Rahimi, Alireza; Liaw, Siaw-teng; Ray, Pradeep; Taggart, Jane; Yu, Hairong

Article

Ontological specification of quality of chronic disease data in EHRs to support decision analytics: A realist review

Decision Analytics

Provided in Cooperation with:
Springer Nature

Suggested Citation: Rahimi, Alireza; Liaw, Siaw-teng; Ray, Pradeep; Taggart, Jane; Yu, Hairong (2014) : Ontological specification of quality of chronic disease data in EHRs to support decision analytics: A realist review, Decision Analytics, ISSN 2193-8636, Springer, Heidelberg, Vol. 1, Iss. 1, pp. 1-31, https://doi.org/10.1186/2193-8636-1-5

This Version is available at:
https://hdl.handle.net/10419/161812

Decision Analytics
a SpringerOpen Journal

RESEARCH Open Access

Ontological specification of quality of chronic disease data in EHRs to support decision analytics: a realist review

Alireza Rahimi[1,2,3], Siaw-Teng Liaw[1,3,4,6*], Pradeep Ray[5], Jane Taggart[1,3] and Hairong Yu[3]

* Correspondence: siaw@unsw.edu.au
[1]UNSW School of Public Health & Community Medicine, Sydney, Australia
[3]UNSW Centre for Primary Health Care & Equity, Sydney, Australia
Full list of author information is available at the end of the article

Abstract

This systematic review examined the current state of conceptualization and specification of data quality and the role of ontology based approaches to develop data quality based on "fitness for purpose" within the health context. A literature review was conducted of all English language studies, from January 2000-March 2013, which addressed data/information quality, fitness for purpose of data, used and implemented ontology-based approaches. Included papers were critically appraised with a "context-mechanism-impacts/outcomes" overlay. We screened 315 papers, excluded 36 duplicates, 182 on abstract review and 46 on full-text review; leaving 52 papers for critical appraisal. Six papers conceptualized data quality within the "fitness for purpose" definition. While most agree with a multidimensional definition of DQ, there is little consensus on a conceptual framework. We found no reports of systematic and comprehensive ontological approaches to DQ based on fitness for purpose or use. However, 16 papers used ontology-specified implementations in DQ improvement, with most of them focusing on some dimensions of DQ such as completeness, accuracy, correctness, consistency and timeliness. The majority of papers described the processes of the development of DQ in various information systems. There were few evaluative studies, including any comparing ontological with non-ontological approaches, on the assessment of clinical data quality and the performance of the application.

Keywords: Data quality; Fitness for purpose; Data model; Ontology development methodology

Background

The growing use of electronic health records (EHRs) raises issues of semantic interoperability and the quality management/improvement of large datasets derived from multiple EHRs. Improved data quality in EHRs can improve the quality of decisions and lead to better policy that actually meet needs, strategies, evidence-based care and patient outcomes.

The acceptable level of data quality is not fixed in the system. Rather health professionals can provide it at different times and data users need to assess that quality contextually, based on the fitness for research, audit and quality assurance purposes (Devillers et al. 2007). It is important to take a user view point of quality because it is the end user who evaluate whether or not data is fit for use. A focus is the quality of patient or disease registers derived from EHRs to support policy and

practice. Patients registers need to have a level of completeness and the information contained, need a level of correctness and consistency to be useful for clinical, quality improvement and research purposes (Liaw et al. 2011).

DQ was conceptualised in terms of its "fitness for purpose/use" in a few papers (Wang 1998; Wang et al. 1996). DQ can be described from two perspectives: (1) intrinsic quality of data elements and set of data elements (data set) and (2) how the set meets the user's needs i.e. fitness for purpose. The commonly approved definition of DQ has been epitomized in the International Standards Organisation definition: "*the totality of features and characteristics of an entity that bears on its ability to satisfy stated and implied needs*" (ISO 8402-1986, Quality Vocabulary). DQ also can be specified in terms of its "fitness for purpose/use" (Wang 1998; Wang et al. 1996).

Intrinsic DQ refers to the extent that data is free of defects as measured by specific DQ dimensions, including "accuracy, perfection, freshness and uniformity" (Redman 2005) and "completeness, unambiguity, meaningless and correctness" (Choquet et al. 2010; Orme et al. 2007; Wand and Wang 1996; Yao et al. 2005). The Canadian Institute for Health Information recommendations were the basis for an information quality framework comprising 69 quality criteria grouped into 24 quality characteristics, which was further grouped into 6 quality dimensions: accuracy, timeliness, comparability, usability, relevance and privacy & security (Kerr et al. 2007). Research in DQ has tended to focus on the identification of generic quality characteristics such as accuracy, currency and completeness (Orme et al. 2007; Wang et al. 1996) or completeness, correctness, consistency and timeliness (Liaw et al. 2011) as core dimensions of DQ that are relevant across application domains. However, our pervious review shows there is a lack of consensus conceptual framework and definition for DQ (Liaw et al. 2013).

Many studies regularly report a range of deficiencies in the collected information for professional practice (Devillers et al. 2007; Kahn et al. 2002), clinical (Azaouagh and Stausberg 2008; de Lusignan et al. 2010; Mitchell and Westerduin 2008; Moro and Morsillo 2004) and health promotion (Gillies 2000b) purposes. Similar deficiencies exist with information in geographic (Devillers et al. 2007; Ivanova et al. 2013), hospital and general practice (Liaw et al. 2012) information systems, where the lack of coding rules meant that much of the data are often incomplete or in relatively inaccessible text format. The evidence is more encouraging for data for administrative purposes (Lain et al. 2008; Quan et al. 2008). Hybrid record keeping systems in primary care are believed to be more complete than computer-only or paper-only systems (Hamilton et al. 2003).

Relational database models have been prevalent in last few decades, enabling information to be efficiently stored and required within a hierarchical database architecture. On the other hand, ontologies, usually with non-hierarchical databases, have been used in applications that required more flexibility in capturing more semantic meanings. However, there is no well-documented evidence or experiments that suggest that one is better than the other in terms of outputs, data quality and fitness for purpose.

In contrast to our previous review (Liaw et al. 2013), this systematic review will examine the breadth and depth of research into the conceptualization of data quality based on the "fitness for purpose" paradigm, methodologies to specify data quality for implementation, some advantages of ontology-based approaches to develop data quality,

and semantic interoperability. This study aims to examine the role of ontology-based approaches to develop data quality based on "fitness for purpose" whereas the previous review focused on data quality as a general concept in health context. This study was broader in the databases searched and the search terms and produced results built on the previous literature review (Liaw et al. 2013) to address the following questions:

1. How is data quality being conceptualized within the "fitness for purpose" definition for a range of uses?
2. What specification methodologies are being used to specify data quality for implementation?
3. What ontology-specified implementations are being used and how do they compare with other methods? and
4. How is the impact of implementing ontology-based specifications for data quality in chronic disease management being measured and evaluated?

Methods

A literature review was conducted of all English language studies, from January 2000-March 2013, which addressed data/information quality, fitness for purpose, used ontology-based approaches and involved healthcare/chronic disease. Inclusion criteria were: (a) conceptualises data quality based on "fitness for purpose"; (b) formal methodologies used to specify data quality for implementation; (c) involved some form of data models and ontologies to improve quality of clinical data in EHRs and patient registers; and (d) used data models and ontology-based approaches in CDM. These papers were screened by title and abstract content for inclusion. The references of the included papers were hand-searched for other eligible papers.

Included papers were critically appraised with a "context-mechanism-impacts/outcomes" framework. Appraised papers were summarized using specifically developed templates and discussed to achieve the final consensus on how it addressed the review questions. The conceptual framework developed for the literature review included:

- **Context:** integrated CDM, evidence based practice, evidence-based policy patient or disease registers, "decision analytics";
- **Mechanisms:** methods to assess and manage quality of the register/EHR and data quality based on "fitness for purpose", ontology-based approaches;
- **Impacts/outcomes:** Measurable impacts outcomes based on improved quality of the register, data quality, "fitness for purpose", "decision analytics".

The search strategy and keywords were organised around the three broad concepts:

- Context: Diseases (chronic diseases, chronic illnesses, chronic disease management, chronic illness management, electronic health records (EHRs), registers);
- Mechanisms: Data models and ontology (ontological based models, ontology approaches, ontology based multi agent systems (OBMAS), and ontological framework);
- Impacts: Data Quality (data quality, information quality, data quality management, data quality assessment, quality of register, fitness for purpose).

The search was repeated three times with the following phrases:

(data quality OR information quality) AND ("fitness for purpose" OR "fitness for use") AND (quality of register* OR quality of electronic health records) AND (decision analytics) in Title, Abstract or Keywords, Subject or MESH

(ontology OR data model*) in Title, Abstract or Keywords, Subject or MESH AND (data quality OR information quality OR quality of register) in Title, Abstract or Keywords, Subject or MESH AND (fitness for purpose OR fitness for use) AND (decision analytics) in Title, Abstract or Keywords, Subject or MESH

((ontology AND traditional data model*) in Title, Abstract or Keywords, Subject or MESH OR (ontology AND SQL) in Title, Abstract or Keywords, Subject or MESH) AND (chronic diseases OR chronic illnesses) in Title, Abstract or Keywords, Subject or MESH AND (data quality OR information quality OR quality of register) in Title, Abstract or Keywords, Subject or MESH.

The initial screening of the articles was based on their abstracts. AR read all abstracts independently and studies without electronic abstracts were excluded. Selection of relevant articles was based on the information obtained from the abstracts and was agreed upon in discussion with co-authors. In the case of differences, the original paper was obtained and agreement was achieved after it was read. We hand-searched the references of the included papers to ensure completeness of the search. Papers that satisfied the inclusion criteria were independently examined by authors and any disagreements resolved by consensus. AR appraised all 52 papers using the realist "context-mechanism-impacts/outcomes" approach using extraction template (see Additional file 1: Figure S1).

The template kept the extracted information consistent and focused on the analysis and synthesis of the literature review by study types, methods, tools, outputs and impacts in terms of: requirements analysis, design and tools development, implementation, deployment and testing, evaluation: descriptive evaluation, comparative and/or contemporary control. The quality appraisal uses traditional methods of critical appraisal for validity (internal and external), reliability, generalizability and relevance of the research methods, tools and measurements. We also classified a paper as having addressed "fitness for purpose" if it a) defined a purpose for the project or dataset and b) assessed whether the data or dataset was fit for the specified purpose.

Results

The main medical, computer and business sciences online databases were searched: MEDLINE (67 papers), the Cochrane Library (18 papers), ISI Web of Knowledge (35 papers), Science Direct (75 papers), Scopus (76 papers), IEEE Xplore (25 papers), and Springer (19 papers). All search strategies have been expanded in the following business databases consisting of (Emerald Fulltext, Business Source Premier, Biotechnology and Bioengineering Abstracts, British Humanities Index: BHI, Proquest Asian Business and Reference) to find more business analytics papers however the result demonstrated insufficient studies and no more paper in this area. Table 1 summarised the sources of the 315 papers found.

In the first iteration, searches using a combination of keywords and controlled vocabulary term searches (specifically in Titles and Subjects fields of all papers) were

Table 1 Online databases used and papers found

Database	Subjects	Field	Document type	# papers
Pubmed	Medicine, Health Science, Medical Informatics and Bioinformatics	Title, Mesh and Abstract	Journal articles and Proceeding	**67**
Cochrane Central Databases	Medicine and Health Science	Title, Mesh and Abstract	Journal articles	**18**
ISI Web of Sciences	Computer Science, Information Technology, Medical Informatics, Bioinformatics and Health Science	Title, Subject and Abstract	Journal articles	**35**
ScienceDirect	Computer Science, Medical Informatics, Engineering,	All fields	Journal articles	**75**
	Decision Science, Engineering, Mathematics, Psychology, Social Sciences, and Medicine			
Scopus	Computer Science, Health Science, Medical Informatics, Bioinformatics, Information Technology, Psychology,	Al fields	Journal articles	**76**
	Social and Behavioural Sciences			
IEEE Xplore	Computing and Processing, Medical Informatics, Bioinformatics, Communication Networking and Cybernetics	Title, Subject and Abstract	Journal articles	**25**
SpringerLink	Computer Science, Medical Informatics, Bioinformatics, information science and Engineering	Title, Subject and Abstract	Journal articles	**19**
Business data bases	Emerald Full text, Business Source Premier, Biotechnology and Bioengineering Abstracts, British Humanities Index: BHI, Proquest Asian Business and Reference	Title, Subject and Abstract	Journal articles	**0**
Total				**315**

conducted. The application of Titles and Subjects fields in a user's search strategy and search limitation in each database has been shown to increase relevance, precision and recall (McJunkin 1995). We screened 315 papers, excluded 36 duplicates, 182 on abstract review and 46 on full-text review; leaving 52 papers for critical appraisal. Of these 6 papers conceptualized data quality within the fitness for purpose definition for a range of uses, 16 used a defined process to specify data quality for implementation, 2 papers used the ontology-specified implementation in DQ improvement compare with other non-ontological approaches, and 28 demonstrated how the impact of implementing ontology-based specifications for data quality in chronic disease management is being measured and evaluated.

It can be seen from the results of the field of publications in Table 1 that 85 papers (26.98%) in the medicine and health areas, 44 papers (13.97%) in computer and IT sciences and also 186 papers (59.05%) in the multi-disciplinary areas which is significantly more than the other two groups.

Figure 1 shows how other eligible papers were included in the second iteration using hand-searching process. The references were retrieved from the papers included in the first iteration. The keywords of references that matched with the search keywords were chosen. Based on their title, keywords, abstract and full text, 7 papers were included from the hand-searching.

It can be seen from the data in Table 2 that most of the papers (54%) show the various roles and impacts of ontology based approaches in CDM and how those approaches can be evaluated.

Table 3 presents the analysis of papers by study type and how they contributed to the review questions. The majority (83%) of studies involved design and tools development;

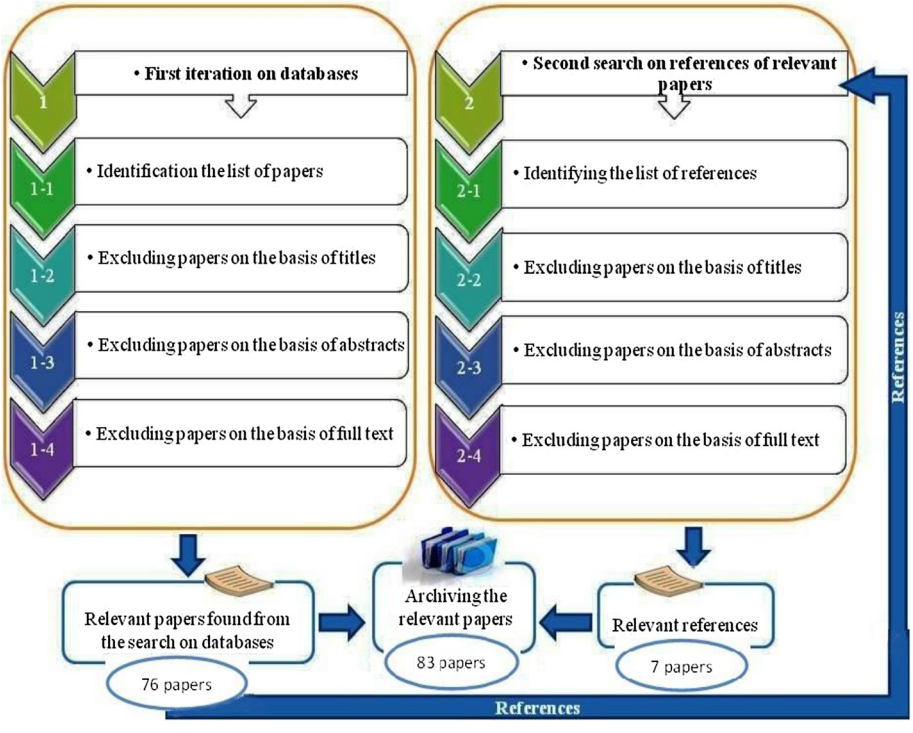

Figure 1 Paper selection process.

Table 2 Distribution of papers by review questions

Review questions	Number	%
1. How is data quality being conceptualized within the "fitness for purpose" definition for a range of uses?	6	12%
2. What specification methodologies are being used to specify data quality for implementation?	16	31%
3. What ontology-specified implementations are being used and how do they compare with other methods?	2	4%
4. How is the impact of implementing ontology-based specifications for data quality in chronic disease management being measured and evaluated?	28	54%

Note: Total papers >52 because each paper may be classified as two or more study types, or may address two or more review questions.

38% implemented/deployed and tested implementations; and 20% conducted a descriptive evaluation. A considerable number of studies (42 papers) demonstrate that the ontological approach was used to address semantic interoperability, data linkage, data integration, remote patient monitoring and reduce complexity of information models and networks. The majority of ontology-specified implementations in this category did not compare the performances and processes between ontology and non-ontology approaches. There were few attempts to conceptualize data quality based on "fitness for purpose" definition in a range of uses and purposes.

Figure 2 shows an increase in papers on ontology in CDM and DQ from 2006. There is an increase in studies reporting on the use of "fitness for purpose" when dealing with data quality from 2010 (probably started with the small spike in 2007). This suggests that researchers may be starting to take a more realistic approach to the quality of "big data": the intrinsic data quality is important but it does not need to be prefect to be "fit for purpose".

Figure 3 gives a breakdown of the frequency of the studies conducted in different continents 2006 based on the setting of the studies. Europe is the most profile with 42.6% of the authors affiliated with European universities and institutions. North America is next with 21.3% of the studies followed by Oceania (18%), Asia (13.1%), South America (3.3%) and Africa with 1.7%. Although a paper being affiliated to a particular university in a country does not necessarily mean that the context under study has been in the same country or even continent, it might provide insights to a limited extent. For example, data quality research and ontological frameworks proposed seem to be much higher in the

Table 3 Distribution of papers by study types and review questions

Study type	Study type		Review questions							
			Q1		Q2		Q3		Q4	
	n	%	n	%	n	%	n	%	n	%
1. Formal requirements analysis e.g. literature reviews, qualitative research	29	34%	4	5%	10	12%	9	11%	36	43%
2. Design & tools development: including data/information models & ontologies	69	83%	4	5%	4	5%	13	15.5%	41	49%
3. Implementation, deployment and testing of information systems	32	38%	2	2.5%	3	3.5%	5	6%	22	26.5%
4. Evaluation: descriptive evaluation of DQ or ontology in health area	17	20%	1	1%	2	2.5%	2	2.5%	12	14.5%
5. Evaluation: comparative +/− contemporary control (e.g. RCT)	2	2.5%	0	0	0	0	1	1%	1	1%

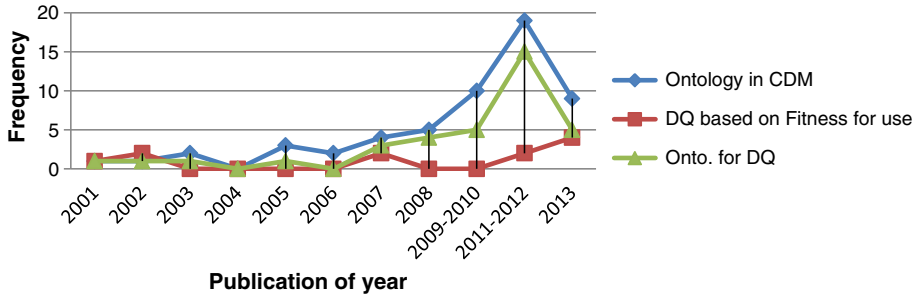

Figure 2 Distribution of papers from each category by year.

European countries. That might be because of a greater concern with DQ and/or ontologies in Europe. North America, Oceania and Asia stand in the second, third and fourth spot after Europe in terms of the number of studies that have been conducted. South America and Africa have a relatively lower rate of papers than the other continents, which is consistent with the general trends. The distribution of papers by continent might suggest that the topic has grabbed the attention of academics as well as health professionals as a major concern for patients registers.

The drivers of ontological approaches for DQ and/or CDM include better software for: (1) quality of care and/or health care issues and (2) the description, assessment and management of DQ in health (e.g. role of clinical guidelines in DQ, effects of quality of information in CISs and networking, defining and describing various attributes of DQ) as well as individual dimensions of DQ (e.g. accuracy, completeness, correctness, and consistency).

Conceptualization of data quality within the "fitness for purpose" paradigm

Table 4 shows a few studies have conceptualized and implemented data quality based on the "fitness for purpose" definition in their data models for a range of uses in health and non-health areas including improved searches for spatial data resources, including in languages other than English (Ivanova et al. 2013), support expert users in the assessment of the fitness for purpose of a given dataset (Devillers et al. 2007), better decision making (Chen 2009), support analyses in comparative effectiveness research (Kahn et al. 2012), support agents to choose how much information to gather (Chen 2009), and for research and clinical purposes (Liaw et al. 2011).

Many studies regularly report a range of deficiencies in the collected information for professionals requirements (Devillers et al. 2007; Kahn et al. 2002), clinical

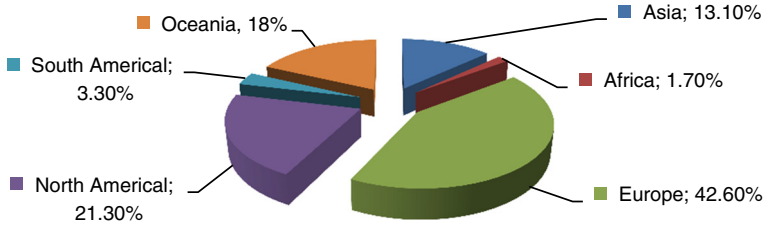

Figure 3 Distribution of papers found by continent.

Table 4 Papers where data quality was conceptualized within fitness for purpose paradigm

Author reference	Context	Aims of project	Methods/tools used in project	Results
(Ivanova et al. 2013)	Geo-spatial datasets in the national geo-information repositories in Netherlands	To suggest a system for guided search for spatial data resources called GUESS	-Use of popular search engines like OpenSearch to help in assessing fitness for purpose	Defined fitness for purpose of data based on users (experts and non-experts in geo-informatics) satisfaction from search results
			-Use metadata (information that helps users to assess the usefulness of a dataset relative to their problem) as a tool to evaluate fitness for purpose of datasets	
			-Their approach is based on a 3-part data model (user profile, spatial data profiles and interaction profiles)	Allowed users without specific expertise to conduct free form search requests in their own language
			-Theoretical discussion on accuracy and completeness of data	
(Devillers et al. 2007)	Spatial On-Line Analytical Processing (SOLAP) as a GIS data repository	To manage heterogeneous data quality and provide functions to support expert users in the assessment of the fitness for purpose of a given dataset	-Use the Quality Information Management Model = QIMM	Defined fitness for purpose as the closeness of the agreement between data characteristics and the explicit and/or implicit needs of a user for a given application in a given area
			-Focus on intrinsic data quality indicators such as completeness, correctness and accuracy underpins a prototype	
			-Apply data quality analysis tool which is the Multidimensional User Manual (MUM) prototype	Researchers attempt to provide data quality indicators to help users determine a dataset's fitness for purpose and better assess the fitness of data based on quality indicators/experts in GIS
			-Validate the QUMM of through demonstrations of the prototype to different users (GIS scientists, specialists in data quality issues, consultants in GIS, data producers, governmental agencies, typical GIS users, etc.)	
(Kahn et al. 2012)	Clinical dataset in US	To develop the efficacy of their data model in three large healthcare organizations	-Use a two-by-two conceptual model (PSP/IQ) for describing IQ	This is a well-grounded, logical approach and a case study to indicate health organizations

Table 4 Papers where data quality was conceptualized within fitness for purpose paradigm *(Continued)*

			-Focus on 8 dimensions of data quality (completeness, correctness, flexibility, etc.)	need to use "fitness of use" to determine IQ (specifically soundness, dependable, useful and usable information) for analytical purposes
			-Surveyed 45 professionals to determine which IQ dimensions belong in each quadrant of the model -Use case study method in 3 healthcare organizations that 75 people in each organization completed a 70-item questionnaire for assessing the quality of their patients information on the IQ dimensions	This assessment of DQ provides a reasonable baseline for determining what improvements should be made in DQ based on fitness for purpose for analytical purposes
(Chen 2009)	Infectious diseases dataset in US	To investigate the effect of 'quality' of information and 'amount' of information are used in the health behaviour	-Use mathematical modelling of infectious disease transmission, seeks to analyse how the amount of information about disease prevalence affects individuals' incentives	Demonstrated "fitness for purpose" of data for agents to choose how much information to gather from others (personal communication from an anonymous reviewer)
			-More focus on data timeliness -Use of mathematics software	This is a theoretical paper using several mathematical models to show that information quality affects health behaviour i.e. better information leads to better decision making
(Liaw et al. 2011)	An electronic Practice Based Research Network (ePBRN) with a data repository of routinely data from multiple EHRs	To develop a matrix for assessment and management the quality of data	Their methods include 3 phases: (1) requirements specification based on the conceptual framework, (2) design and establishment of the ePBRN, and	They used a well-designed framework to describe the intrinsic DQ (correctness and consistency) and fitness for purpose (completeness) for research and clinical purposes

Table 4 Papers where data quality was conceptualized within fitness for purpose paradigm *(Continued)*

			(3) evaluation of the data quality and fitness for research.	
			-Use Microsoft Structured Query Language (SQL) to manage the extracted data and SAS used for datacleansing and analysis	This study raised the theoretical dependence of the SQL/SAS approach on the lack of a transparent and explicit data model, metadata and process within proprietary EHRs
			-Focus on correctness, completeness and consistency of clinical data	
(Hamilton et al. 2003)	Eighteen general practices in the Exeter Primary Care Trust in UK	To compare computer-only record keeping to paper-only and hybrid systems	-Use case control study of cancer patients aged over 40 years	Defined completeness as fitness for consultation in primary care
			-Classify records as paper, computer, or hybrid, depending on which medium stored the clinical information from consultations by descriptive statistics	Hybrid systems of primary care record keeping document higher numbers of consultations than computer-only or paper-only systems
			-Focus on completeness of data	

(Azaouagh and Stausberg 2008; de Lusignan et al. 2010; Mitchell and Westerduin 2008; Moro and Morsillo 2004) and health promotion (Gillies 2000b) purposes. Similar deficiencies exist with information data in geographic (Devillers et al. 2007; Ivanova et al. 2013), hospital and general practice information systems (Liaw et al. 2012), where the lack of coding rules meant that much of the data are often incomplete or in relatively inaccessible text format. The evidence is more encouraging for data for administrative purposes (Lain et al. 2008; Quan et al. 2008). Hybrid record keeping systems in primary care are believed to be more complete than computer-only or paper-only systems (Hamilton et al. 2003).

Methodologies to specify data quality for implementation

Table 5 shows that the majority of studies (81%) reported the design and development of tools to specify data quality for implementation; requirements analysis e.g. literature reviews and qualitative research methodologies (75%); system implementation, deployment and testing of information systems (25%), and descriptive evaluation (12%). There were no outcomes or comparative evaluation of the methodologies used.

Various qualitative methods such as interview and reports analysis, usually interpreted using grounded theory have been implemented to evaluate usability (Kerr et al. 2007), privacy (Stvilia et al. 2009), comparability (Kerr et al. 2007) and relevance (Kerr et al. 2007). Consistency (Chen et al. 2009) of data has been assessed with concept mapping in non-health contexts. Timeliness (currency) (Huaman et al. 2009; Kerr et al. 2007), accuracy (precision) (Stvilia et al. 2009), reliability (Britt et al. 2007), representativeness (Britt et al. 2007), correctness (Gillies 2000a) and completeness (Kiragga et al. 2011) were assessed with quantitative statistical methods.

Ontology-specified implementation to develop data quality and compare with other models

Table 6 shows two papers found that used ontological and non-ontological approaches to DQ in clinical information systems (CIS). Both papers suggested that ontology-based models had more advantages than other data models in the health domain. For example, Mabotuwana and Warren (2009) showed the ontology driven approach to determining patients who needed a follow-up in hypertension management provided more advantages than SQL. They listed the limitations of the traditional SQL-based approach as i) lack of abstract, domain-level query support; ii) lack of the notion of a hierarchy and iii) nature of temporal SQL queries (Mabotuwana and Warren 2009). They used SWRL rules which allow user to write rules to reason about individuals and to infer new knowledge about these individuals. The ontology based approach was sufficiently flexible to enable new audit criteria to be easily added as required, easy visualization of the knowledge base and standardized ways of querying the knowledge based. However, the paper was not explicit about whether was a formal outcome-based comparison of ontological and non-ontological approaches was conducted.

Maragoudakis et al. (2008) developed an ontology with 5 domains for a clinical Decision Support System (CDSS) for management of Chronic Obstruction Pulmonary Disease (COPD). The ontology, based on hierarchical Bayesian networks, encoded a domain (COPD) and compared the predictive accuracy of this ontology-based hierarchical

Table 5 Methodologies used to specify data quality for implementation

Study types	1	2	3	4	5	Summary and results of methodologies	Contexts
Reference							
(Gillies 2000a)		√				Represent a tool to assist with continuous improvement of the use of information systems in general practice based on their requirements which is accurate information	Health information
						Shows how the model can be practically used to improving the use of coding (external consistency of data) and accurate information (data correctness) within a general practice in a systematic way	
(Kahn et al. 2012)	√	√				This is a well-grounded, logical approach and a case study to indicate health organizations need sound, dependable, useful and usable information for analytical purposes.	Clinical data
						However, there is need to some details of their participants, sampling and why focus on only 16 dimensions of Information Quality (IQ).	
						This approach could be applicable way for the assessment of DQ in CDM because such an assessment provides a reasonable baseline for determining what improvements should be made in DQ based on fitness for purpose for analytical purposes	
(Liaw et al. 2011)	√	√	√			They used a well-designed framework to describe the intrinsic DQ (correctness and consistency) and fitness for purpose (completeness) for research and clinical purposes	Clinical data
						However, this study raised the theoretical dependence of the SQL/SAS approach on the lack of a transparent and explicit data model, metadata and process within proprietary EHRs	
(Arts et al. 2003)	√	√				Their approach demonstrates that after physicians' training, completeness, correctness and adherence to data definitions increased in ICUs significantly	Clinical data
(Arts et al. 2002b)	√	√				Demonstrate a list of procedures for high data quality assurance in medical registry based on causes of insufficient data quality	Health information
(Arts et al. 2002a)	√	√	√			Show that the overall DQ of medical registries has good quality (focusing on accuracy and completeness) and also explain their positive results as compared with earlier reports from the literature.	Clinical data
						However, they did not compare data quality before and after the implementation of procedures to improve the accuracy of data	

Table 5 Methodologies used to specify data quality for implementation *(Continued)*

(Stvilia et al. 2009)	√	√	Use a mixed methodology with multiple data sources: 1. The analysis of 150 Web pages and related web sites identified the major approaches the providers use to define their	Health web pages
			IQ criteria set: a. centrally defined, b. community constructed, and c. outsourced to third-party raters. 2. The researchers surveyed a convenience sample of 108 healthcare information consumers to gain better	
			insight into the health IQ evaluation behaviour of consumers. 3. Semi structured in-depth interviews with a sample of 20 survey participants	
			Use a sample of the IPL's Q&A communication archives to identify the healthcare IQ criteria used by consumers and information intermediaries	
			Results show that consumers may lack the motivation or literacy skills to evaluate the information quality of health	
(Kahn et al. 2002)	√		Developing a two-by-two conceptual model for describing IQ (PSP/IQ)	Health information
			Mapping the 16 IQ dimensions into their model	
			Survey 45 professionals to determine which IQ dimensions belong in each quadrant of the model	
			Case study in 3 healthcare organizations that 75 people in each organization completed a 70-item questionnaire (a 10-point Likert scale) for assessing the quality of their patients information on	
			Provide a reasonable baseline for determining what improvements should be made in DQ (soundness, dependable useful and usable information) based on fitness for purpose for professionals analytical purposes.	
			Demonstrating the efficacy of the PSP/IQ model in three large healthcare organizations	
(Britt et al. 2007)	√	√	Use statistical methods to manage data quality using SAS as a computer program in statistical package	Clinical data
			Measure representativeness, reliability, validity and accuracy of BEACH data eg. Reliability of coding of reasons for encounters and issues validity of ICPC to categorizing data. Accuracy of problem labels recorded by GPs (About 1000 GPs participate yearly)	
(Chen 2009)	√	√	Focus on a full mathematical analysis (mathematical software)	Infectious diseases

Table 5 Methodologies used to specify data quality for implementation *(Continued)*

				Investigate the effect of quality of information and amount of information are used interchangeably in the health behaviour e.g. decision making	
(Choquet et al. 2010)		√		Use Talend Open Studio open source software as well as developed stored procedures in SQL for the object quality criteria	Hospital dataset
				Use the 6 HL7 information models for modelizing their domain	
				Apply the TDQM 4 steps approach to score quality of each vertex of IQT	
				Use two consensual resources to standardize the EHR vocabulary, include: 1) ATC: The WHO drugs and substances international classification and 2) NEWT: organisms taxonomy database	
				Propose methods and measures to assess data quality (focus on data accuracy)	
				Propose 3 dimensions to classify the quality measures proposed (objects, concepts, and terms) as vertexes of their model Information Quality Triangle = IQT)	
				Measure the distance between standardized information models and reference terminologies against its CIS	
				Allow building pertinent and coherent monitoring trends	
				Present that controlled vocabularies are a necessity to share data	
(Cunningham-Myrie et al. 2008)			√	Use ICD-10 for coding various collected data and to facilitate comparability of standardized data	Health information
				Use Two broad categories of information were sought: a) epidemiological data and b) health service utilization data	
				Show that data management systems in hospitals were not linked to facilitate generation of cost-effectiveness estimates and other information required to compare options for health investment	
				Show methodological way for improvement health information quality for the economic analysis	
(Huaman et al. 2009)	√	√		Timeliness and data quality were assessed by calculating the percentage of reports sent on time and percentage of errors per total number of reports, respectively	Infectious disease surveillance
				Use training program: 12 week prospective study with training program for reporting personnel.	

Table 5 Methodologies used to specify data quality for implementation *(Continued)*

				Randomised selection to phone, visit or control for their supervisions	
				The training improved report timeliness but did not have such impact on data quality.	
(Kiragga et al. 2011)	√	√	√	Use the Research Cohort database as the reference "gold standard" for the assessment of data accuracy	Infectious diseases
				Use statistical test e.g.: Categorical variables were compared using Chi-square test, the Mann–Whitney test was used for the continuous variables	
				Compare 2 databases, one from a clinic and one from a research team to assess the quality of data (completeness and accuracy)	
				Results show that there is a high rate of underreporting of OIs in a routine HIV clinic database and demonstrate high rates differences between clinic and research databases	
				Their findings have important implications for the use and interpretation of data derived from routine HIV observational databases for research and audit, and they highlight the need for ongoing regular validation of key data items in these databases	
(Lima et al. 2010)		√	√	Use a decision support example around a hypothetical patient called John who experiences an exacerbation of his COPD	Clinical Guidelines (CG) for COPD
				Use the Clinical Guideline for COPD that there are 16 criteria that suggest the patient should be admitted and the model takes into account answers to each criterion	
				Present a model for the prediction and evaluation of quality of information to a multi criteria decision making process	
				Model describes a decision support tool for use in the management of COPD	

Notes for study types: See Table 2 for legend.

Table 6 Studies that compared ontologies and other data models in specification and implementation

References	Research findings	Results of ontology implemented for data quality	Compare with non-ontology	Context
(Maragoudakis et al. 2008)	A tool in hierarchical Bayesian networks which can encode a domain and make prediction	Data mining classification No DQ	By using precision and recall metrics, show ontology approach is more accurate than Linear Programming in the monitoring of patients	COPD
(Mabotuwana and Warren 2009)	Enhance and facilitate temporal querying requirements in general practice medicine	Facilitate temporal querying requirements	Represent only some limitations of traditional SQL-based approach to show flexibility of ontology in easily add any requirement in ontology queries	CVD (hypertension)

Bayesian network method with linear programming and artificial neural network methods (Maragoudakis et al. 2008).

By using 10-fold cross validation and precision and recall metrics, they concluded that the Hierarchical Bayesian method is comparable to Artificial Neural Network (ANN) and far more accurate than linear programming approaches. In addition, their ontology can be easily updated with new elements, while using ANN to do this would be a painstaking laborious process. The most important advantage of such an approach, however, is the ability to shift this model to other domains, incorporating new mobile network appliances - such as GPS - and new hospitals and other health institutes, in an attempt to effectively monitor a patient in different locations.

The impact of ontologies for data quality in CDM and their evaluation

As Table 7 shows, a considerable amount of studies in this category have been published on the application of ontologies in both health and non-health areas. However, they do not compare ontologies with other data models. Studies to demonstrate the impact of ontology-based implementations included clinical decision support systems (Brüggemann and Grüning 2009; Min et al. 2009; Topalis et al. 2011) for information management (O'Donoghue et al. 2009; Young et al. 2009), diagnosis (Nimmagadda et al. 2008), clinical data analysis and management (Li and Ko 2007). A few studies examined ontology-based approaches to support data consistency (Esposito 2008a) and accuracy. However, we found no reports on any systematic and comprehensive ontological approaches to DQ issues or evaluation in the various contexts.

The application of ontological approaches to data quality management addressed the following issues: data quality problems and errors (Brüggemann and Grüning 2009), data heterogeneity problem (Min et al. 2009), semantic decision making (Lee et al. 2009), efficient services (Li and Ko 2007), procedures concerning the acquisition of data (Nimmagadda et al. 2008), classification and identification of specific patients types (Lee et al. 2009; Wang et al. 2007), data collection, data sharing and data integration (Min et al. 2009; O'Donoghue et al. 2009; Perez-Rey et al. 2006; Young et al. 2009). There were no studies that examined efficiency or effectiveness of ontology-based models in DQ management.

As Table 8 represents, the second application is the use of domain ontologies for the assessment of data quality in the querying requirements (Mabotuwana and Warren 2009), extracting knowledge from natural language documents (Valencia-Garcia et al. 2008), and data expression (Preece et al. 2008). The majority of these studies used precision and recall as metrics to assess the accuracy and validate the ontological approaches (Brank et al. 2005; Brewster et al. 2004; Euzenat 2007; Gangemi et al. 2006; Li 2010; Min et al. 2009; Pathak et al. 2012a, 2012b; Spasic and Ananiadou, 2005; Stvilia et al. 2009; Valencia-Garcia et al. 2008; Wang et al. 2007).

Despite a growing body of literature on ontology-based approaches in assessing the accuracy of the retrieval of clinical data, none of them have attempted to compare the performance between ontology-based and other (non-ontological) approaches. Most studies have used precision and sensitivity (recall) to assess the accuracy of ontology-based approaches in health domains (Brewster et al. 2004; Euzenat 2007; Gangemi et al. 2006; McGarry et al. 2007; Min et al. 2009; Pathak et al. 2012a,

Table 7 The impact of implemented ontologies for the management of data quality

Ontology functions	References	Defined purpose	Assessed of fitness for purpose using DQ and findings	Context
Management (*9 papers*)	(Li and Ko 2007)	To develop automated ontology approach to manage nutrients in a diabetes diet care knowledge management	-Used expert opinions to decide which are the important nutrients to include in the diabetes diet and therefore the ontology -This is face validity and consistency of the data -Authors suggested that there is a further step using ontology approach for more efficient diet knowledge management	Diabetes diet care in Taiwan
	(Esposito 2008a)	To detect abnormalities and malformations due to heart diseases	-Use as an ontology approach and rules to perform the instance and consistency checking and verifies that patient information violates the normal cardiovascular model loaded based on the SNOMED vocabulary -Theoretical discussion on data consistency -Researchers show applicability of ontology to define either the anatomy of the cardiovascular system in normal patients or the anatomy characterized by malformations or abnormalities in CHD patients to support cardiologist in the identification of diseases	Congenital Heart Disease (CHD) dataset in Italy
	(Nimmagadda et al. 2008)	To provide a solution to problems around handling increasing amounts of clinical information and solves some issues related to managing large	-Simulate human body disorders into metadata through ontology based data warehouse modelling -Theoretical discussion on managing accuracy and correctness of data -Authors states ontology can facilitate logic processes and semantics for data quality management and decision support for health care providers and clinicians	Human body anatomy and pathology dataset in Australia
	(Min et al. 2009)	To collect/retrieve information intelligently and address the semantic heterogeneity problem from the integration of data from multiple information resources	-Apply ontology mapped with medical thesaurus to integrate and retrieve the data from two independent database systems -Theoretical discussion about data consistency -Authors state that ontology can solve the semantic heterogeneity problem from the integration of two databases by recognition of inconsistence data	3000 records registered for the prostate cancer patients and Tumour Registry in US

Table 7 The impact of implemented ontologies for the management of data quality *(Continued)*

(Brüggemann and Grüning 2009)	To improve the outcome of data quality management (DQM)	-Use an algorithm and data model for consistency checking, an algorithm for detecting duplicates and give three examples of DQM-specific metadata tasks (data provenance, data quality annotations at schema and instance level and an ontology for the DQM domain) -Authors mentioned the usefulness of their ontology approach to define a shared vocabulary for improved interoperability, and performing DQM include consistency checking, data duplicate detecting and metadata management	Cancer registries in Germany
(Topalis et al. 2011)	To retrieve data and information extraction	-Use ontology based model to integrate and capture the right terms (variables) and the relationships between such concepts in a disease map -Theoretical discussion about data accuracy in multiple information sources -Authors demonstrate the importance of capturing the right terms in ontologies to use both in the development of specific databases and, in the construction of decision support systems to control diseases for biologists, and epidemiologists	Neurological disease, malaria, vector-borne diseases in Greece
(Perez-Rey et al. 2006)	To develop a method and tool for database integration from remote sources	-Test the implemented ontology on eight different private databases with biomedical data stored in different database management packages such as MySQL, PointBase, Access, and others and provide integrated access to their data -Use case study to retrieve information in three sources using queries and theoretical discussion on data consistency -Authors believe that ontologies are the most suitable representation formalism for schemas in database integration system	Public genomic and clinical databases in Spain
(Lee et al. 2009)	To classify a person as a diabetic patient	-Represent new ontology methods for fuzzy medical relationship using taxonomical knowledge	Diabetes domain in Taiwan

Table 7 The impact of implemented ontologies for the management of data quality *(Continued)*

		-Manage accuracy of data -Authors state that fuzzy ontology can effectively develop semantic decision making and reduce uncertainty (inaccurate data) to classify patients for medical staffs	
(O'Donoghue et al. 2009)	To demonstrate the data quality benefits of integrating remote patient monitoring solutions	-Use a Body Area Network (BAN) datasets within patient EHR solutions -Use Jade Content Ontology classes for their the Medical Knowledge Base agent -Use 2 experiments (with/without knowledge base) for effect on risk prediction accuracy -Focus on data accuracy and correctness -Authors states that ontology can improve patient management through the reduction of false alarm generations and facilitate the categorisation of the data to indicate risk categories for decision support	Three patient types are identified 1) Non-Athletic Adult, 2) Athletic Adult and 3) Child from Ireland

Table 8 The impact of implemented ontologies for the assessment of data quality

Ontology functions	References	Defined purpose	Assessed of fitness for purpose using DQ	Context
Assessment (*7 papers*)	(Jacquelinet et al. 2003)	To develop semantic data interoperability	-Apply an ontological tool to develop semantic data interoperability through domain terminologies using quantitative analysis of the existing coding information system and a qualitative analysis checking completeness, consistency, ambiguity and implicitness of terms	Failure, dialysis and transplant datasets from National information system in France
			- Represent DQ factors such as completeness of data, appropriated terms, structured thesaurus, and terminology standard	
			-Authors state usefulness of ontology based approach to support the processing of texts, and extending a terminological basis for medical experts	
	(Maragoudakis et al. 2008)	To develop decision support system	-Use 25 patients records from various networking appliances such as mobile phones and wireless medical sensors to establish a ubiquitous environment for medical treatment of pulmonary diseases	Mobile sensor data from 25 patients in Artificial Neural Network (ANN) in GREECE
			-Use ontology approach based on hierarchical Bayesian networks which can encode a domain and make prediction	
			-Focus on data timeliness	
			-Authors states the importance of ontology based model as an ubiquitous platform to improve patient monitoring and health services in real time treatment decision	
	(Wang et al. 2007)	To classify diabetic patients	Use measuring precision and recall of results to show accuracy of clinical data achieved from an ontology-based fuzzy inference agent, including a fuzzy inference engine, and a fuzzy rule base, for diabetes classification	Retrieve 392 cases from the Pima Indians diabetes database in US
			-Authors state that ontology approach can classify effectively classify a person as a diabetic patient for medical staff	
	(Valencia-Garcia et al. 2008)	To develop retrieval and extract clinical information	-Represent multiple semantic relationships among concepts with UMLS ancestors through MESH descriptors in the ontology to enrich the ontology extracted from the text	Use breast cancer domain in the system with a Spanish corpus of 8649 words in Spain

Table 8 The impact of implemented ontologies for the assessment of data quality *(Continued)*

		-Use an experiment (4 PhD students were asked to use the system with a Spanish corpus) to analyse a software tool by measuring precision and recall of the result (accuracy of data)	
		-Solve semantic clinical data issues and develop accuracy of retrieval information through ontologies	
(Mabotuwana and Warren, 2009)	To identify hypertensive patients in the context of quality use of medicines	-Use the querying capabilities of one GP database in the context of quality use of medications in the management of hypertension over time	CVD in practice management system in NZ
		-Use 8 criteria and 4 scenarios to identify hypertensive patients	
		-Focus on semantic interoperability and also data completeness and timeliness, consistency	
		-Authors show the importance of ontology based approach to enhance temporal querying requirements and identify patient data, semantically	
(Young et al. 2009)	To develop semantic data collection and integration	-Use the modelling of terms to conform to and extend the existing ontologies development framework	Data on Autism in the National Database for Autism Research in US
		-Theoretical discussion on completeness of data, data availability and accessibility	
		-Authors state that ontology help to extract, query, integrate and federate data for clinical researcher	
(Preece et al. 2008)	To manage information quality (IQ) in a real-life example of gene expression research	- Implication of viewing high IQ as 'fitness for purpose' for providers and consumers, in which users state their quality requirements in terms of domain concepts (such as accuracy, currency and completeness)	Gene expression data which involve the use of microarrays in UK
		- Guide the development and use of metrics to measure the complexity and cohesion of ontologies	
		-Authors state that ontology helps to allow a practical division of the work between providers and consumers, in order to minimize the costs to all concerned	

2012b; Spasic and Ananiadou 2005; Stvilia et al. 2009; Valencia-Garcia et al. 2008; Wang et al. 2007).

Table 9 illustrates various definitions to identify the most common criteria to assess validity of ontologies and data models. Studies have attempted to define criteria such as Flexibility, Reusability, Cohesiveness, Precision, and Recall. However, there are less coordinated attempts to define other criteria such as Scalability, Completeness, Correctness, Extensibility, and Adaptability.

There are overlaps in the definition of criteria such as Flexibility, Scalability, Completeness, Correctness, Extensibility, and Adaptability in both ontological and non-ontological approaches. There were no guidance on the definition and scope of Reusability, Cohesiveness, Precision, and Recall in the data model approaches in the literature. Standardising these metrics can help to standardise the specification of ontologies and data models. This can then standardise the comparison of ontology and non-ontology approaches.

Discussion

This review examined the role of ontology-based approaches to develop data quality based on "fitness for purpose" in the health context. The findings updated and corroborated much of our previous work in this field and added new knowledge to ontology-based approaches to data quality and "fitness for purpose" of information systems.

How is data quality being conceptualized within the "fitness for purpose" definition for a range of uses?

We found few papers on DQ used within the definition of fitness for purpose. There are more studies on the ontologies for management of DQ (26 papers) and assessment of DQ in all contexts (11 papers). These findings support the current perception of DQ as a complex concept with many dimensions, often overlapping conceptually (Wand and Wang 1996). Liaw et al. (2011) developed a conceptual framework for DQ that include intrinsic DQ (correctness and consistency) of data elements and fitness for purpose (completeness) of data set for research and clinical purpose.

What specification methodologies are being used to specify data quality for implementation?

The literature on the specification of data quality for implementation is fragmentary and there is not a comprehensive approach. The findings of the current study are consistent with our previous review (Liaw et al. 2013) that the ontological approach to develop DQ is poorly evaluated. However, most agreed that DQ is a multidimensional construct (Devillers et al. 2007; Nimmagadda et al. 2008); with completeness, accuracy, correctness, consistency and timeliness being the most commonly used dimensions. A few studies examined ontology-based approaches to support data consistency and accuracy. However, no research was found that formally and systematically assessed the association between ontologies for DQ and fitness for purpose in various contexts.

Table 9 Metrics to evaluate and compare ontology and traditional data model approaches

Criteria	Metrics for ontology evaluation	References	Metrics for data model evaluation	References
Flexibility	Easily adapted to multiple views in terms of parameters such as modularity, partitioning, context-boundedness	Gangemi et al. 2006	Ability to deal with changes in business and/or regulatory rules/context?	Moody and Shanks 2003
	Ability to accept input of new data from various research groups and disciplines	Maiga and Williams 2008	Ability to add new data elements and relationships if project scope or regulatory rules (e.g. patient identification) change	Kahn et al. 2012
	Easily re-define the extraction procedure logics and adapt it to user needs	Pannarale et al. 2012	Flexibility of data models include "extensibility", "scalability", and "adaptability" as defined operationally below.	Kahn et al. 2012
	Easily manage the changes of the database schema or the ontology	Pannarale et al. 2012		
Reusability	Ability to integrate data so that it is useful to different users and disciplines	Maiga and Williams 2008		
	Ability to match user requirements across different disciplines	Pinto 2004		
Scalability			Can data model be sized in smaller or larger data sets?	Kahn et al. 2012
Completeness			Does the data model contain all user requirements?	Moody and Shanks 2003
			Can the data model store and retrieve data to meet investigator needs?	Kahn et al. 2012
Correctness			Does the data model conform to the rules of the data modelling techniques?	Moody and Shanks 2003
			Does the model conform to good data modelling practices such as limited data storage redundancy?	Kahn et al. 2012
Extensibility			Can the data model expand data elements, data types and include new data domains?	Kahn et al. 2012
Adaptability			Can the data model represent a broad data domain?	Kahn et al. 2012
Cohesiveness	A measure of the separation of responsibilities and independence of components of ontologies	Yao et al. 2005		
Precision	A measure of the amount of knowledge correctly identified in the ontology w.r.t. the whole domain knowledge available	Brewster et al. 2004		

Table 9 Metrics to evaluate and compare ontology and traditional data model approaches *(Continued)*

Recall	A measure of the amount of knowledge correctly identified with respect to all the knowledge that it should identify	Brewster et al. 2004		
Fitness for purpose	Can the ontology define and assess if routinely collected EHR data is fit for purpose?	Wand and Wang, 1996; Liaw et al. 2011	Can the data model store and retrieve data to meet investigator needs correctly? (Note: Kahn defined this as completeness of the data model)	Kahn et al. 2012

What ontology-specified implementations are being used and how do they compare with other methods?

There were few comparative and evaluative studies on assessment of data quality or compared ontological and non-ontological approaches to representing knowledge in clinical information systems. This literature review suggests that, compared to non-hierarchical data models, there may be more advantages and benefits in the use of ontologies to solve semantic clinical data quality issues and improve the validity and reliability of data retrieval, collection, storage, extraction and linkage algorithms and tools. Formal ontological approaches enable the systematic development of automated, valid and reliable methods to assess and manage the DQ and semantic interoperability issues (Lee et al. 2009; Valencia-Garcia et al. 2008; Verma et al. 2009, 2008). The expressiveness of ontology based models can facilitate accuracy and precision compared to non-ontology models and approaches (Esposito 2008a, 2008b; Preece et al. 2008).

Current ontological approaches have limited evaluation. There are little comparative studies in the chronic disease management domain and even less examining data quality. The challenges to the development and validation of an ontology-based model to the assessment and management of DQ include methodological immaturity, an immature knowledge base, and a lack of tools to support ontology-based design of information systems, evaluation of ontological approaches, and engagement of users in design and implementations. There are insufficient studies to define ontology evaluation metrics comprehensively and show practical techniques to evaluate ontological approaches in terms of flexibility, scalability and reusability versus non-ontology based models.

How is the impact of implementing ontology-based specifications for data quality in chronic disease management being measured and evaluated?

Current evidence demonstrates there is a lack of valid and reliable data quality assurance (Arts et al. 2003, 2002b) to ensure fitness for a range of uses by consumers, patients, health providers and professionals. This study has added to our understanding of ontology-based approaches to improve the quality of the data so it is useful for the various purposes such as clinical research, teaching, audit and evaluation. (e.g. quality assurance and clinical decision making). The main advantages of building ontologies for data quality in health are to automate the extraction of data from EHRs into clinical data warehouses; assessment and management of the intrinsic quality and completeness of this "big data" so that they are fit for purposes such as research, quality improvement and health information exchange and sharing; management of controlled vocabularies and optimising semantic interoperability; curation of data for use by human users and applications such as electronic decision support systems; mining of data to discover relationships between the concepts; discovery of new knowledge; and reuse of knowledge in the management of chronic diseases (Abidi 2011; Buranarach et al. 2009; Gedzelman et al. 2005; Gupta et al. 2003; Jara et al. 2009).

Limitations of the review

The majority of studies involved design and tools development for data models and ontologies in health area and chronic diseases rather than implementation, deployment and evaluation of the relevant procedures and tools. The trends are encouraging for

ontological approaches. However, there are no formal large scale studies to systematically compare the quality of outputs of ontological to non-ontological approaches to the assessment and management of data quality and fitness for purpose of the implementations. We did not search the grey literature, an important source in this relatively immature field. However, there were also limitations of access to proprietary materials. In future investigations it might be possible to use an ontological approach to develop data quality in different administrative, financial and clinical information systems.

Managerial implication

The findings of this study have several important practical implications for developing enterprise information systems. For instance, a health organisation can determine the current status of advancement of their ontology and information model, to guide the further design of a semantic strategy and to achieve specific goals, given the current data quality in their clinical information systems (CIS). The findings of this study and our previous review may serve as a benchmark for developing an ontology model as a tool for assessing and managing data quality in clinical information systems.

Also, for the development of CIS and clinical data warehouses managers can determine which features or functions of ontology based approaches could support their health professionals and patients better. Additionally, managers can use the ontology model to develop their information system in terms of all dimensions of data quality: it can show them the major strengths and weaknesses of their quality of information in terms of supporting end users in their decision making process. This is the fitness for purpose paradigm.

Conclusion

The understanding of data quality, as a multidimensional concept applied to the data elements (intrinsic DQ) and the set of data elements (extrinsic DQ) is progressing. Ontological approaches are emerging and theoretically important to address the complex relationships among overlapping concepts in this complex area. This review has described the current published literature in this domain and points to number of directions for ongoing research into the use of ontological approaches to managing the fitness for purpose of "big data" from multiple EHRs.

Additional file

Additional 1: Figure S1. Template used to analyse papers.

Abbreviations

CDM: Chronic disease management; DQ: Data quality; DQM: Data quality management; ePBRN: The electronic practice based research network; EHR: Electronic health records; CIS: Clinical information system; GPS: General practice system; OBMAS: Ontology based multi-agent system; SNOMED CT AU: Systematised nomenclature of medicine clinical term Australian release; MESH: Medical subject headings; COPD: Chronic obstructive pulmonary disease; ANN: Artificial neural network; SWRL: Semantic web rule language; SPARQL: Semantic protocol and RDF query language.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

STL, AR and PR developed the conceptual framework and templates for the literature review. AR managed the review and appraised all included papers as part of his PhD studies. All authors discussed their appraisals with AR and STL to achieve consensus. AR prepared this paper iteratively with input from all co-authors prior to submission. All authors read and approved the final manuscript.

Acknowledgment
The authors would like to thank Dr Sarah Dennis and Dr Sanjyot Vagholkar for their previous and ongoing contributions in this study.

Author details
[1]UNSW School of Public Health & Community Medicine, Sydney, Australia. [2]Isfahan University of Medical Sciences, Faculty of Management and Medical Information Sciences, Health Information Technology Research Centre, Isfahan, Iran. [3]UNSW Centre for Primary Health Care & Equity, Sydney, Australia. [4]SW Sydney Local Health District (SWSLHD) General Practice Unit, Sydney, Australia. [5]UNSW Asia Pacific Research Centre for Ubiquitous Healthcare, Sydney, Australia. [6]UNSW/SWSLHD General Practice Unit,
PO Box 5, Fairfield, NSW 1860 Sydney, Australia.

Received: 20 November 2013 Accepted: 12 December 2013
Published: 19 February 2014

References

Abidi, SR. (2011). *Ontology-based knowledge modeling to provide decision support for comorbid diseases*. Paper presented at the 19th European Conference in Artificial Intelligence. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-79952016090&partnerID=40&md5=d6e8e7441e3e9118fa395e5fc0b77b95.

Arts, D, de Keizer, N, Scheffer, GJ, & de Jonge, E. (2002a). Quality of data collected for severity of illness scores in the Dutch National Intensive Care Evaluation (NICE) registry. *Intensive Care Medicine, 28*(5), 656–659.

Arts, DG, de Keizer, NF, & Scheffer, GJ. (2002b). Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *Journal of the American Medical Informatics Association, 9*(6), 600–611.

Arts, DG, Bosman, RJ, de Jonge, E, Joore, JC, & de Keizer, NF. (2003). Training in data definitions improves quality of intensive care data. *Critical Care, 7*(2), 179–184.

Azaouagh, A, & Stausberg, J. (2008). Frequency of hospital-acquired pneumonia–comparison between electronic and paper-based patient records. *Pneumologie, 62*(5), 273–278.

Brank, J, Grobelnik, M, & Mladenić, D. (2005). *A survey of ontology evaluation techniques*. Paper presented at the Proc. of 8th Int. Multi-Conf. Information Society.

Brewster, C, Alani, H, Dasmahapatra, S, & Wilks, Y. (2004). *Data Driven Ontology Evaluation*. Paper presented at the International Conference on Language Resources and Evaluation. Retrieved from http://eprints.soton.ac.uk/259062/.

Britt, H, Miller, G, & Bayrarn, C. (2007). The quality of data on general practice - a discussion of BEACH reliability and validity. *Australian Family Physician, 36*(1–2), 36–40.

Brüggemann, S, & Grüning, F. (2009). Using ontologies providing domain knowledge for data quality management. *Studies in Computational Intelligence, 221*, 187–203.

Buranarach, M, Chalortham, N, Chatvorawit, P, Thein, Y, & Supnithi, T. (2009). An ontology-based framework for development of clinical reminder system to support chronic disease healthcare. Retrieved from http://text.hlt.nectec.or.th/ontology/sites/default/files/reminder_isbme09_cr_0.pdf.

Chen, FH. (2009). Modeling the effect of information quality on risk behavior change and the transmission of infectious diseases. *Mathematical Biosciences, 217*(2), 125–133.

Chen, WL, Zhang, SD, & Gao, X. (2009). Anchoring the Consistency Dimension of Data Quality Using Ontology in Data Integration. In *2009 Sixth Web Information Systems and Applications Conference, IEEE*.

Choquet, R, Qouiyd, S, Ouagne, D, Pasche, E, Daniel, C, Boussaïd, O, et al. (2010). *The information quality triangle: A methodology to assess clinical information quality*. Paper presented at the 13th World Congress on Medical and Health Informatics, Medinfo 2010, Cape Town.

Cunningham-Myrie, C, Reid, M, & Forrester, TE. (2008). A comparative study of the quality and availability of health information used to facilitate cost burden analysis of diabetes and hypertension in the Caribbean. *West Indian Medical Journal, 57*(4), 383–392.

de Lusignan, S, Khunti, K, Belsey, J, Hattersley, A, van Vlymen, J, Gallagher, H, et al. (2010). A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: a pilot and validation study of routinely collected data. *Diabetic Medicine, 27*, 203–209.

Devillers, R, Bedard, Y, Jeansoulin, R, & Moulin, B. (2007). Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. *International Journal of Geographical Information Science, 21*(3), 261–282.

Esposito, M. (2008a). Congenital Heart Disease: An ontology-based approach for the examination of the cardiovascular system. In I Lovrek (Ed.), *Knowledge - Based Intelligent Information and Engineering Systems, Pt 1, Proceedings Vol. 5177* (pp. 509–516).

Esposito, M. (2008b). *An ontological and non-monotonic rule-based approach to label medical images*. Los Alamitos: IEEE Computer Soc.

Euzenat, J. (2007). *Semantic Precision and Recall for Ontology Alignment Evaluation*. Paper presented at the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07).

Gangemi, A, Catenacci, C, Ciaramita, M, & Lehmann, J. (2006). *Modelling ontology evaluation and validation*. Paper presented at the Proceedings of the 3rd European conference on The Semantic Web: research and applications.

Gedzelman, S, Simonet, M, Bernhard, D, Diallo, G, & Palmer, P. (2005). Building an ontology of cardio-vascular diseases for concept-based information retrieval. *Computers in Cardiology, 32*, 255–258.

Gillies, A. (2000a). Assessing and improving the quality of information for health evaluation and promotion. *Methods of Information in Medicine, 39*(3), 4.

Gillies, A. (2000b). Assessing and improving the quality of information for health evaluation and promotion. *Methods of Information in Medicine, 39*(3), 208–212.

Gupta, A, Ludäscher, B, Grethe, JS, & Martone, ME. (2003). Towards a formalization of disease-specific ontologies for neuroinformatics. *Neural Networks, 16*(9), 1277–1292.

Hamilton, WT, Round, AP, Sharp, D, & Peters, TJ. (2003). The quality of record keeping in primary care: a comparison of computerised, paper and hybrid systems. *The British Journal of General Practice, 53*(497), 929–933. discussion 933.

Huaman, MA, Araujo-Castillo, RV, Soto, G, Neyra, JM, Quispe, JA, Fernandez, MF, et al. (2009). Impact of two interventions on timeliness and data quality of an electronic disease surveillance system in a resource limited setting (Peru): a prospective evaluation. *BMC Med Inform Decis Mak., 9*, 16.

Ivanova, I, Morales, J, de By, RA, Beshe, TS, & Gebresilassie, MA. (2013). Searching for spatial data resources by fitness for use. *Journal of Spatial Science, 58*(1), 15–28.

Jacquelinet, C, Burgun, A, Delamarre, D, Strang, N, Djabbour, S, Boutin, B, et al. (2003). Developing the ontological foundations of a terminological system for end-stage diseases, organ failure, dialysis and transplantation. *International Journal of Medical Informatics, 70*(2–3), 317–328. doi: 10.1016/S1386-5056(03)00046-7.

Jara, AJ, Blaya, FJ, Zamora, MA, & Skarmeta, AFG. (2009). *An Ontology and Rule Based Intelligent Information System to Detect and Predict Myocardial Diseases*. New York: IEEE.

Kahn, BK, Strong, DM, & Wang, RY. (2002). Information quality benchmarks: product and service performance. *Communications of the ACM, 45*(4), 8.

Kahn, MG, Batson, D, & Schilling, LM. (2012). Data model considerations for clinical effectiveness researchers. *Medical Care, 50 Suppl*, S60–S67.

Kerr, K, Norris, A, & Stockdale, R. (2007). *Data quality, information and decision making: a healthcare case study*. Paper presented at the 18th Australasian Conference on Information Systems, Toowoomba, Australia.

Kiragga, AN, Castelnuovo, B, Schaefer, P, Muwonge, T, & Easterbrook, PJ. (2011). Quality of data collection in a large HIV observational clinic database in sub-Saharan Africa: implications for clinical research and audit of care. *Journal of the International AIDS Society, 14*(1).

Lain, SJ, Roberts, CL, Hadfield, RM, Bell, JC, & Morris, JM. (2008). How accurate is the reporting of obstetric haemorrhage in hospital discharge data? A validation study. *Australian and New Zealand Journal of Obstetrics and Gynaecology, 48*(5), 481–484.

Lee, CS, Wang, MH, Acampora, G, Loia, V, & Hsu, CY. (2009). *Ontology-based Intelligent Fuzzy Agent for Diabetes Application*. New York: IEEE.

Li, Z. (2010). *An ontology-driven concept-based information retrieveal approach for Web documents*. Edmonton, Alberta: University of Alberta.

Li, HC, & Ko, WM. (2007). *Automated food ontology construction mechanism for diabetes diet care*. New York: IEEE.

Liaw, S, Taggart, J, Dennis, S, & Yeo, A. (2011). Data quality and fitness for purpose of routinely collected data – a general practice case study from an electronic Practice-Based Research Network (ePBRN). In *AMIA 2011 Annual Symposium Improving Health: Informatics and IT Changing the World; October 22-26, 2011* (pp. 785–94). Washington DC, US: AMIA.

Liaw, ST, Chen, HY, Maneze, D, Taggart, J, Dennis, S, Vagholkar, S, & Bunker, J. (2012). Health reform: is routinely collected electronic information fit for purpose? *Emergency Medicine Australasia, 24*(1), 57–63.

Liaw, ST, Rahimi, A, Ray, P, Taggart, J, Dennis, S, de Lusignan, S, et al. (2013). Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. *International Journal of Medical Informatics, 82*(2), 139.

Lima, L, Novais, P, Costa, R, Cruz, J, & Neves, J. (2010). Decision Making Based on Quality-of-Information a Clinical Guideline for Chronic Obstructive Pulmonary Disease Scenario. In A de Leon, F de Carvalho, S Rodríguez-González, J De Paz Santana, & J Rodríguez (Eds.), *Distributed Computing and Artificial Intelligence Vol. 79* (pp. 417–424). Berlin/Heidelberg: Springer.

Mabotuwana, T, & Warren, J. (2009). An ontology-based approach to enhance querying capabilities of general practice medicine for better management of hypertension. *Artificial Intelligence in Medicine, 47*(2), 87–103.

Maiga, G, & Williams, D. (2008). A flexible approach for user evaluation of biomedical ontologies. *International Journal of Computing and ICT Research, 2*(2), 62–74.

Maragoudakis, M, Lymberopoulos, D, Fakotakis, N, & Spiropoulos, K. (2008). A Hierarchical, Ontology-Driven Bayesian Concept for Ubiquitous Medical Environments- A Case Study for Pulmonary Diseases. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vols 1–8* (pp. 3807–3810). New York: IEEE.

McGarry, K, Garfield, S, & Wermter, S. (2007). Auto-extraction, representation and integration of a diabetes ontology using Bayesian networks. In P Kokol, V Podgorelec, D MiceticTurk, M Zorman, & M Verlic (Eds.), *Twentieth IEEE International Symposium on Computer-Based Medical Systems, Proceedings* (pp. 612–617).

McJunkin, MC. (1995). Precision and recall in title keyword searches. *Information Technology and Libraries, 14*(3), 161–171.

Min, H, Manion, FJ, Goralczyk, E, Wong, YN, Ross, E, & Beck, JR. (2009). Integration of prostate cancer clinical data using an ontology. *Journal of Biomedical Informatics, 42*(6), 1035–1045.

Mitchell, J, & Westerduin, F. (2008). Emergency department information system diagnosis: how accurate is it? *Emergency Medicine Journal, 25*(11), 784.

Moody, DL, & Shanks, GG. (2003). Improving the quality of data models: empirical validation of a quality management framework. *Information Systems, 28*(6), 619–650.

Moro, ML, & Morsillo, F. (2004). Can hospital discharge diagnoses be used for surveillance of surgical-site infections? *Journal of Hospital Infection, 56*(3), 239–241.

Nimmagadda, SL, Nimmagadda, SK, & Dreher, H. (2008). Ontology based data warehouse modeling and managing ecology of human body for disease and drug prescription management. In *2008 2nd IEEE International Conference on Digital Ecosystems and Technologies* (pp. 465–473).

O'Donoghue, J, Herbert, J, O'Reilly, P, & Sammon, D. (2009). Towards Improved Information Quality: The Integration of Body Area Network Data within Electronic Health Records. In M Mokhtari, I Khalil, J Bauchet, D Zhang, & C Nugent (Eds.), *Ambient Assistive Health and Wellness Management in the Heart of the City, Proceeding Vol. 5597* (pp. 299–302).

Orme, AM, Yao, H, & Etzkorn, LH. (2007). Indicating ontology data quality, stability, and completeness throughout ontology evolution. *Journal of Software Maintenance and Evolution-Research and Practice, 19*(1), 49–75.

Pannarale, P, Catalano, D, De Caro, G, Grillo, G, Leo, P, Pappada, G, et al. (2012). GIDL: a rule based expert system for GenBank intelligent data loading into the molecular biodiversity database. *BMC Bioinformatics, 13*(Suppl 4), S4.

Pathak, J, Kiefer, RC, Bielinski, SJ, & Chute, CG. (2012a). Mining the human phenome using semantic web technologies: a case study for type 2 diabetes. *AMIA Annual Symposium Proceedings, 2012*, 699–708.
Pathak, J, Kiefer, RC, & Chute, CG. (2012b). Using semantic web technologies for cohort identification from electronic health records for clinical research. *AMIA Summits on Translational Science Proceedings, 2012*, 10–19.
Perez-Rey, D, Maojo, V, Garcia-Remesal, M, Alonso-Calvo, R, Billhardt, H, Martin-Sanchez, F, et al. (2006). ONTOFUSION: ontology-based integration of genomic and clinical databases. *Computers in Biology and Medicine, 36*(7–8), 712–730.
Pinto, HS. (2004). Ontologies: how can they be built? *Knowledge and Information Systems, 6*(4), 441–464.
Preece, A, Missier, P, Ernbury, S, Jin, B, & Greenwood, M. (2008). An ontology-based approach to handling information quality in e-science. *Concurrency and Computation-Practice and Experience, 20*(3), 253–264.
Quan, H, Li, B, Saunders, LD, Parsons, GA, Nilsson, CI, Alibhai, A, et al. (2008). Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. *Health Services Research, 43*(4), 1424–1441.
Redman, T. (2005). Measuring data accuracy. In W Rea (Ed.), *Information Quality* (p. 21). Armonk NY: ME Sharpe Inc.
Spasic, I, & Ananiadou, S. (2005). A flexible measure of contextual similarity for biomedical terms. *Pacific Symposium on Biocomputing, 10*, 197–208.
Stvilia, B, Mon, L, & Yi, YJ. (2009). A model for online consumer health information quality. *Journal of the American Society for Information Science and Technology, 60*(9), 1781–1791.
Topalis, P, Dialynas, E, Mitraka, E, Deligianni, E, Siden-Kiamos, I, & Louis, C. (2011). A set of ontologies to drive tools for the control of vector-borne diseases. *Journal of Biomedical Informatics, 44*(1), 42–47.
Valencia-Garcia, R, Fernandez-Breis, JT, Ruiz-Martinez, JM, Garcia-Sanchez, F, & Martinez-Bejar, R. (2008). A knowledge acquisition methodology to ontology construction for information retrieval from medical documents. *Expert Systems, 25*(3), 314–334.
Verma, A, Kasabov, N, Rush, A, & Song, Q. (2008). *Ontology based personalized modeling for chronic disease risk analysis: an integrated approach*. Paper presented at the The 15th international conference on Advances in neuro-information processing.
Verma, A, Fiasché, M, Cuzzola, M, Iacopino, P, Morabito, P, & Kasabov, N. (2009). Ontology based personalized modeling for type 2 diabetes risk analysis: An Investigated Approach. In CS Leung, M Lee, & JH Chan (Eds.), *ICONIP 2009, Part II* (pp. 360–366). Berlin: Springer-Verlag.
Wand, Y, & Wang, Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM, 36*(11), 86–95.
Wang, R. (1998). A product perspective on total data quality management. *Communications of the ACM, 41*(2), 58–65.
Wang, R, Strong, D, & Guarascio, L. (1996). Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33.
Wang, MH, Lee, CS, Li, HC, & Ko, WM. (2007). *Ontology-based fuzzy inference agent for diabetes classification*. New York: IEEE.
Yao, H, Orme, A, & Etzkorn, LH. (2005). Cohesion metrics for ontology design and application. *Journal of Computer Science, 1*(1), 107–113.
Young, L, Tu, SW, Tennakoon, L, Vismer, D, Astakhov, V, Gupta, A, et al. (2009). Ontology Driven Data Integration for Autism Research. In *2009 22nd IEEE International Symposium on Computer-Based Medical Systems* (pp. 54–60). New York: IEEE.

doi:10.1186/2193-8636-1-5
Cite this article as: Rahimi *et al.*: Ontological specification of quality of chronic disease data in EHRs to support decision analytics: a realist review. *Decision Analytics* 2014 **1**:5.