A	1.018	-	Scopus (Elsevier, 2016b)	0	0
8	BMC Medical Informatics and Decision Making	B	2.042	33	Scopus (Elsevier, 2016b)	10	7
9	Computers in Biology and Medicine	B	1.521	32	ScienceDirect (Elsevier, 2016a)	7	1
10	Journal of the American Medical Informatics Association	B	3.428	61	Scopus (Elsevier, 2016b)	8	1
11	International Journal of Bioinformatics Research and Applications	B	-	8	Scopus (Elsevier, 2016b)	0	0
12	Journal of Biomedical Informatics	B	2.447	44	Scopus (Elsevier, 2016b)	12	2
	Total number of papers					99	24

Notes: " - " indicates that the Journal Citation Reports have not provided the impact factor for that journal or Google Scholar has not provided the number of citations for that journal in the last 5 complete years.

[2] Australian and New Zealand standard Fields of Research (FoR) code(s) assists in research classification (CORE, 2016b).

In order to generate the search keywords, we conducted a preliminary literature survey about data quality on those journals concerning the field of research in Information Systems and Computer Science. We identified several terms used to describe completeness in reviewed publications as our search terms, such as "availability", "coverage", "presence", "missingness", "omission" and "commission". These terms related to completeness have been frequently adopted in the assessment of data quality. Our search keywords developed for this literature survey contained: (1) data quality; (2) completeness; (3) healthcare. Additionally, we adopted alternative terms of "completeness" by using those six terms highlighted. Therefore, our search began with those keywords by using the Boolean operator as the following search string: 'data quality' AND ('completeness' OR 'availability' OR 'coverage' OR 'presence' OR 'missingness' OR 'omission' OR 'commission') AND 'healthcare' in all fields in the corresponding databases based on the title of each journal.

Relevance Appraisal

The inclusion criteria entail: firstly, the publications were in English; secondly, the papers were on the publication date between January 2011 and April 2016; thirdly, the literature contained an investigation on data completeness in healthcare. In addition, our search phrases included the search term referring to "completeness".

The exclusion criteria for removing a paper from further analysis is that the study only mentioned about completeness but without any explicit interpretation or application to healthcare.

We identified 99 publications from the initial search resources (12 journals), eliminated 75 papers by abstract and full-text review depending on exclusion criteria, thus 24 papers were selected for further study (in Table 1). Figure 1 depicts our search process in this review.

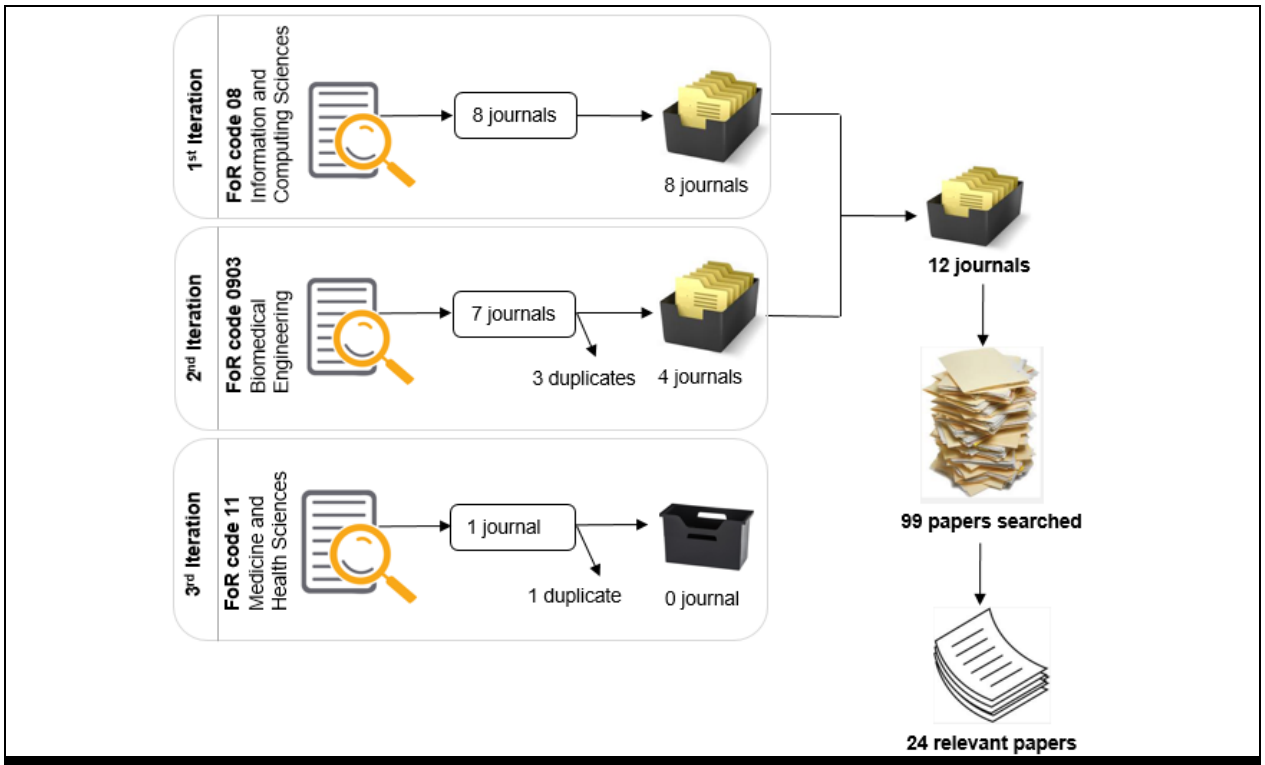


Figure 1 - The process of identifying relevant papers in this review

Data Extraction and Analysis

After identification of the relevant papers, we designed a form to extract data from included studies based on our research questions (see Appendix A). There are three criteria we applied in our form for data extraction. These criteria were derived from our three RQs.

Healthcare records – We classified the forms of healthcare records in the included studies into three groups: electronic records, paper-based records, and hybrid records. We assigned each paper to the corresponding group after identification of the form of healthcare records.

Definition and measure – We summarized several examples from those papers that explicitly provided definitions and measures of data completeness in their studies. In this way, we could address our fundamental questions about definition and measure of data completeness raised in the section of Introduction.

Study themes – In order to identify the study themes per paper, we emphasized the objectives and content of the study to determine the theme. If the objective of one study is to design or develop tools or methods to address data completeness, this kind of study could be categorised into 'design and development' theme. If the studies conducted an assessment of data completeness, they could be invited into the theme of 'evaluation'. While other studies investigated the challenges or barriers to achieving data completeness in healthcare, they could be classified into 'determinants' theme, since these difficulties encountered by practitioners in achieving data completeness also can be seen as factors impacting the achievement of completeness. Some paper may have multiple objectives and therefore, such research could address two or more than two study themes.

We reviewed each paper thoroughly keeping the three main criteria in focus. We applied thematic analysis (Guest et al., 2011) and frequency analysis (Onwuegbuzie and Combs, 2011) on the text of each paper to extract and record explicit response from publications for each criterion (healthcare records, definition and measure, and study themes). After finishing the data extraction, the first author labelled the dataset content with explicitly mentioned terminology in the literature and grouped them into several categories based on the criterion in the first round of classification. In the second and third rounds, we revised, merged or split some of the categories. In order to resolve discrepancies and achieve the final classification, we acted in two groups: literature survey and assessment. In this manner, the extracted data was assessed and re-examined against the literature for the validity of classification.

Results

This section presents the results of our review to address our research questions regarding healthcare records, definition and measure of completeness, and study themes of completeness. The studies reported in this section were selected from our sample based on their specific coverage of the concepts being investigated.

Healthcare Records

Table 2 presents the distribution about the forms of healthcare records in this review as following.

We can see that the included studies addressed the quality of the data mainly derived from paper-based records or electronic records. For instance, Adeleke et al. (2012) investigated data completeness in the first category of healthcare records (paper-based records). Their study focused on the documentation of inpatient admitted and discharged in paper charts. They assessed data completeness in paper-based records, standing at a set of specific documentation standards. They revealed that the discharge summary in psychiatry ward was underutilized because this

summary suffered from incompleteness in documentation.

The majority of studies applied the second category of healthcare records for data quality assessment. In this category, the healthcare records for investigation were derived from various sources such as EHR, EMR and EPR (as shown in Table 2). For example, Puttkammer et al. (2016) evaluated data quality of a retrospective dataset derived from the multisite EMR system in Haiti based on a set of health indicators. These indicators related to data completeness were specified into seven data elements of the patient's records such as age, height, weight, pregnancy status, HIV antiretroviral therapy eligibility, tuberculosis status and program discontinuation. By comparing data quality in EMR between two periods of time (2005-2012 and June-July 2013), they reported that there was a significant improvement in data quality at healthcare facilities in Haiti after the adoption of the EMR system.

Table 2 - Distribution on the forms of healthcare records in this review

Forms of healthcare records			Data context
Electronic records	EHR/Electronic medical records (EMR)/Electronic patient records (EPR)	Puttkammer et al. 2016	HIV
		Landis-Lewis et al. 2015	HIV
		Bruland et al. 2014	Pruritic dermatoses
		Haskew et al. 2015	Maternal and child health
		van Engen-Verheul et al. 2016	Cardiac rehabilitation
		Taggart et al. 2015	Cardiovascular disease and diabetes
		Hoffer et al. 2012	Kidney cancer
		Rahimi et al. 2014	Type 2 Diabetes Mellitus
		Heidebrecht et al. 2014	Public health (immunization)
		Köpcke et al. 2013	Clinical trial
		Tu et al. 2015	Not mentioned
		García-de-León-Chocano et al. 2015	Maternal and child health
		Weiskopf et al. 2013	Not mentioned
	Registry	Adolfsson and Rosenblad 2011	Diabetic
		Rousseau et al. 2014	Population health(vaccination)
	Reporting systems	Hirdes et al. 2013	Not mentioned
	Hospital information systems	Cohen et al. 2016	Not mentioned
		Herzberg et al. 2011	Medical history forms and stress injection protocols
		Breil et al. 2011	Oncology (urology and haematology)
		Cruz-Correia et al. 2013	Audit trail
	Distributed e-healthcare information environment	Wu et al. 2012	Breast cancer
	RFID systems	van der Togt et al. 2011	Blood products
Paper-based records		Adeleke et al. 2012	Inpatient health
Not mentioned		Liaw et al. 2013	Integrated chronic disease

Definition and Measure of Completeness

Table 3 lists several examples from the selected studies that have explicitly defined and measured data completeness for a dataset generated from healthcare records. In addition, these studies investigated the quality of the data extracted from electronic records, so we could easily compare their definitions and measures of completeness.

Table 3 - Definition and measure of completeness

Author and year	Definition of completeness	Measure of completeness
Puttkammer et al., 2016	Mandatory data related to a specific patient	Proportion of the number of flags with incomplete data observed by healthcare facilities monthly to the total number of flags by healthcare facilities monthly
Adolfsson and Rosenblad, 2011	Patients with non-missing data for a particular variable	Proportion of the number of patients with non-missing data in a required variable by the total number of patients reported to the Swedish National Diabetes Register
Taggart et al., 2015	A patient with at least one record for a specific attribute	Proportion of the number of patients with at least one record for a specific attribute to the total number of patients who had three or more visits in the two years
Rahimi et al., 2014	Two level definition: (1) availability of at least one record per patient; (2) availability of information required on a clinical decision making.	Proportion of the number of Type 2 Diabetes Mellitus (T2DM) attributes identified by both manual audit and ontology-based algorithm to the total number of numerator together with patients with T2DM attributes identified by the algorithm but not the manual audit
Weiskopf et al., 2013	Four perspectives: (1) documentation: all required observations are recorded during a clinical encounter; (2) breadth: availability of required multiple types of data; (3) density: availability of sufficient numbers of data points over time; (4) predictive: availability of sufficient information to predict an outcome.	(1) Proportion of the overall number of at least one visit recorded with a free-text note or report to the total number of visits; (2) Proportion of the overall number of at least one visit with multiple types of information to the total number of visits; (3) Proportion of the overall number of visits at least 1 day with multiple types of information to the total number of visits; (4) Using the logistic regression model.
Cruz-Correia et al., 2013	Sufficient information in depth, breadth, and scope to be used as an audit trail	Percentage of non-missing values in the essential data fields
Köpcke et al., 2013	Two conditions: (1) Data elements need to exist; (2) Fill in at least one of required data elements.	(1) Fraction of patient characteristics with at least one relevant data element (2) Fraction of patients with any data in at least one of required data elements in each characteristic

In this literature survey, the definitions of data completeness differed from each other. Basically, data completeness can be defined as the availability of at least one record for a specific attribute (Taggart et al., 2015; Rahimi et al., 2014; Köpcke et al., 2013; Adolfsson and Rosenblad, 2011;Liaw et al., 2013; Cruz-Correia et al., 2013; Weiskopf et al., 2013), and it also refers to element presence in core data fields (Rahimi et al., 2014; Köpcke et al., 2013; Wu et al., 2012; Puttkammer et al., 2016; Bruland et al., 2014; Heidebrecht et al., 2014; Hirdes et al., 2013; García-de-León-Chocano et al., 2015; Liaw et al., 2013; Cruz-Correia et al., 2013; Weiskopf et al., 2013). In addition, data completeness can be used to describe the availability of at least one record for an entity (Herzberg et al., 2011; van Engen-Verheul et al., 2016; Breil et al., 2011; Rousseau et al., 2014; Adeleke et al., 2012; van der Togt et al., 2011; Liaw et al., 2013; Cruz-Correia et al., 2013; Weiskopf et al., 2013). We can see that on the one hand, completeness implies that a patient has at least one record during an encounter. On the other hand, completeness presents that there are non-missing data elements in the crucial data fields for the context of use.

From 24 papers that was reviewed, 22 of them adopted a ratio scale to measure data completeness. Only one paper used regulation model for measuring data completeness (Weiskopf et al., 2013). The remaining two papers did not mention what methods they utilised to measure data completeness (Landis-Lewis et al., 2015; García-de-León-Chocano et al., 2015).

Study Themes of Completeness

We uncovered three main research themes including (i) Design and development; (ii) Evaluation; (iii) Determinants, as shown in Table 4.

Design and Development

This theme concerns design and development for data completeness maintenance including: (i) development of tools for data extraction, (ii) design of reminder systems, and (iii) construction of care data repositories. There are 4 studies in this category:

(i) Wu et al. (2012) integrated intelligent agents into the service-oriented architecture on the web to monitor data extraction process in the distributed e-healthcare information system environment. The agents could provide an optimal composition of task sequence by a quality of service based algorithm to preserve data completeness in data extraction according to different data tasks (Wu et al., 2012). On the other hand, Rahimi et al. (2014) adopted an ontology-based approach to identifying Type 2 Diabetes Mellitus (T2DM) patients in the EHR. The ontology-based algorithm added more constrains in the process of data extraction with domain knowledge of symptom and requirements of data quality to query the structured fields in the data repository (Rahimi et al., 2014). Completeness of the data extracted by evolutionary algorithms in both studies was good enough for their research purposes.

Table 4 - Study themes related to data completeness in healthcare in this review

No.	Author and year	Design and development			Evaluation						Determinants
		Data extraction tools	Reminder systems	Data repositories	For a single source only once	With a gold standard	Comparison between different tools, information systems or data tasks	Comparison among healthcare institutions	Before and after data quality interventions	During time intervals	
1	Wu et al., 2012	x					x				
2	Puttkammer et al., 2016									x	x
3	Cohen et al., 2016				x						x
4	Landis-Lewis et al., 2015										x
5	Herzberg et al., 2011		x						x		
6	Bruland et al., 2014								x		
7	Haskew et al., 2015							x		x	
8	van Engen-Verheul et al., 2016								x		x
9	Adolfsson and Rosenblad, 2011						x				
10	Taggart et al., 2015									x	x
11	Hoffer et al., 2012						x				
12	Liaw et al., 2013										x
13	Rahimi et al., 2014	x				x					x
14	Breil et al., 2011						x				
15	Cruz-Correia et al., 2013							x			
16	Heidebrecht et al., 2014						x				
17	Hirdes et al., 2013										x
18	Köpcke et al., 2013							x			
19	Rousseau et al., 2014					x					
20	Tu et al., 2015									x	
21	García-de-León-Chocano et al., 2015			x							
22	Adeleke et al., 2012				x						x
23	van der Togt et al., 2011					x					x
24	Weiskopf et al., 2013						x				
Number of papers in the category of Design and development: 4 Number of papers in the category of Evaluation: 19 Number of papers in the category of Determinants: 10											

Note: The total number of study themes is greater than the number of included studies, since each study may address two or more study themes.

(ii) Currently, healthcare practitioners are playing the role both in their main duties and data recording tasks. Sometimes heavy workload in the main duties detracts personnel's attention from data recording tasks (Odega et al., 2010). For example, in a clinic operation, the clinic information has to be recorded at the beginning and the end of anesthesia (e.g., operation time), while anesthetists concentrate on their clinical work rather than the recording task (Wrightson, 2010). Thus, part of the clinic information could not be recorded in time so that the data is inevitably missing. Herzberg et al. (2011) designed and developed a computer-based reminder system for the improvement of documentation completeness. This system could automatically identify incomplete forms, provide a schedule about due records, and notify the responsible personnel by an email after a certain grace period. They reported that completeness in clinical documentation increased highly significantly with 100% complete forms after implementation of the reminder system. Hence, a reminder system could help healthcare practitioners to achieve data completeness in their routine work within a balance between the main duties and recording tasks.

(iii) The implementation of the EHR systems provides valuable data for monitoring the clinical process and activating research studies. Due to a lack of control mechanisms about the deployment of the Baby Friendly Hospital Initiative for the protection, promotion and support of breastfeeding purposes, García-de-León-Chocano et al. (2015) constructed care data repositories for infant feeding from the EHR. They ensured the quality of input and output data in each step of the construction by using semantic integrity of clinical concepts in the dataset content to achieve high-quality data (such as effective data completeness) for their specific tasks (García-de-León-Chocano et al., 2015).

Evaluation

In this theme, we focused on data completeness evaluation. 19 papers addressed the assessment of data completeness in healthcare records for a given task. Figure 2 presents the distribution of various methods adopted to assess data completeness in this category.

The data quality assessment could be performed for a single source only one time (Adeleke et al., 2012; Cohen et al., 2016). Some studies assessed data completeness by using another source (a gold standard) such as manual audit of the EHR (Rahimi et al., 2014), paper lists (Rousseau et al., 2014), and real entities (van der Togt et al., 2011). Other research evaluated the differences in completeness between various methods of data collection, utilizing different task sequence related to data quality (Wu et al., 2012), information systems (Adolfsson and Rosenblad, 2011; Heidebrecht et al., 2014), documentation tools (Hoffer et al., 2012), documentation forms (Breil et al., 2011), or models to address different definitions of completeness (Weiskopf et al., 2013). This sub group has received the most attention at 32% of studies reviewed in the category of evaluation. The rest of the studies assessed data completeness in different healthcare institutions (Cruz-Correia et al., 2013; Haskew et al., 2015; Köpcke et al., 2013), pre- and post-implementation of data quality interventions regarding introduction of advanced information systems (Bruland et al., 2014; Herzberg et al., 2011; van Engen-Verheul et al., 2016) and during different time points (Haskew et al., 2015; Puttkammer et al., 2016; Taggart et al., 2015; Tu et al., 2015).

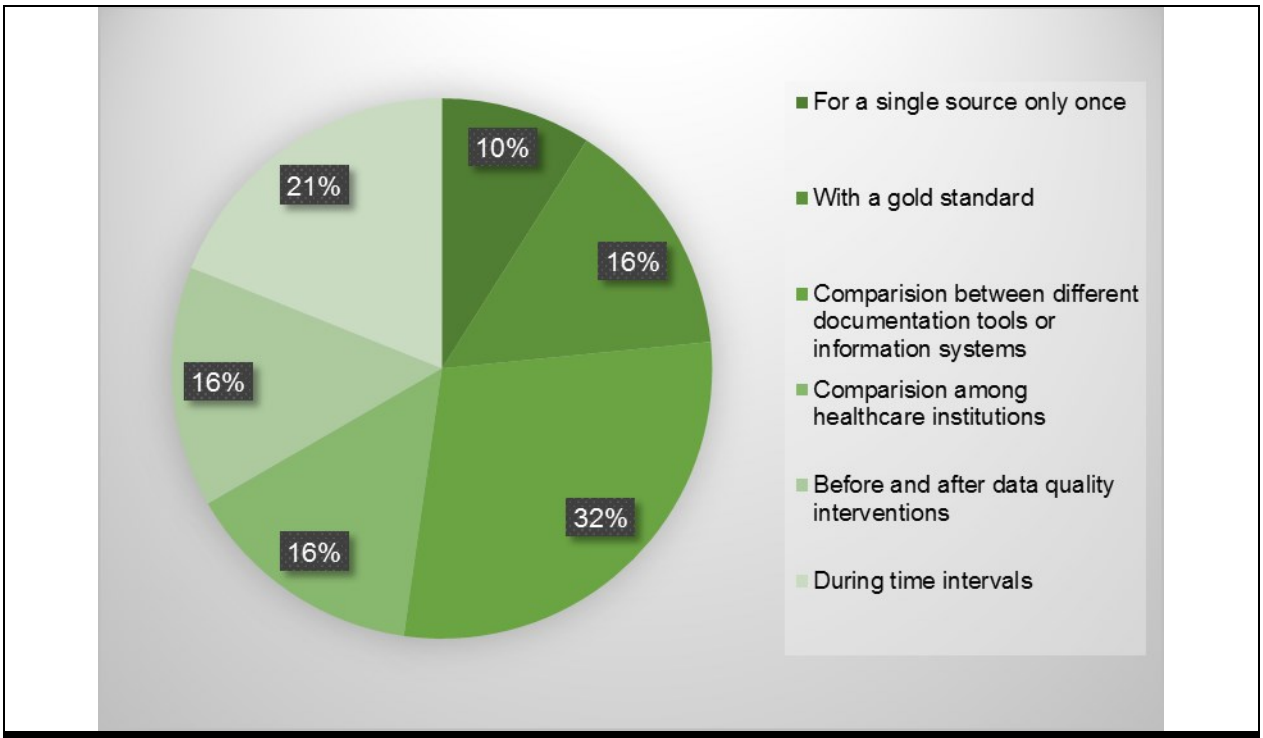


Figure 2 - Distribution of study themes about evaluation in this review

Determinants

In terms of this study theme, we concentrated on factors affecting data completeness in healthcare records. For paper-based records, a lack of standards for documentation could result in incomplete data (Adeleke et al., 2012). Furthermore, when transcribing a patient's information from hard copies into information systems, data entry personnel might record insufficient or biased information due to legibility issues. For electronic records, data completeness suffers from (i) human, (ii) technical, and (iii) environmental challenges.

(i) User acceptance of eHealth impacts an individual's attitudes and intentions towards eHealth and his or her adoption of eHealth. van Engen-Verheul et al. (2016) revealed that most clinical officers do not use the EHR much. This could cause incomplete data entered into the system. While Taggart et al. (2015) found that poor data completeness could be due to patients' uncooperative attitude in providing their information. In addition, their study uncovered that time constraints on recording tasks and difficulties in achieving balance between recording tasks and clinical work resulted in incompleteness in healthcare records. Moreover, personnel do not follow the organisational data protocols in data management that could lead to ineffective data completeness (Liaw et al., 2013). Additionally, manual errors in data extraction due to insufficient experience could cause incomplete data aggregated (Liaw et al., 2013).

(ii) Ineffective information technology affects data completeness. Liaw et al. (2013) presented that poor coding rules and corruption of the database architecture or information systems could cause incompleteness in managing routinely collected healthcare data. One of the problems caused by poor coding rules is the warp of observations to be the equivalent of incomplete data. For instance, if two responses cannot logically both be true, it is not obvious which one is true (Hirdes et al., 2013). Hence, these problems could lead to missing values. Furthermore, corruption of

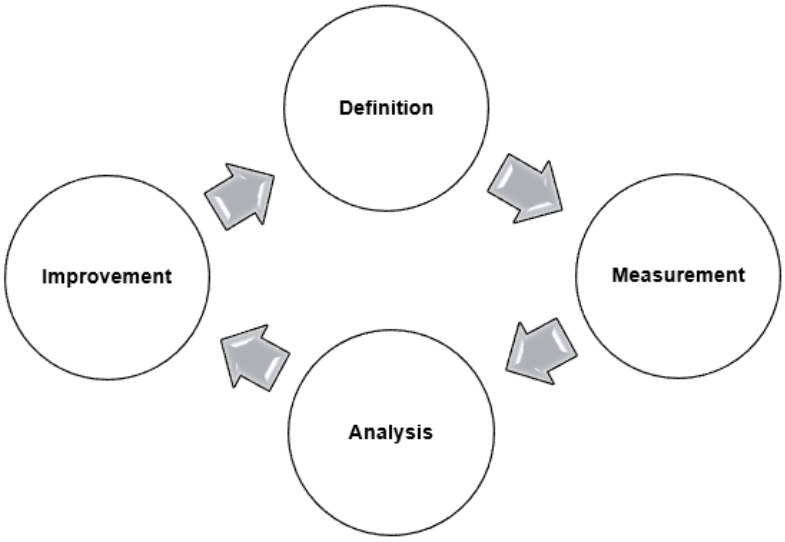
the database architecture or information systems could not meet users' requirements to manage daily patient-generated health data in clinical care that undoubtfully affect data completeness. Specifically, two studies employed importance-performance analysis (Cohen et al., 2016) and think-aloud usability testing (van Engen-Verheul et al., 2016) separately to analyze attributes of healthcare systems. Both studies collected a feedback from those practitioners who used the systems and gave insights into what are barriers to achieving data completeness. Cohen et al. (2016) revealed that the limited input space of user interface could result in incomplete data in patient records during the data entry, while van Engen-Verheul et al. (2016) disclosed three usability problems regarding incomplete data collection as violations of the heuristic match between system and world, such as consistency and standards, flexibility and efficiency, and visibility of system state.

(iii) Data completeness also suffers from the external environment, especially for sensor data. van der Togt et al. (2011) addressed quality testing of the data generated by Radio Frequency IDentification (RFID). This system could read and write physical objects. With automatic identification and tracking technology, the RFID system could generate location, time and temperature data of tagged blood products left in the blood transfusion laboratory. Then, they tested the system's performance with these blood products left in the blood transfusion chain. They found that the blood bag blocked the signals of tags that could disturb a good performance of the system to identify the location, time and temperature data of the blood products and reduce completeness of the sensor data. Furthermore, frequency of power interruptions could lead to interruptions of recording tasks in care delivery (Puttkammer et al., 2016). Hence, complete patient's information cannot be well documented.

Discussion

Figure 3 presents a lifecycle of data completeness management in healthcare that this work has derived from the literature as inspirited by prior study (Batini et al., 2009). Our findings are in the agreement with Wang (1998)'s TDQM methodology. Based on the existing healthcare studies about data completeness, this study has summarized the definitions and measures of data completeness concerning the fit to improving the quality of care and clinical decisions. Then, determinants influencing practitioners in achieving data completeness could be identified after the analysis of relevant issues derived from the data completeness assessment. In this way, we could determine the core areas for improvement via a variety of data completeness interventions (such as implementation of reminder systems and ontology technology). This also facilitates the design and development of techniques (or tools) for improving data completeness in healthcare records in order to achieve the quality of clinical decision and quality of care.

In this literature survey, we investigated the forms that healthcare records take in the included studies in order to address data completeness. Additionally, we described several examples that explicitly defined and measured data completeness in this survey. Furthermore, we summarized three study themes regarding data completeness in healthcare: design and development, evaluation, and determinants. In this section, we answer each of RQs and outline areas that require further exploration below.



Phase 1: Definition
Input: Data completeness requirements from data stakeholders
Data analysis: Definition of data completeness at three levels: (1) high level description of every instance of data in a dataset that represents an entity in the real world: (2) low level description of a state where there is a non-null value in every field of a dataset; and (3) predictive level description of availability of sufficient information to predict an outcome.
Output: Design of information product with specific completeness requirements

Phase 2: Measurement
Input: Information products
Evaluation: Assessment of data completeness within appropriate models
Output: Data completeness related issues

Phase 3: Analysis
Input: Data completeness related issues
Identification of determinants: Exploration of difficulties in achieving data completeness
Output: Strategies for improvement of data completeness

Phase 4: Improvement
Input: Data completeness matrix of information products
Design and development: Selection of strategies and techniques
Output: Data completeness intervention techniques

Figure 3 - Lifecycle of data completeness management in healthcare in this review

Healthcare Records

RQ1: What are the forms of healthcare records investigated in the literature to address data completeness?

Answer: There are two main forms: paper-based records and electronic records.

In this review, most studies concentrated on electronic form of healthcare records, and a few studies investigated paper-based records or hybrid records. Due to incentives of the HITECH ACT, many healthcare organizations have decided to make the move from paper-based records to electronic records. The implementation of EHR systems is an expected result, not overnight, but also not open-ended time-wise. Therefore, the meaningful use of EHR and supporting technology is at a steady rate (Jamoom et al., 2012). During the process of transition from paper-based records to EHRs, the complexities of this

migration raise top concerns about the quality of care and patient safety (Dolezel and Moczygemba, 2015). For instance, a hybrid record emerges when the paper-based diagnostic results (e.g., lab reports) are scanned into the EHR. Furthermore, hard copies of healthcare information (e.g., discharge summary) are remained as well. This is challenging the migration of the legacy data selected from paper-based records to the EHR such as what data is needed to transit and how ensure complete and accurate transition (Dolezel and Moczygemba, 2015). In other words, practitioners could meet difficulties of achieving completeness in data transition from paper-based records to the EHR for patient safety.

It is evidential that some small hospitals are falling behind in the implementation of EHR systems due to unconvinced return on investment in the short term but large investment in technology (Thakkar and Davis, 2006). Furthermore, the EHR technology changes the existing systems and processes in clinical practice, which introduces barriers to its adoption (Moreno, 2005). Accordingly, healthcare organizations still need to manage the increasing amount of data generated from paper-based records and electronic records at the same time. Preserving data completeness in the combination of paper-based records and electronic records is expected to achieve the availability of necessary data for assessing quality of patient care. The focus of the existing studies is on data completeness in the EHR, while an investigation of data completeness in the data transition from paper-based records to the EHR system could share valuable experience for data migration and facilitate the implementation of the EHR in healthcare organizations. Furthermore, the integrated use of paper-based records and electronic records is still needed to address data completeness for current routine patient care in order to ensure patient safety.

Definition and Measure of Completeness

RQ2: How has data completeness in healthcare records been defined and measured in the literature?

Answer: (i) The reviewed studies presented many different ways to define data completeness in electronic records, each focussing on specific context or purposes. (ii)The most frequently mentioned method adopted to measure data completeness is a ratio scale.

The definition of completeness could determine the number of complete records for investigation (Weiskopf et al., 2013). For documentation completeness of healthcare records, required attributes per record must be present. Concerning breadth or depth, the definitions of completeness become more complicated. For instance, to identify a patient with T2DM, we could define data completeness as at least one patient record matching the criteria in a specific attribute such as in reason for visit, medication, and pathology (Rahimi et al., 2014). In this case, as long as a patient's record met one characteristic among these attributes, we could determine this patient suffering from T2DM. From the predictive perspective to define data completeness, we could predict an outcome with sufficient information (Weiskopf et al., 2013). The definitions of data completeness in the included studies differ from one to another, while completeness is defined for purposes. From a high-level viewpoint, data completeness implies that every instance of data in a dataset represents an entity in the real world. A more fine-grained definition of data completeness refers to a state where there is a non-null value in every field of a dataset. The aforementioned two viewpoints addressing completeness focus on the existing evidence, while predictive completeness is based on the existing information for trend analysis, estimating a consequence in the future and extending the concept of data completeness. Thus, information products based upon predictive

definition of data completeness could address disease prevention and prediction.

In this review, methods adopted to measure completeness in the literature emphasize on percentage calculation. In other words, data completeness is generally quantified by a ratio. For predictive completeness, we require more complex mathematical models to measure data completeness such as regulation models. The preferences for measures of completeness were on the ground of missing data by counting nulls or empties in a dataset. However, not all nulls in a dataset mean that they are missing, since some null values could be implied by the relationships among the attributes such as functional dependencies in a relational dataset (Liu et al., 2016). If we ignore the relations among the attributes in a dataset, the result of measuring data completeness may be biased. Further, a non-null value does not mean that the data is correct and up-to-date. Since data completeness implies that the data describes all the facts in the real world, the complete data should be accurate and up-to-date as well. That is to say, there is an overlap among completeness and other dimensions of data quality, such as accuracy and timeliness. Thus, the measures of data completeness should go beyond only counting the overall number of nulls or empties.

Study Themes of Completeness

RQ3: What are the themes of data completeness studied in healthcare records?

Answer: There are three main research themes related to data completeness: (i) design and development, (ii) evaluation, and (iii) determinants.

Design and Development

In this review, we identified four studies that addressed design and development of methods to improve data completeness such as a reminder system for improvement of clinical documentation completeness, a quality-assured construction of care data repositories, and evolutionary tools for extracting high-quality data. Recently, more and more intelligent agents are employed to support decision-making for healthcare practitioners. The evolutionary algorithm resting on data quality has been integrated into intelligent agents in data extraction, enabling the aggregation of high-quality data (Wu et al. 2012). Preview research examined ontology-based algorithms in the tasks of identifying T2DM patients which can address semantical interoperability across clinical systems and repositories and at the same time achieve reliable data from integrated healthcare records for the clinical decision about a T2DM patient (Rahimi et al., 2014).

An ontology provides a means for its users to consistently and accurately utilise uniform terminology about the same entities in some domain. Ontologies now underpin almost all aspects of healthcare such as clinical research and patient care, within numerous uses for data management, including data entry (Sahoo et al., 2014a), integration (Wu et al., 2015), access (Sahoo et al., 2014b), collection (Choquet et al., 2015; Klann et al., 2015), and reasoning (El-Sappagh et al., 2015). For example, a healthcare system could take advantages of an ontology to notify users with inconsistent terminology during the entry of patient data. In this manner, data quality in healthcare records could be improved through integrating identified data, decreasing incorrect data or adding missing data (Vandenbussche et al., 2013).

In reality, healthcare practitioners need to manually extract data from narrative fields in the EHR for specific purposes. At this time if staff members lack sufficient domain knowledge or specialist experience, they could not parse and locate the data. Moreover, manual data extraction could not guarantee the consistency in the dataset context. As a result, the data extracted from the unstructured text could be incomplete. However, ontology-based approaches could support automated processes to deal with completeness in data extraction. This is the potential of ontologies for quantifying care

coordination from the narrative notes (Popejoy et al., 2014).

Nowadays, the health services are moving to personalized care from organization-centered care (Kim et al., 2015). The personal health paradigm shared between health professionals and the patient requires a bridge to deal with a broad range of domain terminologies and concepts. Furthermore, as mentioned in the Introduction section, "meaningful use" regulations are highlighted in using health information technology. One of the "meaningful use" regulations is imposing healthcare organizations and systems on using standardized vocabularies and ontologies (Blumenthal and Tavenner, 2010). As a key element of healthcare, ontologies enable the development of automated methods to manage data and measure data completeness in the EHR in order to achieve the quality of health services and patient safety in the personalized care.

Evaluation

For data quality evaluation, the included studies applied various methods to assess data completeness in healthcare. Furthermore, we identified five common issues encountered in assessing data completeness in this review. First, a small size of data collection could limit the generalization of analysis results. Second, the data collected in a short period of time could lead to incomplete data, because some data is recorded outside the time frame established for the study. Third, data collection could meet bias if the data is collected only based on the viewpoint of data collectors themselves, because the extracted data may not achieve the requirements of data completeness in practice. Fourth, a few studies took the external factors into account that could bias the analysis results of data completeness measurement. For example, if there is a data quality intervention (example of power interruptions) during the time period of study, the quality of the data is significantly affected. Last, when assessing data completeness for a specific dataset, a few studies applied a gold standard to validate the result of completeness measurement for the given dataset. As a result, it was difficult for data consumers to determine whether the result of measurement was accurate.

From the literature, we also summarized some implied methods proposed to address those issues, as shown in Table 5.

Table 5 - Data completeness assessment: issues and methods in this review

Issues in assessing data completeness	**Methods proposed to address the issues**			
	Qualitative methods (Delphi processes)	Control group	Gold standard	Examination of data accuracy
A small size of data collection	x			
A short period of time in data collection				
Bias in data collection	x			
Without consideration of external factors		x		
Lack of validation about the measurement results			x	x

Qualitative methods can help to achieve an agreement on data quality priorities for solution, such as Delphi processes (Puttkammer et al., 2016). Because the feedback from data stakeholders represent the requirements on completeness, the data collected depending on such criteria could reduce the bias in data collection and

ascertain the scope of data collection. The study should be under a control group for exclusion of external factors impacting on data quality (Taggart et al., 2015). The reliability of the result of measuring data completeness could be addressed in two ways: (i) a dataset derived from another source or multiple sources (a gold standard) is used to compare with the given dataset; and (ii) an examination of the accuracy for the dataset needs to address when assessing data completeness (Weiskopf et al., 2013). As indicated in Table 5, there were a few studies presenting how to solve the problem of data collection in a short period of time in data quality assessment. Poor assessment practices could affect reliability and validity of the evaluation results. Since practitioners would encounter disparate problems in assessing data completeness both in research and practice, we need a comprehensive and systematic review to summarize the issues in assessing data completeness and methods used to address those issues.

Determinants

We highlighted the factors affecting data completeness in healthcare in the section of Results regarding Determinants. In addition to these difficulties in achieving data completeness identified in reviewed studies, there were some methods used in the literature to address those difficulties, as presented in Table 6.

Table 6 - Determinants of data completeness in healthcare: difficulties and methods in this review

Category	Difficulties	Methods			
		Ontology-based algorithms	Structured data quality reports	Clinical processes embedded in eHealth workflow	Reminder systems
Human	User acceptance of eHealth (Landis-Lewis et al., 2015)		x	x	
	Time constraints on recording tasks (Taggart et al., 2015)			x	x
	Balance between recording tasks and clinic work (Taggart et al., 2015)			x	x
	Compliance to organisational data protocols (Adeleke et al., 2012; Liaw et al., 2013; Taggart et al., 2015)		x		
	Capability of manual data extraction (Liaw et al., 2013; Rahimi et al., 2014)	x			
Technical	Poor design of user interface (Cohen et al., 2016; van Engen-Verheul et al., 2016)				
	Poor coding rules (Hirdes et al., 2013; Liaw et al., 2013)				
	Corruption of database architecture (Liaw et al., 2013)				
Environmental	Power interruptions (Puttkammer et al., 2016)				
	Local circumstances (van der Togt et al., 2011)				

The data quality interventions and activities adopted in reviewed studies to preserve data completeness in healthcare include: (i) ontology-based algorithms for data extraction, (ii) structured data quality reports, (iii) clinical processes embedded in eHealth workflow, (iv) reminder systems for notification of incomplete documentation.

(i) Ontology-based algorithms could deal with semantic integrity of clinical concepts and data quality in data extraction that reduce manual errors due to biased or insufficient domain knowledge related to healthcare (Liaw et al., 2013; Rahimi et al., 2014).

(ii) A formal feedback of data quality assessment could engage healthcare practitioners in standardized use of eHealth and thereby improve data completeness (Taggart et al., 2015).

(iii) The clinical documentation processes embedded in eHealth could reduce manual transcription steps in patient care and make records well documented (Bruland et al., 2014). The workflow of eHealth accommodating with clinical processes could help to address difficulties in achieving balance between recording tasks and clinical work under the pressure of time (Taggart et al., 2015).

(iv) A computer-based reminder system automatically identifies incomplete records and notifies the responsible person with an email after a certain grace period. This could assist in recording tasks to achieve documentation completeness, especially in busy operations (Herzberg et al., 2011).

Table 6 appears that most difficulties in achieving data completeness in healthcare records have moved from technical perspective to human perspective. The technical factors impact implementation and adoption of eHealth in physician practices, and at the same time users' acceptance of eHealth affects the application of technologies in their routine clinical work. Due to a lack of awareness and knowledge base of data completeness, more and more missing items are aggregated during the course of care and then disturb data completeness of care data repositories in the long run. It is difficult to recover these errors without any comparable data sources. As a result, medical errors could occur. In late 1999, the Institute of Medicine released a landmark report, To Err Is Human (Donaldson et al., 2000), and stated that preventable medical errors are a leading cause of death in hospitals. In this review, we can see that individual knowledge and attitude could trigger human errors when dealing with data completeness. In fact, human error has been studied in many human cognitive domains for more than 100 years. This is the potential of use of cognitive theory to explain why errors occur in achieving data completeness. In this manner, we could generate possible mechanisms to preserve data completeness or systematically reduce incomplete data.

The methods adopted to deal with difficulties in achieving data completeness from human perspective as identified in Table 6. However, the included papers might not reveal the methods that have been used to address poor design of user interface, inconsistent coding rules, corruption of the database architecture, power interruptions, and local circumstances with empirical evidence. It should go without saying that data completeness is a topic of prolonged interest for academics and practitioners, since incomplete data could introduce bias into the dataset that results in invalid conclusions and poor decision-making. Even if we cannot know the purpose of a dataset for a users' task in advance, we could make our every effort to preserve completeness at early stages.

Our study only covered the existing literature addressing data completeness in healthcare in the last five years. There is a lack of systemic review of challenges in achieving data completeness in healthcare and solutions used to address those challenges. Moreover, in this literature survey, most studies were interested in the

determinants of poor data completeness, while a few discussed how data completeness impacts on clinical decision-making, disease management, health services settings, cost-efficiency, and quality of care. The main purpose of investigation on data quality for a dataset is to determine whether the data at hand can be used for a valid analysis in order to make a right decision and an effective planning.

Limitations

Any literature survey conducted present some limitations, this one is not an exception. Firstly, the initial search sources we employed were limited. We only selected 24 papers from 12 journals and focused on the publications between 2011 and 2016, which limited the number of papers collected. Secondly, the articles screened were based on our inclusion and exclusion criteria so that the results of articles selected together with data analysis were subjective. Hence, a systematic literature review of data completeness in healthcare is encouraged, which provides a methodical process of collecting and collating the published empirical studies with systematic criteria for section and quality assessment to reduce bias and provide transparency to the process. Finally, thematic analysis is conducted to identify emerging themes are subject to bias, even though the themes and categories selected were reviewed and discussed by all authors.

Conclusion

This study provides an improved understanding of data completeness in current state of the art for healthcare records. We have analysed and synthesized 24 studies to address 3 research questions, different aspects of data completeness in healthcare were summarized, with a description about health records, definition and measure, and study themes. Firstly, what forms of healthcare records could be investigated to address data completeness? The data can be derived from paper-based records or electronic records. Secondly, how data completeness is defined and measured? It is necessary to give appropriate definitions to measure data completeness on the context of use. Lastly, what the themes related to data completeness could be studied? There are three main themes concerning data completeness: (i) design and development, (ii) evaluation, and (iii) determinants. In addition, we have explored future research directions in terms of data completeness in healthcare and indicated potential solutions to address related issues.

References

Adeleke, I. T., Adekanye, A. O., Onawola, K. A., Okuku, A. G., Adefemi, S. A., Erinle, S. A., & James, J. A. (2012). "Data Quality Assessment in Healthcare: A 365-day chart review of inpatients' health records at a Nigerian tertiary hospital," *Journal of the American Medical Informatics Association, 19*(6), pp.1039-1042.

Adolfsson, E. T., & Rosenblad, A. (2011). "Reporting Systems, Reporting Rates and Completeness of Data Reported from Primary Healthcare to a Swedish Quality Register–the National Diabetes Register," *International Journal of Medical Informatics, 80*(9), pp.663-668.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). "Methodologies for Data Quality Assessment and Improvement," *ACM Computing Surveys, 41*(3), pp.1-52.

Blumenthal, D., & Tavenner, M. (2010). "The “Meaningful Use” Regulation for Electronic Health Records," *New England Journal of Medicine, 2010*(363), pp.501-504.

Breil, B., Semjonow, A., Müller-Tidow, C., Fritz, F., & Dugas, M. (2011). "HIS-based Kaplan-Meier Plots-a Single Source Approach for Documenting

and Reusing Routine Survival Information," *BMC Medical Informatics and Decision Making, 11*(1), p.1.

Bruland, P., Forster, C., Breil, B., Ständer, S., Dugas, M., & Fritz, F. (2014). "Does Single-source Create an Added Value? Evaluating the impact of introducing x4T into the clinical routine on workflow modifications, data quality and cost–benefit," *International Journal of Medical Informatics, 83*(12), pp.915-928.

Chapman, A. D. (2005). *Principles of Data Quality.* Report for the Global Biodiversity Information Facility. Copenhagen.

Choquet, R., Maaroufi, M., de Carrara, A., Messiaen, C., Luigi, E., & Landais, P. (2015). "A Methodology for a Minimum Data Set for Rare Diseases to Support National Centers of Excellence for Healthcare and Research," *Journal of the American Medical Informatics Association, 22*(1), pp.76-85.

Cohen, J. F., Coleman, E., & Kangethe, M. J. (2016). "An Importance-performance Analysis of Hospital Information System Attributes: A nurses' perspective," *International Journal of Medical Informatics, 86*, pp.82-90.

CORE (2016a). "CORE Rankings Portal," Retrieved from http://portal.core.edu.au/jnl-ranks/ on May 12, 2016.

CORE (2016b). "About the DB," Retrieved from http://www.core.edu.au/conference-portal/About-the-db on May 15, 2016.

Cruz-Correia, R., Boldt, I., Lapão, L., Santos-Pereira, C., Rodrigues, P. P., Ferreira, A. M., & Freitas, A. (2013). "Analysis of the Quality of Hospital Information Systems Audit Trails," *BMC Medical Informatics and Decision Making, 13*(1), p.84.

Dixon, B. E., McGowan, J. J., & Grannis, S. J. (2011). "Electronic Laboratory Data Quality and the Value of a Health Information Exchange to Support Public Health Reporting Processes," *AMIA Annual Symposium Proceedings, 2011* (2011), pp. 322–330.

Dolezel, D., & Moczygemba, J. (2015). "Implementing EHRs: An exploratory study to examine current practices in migrating physician practice," *Perspectives in Health Information Management, 12*(Winter).

Donaldson, M. S., Corrigan, J. M., & Kohn, L. T. (2000). *To Err is Human: Building a safer health system.* DC National Academy Press: Washington.

El-Sappagh, S., Elmogy, M., & Riad, A. (2015). "A Fuzzy-ontology-oriented Case-based Reasoning Framework for Semantic Diabetes Diagnosis," *Artificial Intelligence in Medicine, 65*(3), pp.179-208.

Elsevier (2016a). "ScienceDirect," Retrieved from http://www.sciencedirect.com/ on May 26, 2016.

Elsevier (2016b). "Scopus," Retrieved from https://www.scopus.com/ on May 26, 2016.

Gamble, M., & Goble, C. (2011). "Quality, Trust, and Utility of Scientific Data on the Web: Towards a joint model," *Paper presented at the Third International Web Science Conference.*

García-de-León-Chocano, R., Sáez, C., Muñoz-Soler, V., & García-Gómez, J. M. (2015). "Construction of Quality-assured Infant Feeding Process of Care Data Repositories: Definition and design (Part 1)," *Computers in Biology and Medicine, 67*, pp.95-103.

Google Scholar. (2017). Google Scholar Matrics. Retrieved from https://scholar.google.com.au/citations

?view_op=top_venues&hl=en on April 14, 2017

Guest, G., MacQueen, K. M., & Namey, E. E. (2011). *Applied Thematic Analysis.* Sage Pulications: Thousand Oaks, Califonia.

Haskew, J., Rø, G., Saito, K., Turner, K., Odhiambo, G., Wamae, A., & Sugishita, T. (2015). "Implementation of a Cloud-based Electronic Medical Record for Maternal and Child Health in Rural Kenya," *International Journal of Medical Informatics, 84*(5), pp.349-354.

Heidebrecht, C. L., Kwong, J. C., Finkelstein, M., Quan, S. D., Pereira, J. A., Quach, S., & Deeks, S. L. (2014). "Electronic Immunization Data Collection Systems: Application of an evaluation framework," *BMC Medical Informatics and Decision Making, 14*(1), p.1.

Herzberg, S., Rahbar, K., Stegger, L., Schäfers, M., & Dugas, M. (2011). "Concept and Implementation of a Computer-based Reminder System to Increase Completeness in Clinical Documentation," *International Journal of Medical Informatics, 80*(5), pp.351-358.

Hirdes, J. P., Poss, J. W., Caldarelli, H., Fries, B. E., Morris, J. N., Teare, G. F., & Jutan, N. (2013). "An Evaluation of Data Quality in Canada's Continuing Care Reporting System (CCRS): Secondary analyses of Ontario data submitted between 1996 and 2011," *BMC Medical Informatics and Decision Making, 13*(1), p.1.

Hoffer, D. N., Finelli, A., Chow, R., Liu, J., Truong, T., Lane, K., & Kurban, G. (2012). "Structured Electronic Operative Reporting: Comparison with dictation in Kidney cancer surgery," *International Journal of Medical Informatics, 81*(3), pp.182-191.

IEEE (2016). "IEEE Xplore Digital Library," Retrieved from http://ieeexplore.ieee.org/ on 26 May, 2016

Jamoom, E., Beatty, P., Bercovitz, A., Woodwell, D., Palso, K., & Rechtsteiner, E. (2012). "Physician Adoption of Electronic Health Record Systems: United States, 2011," *NCHS Data Brief, 98,* pp. 1-8.

Kim, H. H., Lee, S. Y., Baik, S. Y., & Kim, J. H. (2015). "MELLO: Medical lifelog ontology for data terms from self-tracking and lifelog devices," *International Journal of Medical Informatics, 84*(12), pp.1099-1110.

Klann, J. G., Mendis, M., Phillips, L. C., Goodson, A. P., Rocha, B. H., Goldberg, H. S., & Murphy, S. N. (2015). "Taking Advantage of Continuity of Care Documents to Populate a Research Repository," *Journal of the American Medical Informatics Association, 22*(2), pp.370-379.

Köpcke, F., Trinczek, B., Majeed, R. W., Schreiweis, B., Wenk, J., Leusch, T., & Röhrig, R. (2013). "Evaluation of Data Completeness in the Electronic Health Record for the Purpose of Patient Recruitment into Clinical Rrials: A retrospective analysis of element presence," *BMC Medical Informatics and Decision Making, 13*(1), p.1.

Kovac, R., Lee, Y. W., & Pipino, L. (1997). *Total Data Quality Management: The Case of IRI. Paper presented at the Conference on Information Quality.*

Landis-Lewis, Z., Manjomo, R., Gadabu, O. J., Kam, M., Simwaka, B. N., Zickmund, S. L., & Jacobson, R. S. (2015). "Barriers to Using eHealth Data for Clinical Performance Feedback in Malawi: A case study. *International Journal of Medical Informatics, 84*(10), pp.868-875.

Liaw, S.-T., Rahimi, A., Ray, P., Taggart, J., Dennis, S., de Lusignan, S., & 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), pp.10-24.

Liu, Y.-N., Li, J.-Z., & Zou, Z.-N. (2016). "Determining the Real Data Completeness of a Relational Dataset," *Journal of Computer Science and Technology, 31*(4), pp.720-740.

Menachemi, N., & Collum, T. H. (2011). "Benefits and Drawbacks of Electronic Health Record Systems," *Risk Management and Healthcare Policy, 4*, pp.47-55.

Moreno, L. (2005). *Electronic Health Records: Synthesizing Recent Evidence and Current Policy*. Mathematica Policy Research: New Jersey.

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), pp.12-21.

Nobles, A. L., Vilankar, K., Wu, H., & Barnes, L. E. (2015). "Evaluation of Data Quality of Multisite Electronic Health Record Data for Secondary Analysis," *Paper presented at 2015 IEEE International Conference on Big Data*.

Odega, C., Fatiregun, A., & Osagbemi, G. (2010). "Completeness of Suspected Measles Reporting in a Southern District of Nigeria. *Public Health, 124*(1), pp.24-27.

Onwuegbuzie, A. J., & Combs, J. P. (2011). "Data Analysis in Mixed Research: A primer," *International Journal of Education, 3*(1), pp.13.

Popejoy, L. L., Khalilia, M. A., Popescu, M., Galambos, C., Lyons, V., Rantz, M., & Stetzer, F. (2014). "Quantifying Care Coordination Using Natural Language Processing and Domain-specific Ontology," *Journal of the American Medical Informatics Association*, *22* (E1), pp. E93–E103.

Puttkammer, N., Baseman, J., Devine, E., Valles, J., Hyppolite, N., Garilus, F., & Yuhas, K. (2016). "An Assessment of Data Quality in a Multi-site Electronic Medical Record System in Haiti," *International Journal of Medical Informatics, 86*, pp.104-116.

Rahimi, A., Liaw, S.-T., Taggart, J., Ray, P., & Yu, H. (2014). "Validating an Ontology-based Algorithm to Identify Patients with Type 2 Diabetes Mellitus in Electronic Health Records," *International Journal of Medical Informatics, 83*(10), pp.768-778.

Ray, M. N., Houston, T. K., & Birmingham, A. (2005). "Data Quality in the Outpatient Setting: Impact on clinical decision support systems," *AMIA Annual Symposium Proceedings, 66*(44), pp.41-45.

Rousseau, M.-C., Conus, F., Li, J., Parent, M.-É., & El-Zein, M. (2014). "The Québec BCG Vaccination Registry (1956–1992): Assessing data quality and linkage with administrative health databases," *BMC Medical Informatics and Decision Making, 14*(1), p.2.

Sahoo, S. S., Lhatoo, S. D., Gupta, D. K., Cui, L., Zhao, M., Jayapandian, C., Zhang, G.-Q. (2014a). "Epilepsy and Seizure Ontology: Towards an epilepsy informatics infrastructure for clinical research and patient care," *Journal of the American Medical Informatics Association, 21*(1), pp.82-89.

Sahoo, S. S., Jayapandian, C., Garg, G., Kaffashi, F., Chung, S., Bozorgi, A., & Zhang, G.-Q. (2014b). "Heart Beats in the Cloud: Distributed analysis of electrophysiological 'Big Data' using cloud computing for epilepsy clinical research," *Journal of the American*

Medical Informatics Association, 21(2), pp.263-271.

Shaw, I., & Norton, M. (2008). "Kinds and Quality of Social Work Research," *British Journal of Social Work, 38*(5), pp.953-970.

Stark, P. (2010). "Congressional Intent for the HITECH Act," *The American Journal of Managed Care, 16*, pp.24-8.

Taggart, J., Liaw, S.-T., & Yu, H. (2015). "Structured Data Quality Reports to Improve EHR Data Quality," *International Journal of Medical Informatics, 84*(12), pp.1094-1098.

Thakkar, M., & Davis, D. C. (2006). "Risks, Barriers, and Benefits of EHR Systems: A comparative study based on size of hospital," *Perspectives in Health Information Management, 3*(5), pp.1-19.

Thiru, K., Hassey, A., & Sullivan, F. (2003). "Systematic Review of Scope and Quality of Electronic Patient Record Data in Primary Care," *BMJ, 326*(7398), p.1070.

Thomson Reuters (2016). "Journal Citation Reports," Retrieved from https://jcr.incites.thomsonreuters.com/ on 26 May, 2016.

Tu, K., Widdifield, J., Young, J., Oud, W., Ivers, N. M., Butt, D. A., & Jaakkimainen, L. (2015). "Are Family Physicians Comprehensively Using Electronic Medical Records such that the Data Can Be Used for Secondary Purposes? A Canadian perspective," *BMC Medical Informatics and Decision Making, 15*(1), pp.1.

UNSW (2013). "Excellence in Research for Australia (ERA) Outlet Ranking," Retrieved from https://research.unsw.edu.au/excellence-research-australia-era-outlet-ranking on 26 May, 2016.

van der Togt, R., Bakker, P. J., & Jaspers, M. W. (2011). "A Framework for Performance and Data Quality Assessment of Radio Frequency IDentification (RFID) Systems in Health Care Settings," *Journal of Biomedical Informatics, 44*(2), pp.372-383.

van Engen-Verheul, M. M., Peute, L. W. P., de Keizer, N. F., Peek, N., & Jaspers, M. W. M. (2016). "Optimizing the User Interface of a Data Entry Module for an Electronic Patient Record for Cardiac Rehabilitation: A mixed method usability approach," *International Journal of Medical Informatics, 87*, pp.15-26.

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, 20*(5), pp.940-946.

Wang, R. Y. (1998). "A Product Perspective on Total Data Quality Management," *Communications of the Association for Computing Machinery, 41*(2), pp.58-65.

Weiskopf, N. G., Hripcsak, G., Swaminathan, S., & Weng, C. (2013). "Defining and Measuring Completeness of Electronic Health Records for Secondary Use," *Journal of Biomedical Informatics, 46*(5), pp.830-836.

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, 20*(1), pp.144-151.

WHO (2016). "World Health Statistics 2015," Retrieved from http://www.who.int/gho/publications/world_health_statistics/2015/en/ on 12 May, 2016.

Wrightson, W. (2010). "A Comparison of Electronic and Handwritten Anaesthetic Records for Completeness of Information," *Anaesthesia and Intensive Care, 38*(6), pp.1052.

Wu, C. S., Khoury, I., & Shah, H. (2012). "Optimizing Medical Data Quality Based on Multiagent Web Service Framework," *IEEE Transactions on Information Technology in Biomedicine, 16*(4), pp.745-757.

Wu, T.-J., Schriml, L. M., Chen, Q.-R., Colbert, M., Crichton, D. J., Finney, R., & Meerzaman, D. (2015). "Generating a Focused View of Disease Ontology Cancer Terms for pan-cancer Data Integration and Analysis," *Database, 2015*, bav032.

Appendix

Appendix A - Data extraction form	
No.	
Title	
Author	
Publish year	
Source	
Healthcare records	
Definition of completeness	
Measure of completeness	
Study themes	
Limitations	

About the Authors

Ms.Caihua Liu is presently enrolled as doctoral student, in the School of Software, at University of Technology Sydney. Her research interest focuses on data completeness in electronic medical records.

Dr. Amir Talaei-Khoei is an Assistant Professor at the Ansari College of Business in the University of Nevada, Reno (UNR) and a visiting scholar at the University of Technology Sydney (UTS). Prior to joining UNR, Amir spent almost five years in Australia as a faculty member. His research mainly focuses on innovative use of technologies in healthcare settings. He has received his PhD in Information Systems from the University of New South Wales (UNSW), Australia. He also holds MSc of Information Technology from Royal Institute of Technology, Sweden. Prior to academia, Dr. Talaei-Khoei worked in industry in Europe.

Dr. Didar Zowghi is a Professor of Software Engineering at University of Technology Sydney. Professor Zowghi's core research focuses on improving the software development processes and the quality of their products. In particular, her research addresses problems and issues in Requirements Engineering and Business Analysis. She is an Associate Editor of IEEE Software, the regional editor of the Requirements Engineering Journal, and on the editorial board of IET Software journal. She has published over 180 articles in prestigious conferences and journals and has co-authored papers with 80 different researchers from 25 countries.

Dr. Jay Daniel is a Lecturer (Assistant Professor) in the School of Systems, Management and Leadership at the University of Technology Sydney. Previously with DB Schenker, Australia, and Alliance International Registrar, Asia Pacific, Jay held positions of Senior Management Consultant, Supply Chain Solution Analyst, Project Manager, Industry Trainer and Lead Auditor. He is actively engaged in teaching, research, and consulting on information systems and supply chain themes. His research is funded by government agencies and industry collaborators, focusing on Data Analytics, Information Systems, Sustainable Supply Chain and Decision Support Systems.