

Ontology-based approaches to identify patients with type 2 diabetes mellitus from electronic health records: development and validation

Author:

Rahimi Khorzoughi, Alireza

Publication Date:

2015

DOI:

<https://doi.org/10.26190/unsworks/17346>

License:

<https://creativecommons.org/licenses/by-nc-nd/3.0/au/>

Link to license to see what you are allowed to do with this resource.

Downloaded from <http://hdl.handle.net/1959.4/54224> in <https://unsworks.unsw.edu.au> on 2024-05-02

Ontology-based Approaches to Identify Patients with Type 2 Diabetes Mellitus from Electronic Health Records: Development and Validation

Alireza Rahimi

A thesis in fulfilment of the requirements for the degree of
Doctor of Philosophy



Faculty of Medicine
School of Public Health and Community Medicine

February 2015

TABLE OF CONTENTS

Prefacing comments and declarations regarding publications used in thesis	i
Acknowledgements	iv
Abbreviations and Symbols	v
List of Figures	vi
List of Tables.....	vii
Thesis Abstract.....	viii
Chapter 1. Introduction.....	1
1.1 Rationale and Problem Statement	1
1.2 Definitions and Description of Concepts	4
1.2.1 Chronic disease management (CDM)	4
1.2.2 Data quality (DQ).....	4
1.2.3 Ontology-based Approach	5
1.2.3.1 Notion of ontology	5
1.2.3.2 Motivations and advantages of the use of ontology	6
1.3 Research Methodology.....	10
1.4 Overview of Thesis	11
1.5 Summary	14
1.6 References	15
Chapter 2. Ontological Specification of Quality of Chronic Disease Data in Electronic Health Records to Support Decision Analytics: A Realist Review	19
2.1 Background	23
2.2 Methods.....	25
2.3 Results	28
2.3.1 Conceptualisation of data quality within the ‘fitness for purpose’ paradigm ..	33
2.3.2 Methodologies to specify data quality for implementation.....	36
2.3.3 Ontology-specified implementation to develop data quality and compare with other models	39
2.3.4 The impact of ontologies for data quality in CDM and their evaluation	41
2.4 Discussion	47
2.4.1 How is data quality being conceptualized within the ‘fitness for purpose’ definition for a range of uses?	48

2.4.2 What specification methodologies are being used to specify data quality for implementation?	48
2.4.3 What ontology-specified implementations are being used and how do they compare with other methods?	48
2.4.4 How is the impact of implementing ontology-based specifications for data quality in CDM being measured and evaluated?	49
2.4.5 Limitations of the review	48
2.4.6 Managerial implications	48
2.5 Conclusion	50
2.6 References	50
Chapter 3. Development of a Methodological Approach for Data Quality	
Ontology in Diabetes Management	59
3.1 Introduction	63
3.2 Background	64
3.3 A Methodology for Data Quality Ontology (MDQO)	66
3.3.1 Knowledge acquisition	69
3.3.1.1 Patient data audit	69
3.3.1.2 GP and nurse consensus meetings	69
3.3.1.3 Literature review	70
3.3.2 Conceptualization	70
3.3.2.1 Task	71
3.3.2.2 Output	71
3.3.3 Semantic modelling	72
3.3.3.1 Task	72
3.3.3.2 Output	72
3.3.4 Knowledge representation	75
3.3.4.1 Task	76
3.3.4.2 Output	76
3.3.4.2.1 Step 1: Analysis of sources and targets	76
3.3.4.2.2 Step 2: Mappings and queries	77
3.3.5 Validation	80
3.3.5.1 Task	80
3.3.5.2 Output	83
3.4 Discussion	83

3.4.1 Usefulness of the ontology-based approach for DQ specification	83
3.4.2 Applicability of the ontology-based approach for DQ specification	84
3.4.3 Evaluation of MDQO	85
3.4.4 Comparison of MDQO and non-ontological approaches in CDM	86
3.5 Conclusion	86
3.6 References	86
Chapter 4. Validating an Ontology-based Algorithm to Identify Patients with Type 2 Diabetes Mellitus in Electronic Health Records	92
4.1 Introduction	97
4.2 Methods.....	100
4.2.1 Establishing the gold standard with a manual audit.....	100
4.2.2 Mapping data using ontopPro 1.8	100
4.2.3 Semantic query of dataset	100
4.2.4 Assessing the accuracy of the algorithm.....	101
4.2.5 Examining reasons for false positives and false negatives	101
4.3 Results	102
4.3.1 Sample size.....	102
4.3.2 The DMO-based algorithm	102
4.4 Discussion	110
4.5 Limitations of the research.....	113
4.6 Conclusion	114
4.7 References	114
Chapter 5. Discussion	122
5.1 Methodology to Develop an Ontology.....	122
5.2 Validation of MDQO	127
5.2.1 Strengths and weaknesses of the semantic queries	128
5.2.2 Contribution of validation to the knowledge base	130
5.3 Summary	133
5.4 References	134
Chapter 6. Conclusion	135
6.1 Future Work	135
Appendixes.....	138-169

PREFACING COMMENTS AND DECLARATIONS REGARDING PUBLICATIONS USED IN THESIS

The three empirical chapters of this thesis are comprised principally of the ‘Authors’ Accepted Manuscripts’ for three journal papers that have been published or accepted to be published for the second and third papers, which are in press.

The papers are:

1. Rahimi, A., Liaw, S.T., Ray, P., Taggart, J., & Yu, H. (2014). Ontological specification of quality of chronic disease data in electronic health records to support decision analytics: A realist review. *Decision Analytics*, 1(1), 1–31. doi: 10.1186/2193-8636-1-5. <http://dx.doi.org/10.1186/2193-8636-1-5>.
2. Rahimi, A., Parameshwaran, N., Ray, P., Taggart, J., Yu, H., & Liaw, S.T. (2014). Development of a methodological approach for data quality ontology in diabetes management. *International Journal of Electronic Health and Medical Communication*, (Accepted May 2014).
3. 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*. (Accepted May 2014).

Each journal paper was co-authored by me and my postgraduate supervisors, Prof S.T. Liaw and Prof. P. Ray. In each case, I made significant contributions (more than 70%) to the papers. In addition to the above, I contributed to the following publications:

4. A first-authored peer-reviewed book chapter: Rahimi, A., Liaw, S.T., Ray, P.K., Taggart, J., & Yu, H. Ontology for data quality and chronic disease management: A literature review. *Chapter xx In Healthcare Informatics and Analytics: Emerging Issues and Trends*. Hershey, PA: IGI Global. (Accepted Nov 2013.)
5. A second-authored peer-reviewed journal paper: Liaw S.T., Rahimi A., Ray P., Taggart J., Dennis S., de Lusignan S., & Talie-Khoei, A. (2013). Towards an

ontology for data quality in integrated chronic disease management: A realist review of the literature. *International Journal of Medical Informatics*, 82(1), 10–24.

6. A first-authored peer-reviewed conference paper: Rahimi A., Liaw S.T., Ray P., & Taggart J. (2012). Developing an ontology for data quality in chronic disease management. *Medical Informatics Europe (MIE) 2012*. Pisa, Italy.
7. Taggart, J., Liaw, S.T., Dennis, S., Yu, H., Rahimi, A., Jalaludin, B. (2012). The University of NSW electronic practice-based research network: Disease registers, data quality and utility. *Studies in Health Technology and Informatics*, 178, 219-227. <http://dx.doi.org/10.3233/978-1-61499-078-9-219>.
8. Yu, H., Liaw, S.T., Taggart, J., & Rahimi, A. (2013). *Using Ontologies to Identify Patients with Diabetes in Electronic Health Records*. Paper presented at 12th International Semantic Web Conference and the 1st Australasian Semantic Web Conference, Sydney, Australia. http://www.iswc2013.semanticweb.org/sites/default/files/iswc_demo_20.pdf.

ACKNOWLEDGEMENTS

I am very grateful to Allah the Almighty for his guidance. Without his generosity, inspiration and infinite bounties, I would be helpless.

Data collection for this thesis was supported by the electronic Practice-based Research Network project (ePBRN) (HERC number: HERC/10/LPOOL/29; SSA Western Zone number: SSA/10/LPOOL/127; Local HERC number: 10/026) in partnership with the South Western Sydney Local Health District, the University of New South Wales (UNSW) and Centre for Primary Health Care & Equity, Sydney, Australia.

First and foremost, I wish to express my deepest gratitude to my supervisors – Prof. Siaw-Teng Liaw, head of the General Practice Unit, Fairfield Hospital, UNSW and Prof. Pradeep Ray, Head of the Asia-Pacific Ubiquitous Health Care Research Centre, UNSW – for their wise advice, unswerving encouragement, support and extraordinary patience as well as valuable guidance to every stage of my research. Their devotion is sincerely appreciated. I will never be able to express sufficiently my gratitude to them.

I would like to thank the staff and scientists in the GPU, Fairfield Hospital who helped me in this research and to express my appreciation to Ms. Jane Taggart in particular, Dr. Hairong Yu, members of the Centre for Primary Health Care & Equity and also Dr. Nandan Parameshwaran, Associate Professor in the School of Computer Science and Engineering, UNSW, Sydney, Australia. Their criticisms, advice and their valuable suggestions regarding my thesis have proven to be extremely valuable.

Words fail me to express my heart-felt tribute to my wife, Zahar, and my son, Mohammad Hussain, whose dedication, love and persistent confidence in me have considerably contributed in my study. Last but not least, I would like to thank my parents, in particular, and also my younger sisters and brothers who have provided much needed support and understanding throughout my research. Finally, I would like to thank the Health Ministry of I.R. IRAN that supported me in all respects during this endeavour.

ABBREVIATIONS & SYMBOLS

ANN	Artificial Neural Network
CDM	Chronic Disease Management
CDW	Clinical Data Warehouse
CIS	Clinical Information System
COPD	Chronic Obstructive Pulmonary Disease
DQ	Data Quality
DMO	Diabetes Mellitus Ontology
DQM	Data Quality Management
EHRs	Electronic Health Records
ePBRN	electronic Practice-based Research Network
GPS	General Practice System
GPU	General Practice Unit, Fairfield Hospital, Sydney
MDQO	Methodology for Data Quality Ontology
MeSH	Medical Subject Headings
OBMAS	Ontology-based Multi-agent System
OWA	Open World Assumption
RDF	Resource Description Framework
SNOMED-CT-AU	Systematised Nomenclature of Medicine – Clinical Term – Australian Release
SWRL	Semantic Web Rule Language
T2DM	Type 2 Diabetes Mellitus
SPARQL	Semantic Protocol and RDF Query Language
UNSW	University of New South Wales

LIST OF FIGURES

Figure 1.1: The definition of Diabetes Management concept with hierarchy and properties.....	8
Figure 1.2: The thesis map	14
Figure 2.1: Template used to analyse papers	27
Figure 2.2: Paper selection process	30
Figure 2.3: Distribution of papers from each category by year	32
Figure 2.4: Distribution of papers found by continent.....	33
Figure 3.1: Five stages approach for the development of the ontology model	69
Figure 3.2: The ontology hierarchical conceptual model with data properties.....	73
Figure 3.3: A sample of object property to show how as an example “hasT2DMRFV” can link two joint classes together	74
Figure 3.4: The data property tab to define various data ranges, types and values for each class.....	74
Figure 3.5: The ‘Context’ as an example class hierarchy shown expanded in a Protégé screenshot. The annotation associated with the subclass “Type 2 diabetes mellitus” describes semantic relationship of this subclass with the reference terminology SNOMED-CT-AU	75
Figure 3.6: The ontology approach OBDA tab to define JDBC connection parameters	77
Figure 3.7: Sample of mapping the Diagnosis table with the ontology approach	79
Figure 4.1: Template to identify T2DM-related information the general practice EHR.....	101
Figure 4.2: Comparing manual an automated identification of T2DM patients from the general practice EHR	102
Figure 4.3: Diabetes mellitus ontology hierarchical conceptual model.....	103
Figure 4.4: Sample of mapping the Diagnosis table with DMO.....	104
Figure 4.5: Sample of SPARQL query to show a semantic way to implement RFV for the identification of T2DM patients	106
Figure 4.6: Sample of SPARQL queries to show a semantic way to implement 3 criteria for the identification of T2DM	107

LIST OF TABLES

Table 2.1: Online databases used and papers found	29
Table 2.2: Distribution of papers by review questions	30
Table 2.3: Distribution of papers by study types and review questions.....	32
Table 2.4: Papers where data quality was conceptualized within fitness for purpose paradigm.....	34
Table 2.5: Methodologies used to specify data quality for implementation	36
Table 2.6: Studies that compared ontologies and other data models in specification and implementation	40
Table 2.7: The impact of implemented ontologies for the management of data quality.....	41
Table 2.8: The impact of implemented ontologies for the assessment of data quality ...	44
Table 2.9: Metrics to evaluate and compare ontology and traditional data model approaches.....	46
Table 3.1: Categories of collected concepts in four different layers.....	71
Table 3.2: Part of SPARQL queries for the research purpose	81
Table 3.3: Accuracy of the model developed.....	2
Table 4.1: SPARQL queries for the research purpose	105
Table 4.2: Sensitivity and Specificity of T2DM using RFV, Rx, Path and in combination.....	107
Table 4.3: Sample of T2DM cases to demonstrate concordance and discordance between results of manual audit and algorithm.....	109

THESIS ABSTRACT

Introduction

Issues around the data quality (DQ) of patient registers are often raised when a data set is used for clinical or research purposes. An ontology-based approach provides a flexible semantic framework and supports the automation of data extraction from electronic health records (EHRs). This research aimed to assess the flexibility of an ontology-based approach to accurately identify patients with type 2 diabetes mellitus (T2DM) in a clinical database. This research also demonstrated the role of an ontology-based approach to assess quality of a register.

Method

A systematic review was conducted, which addressed DQ, ‘fitness for purpose’ of data used and ontology-based approaches. Included papers were critically appraised with a ‘context-mechanism-impacts/outcomes’ overlay. Using a literature review, the Australian National Guidelines for type 2 diabetes mellitus, the Systematised Nomenclature of Medicine – Clinical Term – Australian Release and input from health professionals, a five-stage methodology for DQ ontology (MDQO) was adopted. The methodology consisted of: (1) knowledge acquisition; (2) conceptualisation; (3) semantic modelling; (4) knowledge representation; and (5) validation. Although MDQO can be used in any validation domain, this thesis validated it in the context of T2DM diagnosis and management. The accuracy of the MDQO was validated with a manual audit of general practice EHRs through the diabetes mellitus ontology. Contingency tables were prepared and sensitivity and specificity (accuracy) of the model to diagnose T2DM was determined, using T2DM cases of a general practice, which kept a diabetes register with complete and current reason for visit information, found by manual EHR audit as the gold standard. Accuracy was determined with three attributes – reason for visit, medication and pathology – singly and in combination.

Results

The T2DM ontology included six object properties, 15 data properties, 68 concepts and 14 major themes in four main classes: actor, context, mechanism and impact. The validation showed sensitivity and specificity were 100% and 99.88% respectively with reason for visit; 96.55% and 98.97% with medication; and 15.6% with

pathology test result. This suggests that medication and pathology test result data were not as complete as reason for visit data for the general practice audited. However, the completeness was adequate for the purpose of this thesis, as confirmed by the very small relative deterioration of accuracy (sensitivity and specificity of 97.67% and 99.18%, respectively) when calculated for the combination of reason for visit, medication and pathology test result.

Discussion

Current research shows a lack of comprehensive ontology-based approaches for DQ in chronic disease management and there are few validation studies comparing ontological and non-ontological approaches on the assessment of clinical DQ. The MDQO developed in this thesis provides a semantically flexible mechanism to capture patients' data from EHRs. It is also designed to be generalisable and reusable. This T2DM ontology-based model (constructed using the MDQO) is sufficiently accurate to support a semantic approach, using reason for visit, medication and pathology tests data from EHRs to define patients with T2DM. The accuracy of the T2DM ontology approach was established with respect to the DQ dimensions. The MDQO helps with the implementation of DQ based on “fitness for use” and hence better utilisation of routinely-collected clinical data for research.

Conclusion

This thesis contributes an ontology-based methodology for DQ assessment and management in a diabetes context. It provides new insights into the identification and assessment of patients with T2DM from EHR data. This ontology-based approach can potentially support the assessment of the impact of DQ on a data set in terms of the purpose for which it is used. There is a need for similar ontology-based research in other clinical domains, beyond T2DM, to address DQ in chronic disease management.

CHAPTER 1

INTRODUCTION

1.1 Rationale and Problem Statement

The practice of evidence-based medicine requires access to significant clinical data, collated and presented in a way that the health professional can use at the time of decision making. There is growing recognition of the use of ontology approaches in Electronic Health Records (EHRs) to solve the problem of poor data quality (DQ) and semantic interoperability issues in chronic disease management (CDM), public health services and epidemiological research (Dixon, McGowan, & Grannis, 2011). This is because there is increasing concern that poor DQ is common and affects all industries and health organisations that employ information systems (R. Y. Wang, Strong, & Guarascio, 1996).

The increase in applications of EHRs increases the question of quality control of patients' data for clinical care and research. Various cooperative research using data from multiple sources raises issues of semantic interoperability and the management of large databases (D. Arts, de Keizer, Scheffer, & de Jonge, 2002). Poor DQ has significant economic costs, in terms of poor decisions and planning by organisations and health professionals, and poor quality of care (S. Liaw, Taggart, Dennis, & Yeo, 2011). Also the challenges to improved data fitness for use include poor DQ, increasing DQ and inadequate semantic interoperability (Devillers, Bedard, Jeansoulin, & Moulin, 2007). Complete and accurate information sharing – such as in clinical handover – is vital to maintain continuous and safe patient care across primary and acute services (Cummings et al., 2010).

Many studies in health care regularly report a range of deficiencies in the routinely collected electronic information for clinical (Azaouagh & Stausberg, 2008; de Lusignan et al., 2010; Mitchell & Westerduin, 2008; Moro & Morsillo, 2004) and health promotion (Gillies, 2000) purposes; however there is a lack of a valid and reliable solution to develop DQ (S. T. Liaw et al., 2013). Similar deficiencies in DQ exist with information in hospital and general practice (S. Liaw, Chen, Maneze, Dennis, & Vagholkar, 2011; S. Liaw, et al., 2011) 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 administrative data (Lain, Roberts, Hadfield, Bell, & Morris, 2008; Quan et al., 2008). Electronic health record-keeping systems in primary care are expected to be more complete and accurate than paper-only systems (Hamilton, Round, Sharp, & Peters, 2003). Evident sources of poor data quality include inaccuracies like wrong diagnoses, incomplete or inaccurate data entry, errors in all processes of using data, corruption of the database architecture or management system, and errors in data extraction (Michalakidis et al., 2010).

There are increasing numbers of studies developing ontology- and non-ontology-based approaches to improve quality of clinical data in CDM. These studies aim to address the lack of research and controversy around scientific evidence to develop ontology-based approaches for the improvement of DQ (Esposito, 2008; Kuziemyky & Lau, 2010). The issue of identifying patients with chronic diseases using automated techniques is complex and attempts have been made to understand these issues for over a decade, through methods and tools such as discussion, conceptualisation, models and theories (Liyanage, Liaw, Kuziemyky, & de Lusignan, 2013; Taggart et al., 2012; Verma et al., 2009). Traditional data model-building methods using concept analysis, syntactic and grounded theory development currently do not illuminate how data managers develop quality of data and information to build clinical knowledge (Kahn et al., 2012). Theoretical and conceptual ontology-based approaches may provide the in-depth knowledge required for useful representations of patients' data and registers semantically from EHRs (Appendix 1, page 132).

Improving DQ of EHRs and registers can improve the quality of decisions and lead to better policy, evidence-based care and patient outcomes (Appendix 2, page 143). DQ research has been identified as a priority in medical informatics research. Dixon (2011) and Huaman (2009) in their reviews of the literature for quality of data named research in the quality of clinical data as a critical informatics research priority (Dixon, et al., 2011; Huaman et al., 2009). Data quality research is necessary to improve health care through the translation of research findings into practice (S. Liaw, et al., 2011), national deployment of EHRs (Dixon, et al., 2011; Huaman, et al., 2009), and development of the National Health Information Network (NHIN) (Richesson & Krischer, 2007).

Similar conclusions were drawn for ontology-based approaches at the 24th Conference of European Medical Informatics (MIE, 2012), where informatics expert examined ontology models and priorities at the intersection of medical informatics and bioinformatics. The conference theme was “Quality of Life through Quality of Information” (*MIE Conference Proceeding: Quality of Life through Quality of Information*, 2012) and identified DQ in health areas as a priority necessary to support communication between EHRs and integration of clinical databases. Ontology was also a high priority solution identified in knowledge-based and decision-support systems (Terzi, Vakali, & Hacid, 2003).

Our research on the electronic Practice-based Research Network (ePBRN) has focused on the ontology-based approach to better define and address DQ and semantic interoperability issues. It does this by drawing on our current work on the 3Cs of DQ in diabetes management: completeness, correctness and consistency (Appendix 3, page 147). We identified DQ issues such as incompleteness, incorrectness and inconsistency in the ePBRN repository of data extracted from EHRs of general practices and health services in South Western Sydney. Patient clinical information is extracted from participating general practices and health service information systems, captured in a clinical data warehouse, linked and used for research purposes (S. Liaw, et al., 2011).

An ontology-based model is a data model that exists in the real world, such as that on which we based the design and development of an ontology-based query to identify diabetes and improve the quality of patient registers (Appendix 4, page 156). In health and medical informatics, ontology most often refers to electronic models of real-world phenomena, such as the manner in which quality of patients’ data are represented in EHRs. The term ontology is used to describe the procedure of creating these electronic models, including concepts, properties, relations and decision-making about which aspects of patients’ data will be characterised in detail and which will not, while summarising or omitting other aspects of the context (Appendix 2, page 144).

This literature review suggests that ontological approaches can support the development of accurate automated methods to assess and manage DQ and address semantic interoperability. There is increasing work on ontology-based approaches to the management of chronic disease, but little on ontological approaches to DQ in CDM

specifically or in health generally. This thesis therefore addressed two main research questions:

1. Can an ontology-based approach be developed and used to identify cases of T2DM in a dataset?
 - 1.1 Through a literature review on how ontology-based approaches have been used to identify patients with T2DM; and
 - 1.2 What is the most appropriate methodology to develop an ontology-based approach to identify cases of T2DM?
2. Can an ontology-based approach also contribute to the assessment of DQ required to identify T2DM cases accurately?
 - 2.1 How does the DQ impact on the accuracy of the ontology-based identification of patients with T2DM?
 - 2.2 How accurate is the ontology-based approach to identifying patients with T2DM from an EHR?

Therefore, the present study aims to develop an ontology-based approach to systematically and accurately identify diabetes in an EHR for inclusion in a disease register, as well as to assess the quality of the data in the register. It draws on current work into completeness and correctness of DQ of routinely-collected data from general practice in diabetes. The specific research hypothesis is:

“An ontological approach – as characterised by using clinical concepts, their properties and relations – will be flexible for the identification of patients with type 2 diabetes mellitus in a clinical database within the ePBRN data repository”.

This research also examines the role of ontology to represent and assess quality of the register.

1.2 Definitions and Description of Concepts

1.2.1 Chronic disease management

Chronic disease management (CDM) is “an intervention designed to manage or prevent a chronic condition using a systematic approach to care and potentially employing multiple treatment modalities” (Weingarten et al., 2002). There is growing recognition of the use of EHRs for CDM, public health services and epidemiological

research (Choquet et al., 2010). The large and increasing amount of potentially relevant health and health services data collected as part of routine practice compounds the DQ challenge. In this research we only focused on the development of the ontology based approach in diabetes management.

1.2.2 Data quality (DQ)

The generally accepted definition of DQ is the International Standards Organisation (ISO) 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). Data quality also is defined in terms of its ‘fitness for purpose’ (Richard Y Wang, 1998), in this case, to describe and evaluate the safety and quality of care (Appendix 1, page 132). However, there is no agreement on whether and how DQ should be managed and preserved (Preece, Missier, Ernbury, Jin, & Greenwood, 2008) in clinical information systems (CISs).

In each CIS, improving dimensions of DQ can improve the quality of decisions and lead to better policy, strategies, evidence-based care and patient outcomes. For instance, correctness of data refers to how data confirm with user requirements in reality (Y. Wand & Y. Wang, 1996) for valid and appropriate patients records (S. Liaw, et al., 2011). Also, some studies show that accuracy and completeness of data are aspects of information correctness (Hoorn & Wijngaarden, 2010). Kahn et al. (2002) demonstrate completeness as the extent to which information is not missing and is of sufficient breadth and depth for the task at hand (B. K. Kahn, Strong, & Wang, 2002). Incompleteness of data may cause choice of a wrong information system state during data production, resulting in incorrectness (Y. Wand & Y. Wang, 1996). The long list of dimensions of DQ described has been assessed within the context of DQ and provenance in the data life-cycle (Appendix 1, page 133). For example, multiple dimensions of DQ have been proposed, including “accuracy, perfection, freshness and uniformity” (Redman, 2005) and “completeness, unambiguity, meaningfulness and correctness” (Y. Wand & R. Y. Wang, 1996). One of the significant challenges is lack of consensus in the concepts and definitions of DQ (S. T. Liaw, et al., 2013); there is no agreement on whether, or how, these data should be managed and preserved in CISs.

The most commonly used DQ dimensions – the 3Cs and timeliness – are a potential starting point for a consensus (Appendix 1, page 135). In this thesis, a

definition of the 3Cs related to DQ is based on ‘fitness for our purpose’, which is the identification of patients with chronic diseases. Completeness was defined as the availability of at least 1 record per patient; correctness was defined as a valid and appropriate record – for example, that height is measured in meters and is within range for age; consistency was defined as both internal – using a uniform data type and format (e.g., integer, string, date) with a uniform data label – and external – using codes/terms that could be mapped to the Systematized Nomenclature of Medicine – Clinical Terms Australian version (SNOMED-CT-AU) (S. Liaw, et al., 2011; S. T. Liaw, et al., 2013).

1.2.3 Ontology-based Approach

1.2.3.1 Notion of ontology

In the philosophical domain, ontology is a very old concept first defined around the time of Aristotle. However since the 1990s, ontology has become increasingly attractive to various computing areas such as knowledge engineering, knowledge management, natural language processing, information retrieval and integration, cooperative information systems and agent-based system design (Gamper, Nejdl, & Wolpers, 1999; Ying, Wimalasiri, Ray, Chattopadhyay, & Wilson, 2010).

In computer science, ontologies have been proposed as a data model to represent the meaning of a scientific domain and support the sharing of domain knowledge between human and computer programs (Perez-Rey et al., 2006). In the biomedical informatics literature, ontologies have been described as “collections of formal, machine-process-able and human interpretable representation of the entities, and the relations among those entities, within a definition of the application domain” (Rubin et al., 2006), drawing on the general definition by Gruber - “an explicit, formal specification of a shared conceptualisation” (Gruber, 1995) - Pipino (2002) proposed the most widely accepted definition, considering ontologies as “an explicit specification of a conceptualization” (Pipino, Lee, & Wang, 2002). Explicit concepts and the relationships and constraints are clearly defined and understood by the user.

A formal ontology is computer-readable, allowing a computer to ‘understand’ the relationships – the ‘formal semantics’ – of the ontology. The ontology-based approach provides a vocabulary of terms, their meanings and relationships to be used in various application contexts (Appendix 2, page 145). Ontological methodologies from

computer science and engineering provide the technical means for formalising the depth of knowledge in CDM. The resulting ontology approach prepares the foundation for applications that support general practices and medical research to precisely develop quality of clinical information (Cur, 2012) and enable more accurate decision making (Kuziemytsky & Lau, 2010; Lin, Xiang, & He, 2011; Mabotuwana & Warren, 2009).

1.2.3.2 Motivations and advantages of the use of ontology

The importance of ontology in areas such as knowledge engineering, information retrieval and database design has been widely discussed (Tu, Tennakoon, O'Connor, Shankar, & Das, 2008; Uschold, 2005; Uschold, King, Moralee, & Zorgios, 1998). This research focuses on the importance of an ontology application in the identification of patients and assessment of their DQ in the context of diabetes mellitus type 2 (Appendix 4, page 157). Ontology has been widely recognised for its significant benefits to interoperability and reusability (Gilbert & Ddembe, 2008). One of the major benefits of ontology is that it provides a degree of interoperability (Ying, et al., 2010). Interoperability refers to the ability of heterogeneous components to interact and work with each other to achieve shared or individual goals; it involves not only communication between the heterogeneous components, but also the ability of these components to use exchanged information.

Ontologies have been proposed as a means to assure DQ through representing the meaning of a scientific domain and supporting the sharing of domain knowledge between human and computer programs. An ontology-based approach to DQ assessment and management uses a flexible and modular approach afforded by a constrained model with predefined concepts and relationships, including its own vocabulary versus a traditional data model. For example, Cur and others (2012) showed that one of the main advantages of using an ontology approach is the flexibility of those models to minimise a set of queries by exploiting the reasoning facilities of a description logics reasoner (Cur, 2012). Also, Grenon and others (2004) demonstrated the modularity of their ontology approaches, which allows them to be combined, and provide examples of their applications in biomedicine. Their approach applies the ideas of geospatial dynamics mapped onto the restricted domains of biomedical ontologies. In fact, their ontology-based approach combined an application of basic formal ontology (BFO) to the medical and geographical domain (Grenon et al., 2004).

Metadata for DQ specification and assessment in CDM can be accurately specified and then assessed against a DQ ontology constructed from a consensus framework and dimensions. For instance, Liaw and others showed that ontologies enabled the modelling of a domain and representation of data and metadata requirements to specify a unified context in collaborative environments, such as a clinical data warehouse (CDW) managing data from a number of CISs (Liaw et al., 2011). This allows intelligent software agents to act in spite of differences in concepts and terminology.

An ontology-based approach makes a knowledge base efficient: the definitions and properties of concepts can be extracted semantically to enrich queries. For example, Figure 1.1 illustrates how the concept diabetes management is defined with the help of hierarchies¹.

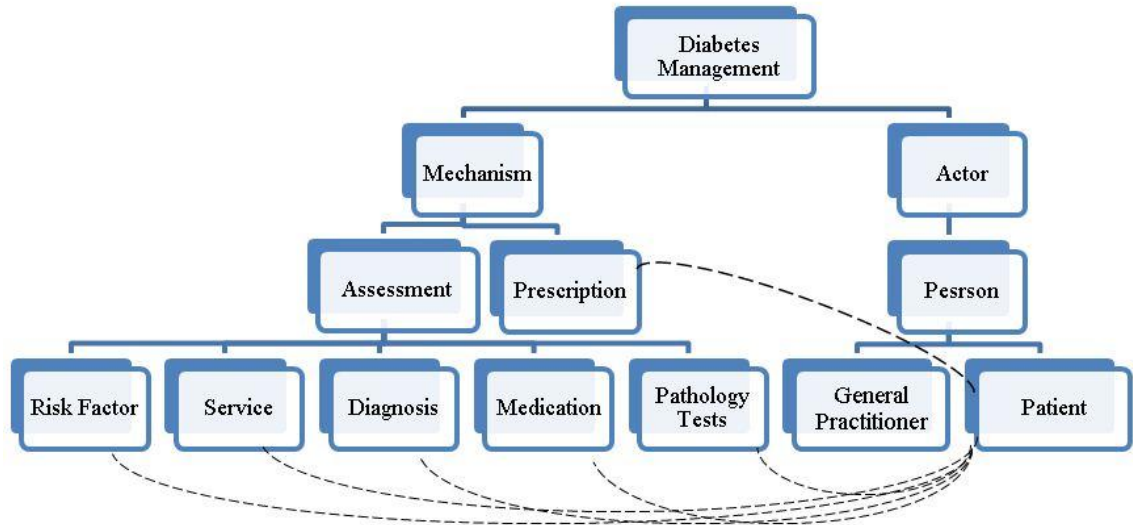


Figure 1.1: The definition of Diabetes Management concept with hierarchy and properties

In Figure 1.1, the concept *Diabetes Management* is defined by two related concepts: *Mechanism* and *Actor*. *Actor* shows some agents that perform as the main role

¹. In Figure 1.1, we list only concepts as an example of our conceptual model and assume that the concrete values (terms) defining them are embedded.

in diabetes management, such as *Person* and *Organisation*. *Mechanism* focuses on the main procedures of assessment and management of diabetes, including *Prescription*, *Referral*, *Services*, and *Assessment*, etc. The concept *Assessment* is further defined by five related concepts: *Risk Factor*, *Service*, *Diagnosis*, *Medication*, and *Pathology Tests*.

Although the example includes the same number of concepts, the knowledge base shown in Figure 1.1 defines not only the concept *Diabetes Management* and concepts *Assessment*, *Prescription* and *Person* but also shows different semantic relationships between *Patient* and other concepts, while the knowledge base in a non-ontological approach - like database schema - only focus on the concept itself without any semantic relationships between their relevant concepts and properties. Therefore, if a query is provided to identify and classify diabetes semantically, the knowledge about the *Patient* in the knowledge base in Figure 1.1 can be directly utilised. Moreover, the hierarchy in Figure 1.1 makes query expansion easier. Because the knowledge base contains hierarchies, it is necessary that queries are in hierarchical structures as well. This ensures that the concepts in queries can be naturally expanded with supplementary knowledge from the knowledge base.

Finally, an ontology provides chances to create more accurate criteria to efficiently identify diabetes from EHRs. For example, after a query that contains the concept *Diabetes Management* is expanded with the knowledge bases presented in Figure 1.1, it is translated into the following query (*q1*), which takes information from Figure 1.1. Let us assume the single concepts listed in the queries (*Patient*, *Diagnosis*, etc.) are already identified by the ontology-based model. To evaluate how well a patient matches a query, we have to aggregate identifications of the single concepts in the EHR. That is, *q1* includes three aggregations:

1. An aggregation of the concepts *Patient* and *General Practitioner* that contributes to the identification of the concepts *Person* and *Actor*;
2. An aggregation of the concepts *Diagnosis*, *Pathology Tests*, *Service*, *Medication* and *Risk Factor* that contributes to the identification of the concepts *Assessment* and *Mechanism*;
3. An aggregation of the concepts *Mechanism*, *Actor* and relationships that contributes to an identification of diabetic patients.

Furthermore, the ontology-based approach uses class attributes (data properties) to capture the correctness and consistency on valid clinical records in the dataset. The ePBRN research team created rules for the 3Cs of DQ, using the Australian National Guidelines and SNOMED-CT-AU to:

- Define data properties
- Use uniform data types and formats (e.g., integer, string, data) for each variable (for internal consistency)
- Define a uniform data format for each concept (e.g., for *Pathology Tests* as a sub-class, “has_HbA1C” is selected as the property of that class, “decimal” is selected as the type and a value range “from ≤ 3 mmol to ≥ 20 mmol” is entered for correctness)
- Select a standard label for each entity (e.g., use “type 2 diabetes mellitus” instead “T2DM” for external consistency).

It was also specified which classes are disjointed, so that an individual (or object) cannot be an instance of more than one of these, as that could lead to more consistency for the specification of DQ in ontology. The ontology-based approach also defined object properties (relationships between different classes and subclasses). The careful modelling of object properties in an ontology environment (like Protégé) can achieve completeness requirements.

However, the ontology-based model integrates with linguistic quantifiers and operators, expanding the possibilities of queries to express more complicated requirements for various purposes. Examples of linguistic quantifiers and operators are *some*, *most*, *filter* and *negations*. In Open World Assumption (OWA) operators that are applied to identify diabetes in our approach, linguistic quantifiers and operators are modelled. For instance, a query can be defined as ‘True Negative’ for diabetes diagnosis based on the reason for the patient visit. In this query, OPTIONAL and FILTER are used to identify non-diabetes without any diabetes reason for visit from the patient’s *Diagnosis* records.

1.3 Research Methodology

The methodology chosen in this thesis can be classified as design science, one of the two core prototypes that characterise much of the research in the health and medical informatics disciplines, the other being behavioural science (Hevner, March, Park, & Ram, 2004; March & Smith, 1995). The behavioural science research prototype seeks to develop and verify theories that explain or predict human/organisational behaviour surrounding the development and use of information systems, while the design science prototype seeks to create innovative artefacts through which the development and use of information can be effectively and efficiently accomplished. Artefacts can be broadly classified as methods (i.e., set of steps, guidelines or algorithms), models (i.e., abstractions and representations), constructs (vocabularies and symbols) and implementation (i.e., prototype systems) (Hevner, et al., 2004). This thesis aims to create two of these artefacts: method and model. The method will be the Methodology for Data Quality Ontology (MDQO) process itself, while the model will be the set of ontology-based query approaches that accompany the MDQO methodology.

March and Smith (1995) identified that a typical design science research should comprise two basic processes: build and validate. Build refers to the constructions of artefacts – i.e., the model and the method of MDQO. The validation process refers to the use of appropriate validation methods to assess the artefacts' performance. The validation of designed artefacts typically uses methodologies available in the knowledge base. The validation method used for this research will be a case study of the MDQO methodology, which will be conducted in both the identification of type 2 diabetes mellitus and the assessment of registers.

1.4 Overview of Thesis

This thesis comprises three published, peer-reviewed journal papers, with an accompanying introduction, discussion and conclusion. The remainder of the thesis is organised as follows.

Chapter 2 contains a published paper providing background and a literature review. This section provides necessary background knowledge for the work; it includes methodology for the systematic literature review, review questions, exclusion criteria for the literature, and categorisation of studies. The chapter systematically reviews

articles, published from 2000 to 2013, that are concerned with ontological approaches for the specification of DQ in health and other areas. The systematic review aimed to produce a framework of the ontology-based approach for DQ in CDM. Hence, included papers were critically appraised with a “context-mechanism-impacts/outcomes” model. The chapter also analyses the trends and the gaps in knowledge, and provides the background knowledge on semantic approaches in the fields of health, non-health and DQ specification in distributed clinical information systems (Rahimi, et al., 2014). This section also discusses some comparison studies between ontological and non-ontological approaches in CDM.

Chapter 3 proposes a rich methodological approach - called MDQO - to develop an ontology-based model by focusing on DQ based on ‘fitness for purpose’, specifically in the identification of patients with T2DM. This approach captures semantics of patients’ EHR data. The notion of a semantic model, as proposed in this chapter, is generalisable and reusable in others domain as well. There are five steps in the framework: knowledge acquisition, conceptualisation, semantic modelling, knowledge representation and validation.

1. ***Knowledge acquisition*** identifies the purpose and scope of the ontology to identify diabetes and develop quality of registers; it includes knowledge acquisition resources.
2. ***Conceptualisation*** describes a conceptual model of the ontology so that it meets the specifications.
3. ***Semantic modelling*** transforms the conceptual model into a formal model. It extends the selective codes by developing hierarchies and relationships for the ontology model. This stage develops three things: a domain ontology, sub-ontologies and problem-solving approaches. The domain and sub-ontologies represent the structure and relationships of the ontology while the problem-solving approaches provide specific solutions to domain problems.
4. ***Knowledge representation*** maps the ontology-based model and our sample data set then implements the formalised ontology in a knowledge representation language and semantic query languages, e.g., SPARQL.

5. **Validation** manually validates the accuracy of the ontology-based model for the identification of diabetes.

Chapter 4 presents findings of the validation of the ontology-based query approach that demonstrate how ontology results can be accurate for our purpose, to identify patients with T2DM in a database. The step-by-step calculations in the process are described in detail, including the ontology-based model to query, using query languages (e.g., SPARQL), the structured fields in the ePBRN data repository were iteratively tested and refined. The accuracy of the final ontology-based model is validated with a manual audit of the general practice EHRs. Tables show sensitivity and specificity (accuracy) of the model to diagnose T2DM, using the T2DM cases found by manual EHR audit as the gold standard. Accuracy was determined with three attributes – Reason for Visit (RFV), Medication (Rx) and Pathology (Path) – singly and in combination.

Chapter 5 discusses the findings and presents the outcomes derived from the work in this thesis. In chapter 6, conclusions are derived and the research contributions to the knowledge base and the field are mentioned, along with recommendations for future work.

The following thesis map shows the interrelationship of the contents and contributions from each chapter to the project and detailed discussions in Chapter 5.

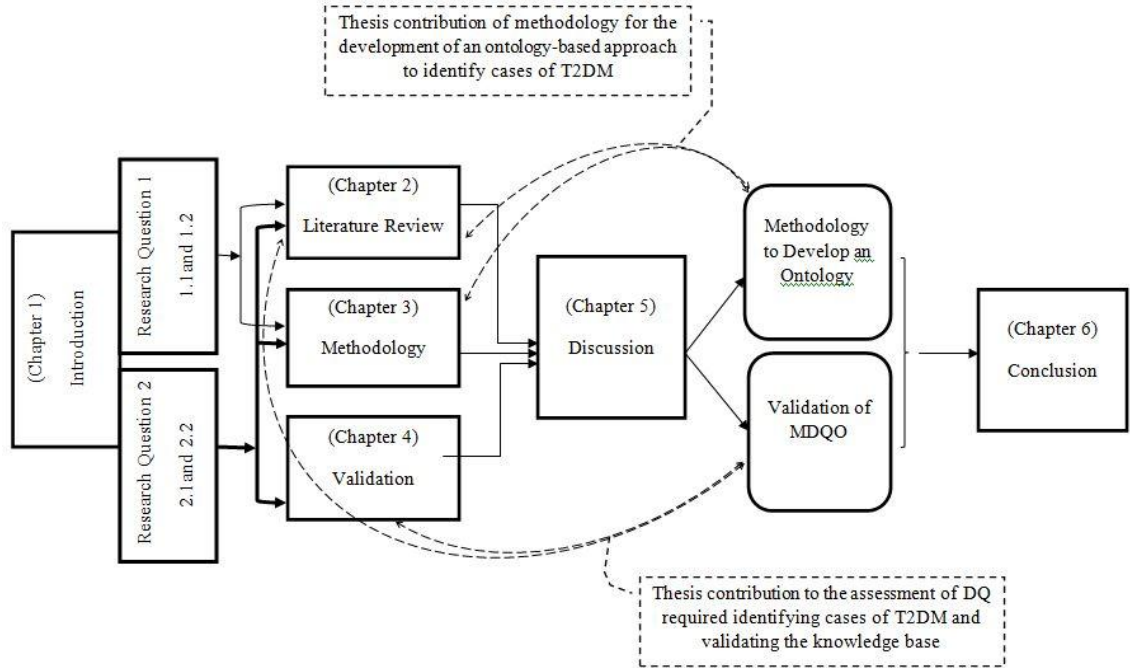


Figure 1.2: The thesis map

1.5 Summary

Data quality is a multidimensional concept, but lacks a consensus framework and definition. Routinely collected data are often incomplete, incorrect and inconsistent, with similar clinical concepts represented in different ways using non-standard terminology. Data quality issues are often identified only when the data are used to provide information in clinical reports from EHRs. Those analysing the data are the first to identify problems that highlight the data are not fit for use for health professionals' research purposes.

Ontologies can support DQ research, particularly because of their inherent ability to address semantic interoperability. Hence, the ontological approach provides a semantic framework into which patients' data for assessment and management of their diseases can be inputted, and risks, profiles and recommendations derived. The ontology-based approach can also support assessment of quality of clinical data used.

An ontology-based approach to identify patients with chronic diseases is a semantic approach with defined constraints, predefined concepts and relationships, including its own vocabulary. This ontology model takes queries, communicates with the knowledge base, and analyses patients' records through semantic annotation, and, in

this instance, identifies patients with diabetes by integrating semantic languages. The ontology-based approach also uses various functions such as terminology management, integration and sharing of data, and knowledge reuse and decision support. Ontologies must represent reality while having a sound theoretical foundation.

This research focuses on the development of an ontology-based approach for the identification of patients with diabetes, assessment of the quality of the registers created and assessment of applicability to diabetes management.

1.6 References

- Arts, D., de Keizer, N., Scheffer, G. J., & de Jonge, E. (2002). 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.
- Azaouagh, A., & Stausberg, J. (2008). [Frequency of hospital-acquired pneumonia--comparison between electronic and paper-based patient records]. *Pneumologie*, 62(5), 273-278.
- Choquet, R., Qouiyyd, 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.
- Cummings, E., Showell, C., Roehrer, E., Churchill, B., Yee, K., Wong, M., et al. (2010). *Discharge, Referral and Admission: A Structured Evidence-based Literature Review*.
- Cur, O. (2012). Improving the Data Quality of Drug Databases using Conditional Dependencies and Ontologies. *J. Data and Information Quality*, 4(1), 1-21.
- 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. *Diabet Med*, 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.
- Dixon, B., McGowan, J., & Grannis, G. (2011). *Electronic Laboratory Data Quality and the Value of a Health Information Exchange to Support Public Health Reporting Processes*. Paper presented at the AMIA 2011 Annual Symposium Improving Health: Informatics and IT Changing the World.
- Eccher, C., Purin, B., Pisanelli, D. M., Battaglia, M., Apolloni, I., & Forti, S. (2006). Ontologies supporting continuity of care: the case of heart failure. *Comput Biol Med*, 36(7-8), 789-801.
- Gamper, J., Nejd, W., & Wolpers, M. (1999). *Combining Ontologies and Terminologies in Information Systems*. Paper presented at the Proceedings of the 5th International Congress on Terminology and Knowledge Engineering, Innsbruck, Austria.

- Gilbert, M., & Ddembe, W. (2008). A Flexible Approach for User Evaluation of Biomedical Ontologies. *International Journal of Computing and ICT Research*, 2(2), 62-74.
- Gillies, A. (2000). Assessing and improving the quality of information for health evaluation and promotion. *Methods Inf Med*, 39(3), 208-212.
- Grenon, P., Smith, B., & Goldberg, L. (2004). Biodynamic ontology: Applying BFO in the biomedical domain. In D. M. Pisanelli (Ed.), *Ontologies in Medicine* (Vol. 102, pp. 20-38).
- Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Human-Comput. Stud*, 43(5-6), 907-928
- Hamilton, W. T., Round, A. P., Sharp, D., & Peters, T. J. (2003). The quality of record keeping in primary care: a comparison of computerised, paper and hybrid systems. *Br J Gen Pract*, 53(497), 929-933; discussion 933.
- Hevner, A. R., March, J., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 31.
- Hoorn, H. F., & Wijngaarden, T. D. (2010). Web Intelligence for the Assessment of Information Quality: Credibility, Correctness, and Readability In Z.-U.-H. Usmani (Ed.), *Web Intelligence for the Assessment of Information Quality* (pp. 305): InTech.
- Huaman, M. A., Araujo-Castillo, R. V., Soto, G., Neyra, J. M., Quispe, J. A., Fernandez, M. F., 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 Medical Informatics and Decision Making*, 9.
- Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information quality benchmarks: product and service performance. *Communications of the ACM*, 45(4), 8.
- Kahn, M. G., Batson, D., & Schilling, L. M. (2012). Data model considerations for clinical effectiveness researchers. *Med Care*, 50 Suppl, S60-67.
- Kuziemsky, C., & Lau, F. (2010). A four stage approach for ontology-based health information system design. *Artificial Intelligence in Medicine* 50, 133-148.
- Lain, S. J., Roberts, C. L., Hadfield, R. M., Bell, J. C., & Morris, J. M. (2008). How accurate is the reporting of obstetric haemorrhage in hospital discharge data? A validation study. *Australian & New Zealand Journal of Obstetrics & Gynaecology*, 48(5), 481-484.
- Liaw, S., Chen, H., Maneze, D., Dennis, S., & Vagholkar, S. (2011). Use of the "principal diagnosis" in emergency department databases to identify patients with chronic diseases (in press). *Electronic Health Informatics Journal*.
- 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)*. Paper presented at the AMIA 2011 Annual Symposium Improving Health: Informatics and IT Changing the World.
- Liaw, S. T., 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. *Int J Med Inform*, 82(1), 10-24.
- Lin, Y., Xiang, Z., & He, Y. (2011). Brucellosis Ontology (IDOBUR) as an extension of the Infectious Disease Ontology. *J Biomed Semantics*, 2(1), 9.
- 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.

- March, S., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, 15, 15.
- Michalakidis, G., Kumarapeli, P., Ring, A., van Vlymen, J., Krause, P., & de Lusignan, S. (2010). A system for solution-orientated reporting of errors associated with the extraction of routinely collected clinical data for research and quality improvement. *Stud Health Technol Inform*, 160(Pt 1)), 724-728.
- MIE Conference Proceeding: *Quality of Life through Quality of Information*. (2012, 29 August 2012). Paper presented at the 24th International Conference of the European Federation for Medical Informatics (MIE), Pisa, Italy.
- Mitchell, J., & Westerduin, F. (2008). Emergency department information system diagnosis: how accurate is it? *Emerg Med J*, 25(11), 784.
- Moro, M. L., & Morsillo, F. (2004). Can hospital discharge diagnoses be used for surveillance of surgical-site infections? *Journal of Hospital Infection*, 56(3), 239-241.
- 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. *Comput Biol Med*, 36(7-8), 712-730.
- Pipino, L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Association for Computing Machinery. *Communications of the ACM* 45(4), 211-218.
- Poulsen, M. K., Henriksen, J. E., Vach, W., Dahl, J., Møller, J. E., Johansen, A., et al. (2010). Identification of asymptomatic type 2 diabetes mellitus patients with a low, intermediate and high risk of ischaemic heart disease: is there an algorithm. *Diabetologia*, 53(4), 659-667.
- 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 & Experience*, 20(3), 253-264.
- Quan, H., Li, B., Saunders, L. D., Parsons, G. A., Nilsson, C. I., 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.
- Rahimi, A., Liaw, S., Ray, P., Taggart, J., & Yu, H. (2014). Ontological specification of quality of chronic disease data in EHRs to support decision analytics: a realist review. *Decision Analytics*, 1(5), 31.
- Redman, T. (2005). Measuring data accuracy. In R. e. a. Wang (Ed.), *Information Quality* (Vol. 1, pp. 21). Armonk NY: ME Sharpe Inc.
- Richesson, R. L., & Krischer, J. (2007). Data standards in clinical research: Gaps, overlaps, challenges and future directions. *J Am Med Inform Assoc*, 14(6), 687-696.
- Rubin, D. L., Lewis, S. E., Mungall, C. J., Misra, S., Westerfield, M., Ashburner, M., et al. (2006). National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *OMICS*, 10(2), 185-198.
- Terzi, E., Vakali, A., & Hacid, M.-S. (2003). Knowledge Representation, Ontologies, and the Semantic Web. In X. Zhou, M. Orłowska & Y. Zhang (Eds.), *Web Technologies and Applications* (Vol. 2642, pp. 382-387): Springer Berlin Heidelberg.
- Tu, S., Tennakoon, L., O'Connor, M., Shankar, R., & Das, A. (2008). *Using an integrated ontology and information model for querying and reasoning about phenotypes: The case of autism*. Paper presented at the AMIA Annu Symp Proc.
- Uschold, M. (2005). An ontology research pipeline. *Applied Ontolog*, 1(1).

- Uschold, M., King, M., Moralee, S., & Zorgios, Y. (1998). The enterprise ontology. *Knowledge Eng Rev* 13(1), 31-89.
- Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11 (Nov)), 86-95.
- Wand, Y., & Wang, Y. (1996). Anchoring Data Quality Dimensions in Ontological Foundations. *COMMUNICATIONS OF THE ACM*, 36(11), 86-95.
- Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, 41(2 (Feb)), 58-65.
- Wang, R. Y., Strong, D. M., & Guarascio, L. M. (1996). Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 5-33.
- Weingarten, S. R., Henning, J. M., Badamgarav, E., Knight, K., Hasselblad, V., & Gano, A. (2002). Interventions used in disease management programmes for patients with chronic illness which ones work? Meta-analysis of published reports. *BMJ* 325(7370), 925-928.
- Ying, W., Wimalasiri, J., Ray, P., Chattopadhyay, s., & Wilson, C. (2010). An Ontology Driven Multi-Agent Approach to Integrated e-Health Systems *International Journal of E-Health and Medical Communications (IJEHMC)*, 1(1), 28-40.

CHAPTER 2

ONTOLOGICAL SPECIFICATION OF QUALITY OF CHRONIC DISEASE DATA IN ELECTRONIC HEALTH RECORDS TO SUPPORT DECISION ANALYTICS: A REALIST REVIEW

Chapter 1 introduced the notion of ontology and DQ as important research topics. Chapter 2 will expand on these notions with a detailed specification of DQ and the role of ontology-based approaches to develop DQ based on ‘fitness for purpose’ within the health context.

Chapter 2 reviews the related literature and assesses the gaps in the domains of interest. The lack of comprehensive ontological approaches for DQ based on the ‘fitness for purpose’ specifically or in health generally is an important research gap and needs to be addressed. Compared with non-hierarchical data models, there may be more advantages and benefits in the use of ontologies to solve clinical DQ issues semantically and improve the validity and reliability of data retrieval, collection, storage, extraction and linkage algorithms and tools.

Theoretically, ontology-based applications could support automated processes to address DQ and semantic interoperability in the health area. The current evidence also supports moving to the ontology-based design of information systems to enable more flexible use of clinical data. Chapter 2 will guide the development of a DQ ontology “fitness for specific purpose” in CDM. The published paper summarises the ontological specification of the quality of data in EHRs to support decision analytics.

NOTICE: This is the final version of the paper that was published in “Decision Analytics”. Changes resulting from the publishing process, such as structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published as:

Alireza Rahimi, Siaw-Teng Liaw, Pradeep Ray, Jane Taggart and Hairong Yu (2014). Ontological specification of quality of chronic disease data in electronic health records to support decision analytics: A realist review. *Decision Analytics*, 1(5), 1-31. doi: 10.1186/2193-8636-1-5. url: <http://dx.doi.org/10.1186/2193-8636-1-5>

Ontological Specification of Quality of Chronic Disease Data in Electronic Health Records to Support Decision Analytics: A Realist Review

Alireza Rahimi^{1,2,5}, alireza.rahimikhorzoughi@unsw.edu.au

Siaw-Teng Liaw^{1,3,4*}, siaw@unsw.edu.au

Pradeep Ray⁵, p.ray@unsw.edu.au

Jane Taggart^{1,3}, j.taggart@unsw.edu.au

Hairong Yu³, hairong.yu@unsw.edu.au

¹ UNSW School of Public Health & Community Medicine, Sydney, Australia

² Isfahan University of Medical Sciences, Faculty of Management and Medical Information Sciences, Isfahan, Iran

³ UNSW Centre for Primary Health Care & Equity, Sydney, Australia

⁴ South Western Sydney Local Health District (SWSLHD) General Practice Unit, Sydney, Australia

⁵ UNSW Asia Pacific Research Centre for Ubiquitous Healthcare, Sydney, Australia

* Corresponding author. UNSW/SWSLHD School of Public Health & Community Medicine, General Practice Unit, PO Box 5, Fairfield, NSW 1860 Sydney, Australia

University of NSW
Authorship Declaration

In the case of the paper “Ontological specification of quality of chronic disease data in EHRs to support decision analytics: a realist review”, contributions to the work involved the following:

Name	Nature of contribution
Alireza Rahimi	Development of the conceptual framework and templates for the literature review, data collection, critical appraisal of studies, analysis and interpretation of papers and drafting of the manuscript
Siaw-Teng Liaw	Development of the conceptual framework and templates for the literature review, critical appraisals of papers and critical review of the manuscript
Pradeep Ray	Development of the conceptual framework and templates for the literature review, and critical review of the manuscript
Jane Taggart	Discussion of her critical appraisals with AR to achieve consensus, and critical review of the manuscript
Hairong Yu	Discussion of her critical appraisals with AR to achieve consensus, and critical review of the manuscript

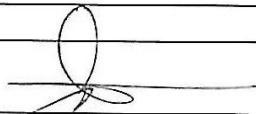
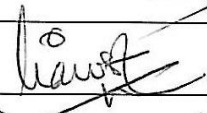
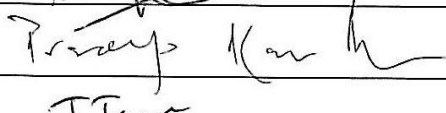
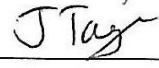
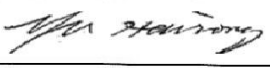
Declaration by co-authors

The undersigned hereby certify that:

- 1) they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field or expertise;
- 2) they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
- 3) there are no other authors of the publication according to these criteria;
- 4) potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
- 5) the original data are stored at the following location(s) and will be held for at least five years from the date indicated below:

Location:

General Practice Unit, Hospital Fairfield, UNSW, Sydney, Australia

Name	Signature	Date
Alireza Rahimi		09/05/2014
Siaw-Teng Liaw		09/MAY/2014
Pradeep Ray		13/05/2014
Jane Taggart		13/5/2014
Hairong Yu		15/5/2014

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 to March 2013, which addressed data/information quality, ‘fitness for purpose’ of data, and 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.

2.1 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 previous 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 professionals 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?

2.2 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) conceptualizes 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 (ontology-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:

1. (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
2. (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
3. ((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 Figure 2.1).

Template: Systematic review of ontological approaches to creating patient registers and assessing their quality

Research questions:

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

	Author / title / reference	Study type ¹	Context & Population studied	Aims of project being reported on	Details of terminology, DQ models & ontology	Methods / Tools used in project	Results / Outputs of project	Critical Appraisal: 1. quality ² of methods & tools 2. relevance to review questions
1								

¹ This classification of study types has been developed to cover the pattern of R&D in this multidisciplinary field. There are 5 types broadly based on the stage in the development lifecycle:

1. Requirements analysis e.g. literature reviews, qualitative research, etc
2. Design and tools development: data/information models and ontologies
3. Implementation, deployment and testing of information systems
4. Evaluation: descriptive evaluation
5. Evaluation: comparative (e.g. pre and post, time series, etc) with/without contemporary control (e.g. RCT)

This matrix (study types X current column headings) will focus the analysis and synthesis of the literature review e.g. by study types, methods, tools, outputs and impacts.

² The quality appraisal will include traditional methods of critical appraisal: validity (internal and external), reliability, generalizability, relevance, etc of the research methods, tools and measurements

Figure 2.1: Template used to analyze papers

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.

2.3 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 Full text, 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 2.1 summarized 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 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 a 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 2.1 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 2.2 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.

Table 2.1: 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, Decision Science, Engineering, Mathematics, Psychology, Social Sciences, and Medicine	All fields	Journal articles	75
Scopus	Computer Science, Health Science, Medical Informatics, Bioinformatics, Information Technology, Psychology, Social and Behavioural Sciences	AI fields	Journal articles	76
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

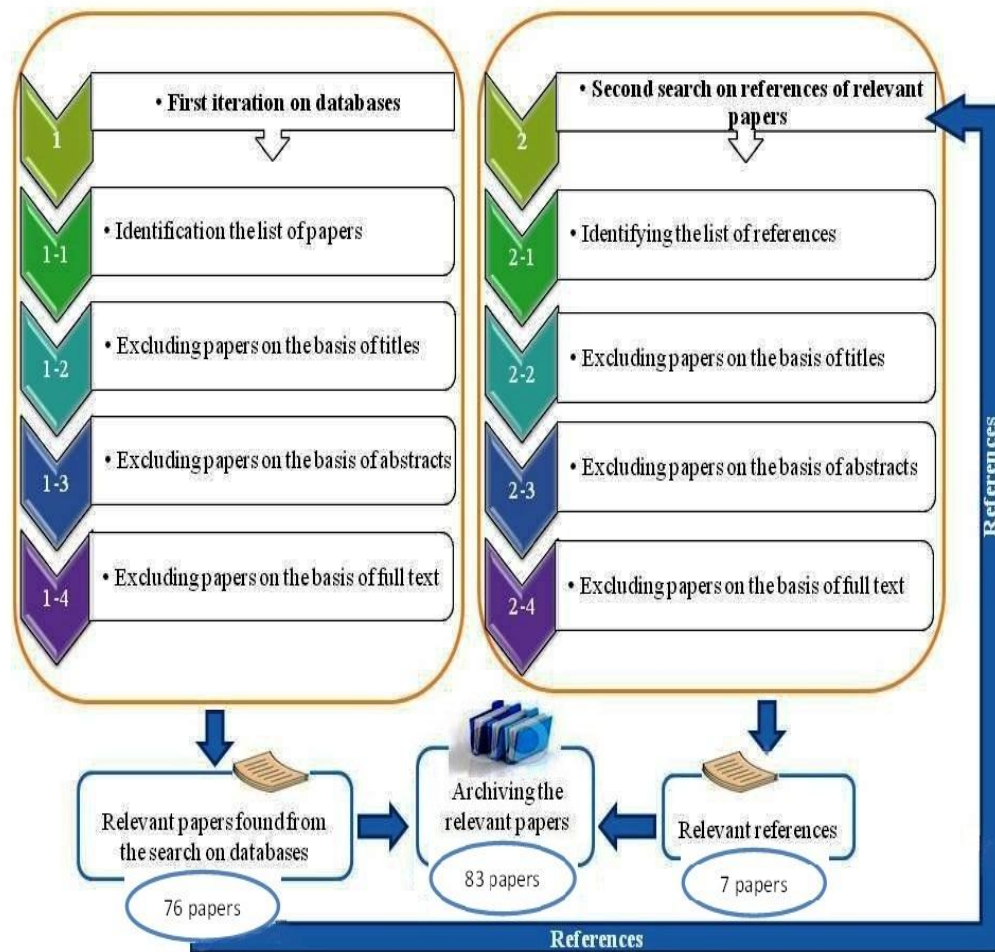


Figure 2.2: Paper selection process

It can be seen from the data in Table 2.2 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 2.2: 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.

Table 2.3 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; 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.

Table 2.3: 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	16%	41	49%
3. Implementation, deployment and testing of information systems	32	38%	2	3%	3	4%	5	6%	22	27%
4. Evaluation: descriptive evaluation of DQ or ontology in health area	17	20%	1	1%	2	3%	2	3%	12	15%
5. Evaluation: comparative +/- contemporary control (e.g., RCT)	2	3%	0	0	0	0	1	1%	1	1%

Figure 2.3 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 perfect to be “fit for purpose”.

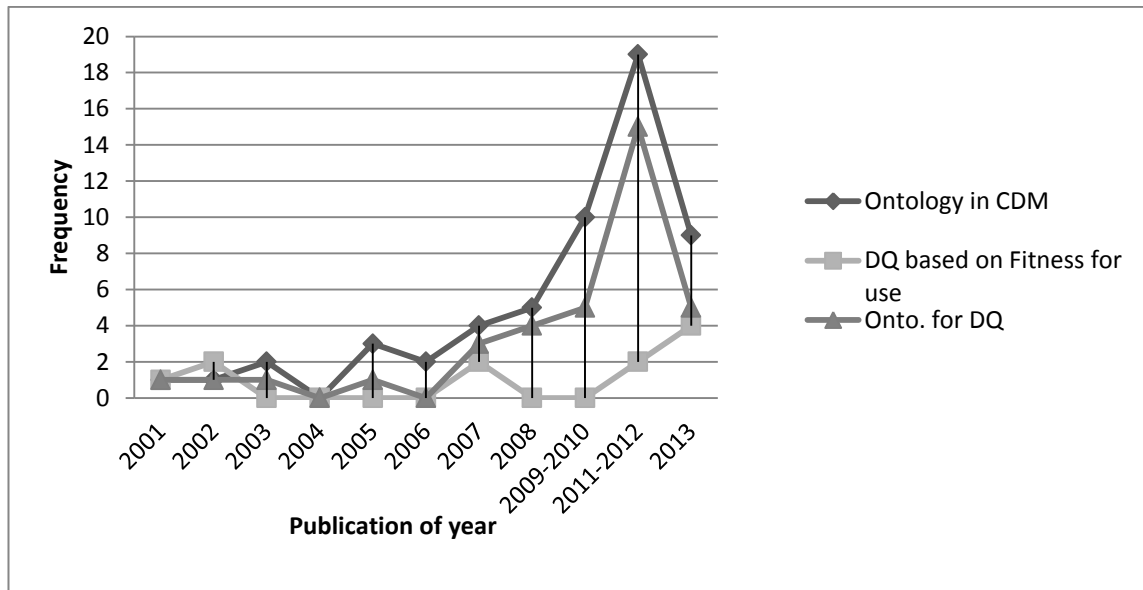


Figure 2.3: Distribution of papers from each category by year

The Figure 2.4 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 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.

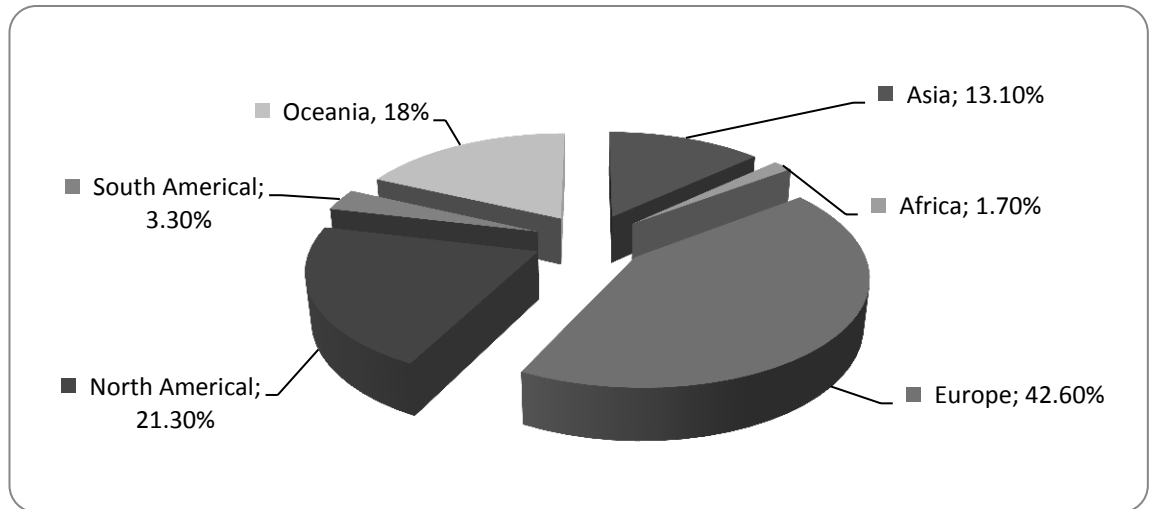


Figure 2.4: Distribution of papers found by continent

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).

2.3.1 Conceptualisation of data quality within the ‘fitness for purpose’ paradigm

Table 2.4 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).

Table 2.4: Papers where data quality was conceptualized within ‘fitness for purpose’ paradigm

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	<ul style="list-style-type: none"> -Use of popular search engines like OpenSearch to help in assessing fitness for purpose -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) -Theoretical discussion on accuracy and completeness of data 	<p>Defined fitness for purpose of data based on users (experts and non-experts in geo-informatics) satisfaction from search results</p> <p>Allowed users without specific expertise to conduct free form search requests in their own language</p>
(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	<ul style="list-style-type: none"> -Use the Quality Information Management Model = QIMM -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 -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.) 	<p>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</p> <p>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</p>
(Kahn, et al., 2012)	Clinical dataset in US	To develop the efficacy of their data model in three large healthcare organizations	<ul style="list-style-type: none"> -Use a two-by-two conceptual model (PSP/IQ) for describing IQ -Focus on 8 dimensions of data quality (completeness, correctness, flexibility, etc.) -Surveyed 45 professionals to determine which IQ dimensions belong in each quadrant of the model 	<p>This is a well-grounded, logical approach and a case study to indicate health organizations need to use “fitness of use” to determine IQ (specifically soundness, dependable, useful and usable information) for analytical purposes</p> <p>This assessment of DQ provides a reasonable baseline for determining what improvements</p>

Reference	Context	Aims of project	Methods/Tools used in project	Results
			-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	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 -More focus on data timeliness -Use of mathematics software	Demonstrated "fitness for purpose" of data for agents to choose how much information to gather from others (personal communication from an anonymous reviewer) 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 (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 data cleansing and analysis -Focus on correctness, completeness and consistency of clinical data	They used a well-designed framework to describe the intrinsic DQ (correctness and consistency) and fitness for purpose (completeness) for research and clinical purposes 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
(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 -Classify records as paper, computer, or hybrid, depending on which medium stored the clinical information from consultations by descriptive statistics -Focus on completeness of data	Define more completeness of data as a fitness for consolation in primary care Hybrid systems of primary care record keeping document higher numbers of consultations than computer-only or paper-only systems

Many studies regularly report a range of deficiencies in the collected information for professionals requirements (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 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).

2.3.2 Methodologies to specify data quality for implementation

Table 2.5 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.

Table 2.5: 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. 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.	Health information
(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. 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	Clinical data
(Liaw et al., 2011)	✓	✓	✓			They used a well-designed framework to describe the intrinsic DQ (correctness and consistency) and fitness for purpose (completeness)	Clinical data

Study types	1	2	3	4	5	Summary and Results of methodologies	Contexts
Reference							
						for research and clinical purposes 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. However, they did not compare data quality before and after the implementation of procedures to improve the accuracy of data	Clinical data
(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 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.	Health web pages
(Kahn, et al., 2002)	✓					Developing a two-by-two conceptual model for describing IQ (PSP/IQ) 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.	Health information
(Britt, et al., 2007)	✓	✓				Use statistical methods to manage data quality using SAS as a computer program in statistical package. Measure representativeness, reliability, validity and accuracy of BEACH data e.g., 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)	Clinical data

Study types	1	2	3	4	5	Summary and Results of methodologies	Contexts
Reference							
(Chen, 2009)	✓	✓				Focus on a full mathematical analysis (mathematical software) Investigate the effect of quality of information and amount of information are used interchangeably in the health behaviour e.g., decision making.	Infectious diseases
(Choquet, et al., 2010)		✓				Use Talend Open Studio open source software as well as developed stored procedures in SQL for the object quality criteria. 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.	Hospital dataset
(Cunningham-Myrie et al., 2008)			✓			Use ICD-10 for coding various collected data and to facilitate comparability of standardized data. 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	Health information
(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. Use training program: 12 week prospective study with training program for reporting personnel. Randomised selection to phone, visit or control for their supervisions. The training improved report timeliness but did not have such impact on data quality.	Infectious disease surveillance
(Kiragga, et al., 2011)	✓	✓	✓			Use the Research Cohort database as the reference “gold standard” for the assessment of data accuracy. 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	Infectious diseases

Study types	1	2	3	4	5	Summary and Results of methodologies	Contexts
Reference							
						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)	✓	✓				<p>Use a decision support example around a hypothetical patient called John who experiences an exacerbation of his COPD.</p> <p>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.</p> <p>Present a model for the prediction and evaluation of quality of information to a multi criteria decision making process.</p> <p>Model describes a decision support tool for use in the management of COPD.</p>	Clinical Guidelines (CG) for COPD

Notes for study types: See Table 2.2 for legend.

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.

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

Table 2.6 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 base. However, the paper was not explicit about whether was a formal outcome-based comparison of ontological and non-ontological approaches.

Table 2.6: Studies that compared ontologies and other data models in specification and implementation

Reference	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 (hyper-tension)

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 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.

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

As Table 2.7 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.

Table 2.7: The impact of implemented ontologies for the management of data quality

References	Defined Purpose	Assessed of Fitness for Purpose using DQ and Findings	Context
(Li and Ko, 2007)	To develop automated ontology approach to manage nutrients in a diabetes diet care knowledge management	<ul style="list-style-type: none"> -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	<ul style="list-style-type: none"> -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	-Simulate human body disorders into metadata through ontology-based data warehouse modelling	Human body anatomy and

References	Defined Purpose	Assessed of Fitness for Purpose using DQ and Findings	Context
	handling increasing amounts of clinical information and solves some issues related to managing large	<ul style="list-style-type: none"> -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 	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	<ul style="list-style-type: none"> -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 inconsistency data 	3000 records registered for the prostate cancer patients and Tumour Registry in US
(Brüggemann and Grüning, 2009)	To improve the outcome of data quality management (DQM)	<ul style="list-style-type: none"> -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	<ul style="list-style-type: none"> -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	<ul style="list-style-type: none"> -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

References	Defined Purpose	Assessed of Fitness for Purpose using DQ and Findings	Context
(Lee et al., 2009)	To classify a person as a diabetic patient	<ul style="list-style-type: none"> -Represent new ontology methods for fuzzy medical relationship using taxonomical knowledge in Taiwan -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 	Diabetes domain
(O'Donoghue, et al., 2009)	To demonstrate the data quality benefits of integrating remote patient monitoring solutions	<ul style="list-style-type: none"> -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

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 2.8 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).

Table 2.8: The impact of implemented ontologies for the assessment of data quality

References	Defined Purpose	Assessed of Fitness for Purpose using DQ	Context
(Jacquelinet et al., 2003)	To develop semantic data interoperability	<ul style="list-style-type: none"> -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 - 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 	Failure, dialysis and transplant datasets from National information system in France
(Maragoudakis, et al., 2008)	To develop decision support system	<ul style="list-style-type: none"> -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 -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 	Mobile sensor data from 25 patients in Artificial Neural Network (ANN) in GREECE
(Wang, et al., 2007)	To classify diabetic patients	<ul style="list-style-type: none"> 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 -Authors state that ontology approach can classify effectively classify a person as a diabetic patient for medical staff 	Retrieve 392 cases from the Pima Indians diabetes database in US
(Valencia-Garcia, et al., 2008)	To develop retrieval and extract clinical information	<ul style="list-style-type: none"> -Represent multiple semantic relationships among concepts with UMLS ancestors through MeSH descriptors in the ontology to enrich the ontology extracted from the text -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) 	Use breast cancer domain in the system with a Spanish corpus of 8649 words in Spain

References	Defined Purpose	Assessed of Fitness for Purpose using DQ	Context
		-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	<ul style="list-style-type: none"> -Use the querying capabilities of one GP database in the context of quality use of medications in the management of hypertension over time -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 	CVD in practice management system in NZ
(Young, et al., 2009)	To develop semantic data collection and integration	<ul style="list-style-type: none"> -Use the modelling of terms to conform to and extend the existing ontologies development framework -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 	Data on Autism in the National Database for Autism Research in US
(Preece, et al., 2008)	To manage information quality (IQ) in a real-life example of gene expression research	<ul style="list-style-type: none"> - 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) - 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 	Gene expression data which involve the use of microarrays in UK

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

Table 2.9 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.

Table 2.9: 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	Kahn, et al., 2012
	Easily manage the changes of the database schema or the ontology	Pannarale, et al., 2012	“adaptability” as defined operationally below.	
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

Criteria	Metrics for ontology evaluation	References	Metrics for data model evaluation	References
			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		
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

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. Standardizing these metrics can help to standardize the specification of ontologies and data models. This can then standardize the comparison of ontology and non-ontology approaches.

2.4 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.

2.4.1 How is data quality being conceptualised 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.

2.4.2 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.

2.4.3 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 studies comparing to addressing chronic disease management 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.

2.4.4 How is the impact of implementing ontology-based specifications for data quality in CDM 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 and extrinsic quality 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).

2.4.5 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 clinical information systems in which applicable by health practicing managers.

2.4.6 Managerial implications

The findings of this study have several important practical implications for developing electronic health records and patients registers. 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.

2.5 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.

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.

Acknowledgments

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

2.6 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, Qouiya, 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. 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*. 2009;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, Castelnovo, 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 retrieval 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). *AMIA 2011 Annual Symposium Improving Health: Informatics and IT Changing the World*; October 22-26, 2011; Washington DC, US. Washington DC: AMIA, 2011. p. 785-94.
- 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, Mitra, 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 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.

CHAPTER 3

DEVELOPMENT OF A METHODOLOGICAL APPROACH FOR DATA QUALITY ONTOLOGY IN DIABETES MANAGEMENT

Chapter 2 found a lack of comprehensive ontology-based approaches to address DQ and semantic interoperability issues. It also defined what is needed in a comprehensive ontology-based approach to DQ. Therefore, Chapter 3 discusses a step-by-step process to develop a novel methodology for data quality ontology (MDQO) to produce a semantic knowledge management approach to identify T2DM and assess the accuracy of ontology.

The published paper in Chapter 3 discusses how T2DM patients are identified. DQ is assessed using the three core dimensions of DQ, namely completeness, correctness and consistency. Chapter 3 will present the intuitions as well as the formalism for a semantically accurate mechanism for capturing DMO-related data from EHRs.

The longer term objective is to develop a flexible, generalisable and reusable semantic approach and mechanism that can be used to design intelligent software agents to identify patient cohorts and the quality of the data.

NOTICE: This is the author's version of a work that was published in the "International Journal of Electronic Health and Medical Communication". Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published as: Alireza Rahimi, Nandan Parameswaran, Pradeep Ray, Jane Taggart, Hairong Yu and Siaw-Teng Liaw (2014). Development of a methodological approach for data quality ontology in diabetes management. *International Journal of Electronic Health and Medical Communication*. 5 (3).

Development of a Methodological Approach for Data Quality Ontology in Diabetes Management

Alireza Rahimi^{1, 2, 3}, alireza.rahimikhorzoughi@unsw.edu.au

Nandan Parameswaran^{3, 4}, paramesh@cse.unsw.edu.au

Pradeep Ray³, p.ray@unsw.edu.au

Jane Taggart⁵, j.taggart@unsw.edu.au

Hairong Yu⁵, hairong.yu@unsw.edu.au

Siaw-Teng Liaw^{*1,3,5}, siaaw@unsw.edu.au

¹ UNSW, School of Public Health & Community Medicine, Sydney, Australia

² Isfahan University of Medical Sciences, Health Information Research Centre, Isfahan, Iran

³ UNSW, Asia-Pacific Ubiquitous Healthcare Research Centre, Sydney, Australia

⁴ UNSW, School of Computer Science and Engineering, Sydney, Australia

⁵ UNSW, Centre for Primary Health Care & Equity, Sydney, Australia

* Corresponding author. UNSW/SWSLHD School of Public Health & Community Medicine, General Practice Unit, PO Box 5, Fairfield, NSW 1860 Sydney, Australia

University of NSW
Authorship Declaration

In the case of the paper "Development of a Methodological Approach for Data Quality Ontology in Diabetes Management", contributions to the work involved the following:

Name	Nature of contribution
Alireza Rahimi	Conception and design, conduct of the studies, analysis and interpretation of data and drafting of the manuscript
Nandan Parameshwaran	Advice on study design, analysis, interpretation and critically reviewed draft of the manuscript
Pradeep Ray	Advice on study design, analysis, interpretation and critically reviewed of the manuscript
Jane Taggart	Advice on study design and critically review of the manuscript
Hairong Yu	Advice on study design and critically review of the manuscript
Siaw-Teng Liaw	Advice on ontology development and testing, study design and critically review of the manuscript


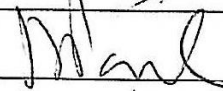
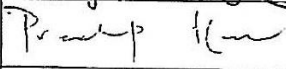
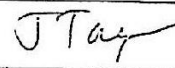
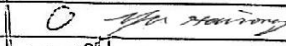
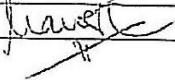
Declaration by co-authors

The undersigned hereby certify that:

- 1) they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field or expertise;
- 2) they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
- 3) there are no other authors of the publication according to these criteria;
- 4) potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
- 5) the original data are stored at the following location(s) and will be held for at least five years from the date indicated below:

Location:

General Practice Unit, Hospital Fairfield, UNSW, Sydney, Australia

Name	Signature	Date
Alireza Rahimi		09, 05/2014
Nandan Parameshwaran		15 May 2014
Pradeep Ray		13/05/2014
Jane Taggart		13/5/2014
Hairong Yu		15/5/2014
Siaw-Teng Liaw		09 MAY 2015

Abstract

The role of ontologies in chronic disease management and associated challenges such as defining data quality (DQ) and its specification is a current topic of interest. In domains such as Diabetes Management, a robust Data Quality Ontology (DQO) is required to support the automation of data extraction semantically from Electronic Health Record (EHR) and access and manage DQ, so that the data set is fit for purpose. A five steps strategy is proposed in this paper to create the DQO which captures the semantics of clinical data. It consists of: (1) Knowledge acquisition; (2) Conceptualization; (3) Semantic modeling; (4) Knowledge representation; and (5) Validation. The DQO was applied to the identification of patients with Type 2 Diabetes Mellitus (T2DM) in EHRs, which included an assessment of the DQ of the EHR. The five steps methodology is generalizable and reusable in other domains.

Keywords: Ontological approach, Semantic model, Data quality, Diabetes management.

3.1 Introduction

Improving data quality (DQ) in health organizations can improve quality of decisions and support better policy, strategies, and evidence-based patient care. DQ can be defined in terms of its *fitness for purpose* (Richard Y Wang, 1998). The most frequently used DQ dimensions are *accuracy*, *completeness*, *consistency*, *correctness* and *timeliness* (S. T. Liaw, et al., 2013). Research in DQ has tended to focus on the identification of generic quality characteristics that are applicable in a wide range of domains (Y. Wand & Y. Wang, 1996).

In the field of healthcare, data is collected routinely and may be used for research. It is becoming apparent that the quality of routinely collected data is not as good as it should be for many research applications. It is still not clear how DQ can be expressed in the context of fitness for purpose. Reference terminologies and ontologies have been used to specify DQ thus influencing data collection and analysis (Brown, Warmington, Laurence, & Prevost, 2003). They also act as benchmarks for assessing DQ (S. Liaw, et al., 2011). An ontological approach can play a major role in the assessment of DQ and specification of fitness for purpose of a dataset (S. T. Liaw, et al., 2013; Rahimi, et al., 2014).

Building robust ontologies for DQ in healthcare helps automation of data extraction from the Electronic Health Records (EHRs) into clinical data warehouses; assessment and management of the quality of big data so that they are fit for purposes such as research, quality improvement, health information exchange and sharing; management of controlled vocabularies and optimizing 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 finally reuse of knowledge in the management of chronic diseases (Y. Wand & Y. Wang, 1996).

In the biomedical informatics literature, ontologies have been described as collections of formal, machine process-able and human interpretable representation of the entities, and the relations among those entities, within a definition of the application domain (Rubin, et al., 2006). Pipino (2002) proposed the most widely accepted definition, where he considers ontologies as an explicit specification of a conceptualization (Pipino, et al., 2002). Ontology provides a vocabulary of terms, their

meanings and relationships to be used in various application contexts (Borst, 1997). This allows intelligent software agents to act more meaningfully in spite of differences in concepts and terminology.

We have previously described and discussed an ontology based approach (S. T. Liaw, et al., 2013; Rahimi, et al., 2014) to assessing the completeness, correctness and consistency (the 3Cs of DQ) of data and datasets. This approach is helpful in modeling the domain and representation of data and metadata requirements to identify diabetes on the data set from the University of NSW electronic Practice-based Research Network (ePBRN). This study used the dataset of 927 active patients from a general practice participating in the ePBRN, hereafter referred to as the General Practice Unit (GPU) dataset.

The ePBRN DQ research and development has focused on the 3Cs of DQ for ongoing ontology-based work to better define and address DQ, examine the issues and challenges for the network of data extraction and linkage, and semantic interoperability of large data sets (S. Liaw, et al., 2011). The ontology based approach can assist the terminology management and decision support to identify and classify different types of diabetes (S. Liaw, et al., 2011). This approach is also helpful in developing automated techniques and tools to extract and semantically link data elements (and concepts) in large data sets derived from multiple EHRs.

The objective of this study is to develop a methodology for the systematic construction of a Data Quality Ontology (DQO), use the ontology to identify patients with Type 2 Diabetes Mellitus (T2DM) in an EHR, and assess the quality of data and its impact on the accuracy of identification.

The paper is organized as follows. Section 3 details the background. Section 4 describes the methodology and different steps in the development of the DQO for T2DM and the materials and tools used for the work. Section 5 discusses perspectives expected from this work. Section 6 draws conclusions from this work.

3.2 Background

DQ is a complex idea with many dimensions, often overlapping conceptually (Devillers, et al., 2007; Nimmagadda, Nimmagadda, Dreher, & Ieee, 2008; Y. Wand & Y. Wang, 1996) with completeness, accuracy, correctness, consistency and timeliness

being the most commonly used dimensions. Liaw et al. (2010) developed a framework for extrinsic (e.g., representation) and intrinsic (e.g., correctness and consistency) concepts of data elements, and fitness for purpose (e.g., completeness) of data set for research and clinical purposes. Talaei-Khoei et al. (2011) examined the consistency and completeness of data in healthcare settings, reporting that these issues may result in disruption for practitioners (Talaei-Khoei, Solvoll, Ray, & Parameshwaran, 2011, 2012).

A previous literature review showed the understanding of DQ, as a multidimensional concept applied to the data elements (intrinsic DQ) and the set of data elements (extrinsic DQ) is progressing (S. T. Liaw, et al., 2013). Ontological approaches are emerging and theoretically important to address the complex relationships among overlapping concepts in this domain (Rahimi, et al., 2014).

The literature on the specification of DQ is fragmentary, lacks a comprehensive approach and is poorly evaluated (Rahimi, et al., 2014). A few other studies have examined ontology based approaches to support data consistency and accuracy (O-Hoon, Jung-Eun, Hong-Seok, & Doo-Kwon, 2008). However, no research was found that formally and systematically assessed the association between ontologies for DQ and fitness for purpose in various contexts. There are also few comparative and evaluative studies on assessment of DQ or that compared ontological and non-ontological approaches to representing knowledge in clinical information systems (Nimmagadda, et al., 2008).

The recent literature review (Rahimi, et al., 2014) also suggested that compared to non-hierarchical data models, there may be more advantages and benefits in the use of ontologies to solve semantic clinical issues and improve the validity and reliability of data retrieval, collection, storage, extraction and linkage algorithms and tools. Formal ontological approaches enable systematic development of automated, valid and reliable methods to assess and manage data and semantic interoperability issues (Lee et al., 2009; Valencia-Garcia, Fernandez-Breis, Ruiz-Martinez, Garcia-Sanchez, & Martinez-Bejar, 2008; Verma et al., 2009; Verma, Kasabov, Rush, & Song, 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 are poorly evaluated, with few comparative studies in chronic disease management or DQ assessment or management. The challenges to the development and validation of ontology-based models to assess and manage DQ include methodological immaturity, immature knowledge base, lack of tools to support ontology-based design of information systems, evaluation of ontological approaches, and engagement of users in design and implementations (Rahimi, Liaw, Ray, & Taggart, 2012; Rahimi, et al., 2014). There have also been several attempts to define ontology evaluation metrics and provide practical techniques to evaluate ontological approaches (in terms of flexibility, scalability and reusability) against non-ontology based models (Cur & #233, 2012; Maragoudakis, Lymberopoulos, Fakotakis, Spiropoulos, & Ieee, 2008). There is a lack of valid and reliable DQ assurance (D. Arts, De Keizer, & Scheffer, 2002; D. G. Arts, Bosman, de Jonge, Joore, & de Keizer, 2003; Peleg, Keren, & Denekamp, 2008) to ensure fitness for a range of uses by consumers, patients, health providers and professionals.

The significance of this work emanates from the fact that DQ research has been identified as a priority in medical informatics. Dixon (2011) and Huaman (2009) in their review of literature identified research in the quality of clinical data as a critical informatics research priority (Dixon, et al., 2011; Huaman, et al., 2009). The authors cited DQ research as necessary for improving health care through the translation of research findings into practice (S. Liaw, et al., 2011), national deployment of EHRs (Dixon, et al., 2011; Huaman, et al., 2009), and development of the National Health Information Network (NHIN) (Richesson & Krischer, 2007).

A recent review (Rahimi, et al., 2014) demonstrated a lack of comprehensive studies on the use of ontology-based tools to assess and manage DQ so that data sets are fit for purpose in healthcare and chronic disease management (CDM). This paper reports on a rich methodological approach to develop a DQO, using the identification of patients with T2DM as a case study to illustrate the important issues, focusing on fitness for purpose along the lines presented in Section 3.

3.3 A Methodology For Data Quality Ontology (DQO)

In this section, we present the details of a 5 step methodology to develop a DQO for application in the domain of Diabetes Mellitus. The DQO was applied to the identification of patients with Type 2 Diabetes Mellitus (T2DM) in an EHR. The

validation of the DQO examined the technical aspects of the model and its accuracy in identifying patients with T2DM.

The purpose and scope of DQO is to identify, in this case, diabetic patients using three core attributes namely Reason for Visit (RFV), Pathology (Path) tests results such as Hemoglobin A1c (HbA1C), Blood Sugar Level (BSL) and Random Blood Glucose (RBG), and medication (Rx) in the GPU dataset. We first determined the scope of the domain, and purpose of the task that the DQO is to be fit for. The ePBRN has selected completeness, correctness, and consistency as the core DQ metrics for demographic and clinical data collected from disparate EHRs, and even within an individual EHR. We now briefly discuss the three core dimensions of DQ.

Completeness

Completeness refers to the extent to which information is not missing and when available it is of sufficient breadth and depth for the task at hand (B. K. Kahn, et al., 2002). In our domain, this requirement means the availability of at least one record for the main patient attributes of RFV, Rx, Path and risk factors to identify Type 2 Diabetes mellitus. At the clinical level, completeness could mean that it include the availability of all information required to make a clinical decision about diabetes. Thus, each patient must have at least 1 record in one of the target attributes which consist of RFV, Path and Rx (S. Liaw, et al., 2011).

Correctness

Correctness refers to data for each attribute being free of any errors (Pipino, et al., 2002) that is, each valid and appropriate clinical record must have the correct unit of measurements and must be within the acceptable clinical ranges. For instance, 'diabetes' is a correct value for the attribute RFV, and there are no errors in the way it has been written. Any other type of value for RFV is incorrect. For pathology tests and risk factors, correct ranges lie between the minimum and the maximum range of ePBRN data while respecting the Australian National Guidelines for T2DM ("Diabetes Management in General Practice Guidelines for Type 2 Diabetes ", 2012). A datum that lies outside this range is considered incorrect. Similarly, for medication, it would be incorrect if there were other attributes recorded for the script and the script name was missing.

Consistency

Consistency refers to representing data values of the attributes following the same schema and format (B. K. Kahn, et al., 2002). It includes values and physical representation of data (Y. Wand & Y. Wang, 1996). External consistency uses a uniform data type, format and standard terminology (S. Liaw, et al., 2011) based on the Systematized Nomenclature of Medicine – Clinical Terms – Australian version (SNOMED-CT-AU) (McBride, Lawley, Leroux, & Gibson, 2012). Internal consistency uses a standard adopted specially for practice. For example, the following issues are relevant: Do doctors record diabetes type 2 the same way or does each doctor record it differently? Also, for internal consistency, at the first level, how are the attributes being recorded? An ePBRN question is whether different GPs and general practices record diabetes the same way? For external consistency, each term e.g., RFV used is externally consistent if it can be coded with or mapped to the same concepts in SNOMED-CT-AU.

Based on the analysis of currently available techniques and commonly adopted conceptual steps (Corcho, Fernandez, & Gomez, 2003; Fernandez, 1999; Hadzic, Dillon, & Dillon, 2009; Kuziemsky & Lau, 2010; Pinto, 2004), we used a five stages methodology to create our ontology to identify T2DM in an EHR and assess DQ of the resulting register (Figure 3.1): (1) Knowledge acquisition; (2) Conceptualization of the domain to create DQO; (3) Semantic modeling; (4) Knowledge representation; and (5) Validation of SPARQL query results and comparison with manual results as our research gold standard.

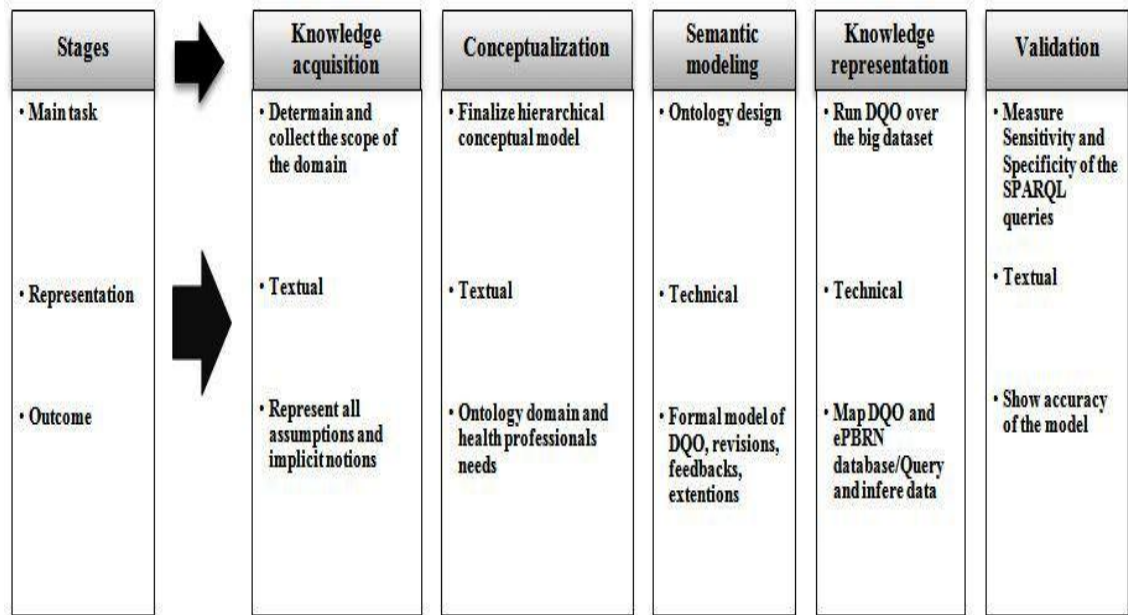


Figure 3.1: Five stages approach for the development of the ontology model

3.3.1 Knowledge acquisition

In this step, we acquired knowledge about the domain and its scope with information from the domain experts and relevant published works. The techniques used include: brainstorming, interviews, questionnaires, text analysis, and inductive techniques (Pinto, 2004).

3.3.1.1 Patient data audit

An audit carried out on 7 patient data tables in the ePBRN includes the following attributes: *Patients*, *Prescription*, *Diagnosis*, *Measure*, *Family History*, *Consultation* and *Patient Referral*. Data collected included: medical information (such as Reason for visit, Medication, Pathology test results), and information about patients' demographics (such as Date of Birth, History, Status and Sex). The ePBRN patient data audit formed the source of the clinical (user) vocabulary for the ontology.

3.3.1.2 GP and nurse consensus meetings

Two types of practice experience data were collected: clinician meetings and clinical observations. Clinician meetings involved one of the authors (AR) attending three meetings with 4 clinicians (2 physicians, 1 nurse and 1 data manager). The clinician meetings took place over 3 months and involved developing different models of diabetes management practices as well as discussion of conceptual models of

diabetes management. Data from the clinician meetings provided a large volume of data which were useful in the design of our ontology. In those meetings there were also discussions about system design considerations for the computer-based diabetes assessment tool. Clinical observations involved the same author (AR) spending 44 hours performing qualitative observation and documentation of diabetes management themes on the clinical flowchart. Those observations were crucial for understanding the clinical workflow.

3.3.1.3 Literature review

A literature review on the automation of identifying diabetes patients, diabetes management, and chronic disease management was carried out. Researching the literature brought in current evidence on diabetes management such as mechanisms for the assessment and management of diabetes, conceptual models on diabetes management and educational resources for primary and secondary care, assessment, diagnosis and management of different types of diabetes. Moreover, the conceptualization of diabetes assessment and management was drawn from evidence-based guidelines based on the Australian National Guidelines for T2DM ("Diabetes Management in General Practice Guidelines for Type 2 Diabetes", 2012). Also, the SNOMED-CT-AU standard guided the specification of the data and domains in the DQO. Current published medical ontologies included the Human Diseases Ontology (DO) (Hadzic & Chang, 2004), Infectious Diseases Ontology (IDO) (Cowell & Smith, 2010), Galen (Rector & Rogers, 2005), and Gene Ontology (GO) (Pan Du et al., 2009). While these are comprehensive and essential models to draw on to develop the DQO prototype, they are not focused on diabetes specifically. The research literature was valuable for contextualization of the ontological concepts and the clinical practices.

3.3.2 Conceptualization

In this step, we identified the key concepts and relationships in the domain and defined terms used to represent these concepts and relationships. Conceptualization denotes the process of turning raw knowledge into clearly established concepts that can be used to create a DQO. It typically includes the identification of the concepts and their relationships within the diabetes domain, taking advice from domain experts.

3.3.2.1 Task

In the current application, the conceptual model was developed through the results of an exhaustive literature review, ePBRN patient data audit, and GPs and nurses meetings.

3.3.2.2 Output

The final consensus meeting of our research team identified 68 concepts to comprehensively model the domain of diabetes management. Table 3.1 shows the categories (along with subcategories) of concepts in four different layers and the concepts relevant to each category.

Table 3.1: Categories of collected concepts in four different layers

Main Categories	Subcategories	Relevant concepts
Actor	Organization	Research Institution, University
	Person	Doctor, Nurse, Patient, Specialist
Context	Problem	Disease
	Setting	Primary care, Secondary care
Impact	Disease Indicator Control	HbA1c, random and fasting glucose levels
	Patients Satisfaction	Patient satisfaction questionnaires
	Quality of Life	QOL questionnaires
Mechanism	Advise	Lifestyle advice
	Assessment	Diagnosis, Family history, Risk Factor
	Billing	Services and supplies
	Consultation	Type of consultations
	Order	Imaging, Medication, Pathology tests
	Prescription	Medication
	Referral	Endocrinologist or general
	Review	Diabetes cycle of care

For example, in the hierarchical conceptual model for *Mechanism* (which is the main class in diabetes management), there are 7 subclasses consisting of *Billing*, *Assessment*, *Review*, *Prescription*, *Referral*, *Advise*, and *Order*. Similarly, the subcategory *Order* includes subclasses *Medication*, *Imaging* and *Pathology tests*.

3.3.3 Semantic modeling

Semantic modeling refers to formalizing the domain ontology. This ontology and the defined rules generate logical inferences and control the relevant objects such as the patient with a diagnosis of diabetes mellitus (DM) and their related properties.

3.3.3.1 Task

In this stage we systematically transform the conceptual models into a formal model through the development of hierarchies and relationships and thus removing any ambiguities in the meanings of the concepts. The semantic model for a concept includes a set of attributes and its relationships with other concepts that characterize the meaning of the concept. The DQO used previously reported definitions of the 3Cs of DQ (S. Liaw, et al., 2011).

3.3.3.2 Output

The formalized ontological model was developed using the Protégé 4.3 ontology editing tool (Gennari et al., 2003) and (Min et al., 2009) with frames as the representational construct. In Protégé, a reference terminology such as SNOMED-CT-AU can be flexibly used with Australian CIS; OntopPro (Rodriguez-Muro, Kontchakov, & Zakharyashev, 2013) can be used as an Ontology-based Data Access (OBDA) plugin for Protégé for querying, inferring and mapping of the ontology approach and GPU datasets; and logic ontology reasoners provide automated support for reasoning tasks in ontology and instance checking and they include Pellet, Racer, Quest as the most popular and effective semantic reasoning engines (Huang, Li, & Yang, 2008).

In Table 3.1, Column describes the main classes and Columns 2 and 3 their subclasses in the ontology approach. The output of this stage is a formalized ontology consisting of 4 main classes (*Actor*, *Content*, *Mechanism* and *Impact*) and 51 subclasses (Figure 3.2) with 8 object properties and 15 data properties. Figures 3.3 and 3.4 provide some illustrations to show the formalization of the ontology approach developed (the hierarchical model and the relations) using the ontology tools and the definition of objects and properties. Protégé was used to add more terms to describe properties and classes within the diabetes domain, viz., relations between classes (e.g., disjointness such as PrimaryCare disjoint_with SecondaryCare), cardinality (e.g., *exactly one*), equality, richer typing of properties, characteristics of properties (e.g., functional for

PatientUUID), and enumerated classes (e.g., MaritalStatus that has several characteristics such as single, married, divorced and widowed).

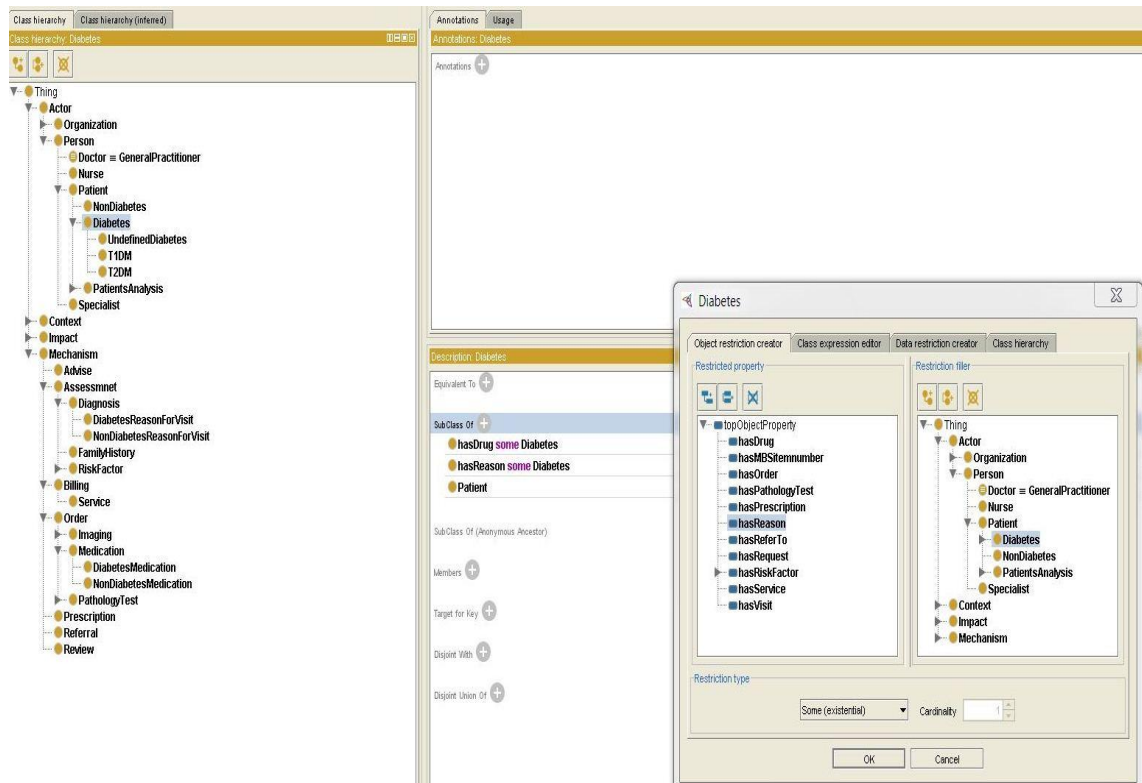


Figure 3.2: The ontology hierarchical conceptual model with data properties

We have specified which classes are disjoint, so that an object cannot be an instance of more than one of these. It ensured consistency in the ontology approach. Figure 3.3, defines (at this stage) object properties and relationships between different classes and subclasses. Careful modeling of object properties in Protégé helped to achieve all patients' data requirements. As figure 3.4 shows, the constraints presented by data properties in Protégé 4.3 are mainly capturing (a) the correctness of valid clinical records in ePBRN (for example, range for HbA1C is between 3.0 and 20 mmol/L), and (b) consistency of patients' data.

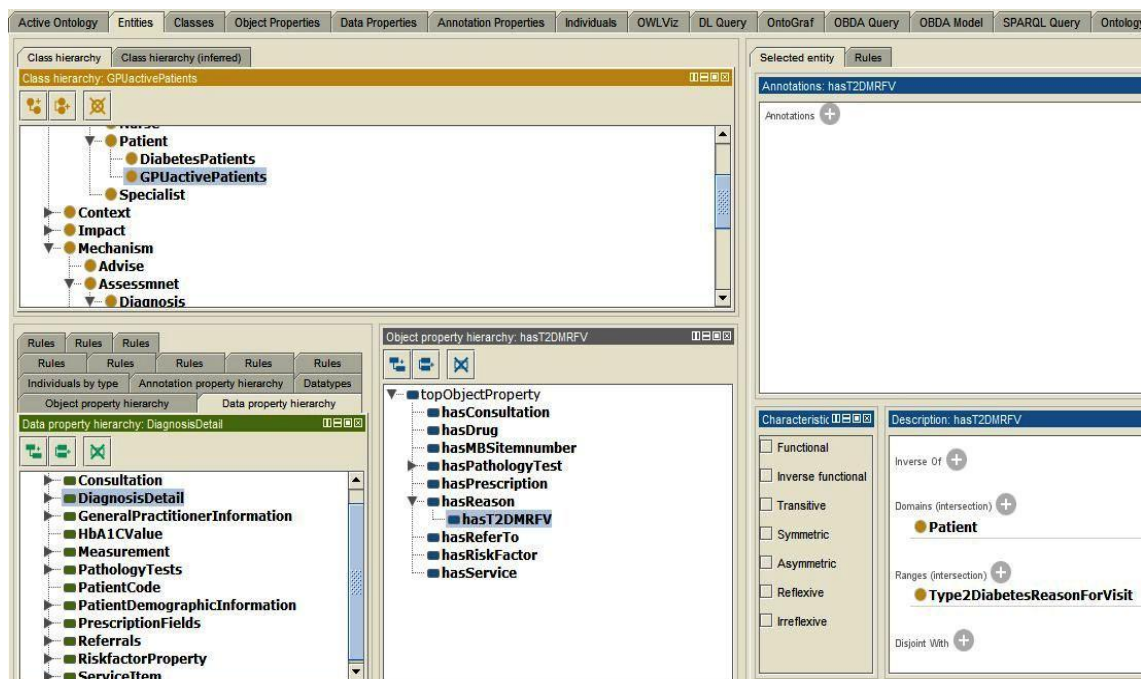


Figure 3.3: A sample of object property to show how as an example “hasT2DMRFV” can link two joint classes together

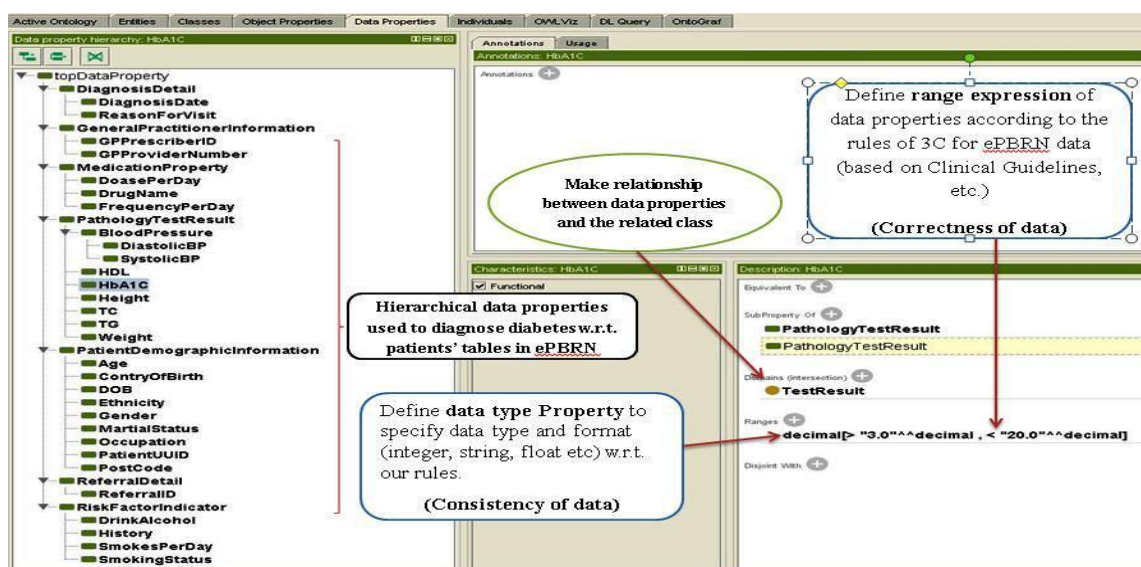


Figure 3.4: The data property tab to define various data ranges, types and values for each class

The end product of this stage is a semantic data model that has been defined as classes, sub-classes and their relationships to assist in identifying diabetic patients by Protégé 4.3 (Chen, Lu, & Liu, 2007). Our formalized ontology consist of 8 object properties, 15 data properties, 68 concepts and 14 major themes in 4 main classes

comprising *Actor*, *Content*, *Mechanism* and *Impact* for improved identification of T2DM patients. Two ontology reasoners (Pellete 3.2.0, RacerPro 1.1.10) (Huang, et al., 2008) were also applied to check internal consistency of the T2DM ontology and the reasoners found no logical inconsistencies in our ontology.

In Figure 3.5, we show that the ontology can be mapped onto the SNOMED-CT-AU Ontology (SCAO) which has more than 300,000 concepts (Yu, Liaw, Taggart, & Rahimi, 2013).

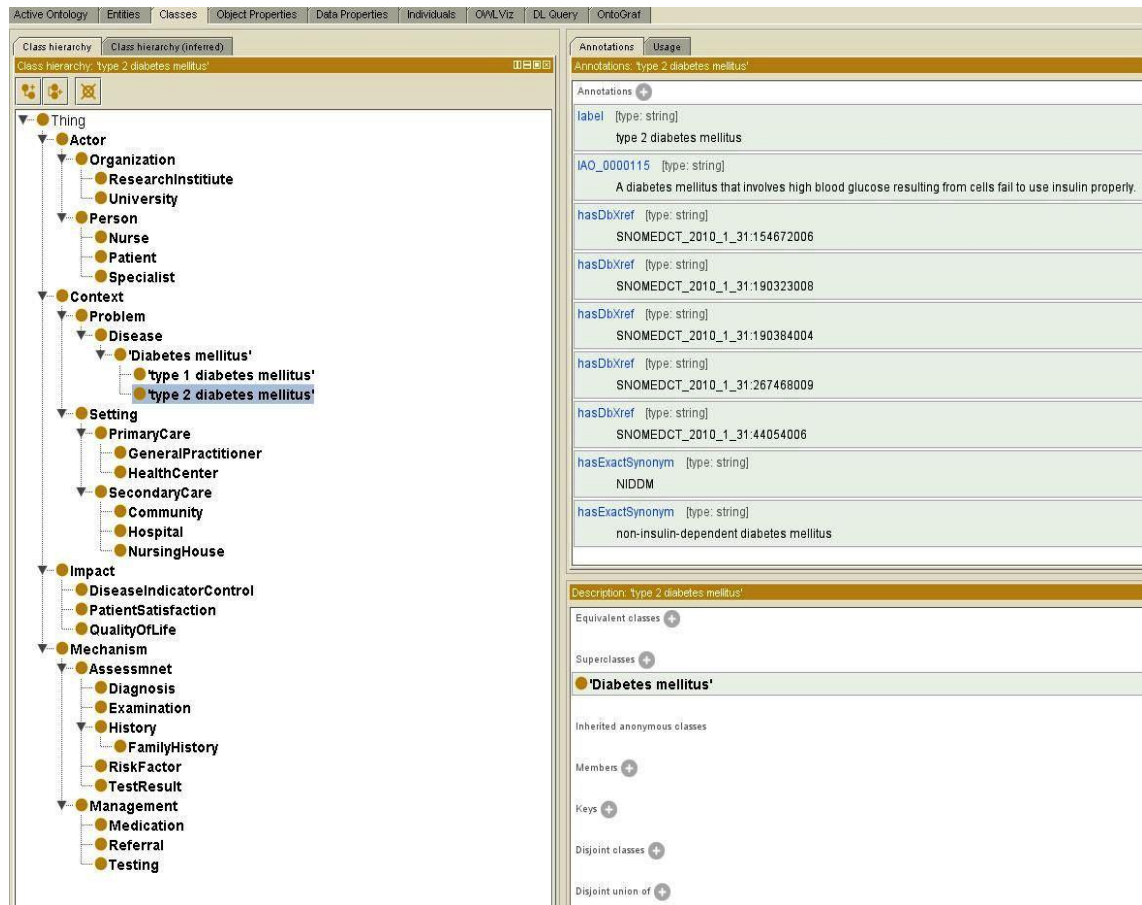


Figure 3.5: The ‘Context’ as an example class hierarchy shown expanded in a Protégé screenshot. The annotation associated with the subclass “Type 2 diabetes mellitus” describes semantic relationship of this subclass with the reference terminology SNOMED-CT-AU

3.3.4 Knowledge representation

In this step, we developed an ontology model initially to represent the domain broadly. The necessary general concepts were included first, followed by the addition of

the necessary constraints. For the Diabetes Mellitus domain, we added the constraints required to assess the DQ of the extracted data.

3.3.4.1 Task

This stage implemented the formalized ontology over the clinical data set extracted from a specific general practice participating in the ePBRN. This data set is a subset of the ePBRN data repository. The DQO has been implemented first to represent the domain broadly. We used it to describe the necessary general concepts from the diabetes management point of view and then added constraints for lower datasets from the database in order to meet our DQ goals.

3.3.4.2 Output

To implement DQO over test data set, data was formalized using Microsoft SQL Server 2008 R2 and the tools like OntopPro (Rodríguez-Muro, et al., 2013) as a plugin for Protégé 4.3, a semantic query language like Simple Protocol and Resource Description Framework Query Language (SPRQL), and a reasoner (Quest) (Rodríguez-Muro & Calvanese, 2012) were then used. We additionally installed drivers to connect the test (relational) database and the ontology approach using Protégé 4.3 preferences tab. For example, in Java, database connections are established using the Java Data Base Connectivity Framework (JDBC) (Calvanese et al., 2009). In our case, Quest and OntopPro use JDBC connections such as MS JDBC Driver for SQL Server to connect to GPU data set, and so, they require JDBC parameters. In particular, for GPU data set, we needed to define 4 parameters: Driver class, JDBC URL, username and password (Figure 3.6).

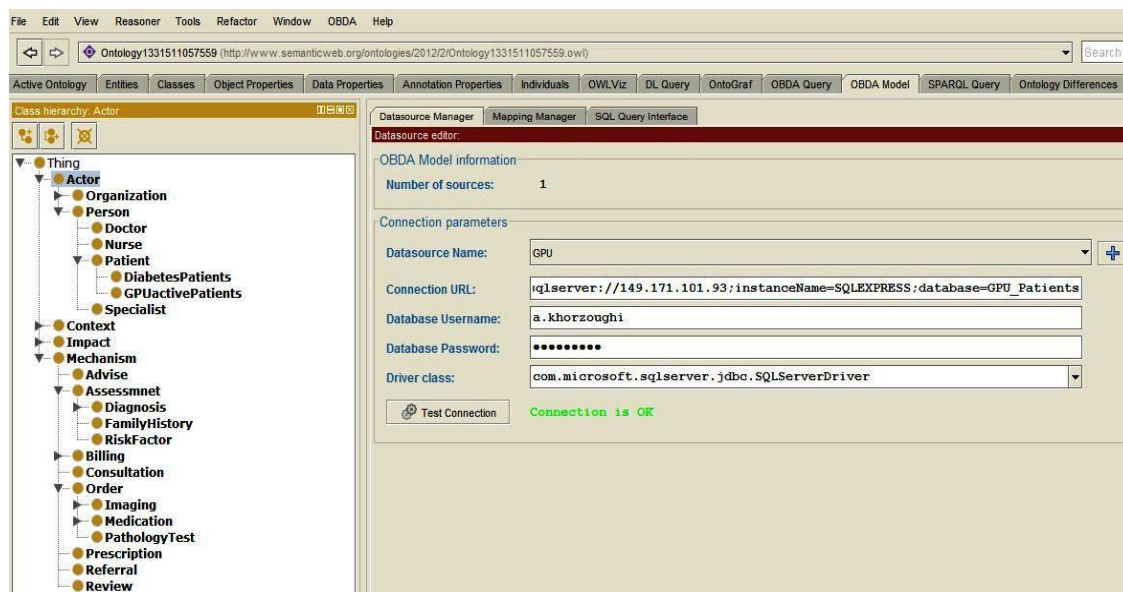


Figure 3.6: The ontology approach OBDA tab to define JDBC connection parameters

Once mappings were created, the plug-in was used to generate Resource Description Framework (RDF) triples for use with OntopPro to query the test dataset, without any imports. A mapping axiom was used to generate RDF triples, and one set of RDF triples for each result row was returned by the source query. The triples were created by replacing the place holders in the target with the values from the row. Each mapping must also contain one or more mapping axioms. A mapping axiom is defined with a source and a target, where the source is an arbitrary SQL query over the database and the target is a triple template that contains placeholders that reference column names mentioned in the source query.

In Figure 3.7, we defined the requirements for this example as follows:

- Example-1: Query for active Patients.
- Example-2: Query for all Patients with the T2DM Reason Item.
- Example-3: Use objects properties to join two tables using Patient_UUID as a unique identifier and identify active patients with T2DMRFV.

3.3.4.2.1 Step 1: Analysis of sources and targets

From DQO, we needed to map the following entities:

- Classes, i.e., Patient and Diagnosis.

- Data properties, i.e., PatientID and ReasonItem.
- Object properties, i.e., hasT2DMRFV.

Analysing our database we find that the following tables can be used to create mappings for these classes and properties: i.e., ePBRN_Active_Patient and ePBRN_DIAGNOSIS. We can see that there is a one to one correspondence between the entities stored in the tables and the classes we wanted to map. Likewise, the columns Patient_UUID from ePBRN_Active_Patient table and REASON from ePBRN_DIAGNOSIS can be used for the data properties. To create the Uniform Recourse Identifiers (URIs) for those entities we could use Patient_UUID as a unique-identifier for these tables.

3.3.4.2.2 Step 2: Mappings and queries

The tables were analysed using the following mappings:

- Example-1: Query for all active Patients.

Source: SELECT Patient_UUID FROM ePBRN_Active_Patients

Target: <patient/{Patient_UUID}> a :GPUactivePatients.

- Example-2: Query for all active Patients with the T2DM Reason Item.

Source: SELECT Patient_UUID, REASON FROM ePBRN_DIAGNOSIS
WHERE (REASON = 'Diabetes Mellitus - NIDDM' OR REASON = 'Diabetes Mellitus - Type II' OR REASON = 'Diabetes Mellitus Type 2' OR REASON = 'NIDDM' OR REASON = 'Diagnosis of Type 2 DM' OR REASON = 'Non-insulin dependent diabetes mellitus' OR REASON = 'Diabetes Mellitus Type II - requiring insulin' OR REASON = 'NIDDM - requiring insulin')

Target: <t2dmrfv/{Patient_UUID}> a :Type2DiabetesReasonForVisit;
:ReasonItem {REASON}.

- Example-3: Use objects property to show how it joins two tables using Patient_UUID as a unique identifier and identify active patients with T2DMRFV.

Source: SELECT 'Patient_UUID' as pid, 'Patient_UUID' as t2dmrfv.

FROM ePBRN_PATIENT_V1, ePBRN_DIAGNOSIS

WHERE ePBRN_PATIENT_V1.Patient_UUID =

ePBRN_DIAGNOSIS.Patient_UUID

Target: <t2dmrfv/{Patient_UUID}> a :Type2DiabetesReasonForVisit ;
:ReasonItem {REASON}.

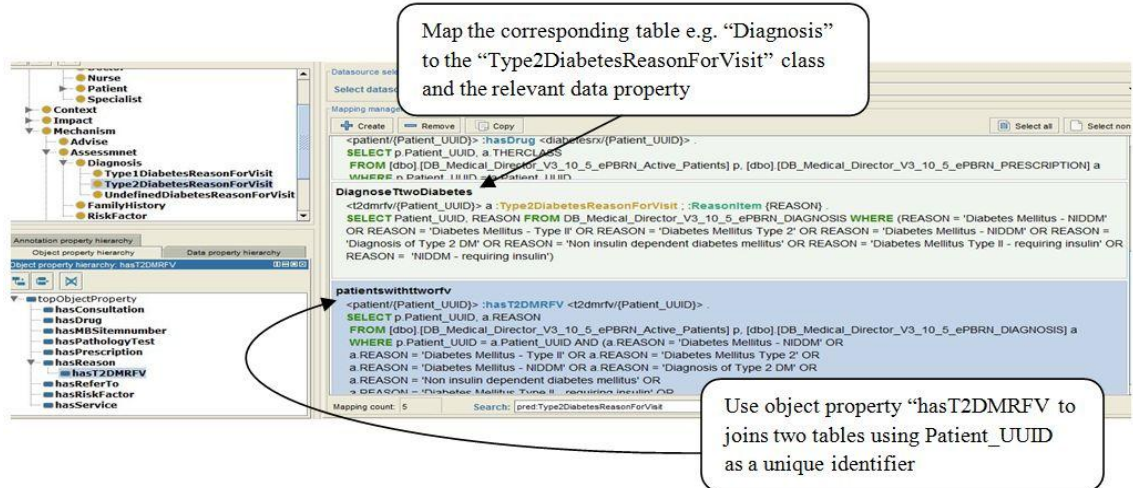


Figure 3.7: Sample of mapping the Diagnosis table with the ontology approach

All mappings required for our criteria to identify T2DM patients were thus complete. Indeed in our knowledgebase, the ABox, associated with instances of ontology classes or properties, was populated through OntopPro. The TBox, related to conceptual terminologies, was built using Protégé. Therefore, once mappings have been created, we were able to use the plug-in to generate RDF triples and use them with OntopPro to query the GPU dataset, without any imports.

Semantic queries were formulated in SPARQL according to requirements from domain experts, and were run using QUEST, a query engine and ontology reasoner. SPARQL used relevant objects, such as T2DM diagnosis, medications and pathology test results, singly and in combination, to construct queries for the identification of patients with T2DM. The sensitivity and specificity of the SPARQL query, and therefore the implementation of the T2DM ontology, were measured (Rahimi, Liaw, Taggart, Ray, & Yu, in-press). The SPARQL queries were validated using SQL over an artificial dataset of 100 patients schematically similar to the ePBRN dataset.

3.3.5 Validation

In this step, the quality of the DQO was assessed as to the correctness and validity of the knowledge encoded in the ontology. The validation of the ontology involves verifying whether the meaning of the concepts and their relationships faithfully model the real world for which the ontology was created. This validation is essential to ensure that the ontology based approach and the DQO developed is fit for purpose.

3.3.5.1 Task

The DQO was tested for its compliance to the requirements domain experts, DQ attributes (3Cs), and accuracy in identifying patients with T2DM over a data set extracted from the EHR of a small general practice participating in the ePBRN. The methodology used was to compare the cases of T2DM identified by the DQO with the cases identified by a manual audit of the EHR from which the data was extracted (GPU dataset). We audited the EHR information of all 908 active patients (i.e. those who have attended the practice at least 3 times in the past 2 years) using a specific template to ensure that we understood all the reasons why the patient might or might not be a T2DM patient. This accuracy the DQO methodology has been reported in another paper (Rahimi, et al., in-press) The following section (outputs) uses information reported in this paper to highlight the compliance of the technical components of the DQO methodology to the requirements of the domain experts and DQ attributes.

3.3.5.2 Output

Explicit and unambiguous queries, using patterns, disjunctions and conjunctions, were built in SPARQL to identify patients with T2DM. The constraints presented by object properties in Protégé 4.3 were used to set up relationships between classes. SPARQL was also used to apply data properties to assess DQ in the patient's attributes. For example, constraints presented by data properties were used to capture: (a) the correctness of valid clinical records in ePBRN (e.g., correct value for HbA1C is \Rightarrow 7%); and (b) consistency of patients' data (e.g., all T2DM patients with uniform, data type and standard value of HbA1C).

Semantic queries in SPARQL were verified by clinicians in the research team (JT, STL) to ensure that they complied with the requirements of the domain experts previously consulted. The DQ requirements were also verified for its fit for

identification of T2DM. Once this was verified, and the queries were run through QUEST, the query engine and OWL reasoner.

The query results met all our expectations regarding the identification of T2DM patients and the assessment of the DQ of the data set (Rahimi et al., 2014, in press). For example in Table 2 it can be seen how a semantically flexible approach uses different object and data properties as well as relevant classes to combine different T2DM attributes (RFV, Rx and Path) for the identification of T2DM patients. The first level of completeness of patients' DQ requirements can be achieved by carefully modeling object properties.

Table 3.2: Part of a SPARQL query using 3 patients' attributes to identify patients with T2DM

Diabetes' attributes	Sample of SPARQL query using combined patients' attributes
T2DM RFV and Rx and abnormal pathology tests	<pre> SELECT DISTINCT ?pid WHERE {{ ?pid a :GPUactivePatients. ?pid :hasT2DMRFV ?r. ?r :ReasonItem ?reason. FILTER(?reason = "Diabetes Mellitus - Type II"^^xsd:String ?reason = "Diabetes Mellitus - NIDDM"^^xsd:String ?reason = "Diagnosis of Type 2 DM"^^xsd:String ?reason = "Diabetes Mellitus Type II - requiring insulin"^^xsd:String ?reason = "Diabetes Mellitus - Type II"^^xsd:String ?reason = "Diabetes Mellitus Type 2"^^xsd:String ?reason = "NIDDM - requiring insulin"^^xsd:String ?reason = "Non insulin dependent diabetes mellitus"^^xsd:String ?reason = "NIDDM"^^xsd:String ?reason = "Diabetes Mellitus - NIDDM"^^xsd:String)} UNION {?pid a :GPUactivePatients. ?pid :hasT2DMHistory ?h. ?h :Condition ?history. FILTER(?history = "Diabetes Mellitus - NIDDM"^^xsd:String ?history = "Diabetes Mellitus - Type II"^^xsd:String ?history = "Non insulin dependent diabetes mellitus"^^xsd:String ?history = "NIDDM"^^xsd:String ?history = "Diagnosis of Type 2 DM"^^xsd:String ?history = "Diabetes Mellitus Type II - requiring insulin"^^xsd:String ?history = "Diabetes Mellitus Type 2"^^xsd:String ?history = "NIDDM - requiring insulin"^^xsd:String)} UNION {?pid a :GPUactivePatients. ?pid :hasDrug ?d. ?d :TherapyClass ?rx. FILTER(?rx = "HDI"^^xsd:String ?rx = "HDO"^^xsd:String ?rx = "HDI"^^xsd:String ?rx = "ODB"^^xsd:String ?rx = "HD"^^xsd:String ?rx = "HDOA"^^xsd:String ?rx = "HDOD"^^xsd:String ?rx = "ODU"^^xsd:String)} UNION {?pid a :GPUactivePatients. ?pid :hasRepeatDrug ?r. ?r :TherapyClass ?rerx. FILTER(?rerx = "HDI"^^xsd:String ?rerx = "HDO"^^xsd:String ?rerx = "HDI"^^xsd:String ?rerx = "ODB"^^xsd:String ?rerx = "HD"^^xsd:String ?rerx = "HDOA"^^xsd:String ?rerx = "HDOD"^^xsd:String ?rerx = "ODU"^^xsd:String)} UNION {?pid a :GPUactivePatients. ?pid :hasT2DMPathologyTest ?p. ?p :TestName ?test. ?p :ResultTest ?result. FILTER(?test = "HbA1C"^^xsd:String && ?result >= "6.5"^^xsd:Integer ?test = "GLUCOSE PLASMA FASTING"^^xsd:String && ?result >= "7.0"^^xsd:Integer ?test = "GLUCOSE Random"^^xsd:String && ?result >= "11.1"^^xsd:Integer ?test = "Glucose Fasting"^^xsd:String && ?result >= "7.0"^^xsd:Integer)}} </pre>

Table 3.2 presents a part of the novel SPARQL query results for a different level of identification of Type 2 diabetic patients. SPARQL queries only referred to classes, object properties and data properties to combine main diabetes criteria for the identification of diabetes semantically. The DQO-based query, partly shown in Table 3.2, identified 105 T2DM using T2DM RFVs, Rx and Path. The query was implemented over the data set using SQL Server 2008 R2. The accuracy of the DQO-based query, as compared to the manual validation as the benchmark, is summarized in Table 3.3. The manual of the EHR that the T2DM RFV was scattered across a number of tables (PAST_HISTORY_TABLE, DIAGNOSIS_Table) and in the progress notes as text unstructured data. Where the RFV were recorded in a structured field, the semantic SPARQL queries identified them accurately. This was similar for the other attributes used to identify diabetes T2DM (using RX and Path).

The Sensitivity and Specificity of the DQO-based queries implemented in SPARQL were calculated and compared with the accuracy of the manual audit. Patients identified as T2DM by the DQO based query and manual audit are true positives (TP); those identified by DQO based query as T2DM but not on manual audit are false positives (FP); the reverse are false negatives (FN); and patients not identified as T2DM by both DQO based query and manual audit are true negatives (TN). Sensitivity, defined as $TP / (TP + FP)$ denotes the ability of the system to accurately identify all those patients who are T2DM patients. Specificity, defined as $TN / (FN + TN)$ measures the model's accuracy in identifying the proportion of all patients without T2DM who are not included in the dataset. As Table 3.3 suggests, identification of T2DM using Path data was not as accurate as that using RFV or Rx.

Table 3.3: Accuracy of the model developed (Rahimi, et al., in-press)

	RFV	Medication	Pathology tests	All attributes
Sensitivity	100%	96.55%	15.6%	97.67%
Specificity	99.88%	98.97%	98.92%	99.18%

This reflects inaccurate Path data due to change in the units for reporting of HbA1c results. However, this level of inaccuracy was acceptable for our purpose as confirmed by the very small relative deterioration of the accuracy (Sensitivity and

Specificity were 97.67% and 99.18%, respectively) when calculated for the combination of RFV, Rx and Path. The completeness and correctness of the RFV and Rx data compensated for the poor completeness and correctness of the Path data in the DQO-based approach. The manual EHR audit suggested that the accuracy of the algorithm was determined by DQ issues such as unavailability of data due to non-documentation or documented in the wrong place, problems with data extraction, encryption and data management errors. The multi-attribute ontological approach to defining a T2DM case, can compensate for poor DQ in one or more of the component attributes and therefore not lose the overall accuracy.

3.4. Discussion

This paper presented a semantic knowledge management approach for identifying T2DM and assessed its DQ using: (a) knowledge acquisition techniques to derive diabetes management strategy from the results obtained in our literature review and evidence-based resources; (b) a conceptualization process to develop a hierarchical data model; (c) a knowledge model to transfer the conceptual model to the formal model with the help of knowledge management tools; (d) knowledge representation techniques to map the data set into the DQO, using OntoPro; and (e) manual validation to confirm the accuracy of the DQO based approach.

The DQO based approach to identify T2DM patients can be modular and generic, enabling the development of intelligent software agents (A. H. Ghapanchi & Aurum, 2011; Amir Hossein Ghapanchi & Aurum, 2012) to act in various semantic contexts to identify patients with a range of diseases (Mabotuwana & Warren, 2009), support decision making in health care (Lezcano, Sicilia, & Rodríguez-Solano, 2011), and conduct audit, evaluation and research on patients with other chronic diseases (Pathak, Kiefer, Bielinski, & Chute, 2012a, 2012b; Pathak, Kiefer, & Chute, 2012).

3.4.1. Usefulness of the ontological based approach for DQ specification

As we presented in the Conceptualization and Semantic modeling section, the ontology based model is particularly useful to enable quick development and testing so that feedback can be cycled back into the development process. For example, the ontology classes and data properties guide research team to ensure fields, records, tables and relationships in the database are appropriately presented.

The ontology-based approach can therefore access and manage the quality of data in a way that is generalizable and reusable, to examine the issues and challenges in data extraction, linkage and semantic interoperability (S. T. Liaw, et al., 2013; Rahimi, et al., 2014).

The DQO based approach implemented here corroborates the belief that ontological approaches have theoretical and practical advantages in developing automated methods for identifying patients with chronic diseases, guiding clinical care, and quality improvement and research (Buranarach, Chalortham, Chatvorawit, Thein, & Supnithi, 2009; Chalortham, Buranarach, & Supnithi, 2009; Colombo et al., 2010; Coltell et al., 2004).

3.4.2. Applicability of the ontology based approach for DQ specification

The suggested ontology based approach can accurately specify metadata for DQ specification and assessment for particular clinical domains. In the semantic modeling stage, it has been shown that DQ can be expressed by constraints and axioms to cope with DQ specification. For example, as we demonstrated in the semantic modeling section, class attributes (data properties) have been defined to capture correctness and consistency of valid clinical records. The ePBRN team has created the rules for quality metrics (3Cs), using Australian National T2DM diagnosis and management Guidelines and SNOMED-CT-AU (S. Liaw, et al., 2011; Rahimi, et al., 2012; Yu, et al., 2013) to:

- Define data properties.
- Use uniform data types and formats (e.g., integer, string and real) for each variable (for Internal Consistency).
- Define uniform data format for each concept (e.g., for Assessment sub-class, hasHbA1C is selected as the property of the class, decimal is selected as the type and a value v where $3 \leq v$ and $v \leq 20$ is entered for Correctness)
- Select standard label for each entity (e.g., use type 2 diabetes mellitus instead T2DM for External Consistency).

The knowledge management tools, such as Protégé, allows specifications of properties of classes, such as disjoint, so that an individual (or object) cannot be an

instance of more than one of the specified classes. This leads to more consistency and correctness, as well as enable an assessment of data set completeness.

In addition to accuracy, a DQO based application to enable automated assessment of patients' data can also be flexible and applicable to other chronic diseases such as COPD and other areas such as population health. Therefore, our model can support other studies that it is applicable to information retrieval and analysis (Valencia-Garcia, et al., 2008), intelligent data mining (seeking concepts and relationships) (Chen, et al., 2007), discover new knowledge, and reuse knowledge for decision support systems and patient decision aids (Abidi, 2011). Our approach fills current gap in the application and applicability of ontological models to assess and manage quality of information in EHRs.

3.4.3. Evaluation of DQO methodology

Our methodology confirmed that the validation of an ontology should and can be done through its use in a concrete application (e.g., the identification of T2DM) (Kuziemsky & Lau, 2010; Rahimi, et al., in-press). The development and deployment of ontologies must include evaluation metrics. Our previous literature reviews have shown that the ontological approach to develop DQ is poorly validated (S. T. Liaw, et al., 2013; Rahimi, et al., 2014) and identified the most common criteria to assess validity of ontologies and data models are Flexibility, Reusability and Scalability (Rahimi, et al., 2014).

The DQO based approach can add more axioms and constraints to the concepts based on the specific purposes of DQ assessment and management. The ontology based approach is more flexible than the non-ontological and non-semantic techniques for solving semantic interoperability and technological issues derived from poor DQ (Gangemi, Catenacci, Ciaranita, & Lehmann, 2006; Gilbert & Ddembe, 2008; Pannarale et al., 2012). It also has the flexibility of being applicable to and therefore reusable in other domains (Gilbert & Ddembe, 2008; Pinto, 2004).

As shown in the Knowledge Representation stage, the DQO approach mapped a small part, a unique general practice with 908 active patients, from the larger ePBRN data repository. This demonstrates the scalability of the ontology based approach (Cur & #233, 2012).

3.4.4. Comparison of DQO and non-ontological approaches in CDM

The proposed methodological approach particularly in the Conceptualization and Semantic Modeling stages reveals that the ontology based approach contains more explicit semantic information compared to non-semantic and non-ontological approaches. Hence, for DQ specification, as opposed to non-ontological approaches, an ontology is a formal, explicit specification of a shared conceptualization that provides a vocabulary of terms, their meanings and relationships to be used in various application contexts so that intelligent agents can act in spite of differences in terminology and their meanings (Pinto, 2004). They enable the modeling of the domain and representation of information requirements to specify the context in collaborative environments (Ganguly, Ray, & Parameswaran, 2005). DQ models and ontologies are being developed to enable the application of ontology-based tools for automated specification, assessment and management of DQ (Ganguly, et al., 2005; Ying, et al., 2010).

3.5. Conclusion

The ontology-based approach to DQ assessment and management in the context of type 2 diabetes mellitus identification has been examined. The traditional five stage methodology - knowledge acquisition, conceptualization, semantic modeling, knowledge representation, and validation was successfully used to develop the DQO. This semantic mechanism to purposefully capture patient data from EHRs is flexible, generalizable and potentially reusable in other domains. The accuracy was validated by a manual audit of active patients from the EHR. This approach can address the challenges in automated data extraction, linkage and assessment of the quality of routinely collected data in EHRs.

Acknowledgments

The authors would like to thank the ePBRN research team for their previous and ongoing contributions in this study.

3.6 References

Abidi, S. R. (2011). *Ontology-based knowledge modeling to provide decision support for comorbid diseases*. Paper presented at the 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, G. J. (2002). Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework. *Journal of the American Medical Informatics Association*, 9(6).
- Arts, D. G., Bosman, R. J., de Jonge, E., Joore, J. C., & de Keizer, N. F. (2003). Training in data definitions improves quality of intensive care data. *Crit Care*, 7(2), 179-184.
- Borst, W. N. (1997). *Construction of Engineering Ontologies*. University of Twente, Enschede, NL.
- Brown, P., Warmington, V., Laurence, M., & Prevost, A. (2003). Randomised crossover trial comparing the performance of Clinical Terms Version 3 and Read Codes 5 byte set coding schemes in general practice. *BMJ*, 326(7399), 1127.
- 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.
- Calvanese, D., Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M., et al. (2009). *Ontologies and Databases: The DL-Lite Approach, Reasoning Web. Semantic Technologies for Information Systems: 5th International Summer School 2009, Brixen-Bressanone, Italy, August 30-September 4, 2009, Tutorial Lectures*: Springer-Verlag, Berlin, Heidelberg.
- Chalortham, N., Buranarach, M., & Supnithi, T. (2009). Ontology Development for Type II Diabetes Mellitus Clinical Support System.
- Chen, X. H., Lu, J., & Liu, Z. Y. (2007). Assistance ontology of quality control for enterprise model using data mining. In M. Helander, M. Xie, M. Jaio & K. C. Tan (Eds.), *2007 Ieee International Conference on Industrial Engineering and Engineering Management, Vols 1-4* (pp. 602-606).
- Colombo, G., Merico, D., Boncoraglio, G., De Paoli, F., Ellul, J., Frisoni, G., et al. (2010). An ontological modeling approach to cerebrovascular disease studies: The NEUROWEB case. *Journal of Biomedical Informatics*, 43(4), 469-484.
- Coltell, O., Arregui, M., Perez, C., Domenech, M. A., Corella, D., & Chalmeta, R. (2004). *Building an ontology on genomic epidemiology of cardiovascular diseases*. Orlando: Int Inst Informatics & Systemics.
- Corcho, O., Fernandez, M., & Gomez, A. (2003). Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data & Knowledge Engineering*, 46, 41-64.
- Cowell, L. G., & Smith, B. (2010). The Infectious Disease Ontology. In S. V. (Ed.), *Infectious Disease Informatics* (Vol. Chapter 19, pp. P373-395). New York: Springer
- Cur, O., & #233. (2012). Improving the Data Quality of Drug Databases using Conditional Dependencies and Ontologies. *J. Data and Information Quality*, 4(1), 1-21.
- 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.
- . Diabetes Management in General Practice Guidelines for Type 2 Diabetes (2012). In D. A. a. R. A. C. o. G. Practitioners (Ed.), (Seventeenth edition 2011/12 ed., Vol. 2011/12): Diabetes Australia.

- Dixon, B., McGowan, J., & Grannis, G. (2011). *Electronic Laboratory Data Quality and the Value of a Health Information Exchange to Support Public Health Reporting Processes*. Paper presented at the AMIA 2011 Annual Symposium Improving Health: Informatics and IT Changing the World.
- 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.
- Fernandez, M. (1999). *Overview Of Methodologies For Building Ontologies*. Paper presented at the IJCAI-99 workshop on Ontologies and Problem-Solving Methods (KRR5), Stockholm, Sweden.
- 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.
- Ganguly, P., Ray, P., & Parameswaran, N. (2005). Semantic Interoperability in Telemedicine through Ontology-Driven Services. *Telemedicine & e-Health*, 11(3), 8.
- Gennari, J. H., Musen, M. A., Ferguson, R. W., Grosso, W. E., Eriksson, H., Noy, N. F., et al. (2003). The evolution of Protege: an environment for knowledge-based systems development. *Int. J. Hum.-Comput. Stud.*, 58(1), 89-123.
- Ghapanchi, A. H., & Aurum, A. (2011, 4-7 Jan. 2011). *Measuring the Effectiveness of the Defect-Fixing Process in Open Source Software Projects*. Paper presented at the System Sciences (HICSS), 2011 44th Hawaii International Conference on.
- Ghapanchi, A. H., & Aurum, A. (2012). The impact of project capabilities on project performance: Case of open source software projects. *International Journal of Project Management*, 30(4), 407-417.
- Gilbert, M., & Ddembe, W. (2008). A Flexible Approach for User Evaluation of Biomedical Ontologies. *International Journal of Computing and ICT Research*, 2(2), 62-74.
- Hadzic, M., & Chang, E. (2004). Role of the ontologies in the context of grid computing and application for the human disease studies. *Semantics of a Networked World: Semantics for Grid Databases*, 3226, 316-318.
- Hadzic, M., Dillon, D. S., & Dillon, T. S. (2009). *Use and Modeling of Multi-agent Systems in Medicine*.
- Huaman, M. A., Araujo-Castillo, R. V., Soto, G., Neyra, J. M., Quispe, J. A., Fernandez, M. F., 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 Medical Informatics and Decision Making*, 9.
- Huang, T., Li, W., & Yang, C. (2008). *Comparison of Ontology Reasoners: Racer, Pellet, Fact++* Paper presented at the American Geophysical Union, Fall Meeting 2008.
- Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information quality benchmarks: product and service performance. *Communications of the ACM*, 45(4), 8.
- Kuziemsky, C., & Lau, F. (2010). A four stage approach for ontology-based health information system design. *Artificial Intelligence in Medicine* 50, 133-148.
- Lee, C. S., Wang, M. H., Acampora, G., Loia, V., Hsu, C. Y., & Ieee. (2009). *Ontology-based Intelligent Fuzzy Agent for Diabetes Application*. New York: Ieee.

- Lezcano, L., Sicilia, M.-A., & Rodríguez-Solano, C. (2011). Integrating reasoning and clinical archetypes using OWL ontologies and SWRL rules. *Journal of Biomedical Informatics*, 44(2), 343-353.
- 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)*. Paper presented at the AMIA 2011 Annual Symposium Improving Health: Informatics and IT Changing the World.
- Liaw, S. T., 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. *Int J Med Inform*, 82(1), 10-24.
- 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.
- Maragoudakis, M., Lymberopoulos, D., Fakotakis, N., Spiropoulos, K., & Ieee. (2008). A Hierarchical, Ontology-Driven Bayesian Concept for Ubiquitous Medical Environments- A Case Study for Pulmonary Diseases 2008 30th Annual International Conference of the Ieee Engineering in Medicine and Biology Society, Vols 1-8 (pp. 3807-3810). New York: Ieee.
- McBride, S. J., Lawley, M. J., Leroux, H., & Gibson, S. (2012). Using Australian medicines terminology (AMT) and SNOMED CT-AU to better support clinical research. *Stud Health Technol Inform*, 178, 144-149.
- Min, H., Manion, F. J., Goralczyk, E., Wong, Y. N., Ross, E., & Beck, J. R. (2009). Integration of prostate cancer clinical data using an ontology. *Journal of Biomedical Informatics*, 42(6), 1035-1045.
- Nimmagadda, S. L., Nimmagadda, S. K., Dreher, H., & Ieee. (2008). Ontology based data warehouse modeling and managing ecology of human body for disease and drug prescription management 2008 2nd Ieee International Conference on Digital Ecosystems and Technologies (pp. 465-473).
- O-Hoon, C., Jung-Eun, L., Hong-Seok, N., & Doo-Kwon, B. (2008). *An Efficient Method of Data Quality using Quality Evaluation Ontology*. Paper presented at the Third 2008 International Conference on Convergence and Hybrid Information Technology.
- Pan Du, P., Feng, G., Flatow, J., Song, J., Holko, M., Kibbe, W., et al. (2009). From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations *Bioinformatics* 25,(12), i63-i68.
- 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, R. C., Bielinski, S. J., & Chute, C. G. (2012a). Applying semantic web technologies for phenome-wide scan using an electronic health record linked Biobank. *J Biomed Semantics*, 3(1), 10.
- Pathak, J., Kiefer, R. C., Bielinski, S. J., & Chute, C. G. (2012b). Mining the human phenome using semantic web technologies: a case study for Type 2 Diabetes. *AMIA Annu Symp Proc*, 2012, 699-708.
- Pathak, J., Kiefer, R. C., & Chute, C. G. (2012). Using semantic web technologies for cohort identification from electronic health records for clinical research. *AMIA Summits Transl Sci Proc*, 2012, 10-19.

- Peleg, M., Keren, S., & Denekamp, Y. (2008). Mapping computerized clinical guidelines to electronic medical records: knowledge-data ontological mapper (KDOM). *J Biomed Inform*, 41(1), 180-201.
- Pinto, H. S. (2004). Ontologies: How can They be Built? *Knowledge and Information Systems*, 6(4), 441-464.
- Pipino, L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Association for Computing Machinery. *Communications of the ACM* 45(4), 211-218.
- 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 & Experience*, 20(3), 253-264.
- Rahimi, A., Liaw, S., Ray, P., & Taggart, J. (2012). *Developing an ontology for data quality in chronic disease management*. Paper presented at the the 24th European Medical Informatics Conference.
- Rahimi, A., Liaw, S., Ray, P., Taggart, J., & Yu, H. (2014). Ontological specification of quality of chronic disease data in EHRs to support decision analytics: a realist review. *Decision Analytics*, 1(5), 31.
- Rahimi, A., Liaw, S. T., Taggart, J., Ray, P., & Yu, H. (in-press). Validating an ontology-based algorithm to identify patients with Type 2 Diabetes Mellitus in electronic health records.
- Rector, A., & Rogers, J. (2005). Ontological & practical issues in using a description logic to represent medical concepts: experience from GALEN *Tech rep CS* (Vol. 35, pp. 1-35). Manchester, England: School of Computer Science, University of Manchester.
- Richesson, R. L., & Krischer, J. (2007). Data standards in clinical research: Gaps, overlaps, challenges and future directions. *J Am Med Inform Assoc*, 14(6), 687-696.
- Rodriguez-Muro, M., & Calvanese, D. (2012). *Quest, a system for ontology based data access*: OWLED.
- Rodriguez-Muro, M., Kontchakov, R., & Zakharyashev, M. (2013). OBDA with Ontop. *Proc. of the OWL Reasoner Evaluation Workshop*.
- Rubin, D. L., Lewis, S. E., Mungall, C. J., Misra, S., Westerfield, M., Ashburner, M., et al. (2006). National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *OMICS*, 10(2), 185-198.
- Talaei-Khoei, A., Solvoll, T., Ray, P., & Parameshwaran, N. (2011). Policy-based Awareness Management (PAM): Case study of a wireless communication system at a hospital. *Journal of Systems and Software*, 84(10), 1791-1805.
- Talaei-Khoei, A., Solvoll, T., Ray, P., & Parameshwaran, N. (2012). Maintaining awareness using policies; Enabling agents to identify relevance of information. *Journal of Computer and System Sciences*, 78(1), 370-391.
- Valencia-Garcia, R., Fernandez-Breis, J. T., Ruiz-Martinez, J. M., 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., 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 C. S. Leung, M. Lee & J. H. Chan (Eds.), *ICONIP 2009, Part II* (pp. 360–366). Berlin Springer-Verlag
- Verma, A., Kasabov, N., Rush, A., & Song, Q. (2008, 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
- Wand, Y., & Wang, Y. (1996). Anchoring Data Quality Dimensions in Ontological Foundations. *COMMUNICATIONS OF THE ACM*, 36(11), 86-95.
- Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, 41(2 (Feb)), 58-65.
- Ying, W., Wimalasiri, J., Ray, P., Chattopadhyay, s., & Wilson, C. (2010). An Ontology Driven Multi-Agent Approach to Integrated e-Health Systems *International Journal of E-Health and Medical Communications (IJEHMC)*, 1(1), 28-40.
- Yu, H., Liaw, S., Taggart, J., & Rahimi, A. (2013). *Using Ontologies to Identify Patients with Diabetes in Electronic Health Records*. Paper presented at the Proceedings of the 12th International Semantic Web Conference and the 1st Australasian Semantic Web Conference, Sydney, Australia.

CHAPTER 4

VALIDATING AN ONTOLOGY-BASED ALGORITHM TO IDENTIFY PATIENTS WITH TYPE 2 DIABETES MELLITUS IN ELECTRONIC HEALTH RECORDS

Chapter 4 reports on the validation of the Diabetes Mellitus Ontology (DMO) developed as described in Chapter 3 using real-world EHR data from the ePBRN in South Western Sydney. The sensitivity and specificity (accuracy) of the algorithm to identify patients with T2DM were bench-marked by a manual EHR audit. Accuracy was determined using Reason for Visit (RFV), Medication (Rx) and Pathology (Path), singly and in combination. The combination was based on the DMO.

Chapter 4 addresses the gap identified in Chapter 2: insufficient practical research on the development and validation of ontology-based approaches in the assessment and management of large patient datasets and insufficient studies on the development and testing of information models based on clinical scenarios to systematically test quality of data in chronic diseases.

The other issue is to examine if the ontology-based approach is semantically flexible and modular to enable the development of intelligent software agents to act in various semantic contexts to identify patients with T2DM, support decision making about diabetes care, and conduct audit and evaluation research into diabetes. The paper published for this chapter summarised the validation of the DMO against a manual audit of EHRs.

NOTICE: This is the author's version of a work that was accepted for publication in the "International Journal of Medical Informatics". Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published as:

Alireza Rahimi, Siaw-Teng Liaw, Jane Taggart, Pradeep Ray and Hairong Yu (2014). Validating an ontology-based algorithm to identify patients with Type 2 Diabetes Mellitus in electronic health records. *International Journal of Medical Informatics* 83 (2014), pages 768-788.

Validating an Ontology-based Algorithm to Identify Patients with Type 2 Diabetes Mellitus in Electronic Health Records

Alireza Rahimi^{1,2,5}, alireza.rahimikhorzoughi@unsw.edu.au

Siaw-Teng Liaw^{1,3,4*}, siaw@unsw.edu.au

Jane Taggart^{1,3}, j.taggart@unsw.edu.au

Pradeep Ray⁵, p.ray@unsw.edu.au

Hairong Yu³, hairong.yu@unsw.edu.au

¹ UNSW, School of Public Health & Community Medicine, Sydney, Australia

² Isfahan University of Medical Sciences, Faculty of Management and Medical Information Sciences, Isfahan, Iran

³ UNSW, Centre for Primary Health Care & Equity, Sydney, Australia

⁴ South Western Sydney Local Health District (SWSLHD), General Practice Unit, Sydney, Australia

⁵ UNSW, Asia-Pacific Ubiquitous Healthcare Research Centre, Sydney, Australia

* Corresponding author. UNSW/SWSLHD School of Public Health & Community Medicine, General Practice Unit, PO Box 5, Fairfield, NSW 1860 Sydney, Australia

**University of NSW
Authorship Declaration**

In the case of the paper “Validating an ontology-based algorithm to identify patients with Type 2 Diabetes Mellitus in electronic health records”, contributions to the work involved the following:

Name	Nature of contribution
Alireza Rahimi	Development of the theoretical framework and templates for the validation of study, managed the study, audited all patients’ information from MD3 and also queried data, prepared this paper iteratively with input from all co-authors prior to submission and drafting of the manuscript
Siaw-Teng Liaw	Development of the theoretical framework and templates for the validation of study, oversight of the study and critical review of the manuscript
Jane Taggart	Development of the theoretical framework and templates for the validation of study, oversight of the study and critical review of the manuscript
Pradeep Ray	Development of the theoretical framework and templates for the validation of study
Hairong Yu	Support to solve technical problems in the database and critical review of the manuscript


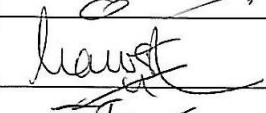
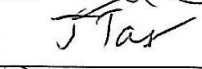

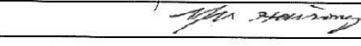
Declaration by co-authors

The undersigned hereby certify that:

- 1) they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field or expertise;
- 2) they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
- 3) there are no other authors of the publication according to these criteria;
- 4) potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
- 5) the original data are stored at the following location(s) and will be held for at least five years from the date indicated below:

Location:

General Practice Unit, Hospital Fairfield, UNSW, Sydney, Australia

Name	Signature	Date
Alireza Rahimi		29/05/2014
Siaw-Teng Liaw		09/MAY/2014
Jane Taggart		13/5/2014
Pradeep Ray		13/5/2014
Hairong Yu		15/5/2014

Abstract

Background: Improving healthcare for people with chronic conditions requires clinical information systems that support integrated care and information exchange, emphasising a semantic approach to support multiple and disparate Electronic Health Records (EHRs). Using a literature review, the Australian National Guidelines for Type 2 Diabetes Mellitus (T2DM), SNOMED-CT-AU and input from health professionals, we developed a Diabetes Mellitus Ontology (DMO) to diagnose and manage patients with diabetes. This paper describes the manual validation of the DMO-based approach using real-world EHR data from a general practice (n=908 active patients) participating in the electronic Practice Based Research Network (ePBRN).

Method: The DMO-based algorithm to query, using Semantic Protocol and RDF Query Language (SPARQL), the structured fields in the ePBRN data repository was iteratively tested and refined. The accuracy of the final DMO-based algorithm was validated with a manual audit of the general practice EHR. Contingency tables were prepared and Sensitivity and Specificity (accuracy) of the algorithm to diagnose T2DM, using the T2DM cases found by manual EHR audit as the gold standard. Accuracy was determined with three attributes - Reason for Visit (RFV), Medication (Rx) and Pathology (Path) – singly and in combination.

Results: The Sensitivity and Specificity of the algorithm were 100% and 99.88% with RFV; 96.55% and 98.97% with Rx; and 15.6% and 98.92% with Path. This suggests that Rx and Path data were not as complete as the RFV for this general practice, which kept its RFV information complete and current for diabetes. However, the completeness is good enough for this purpose as confirmed by the very small relative deterioration of the accuracy (Sensitivity and Specificity of 97.67% and 99.18%) when calculated for the combination of RFV, Rx and Path. The manual EHR audit suggested that the accuracy of the algorithm was influenced by data quality such as unavailable data due to non-documentation or documented in the wrong place or progress notes, problems with data extraction, encryption and data management errors.

Conclusion: This DMO-based algorithm is sufficiently accurate to support a semantic approach, using the RFV, Rx and Path to define patients with T2DM from

EHR data. However, the accuracy can be compromised by incomplete or poor quality data. The extent of compromise requires further study, using ontology-based and other approaches.

Keywords: Ontology; SPARQL; Electronic Health Records; Diabetes Mellitus, Type 2; Validation studies

4.1 Introduction

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 (DQ) in health organizations can improve the quality of decisions in health care (Kerr, Norris, & Stockdale, 2007). It also can lead to better policy that actually meet needs, strategies, evidence-based care and patient outcomes in Chronic Disease Management (CDM) (Dixon, et al., 2011).

Ontologies have been proposed as a method to assure quality of information through representing the meaning of a scientific domain and supporting the sharing of domain knowledge between human and computer programs. In the biomedical informatics literature, ontologies have been described as “collections of formal, machine-process able and human interpretable representation of the entities, and the relations among those entities, within a definition of the application domain” (Rubin, et al., 2006), drawing on the general definition by Gruber: “an explicit, formal specification of a shared conceptualization” (Gruber, 1995). Explicit concepts and the relationships and constraints are clearly defined and understood by the user. A formal ontology is computer-readable, allowing the computer to understand the relationships - the ‘formal semantics’ - of the ontology.

Our previous realist systematic review of the domain highlighted that the major categories of use of ontologies were in semantic data interoperability (Topalis et al., 2011; Ying, et al., 2010); information retrieval, DQ management (Brüggemann & Grüning, 2009), data collection, data sharing and data integration (Min, et al., 2009; O'Donoghue, Herbert, O'Reilly, & Sammon, 2009) in clinical information systems (CIS) for CDM; and regular validation of key data items in clinical data warehouses (CDW) (Nimmagadda, et al., 2008; Perez-Rey, et al., 2006). This review also showed that, while ontology-based approaches to chronic disease management, patient registers, DQ management and semantic interoperability are increasing, they were not systematic or comprehensive in the assessment of the quality of data in CDM (S. T. Liaw, et al., 2013). This is compounded by a lack of studies that evaluated the efficacy of the ontological approach or the relationship to DQ or improved integrated CDM (Rahimi, et al., 2014).

An ontological approach has theoretical and practical advantages in developing automated methods to identify patients with chronic diseases to guide clinical care, quality improvement and research (Lee, et al., 2009; Lezcano, et al., 2011). It may provide the breadth and depth of knowledge required to usefully represent clinical data and to develop type 2 diabetes mellitus (T2DM) registers semantically from EHRs. The application would be flexible and modular, enabling the development of intelligent software agents to act in various semantic contexts to identify patients with T2DM, support decision making about diabetes care, conduct audit and evaluation research into diabetes (Pathak, Kiefer, et al., 2012b; Perez-Rey, et al., 2006; Spasic & Ananiadou, 2005).

However, the automated identification of T2DM cases to create patient registers is complex. Traditional data model building methods, using concept analysis syntactically and grounded theory development, are not able to easily include the assessment and management of quality of data and information to build credible clinical knowledge (Pathak, Kiefer, & Chute, 2012). Ontology-based semantic methodologies, formalised tools in computer science and engineering, can potentially provide the technical solution to represent the required knowledge for effective chronic disease management (CDM) in general practice and primary care. This will also support medical research to assess and manage quality of clinical information (Cur, 2012), leading to more accurate decisions (Kuziemsky & Lau, 2010; Lin, et al., 2011; Mabotuwana & Warren, 2009).

Setting of study: The electronic Practice Based Research Network (ePBRN) in south western Sydney uses GRHANITE™, a privacy-preserving data extraction, aggregation, linkage and management tool, to establish a pseudonymised data repository of multiple EHRs (S. Liaw, et al., 2011). The ePBRN developed and implemented a T2DM identification algorithm, using SQL tools. The key assumption is that the automated identification of T2DM patients is an application of semantic retrieval, i.e. selection criteria are expressed as semantic queries, which are then processed within an ontology to identify eligible patients and extract relevant data and information from the EHR and/or data repository, and to infer implicit knowledge from ontologies simultaneously.

In the ePBRN project, we defined the elements of DQ based on the literature and for our purposes as follows:

Completeness: We defined two levels of completeness. The first was the availability of at least one record the patient reason for visit (RFV), diabetes medication (Rx), and specific diabetes pathology test (Path). The second level was the availability of all information required to make a clinical decision.

Correctness: A valid and appropriate clinical record with correct unit of measurements and within acceptable clinical range.

Consistency: Using a uniform data type, format and standard terminology (S. Liaw, et al., 2011).

This paper is part of ongoing study to develop automated ontology-based approaches to EHR data within CDM. It also implements the ontological query based approach to patient registers, DQ management and semantic interoperability, using the ePBRN “big data”. It aims to develop a Diabetes Mellitus Ontology (DMO), using formal ontology development methodologies to define formal, machine-process-able and human-interpretable representations of the concepts, and their relationships. A unified context is specified to allow the software agent to act in spite of differences in terminology and semantics from different EHRs, support reusability and integration of data, thereby supporting the development of automated systems for data annotation, extraction and linkage, information retrieval, DQ management and natural-language processing (Rubin, et al., 2006). The model was developed through a literature review, implementation of the 17th edition of the Australian National Guidelines for T2DM ("Diabetes Management in General Practice Guidelines for Type 2 Diabetes ", 2012); and refined with input from participating health professionals. Consistency with standard terminologies such as SNOMED-CT AU was requirement.

The DMO was subsequently implemented, using the Semantic Protocol and RDF Query Language (SPARQL) to identify T2DM phenotypes in a EHR-derived dataset. By incorporating defined semantic SPARQL queries, DMO was able to generate logical inferences and control the inclusion/exclusion of relevant objects (Perez-Rey, et al., 2006), such as the patient with a T2DM-specific RFV, Rx or pathology test (Berg, 2003). The validation of the DMO-based algorithm included a

comparison with a manual audit of the EHR from which the data was derived. This study described the validation of the accuracy of an ontology-based semantic query to identify cases of T2DM in a general practice EHR.

4.2 Methods

The following tasks were conducted to validate the DMO-based algorithm to identify T2DM patients.

4.2.1 Establishing the Gold Standard with a manual audit

The gold standard for the general practice dataset was established with a manual audit of the EHR of the smallest participating general practice in the ePBRN. This contained 927 patients with at least 3 visits in a two years period (between 2010 and 2012). This is the definition of an “active patient” by the Royal Australian College of General Practitioners (RACGP). Because the ePBRN dataset is pseudonymised by GRHANITE™ prior to extraction, these 927 RACGP-active patients were re-identified by a reverse process on the same computer in the participating general practice. Re-identification enabled the researcher to manually audit patient records in the EHR, using information in both structured (e.g., RFV, Rx and pathology test results) and text (e.g., family history, clinical notes and scanned paper copies of Path reports) fields to categorize them as T2DM or other DM such as Type 1 DM (T1DM) or Gestational DM (GDM) or not DM at all. A specific template (Figure 4.1) was used to ensure comprehensiveness and consistency as well as systematic cross-checking by another independent records auditor (JT). In addition, the manual audit of the EHR looked for T2DM-related information such as Hemoglobin A1c (HbA1C), Fasting Blood Glucose (FBG) and Random Blood Glucose (RBG), relevant referrals, and BMI to assist with the T2DM categorization. The audit data were stored in an Excel spreadsheet.

4.2.4 Assessing the accuracy of the algorithm

The accuracy of the DMO-based SPARQL queries to identify T2DM in the general practice dataset (=EHR) was compared with the accuracy of the manual audit (gold standard), using the calculated Sensitivity and Specificity of both methods. Sensitivity (Sn) and Specificity (Sp) are easy to understand concepts based on contingency (2x2) tables (Figure 4.2). The true positives (TP) were patients with T2DM attributes found with both manual audit and algorithm; false positives (FP) were patients with T2DM attributes found with the algorithm but not the manual audit; false negatives (FN) were patients with T2DM attributes found by the manual audit but not the algorithm; and true negatives (TN) were patients without T2DM attributes in both manual audit and algorithm. Sensitivity ($TP / (TP + FN)$) is the ability of the algorithm to accurately identify all T2DM patients and Specificity ($TN / (TN + FP)$) is the ability to accurately identify all patients without T2DM.

		Manual check (+)	Manual check (-)
SPARQL query results using RFV, Rx, Path and all 3	+	TP	FP
	-	FN	TN

$$\text{Sensitivity for patient's attributes} = \frac{TP}{TP + FN} \times 100$$

$$\text{Specificity for patient's attributes} = \frac{TN}{TN + FP} \times 100$$

Figure 4.2: Comparing Manual and Automated identification of T2DM patients from the general practice EHR

4.2.5 Examining reasons for false positives and false negatives

We analysed the database and interviewed relevant staff at the participating general practice to understand the possible reasons for the discrepancy between the manual audit and the DMO-based algorithm.

4.3 Results

4.3.1 Sample size: Data cleaning of the initial data extract from the general practice (n=927) revealed that some patients were deceased, inactive or duplicated; 19 patients were excluded, leaving 908 patients in the final dataset for the study.

4.3.2 The DMO-based algorithm: The feedback from participating clinicians suggests that the DMO is a realistic model of the real world of diabetes diagnosis and

management. This underpinned the specific hierarchical conceptual modeling to formalize the ontology. The formalized DMO consists of 68 concepts in 4 main classes (*Actor*, *Content*, *Mechanism* and *Impact*) and 51 subclasses (Figure 4.3) with 8 object properties and 15 data properties. Some of the concepts are map-able to the SNOMED-CT-AU Ontology (SCAO).



Figure 4.3: Diabetes Mellitus Ontology hierarchical conceptual model

To run the DMO-based algorithm over the general practice dataset, 11 mappings were created, linking the relevant structured fields from 7 tables in dataset (“Active_Patients”, “CONSULTATION”, “DIAGNOSIS”, “PRESCRIPTION”, “HISTORY”, “PRESCRIPTION_REPEAT” and “PATHOLOGY”) with the 4 classes of the DMO (“GPUActivePatients”, “Diagnosis”, “Medication” and “PathologyTest”).

As can be shown in Figure 4.4, we presented the requirements for mapping and querying for all patients with the T2DM reason for visits (RFVs). We use “hasT2DMRFV” as an object property to joint two tables using Patient_UUID as a unique identifier and relevant data properties to identify active patients with T2DMRFV through semantic SPARQL queries automatically.

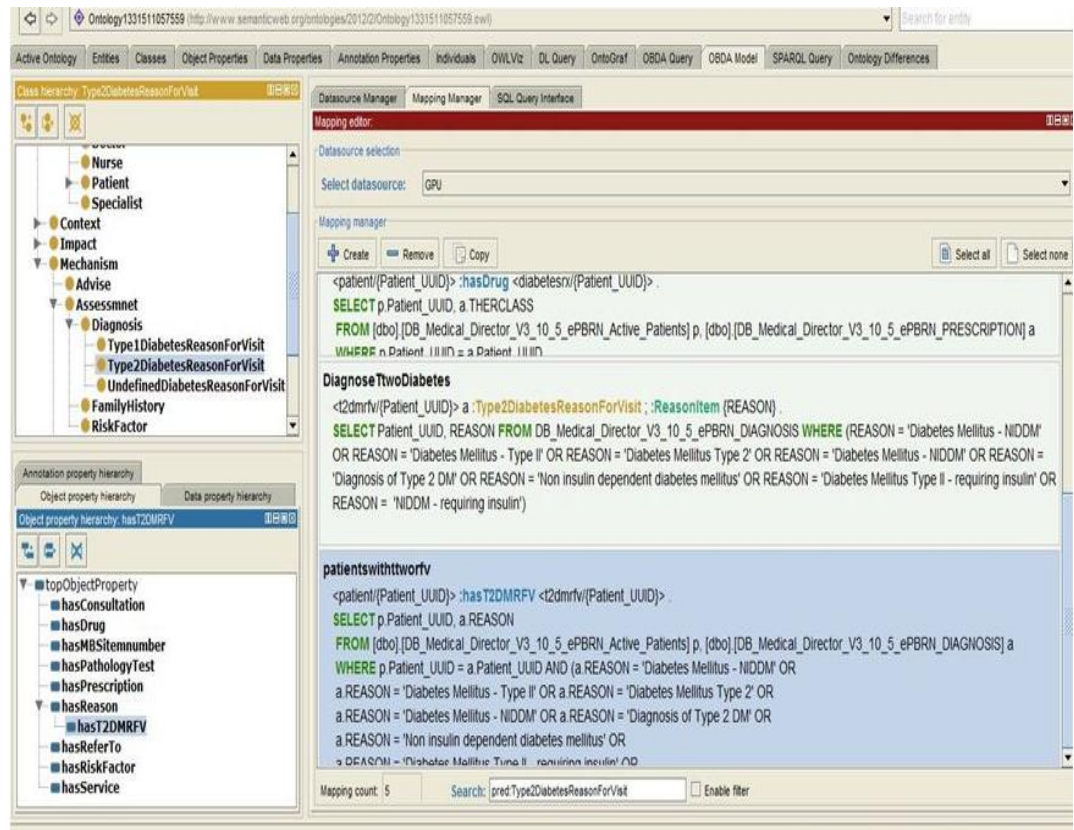


Figure 4.4: Sample of mapping the Diagnosis table with DMO

Table 4.1 shows the results of SPARQL queries for the 3 levels (RFV, Rx, Path) used to identify T2DM patients. The SPARQL queries only used classes, object properties and data properties to combine the main diabetes criteria for the identification of diabetes semantically. At the first level (RFV), the algorithm found 25 patients; at the second level (Rx), 32 patients were identified; and at the third level (abnormal Path: HbA1c \geq 6.5, Fasting Glucose and Random Glucose \geq 7), 73 patients were identified.

Table 4. 1: SPARQL queries for the research purpose

Diabetes Identification level		SPARQL queries	Results
1.	Only T2 DM RFV	<pre>SELECT DISTINCT ?pid WHERE {{?pid a :GPUactivePatients. ?pid :hasT2DMRFV ?r. ?r :ReasonItem ?reason. FILTER(?reason = "Diabetes Mellitus - Type II"^^xsd:String ?reason = "Diabetes Mellitus - NIDDM"^^xsd:String ?reason = "Diagnosis of Type 2 DM"^^xsd:String ?reason = "Diabetes Mellitus Type II - requiring insulin"^^xsd:String ?reason = "Diabetes Mellitus - Type II"^^xsd:String ?reason = "Diabetes Mellitus Type 2"^^xsd:String ?reason = "NIDDM - requiring insulin"^^xsd:String ?reason = "Non insulin dependent diabetes mellitus"^^xsd:String ?reason = "NIDDM"^^xsd:String ?reason = "Diabetes Mellitus - NIDDM"^^xsd:String)}} UNION {?pid a :GPUactivePatients. ?pid :hasT2DMHistory ?h. ?h :Condition ?history. FILTER(?history = "Diabetes Mellitus - NIDDM"^^xsd:String ?history = "Diabetes Mellitus - Type II"^^xsd:String ?history = "Non insulin dependent diabetes mellitus"^^xsd:String ?history = "NIDDM"^^xsd:String ?history = "Diagnosis of Type 2 DM"^^xsd:String ?history = "Diabetes Mellitus Type II - requiring insulin"^^xsd:String ?history = "Diabetes Mellitus Type 2"^^xsd:String ?history = "NIDDM - requiring insulin"^^xsd:String)}}</pre>	25
2.	Only T2 DM Rx	<pre>SELECT DISTINCT ?pid WHERE {{?pid a :GPUactivePatients. ?pid :hasDrug ?d. ?d :TherapyClass ?rx. FILTER(?rx = "HDI"^^xsd:String ?rx = "HDO"^^xsd:String ?rx = "HDI"^^xsd:String ?rx = "ODB"^^xsd:String ?rx = "HD"^^xsd:String ?rx = "HDOA"^^xsd:String ?rx = "HDOD"^^xsd:String ?rx = "ODU"^^xsd:String)}} UNION {?pid a :GPUactivePatients. ?pid :hasRepeatDrug ?r. ?r :TherapyClass ?rerr. FILTER(?rerr = "HDI"^^xsd:String ?rerr = "HDO"^^xsd:String ?rerr = "HDI"^^xsd:String ?rerr = "ODB"^^xsd:String ?rerr = "HD"^^xsd:String ?rerr = "HDOA"^^xsd:String ?rerr = "HDOD"^^xsd:String ?rerr = "ODU"^^xsd:String)}}</pre>	32
3.	Only Abnormal Pathology	<pre>SELECT DISTINCT ?pid WHERE {?pid a :GPUactivePatients. ?pid :hasT2DMPathologyTest ?p. ?p :TestName ?test. ?p :ResultTest ?result. FILTER(?test = "HbA1C"^^xsd:String && ?result >= "6.5"^^xsd:Integer ?test = "GLUCOSE PLASMA FASTING"^^xsd:String && ?result >= "7.0"^^xsd:Integer ?test = "GLUCOSE Random"^^xsd:String && ?result >= "11.1"^^xsd:Integer ?test = "Glucose Fasting"^^xsd:String && ?result >= "7.0"^^xsd:Integer)}</pre>	73

In Figure 4.5, the results at the first level (RFV) shows how a flexible semantic approach use the different object and data properties as well as the relevant classes like “Diagnosis” and “History” to combine the main important attribute (T2DM RFV) to identify 25 T2DM patients.

OBDA query editor

Query Editor

PREFIX quest: <http://obda.org/quest#>
 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?pid WHERE {{?pid a :GPUActivePatients. ?pid :hasT2DMRFV ?r. ?r :ReasonItem ?reason. FILTER(?reason = "Diabetes Mellitus - Type II""xsd:String || ?reason = "Diabetes Mellitus - NIDDM""xsd:String || ?reason = "Diagnosis of Type 2 DM""xsd:String || ?reason = "Diabetes Mellitus Type II - requiring insulin""xsd:String || ?reason = "Diabetes Mellitus - Type II""xsd:String || ?reason = "Diabetes Mellitus Type 2""xsd:String || ?reason = "NIDDM - requiring insulin""xsd:String || ?reason = "Non insulin dependent diabetes mellitus""xsd:String || ?reason = "NIDDM""xsd:String || ?reason = "Diabetes Mellitus - NIDDM""xsd:String)) UNION {?pid a :GPUActivePatients. ?pid :hasT2DMHistory ?h. ?h :Condition ?history. FILTER(?history = "Diabetes Mellitus - NIDDM""xsd:String || ?history = "Diabetes Mellitus - Type II""xsd:String || ?history = "Non insulin dependent diabetes mellitus""xsd:String || ?history = "NIDDM""xsd:String || ?history = "Diagnosis of Type 2 DM""xsd:String || ?history = "Diabetes Mellitus Type II - requiring insulin""xsd:String || ?history = "Diabetes Mellitus Type 2""xsd:String || ?history = "NIDDM - requiring insulin""xsd:String))}}

Execution time: 0.286 sec - Number of rows retrieved: 25

Show: 0 ☒ All ☒ Short RI

pid
patient/070B59B8-DFED-4A9D-8B01-7DF6B40D4C15
patient/07C745A9-53FB-4ED3-B7C8-D54CF10A0BFF
patient/1941EC79-494B-42BA-94F6-E81E3F02A78A
patient/43B9CC64-A116-4B42-A50E-D91E84C9118F
patient/46AAA975-3770-4C0C-BFD0-BD53753D4805
patient/67E33CAE-B272-44F8-9471-0CC043845D07
patient/6B8F3E32-41DB-4BAC-93EA-CE1F62550ABF
patient/8B1F1485-18FE-4617-AB0E-F0FF7CCF51A4
patient/965FB833-B989-4235-8286-4C165C54FE5D
patient/96AF7591-DC1F-46FD-AD48-25E4607E895F
patient/97903F48-DB02-48F9-89F6-6D7A85903950
patient/9D0EFFFF-A847B-4BC8-9980-2ED63E02A10A

Hint: Try to continue scrolling down the table to retrieve more results.

Export to CSV...

The T2DM patients have those RFV conditions based on the various categories of T2DM RFVs from "History" and "Diagnosis" classes

Figure 4.5: Sample of SPARQL query to show a semantic way to implement RFV for the identification of T2DM patients

Figure 4.6 shows how a semantic approach, combining different object and data properties as well as relevant classes, combine all T2DM attributes (RFV, Rx and Path) identified 105 T2DM patients.

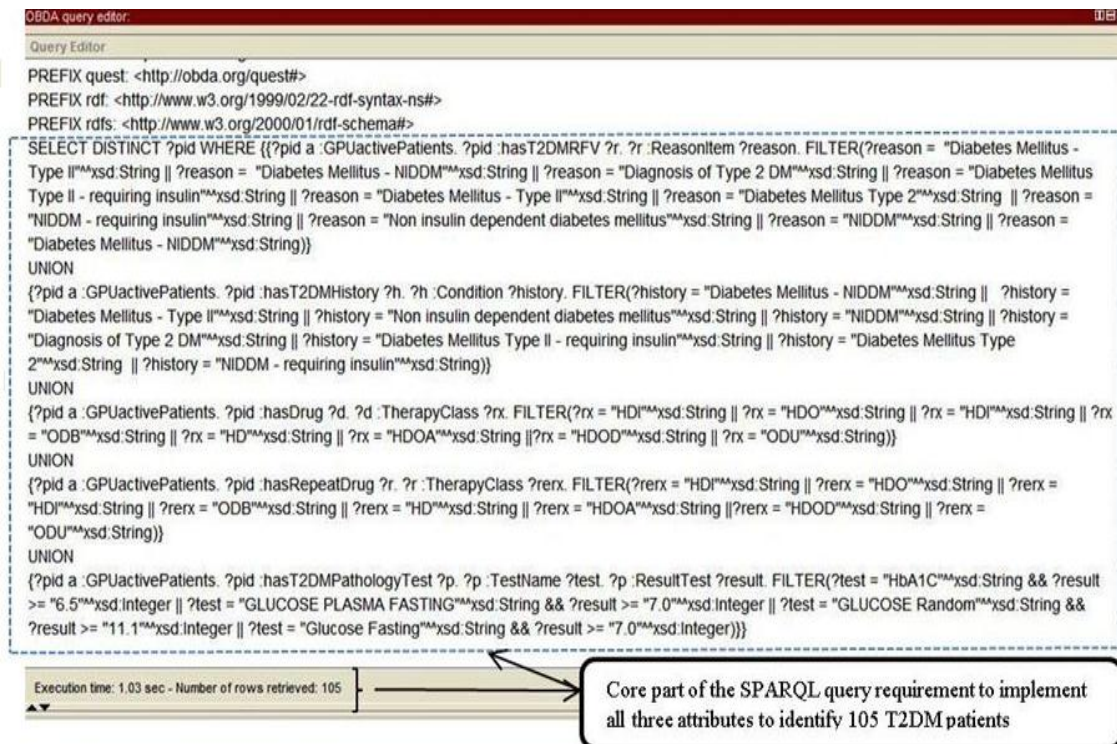


Figure 4.6: Sample of SPARQL queries to show a semantic way to implement 3 criteria for the identification of T2DM

Table 4.2 shows the results of the SPARQL queries compared with those of the manual audit of the EHR, (the gold standard), using the RFV, Rx and Path separately, and then in combination. The accuracy using RFV was 100% sensitive and 99.88% specific. This near-perfect accuracy is due to the fact that RFV data for this general practice was kept complete and current for T2DM RFV as a quality improvement activity.

Table 4.2: Sensitivity and Specificity of T2DM using RFV, Rx, Path and in combination

		Manual validation (RFV)			Manual validation (Rx)			Manual validation (Path)			Manual validation (All)		
		T2DM RFV	Non-T2DM RFV	Total	T2DM Rx	Non-T2DM Rx	Total	T2DM Path	Non-T2DM Path	Total	T2DM All	Non-T2DM All	Total
SPARQL query result	T2DM	25	0	25	28	1	29	11	62	73	41	1	42
	Non-T2DM	1	882	883	9	870	879	9	826	835	7	859	866
	Total	26	882	908	37	871	908	20	888	908	48	860	908

Using Rx, the algorithm was 96.55% sensitive and 98.97% specific. Using T2DM Path, the sensitivity was only 15.6% sensitive and 98.92% specific. The increased FP is due to a change in units for the HbA1C test, which led to an increase in the actual value of the HbA1C by about 10-fold. This revealed that the accuracy of the algorithm was determined by DQ such as unavailable data due to non-documentation or documented in the wrong place, and possibly incorrect ranges of pathology test results. The recent change in standard units for HbA1C posed another problem as demonstrated by Patient 3 in Table 4.3. However, the main reason for this poor quality of pathology data lies in the fact that this general practice EHR does not electronically import the majority of its pathology results because the pathology service it uses does not have the facility to export pathology results electronically. Although unlikely, there may also be a problem with data extraction. Table 4.3 demonstrates examples of patients where the algorithm was not accurate.

Table 4.2 showed improved performance of the algorithm with the combination of all three attributes for T2DM, compared with using each attribute separately. The algorithm was 97.67% sensitive ($41 \text{ true positives} / (41 \text{ true positives} + 1 \text{ false positive})$) and 99.18% specific ($859 \text{ true negatives} / (859 \text{ true negatives} + 7 \text{ false negatives})$). Similar DQ reasons were elucidated for the FN and FP found (Table 4.2). This improvement suggests that using more T2DM attributes in the semantic queries may improve Sensitivity and Specificity, offsetting the effect of poor DQ. While sensitivity and specificity are high, 25 out of 42 patients have a complete RFV record (59.6%). Also, 884 out of 908 patients have a correct record (97.36%).

Further check of the data and interviews with practice staff suggested that the accuracy of the algorithm was affected by the DQ such as unavailable data due to problems with data extraction, encryption and data management errors as illustrated by Patient 3 in Table 4.3. Table 4.3 shows selected examples of T2DM cases to demonstrate the pattern of agreement/disagreement between the results of the manual audit and algorithm.

Table 4.3: Sample of T2DM cases to demonstrate concordance and discordance between results of manual audit and algorithm

	Patient ID	Manual Check / Structured Data					Progress note	SPAE RQL results	Status	Reason for differences
		RFV	Medication windows	HbA1c	Glucose Fasting	Glucose Random				
1.	07C745A9-53FB-4ED3-B7C8-D54CF10A0BFF	✓	-	✓	✓	✓	✓	+	TP	Data structured fields and available to query
2.	52414785-715A-451C-917F-0289876312B2	-	-	-	-	-	-	-	TN	
3.	EDB8C08D-F3A1-43A9-B4E0-4CBFDB1A3328	-	-	-	-	-	-	+	FP	Possible problems with data extraction, encryption, management
4.	01710ED9-17B0-46F6-B91A-4C9D0C35FC7A	-	-	-	-	-	✓	-	FN	Data present only in progress notes, which are not queried
5.	0FCF364A-F1A7-4CA2-97EF-EBD849B4438E	-	-	-	-	-	✓	-	FN	Possibly due to upgrade to MD3
6.	783082B3-F3E1-40EF-89E5-E8FDAC80D8F0	-	✓	-	-	-	✓	-	FN	Poss problem with MD3 saving into script table
7.	9699F98C-0964-4FF0-9532-87E68A63B6F7	-	-	✓	-	-	✓	-	FN	Data entry human error. Path results not imported electronically

Table 4.3 shows a sample of cases to demonstrate agreement/disagreement between the manual audit and algorithm, along with possible reasons for the disagreement. Specific examples include:

- Patient 4 had T2DM RFV in their Progress Notes as free text but did not have any T2DM RFVs in the relevant structured fields or relevant tables.
- Patient 5 has T2DM and was prescribed Metformin back in 2008 as noted in the manual audit; however, the algorithm did not pick up the medication, possibly because it was inaccurately migrated, if at all, after a major software upgrade after the prescription of Metformin and before the data extraction from the EHR.
- Patient 6 has *Diaformin* visible in the Medication window and Progress Notes in the EHRs, but was not found in the extracted data set. This is possibly due to a problem with the data extraction and/or with the EHR not saving the prescription into the Script_table.
- Patient 7 has an abnormal HbA1C pathology test in the “Pathology tests” window, but did not have any T2DM-related tests in the general practice dataset.

This study re-affirms the fact that the documenting of clinical data as text in clinical notes remains a major reason for non-accessible data in EHRs. However, the problem appears to be decreasing across the ePBRN, particularly with eHealth literate general practices like this study practice; clinicians in this practice tend to document the data in both structured and text/narrative fields.

4. Discussion

This study demonstrated the relationship between DQ, as measured by metrics, and fitness for purpose, in this case, finding cases of T2DM in the EHR. Table 4.2 showed that manual validation of the algorithmic queries on the general practice dataset demonstrated nearly 100% accuracy with the use of RFV, slightly less so with Rx and worst with Pathology tests. This confirmed that the accuracy of the identification of T2DM cases in a general practice EHRs is influenced by the completeness of the dataset. This general practice was diligent with its documentation of RFVs for T2DM as part of a quality improvement exercise. However, the Pathology component of the data set was incomplete from not importing pathology results electronically and directly into the EHRs to any significant degree because the general practice’s main pathology

provider did not have the facility to export their reports electronically (Table 4.2). In addition to completeness, the correctness of the data was influenced by the change of units for HbA1c from % to SI units. Thus numerous false positives were identified by the algorithm, leading to reduced accuracy when using Path. Similarly, the accuracy with using Rx was influenced more by correctness as opposed to completeness as indicated by the increased numbers of false positives and false negative. However the effect on accuracy was not as pronounced as with pathology where the completeness was also not good enough.

Table 4.2 demonstrates that the accuracy was acceptable when all three attributes (RFV, RX and Path) were used as guided by the ontological definition of the patient with T2DM as someone with a relevant reason for visit (RFV), is prescribed a relevant medication (Rx) and has a relevant and correct pathology test (Path). It highlights that the ‘fitness for purpose’ definition is more than just the DQ metrics of each attribute. In this case, the ontology-based query to identify patients with T2DM was fit for purpose even when the completeness and correctness of pathology data, and to a certain extent prescribing data, did not appear to be acceptable. This finding confirms previous research that demonstrated the role of ontology-based approaches to assess DQ based on ‘fitness for purpose’ in the health context (Rahimi, et al., 2014).

The DMO included other concepts related to T2DM such as Referrals to a T2DM related service and risk factors such as BMI (obesity), Family history of T2DM and ethnicity. Applying a more complex ontology to identify diabetes (Chalortham, et al., 2009) may increase the accuracy of this semantic approach (McGarry, Garfield, & Wermter, 2007). The more important aspect of the ontological approach as used in this study is that we identified actual and possible cases of diabetes such as pre-diabetes or a tendency to diabetes. Clinicians can then target this vulnerable group with efforts to prevent diabetes (Buranarach, et al., 2009). The program can be run regularly to detect this cohort of patients predisposed to T2DM and how effective the preventive activities have been. Technically, this also demonstrates the flexibility of the ontological approach to model the real world of clinical practice where the patient is usually someone who requires multidisciplinary integrated care for multiple health issues (Cur, 2012).

We highlighted some of the detailed reasons why a T2DM patient may not be identified or, more rarely, why a patient without T2DM is identified as T2DM. Reasons include human errors in data entry; organizational problems like coping with changes in

units of measurements such as HbA1C; system idiosyncrasies like not importing pathology tests and results, or poorly designed system upgrades that loses data; technical problems with data management, and, potentially, technical problems at the data repository level (Poulsen et al., 2010; Valle et al., 2006). This study also confirmed that the documenting of clinical data as text in clinical notes remains a major reason for non-accessible data in EHRs. Incomplete data can occur due to non-documentation, documentation in the wrong place in the EHRs, and problems with data extraction and encryption for some factors (e.g., HbA1C results). This study has updated previous work and added new knowledge to explain non-identification of cases from EHRs (Armstrong, Lavery, Vela, Quebedeaux, & Fleischli, 1998).

The literature review for this study demonstrated a lack of valid and reliable DQ assurance activities (Rahimi, et al., 2014) to ensure fitness for a range of uses by consumers, patients, health providers and professionals. Scaling up from this study, the building of robust ontologies for DQ in health can automate purposeful extraction of data from EHRs into clinical data warehouses; assessment and management of the quality of data so that they are fit for purposes such as research, quality improvement and health information exchange and sharing; management of controlled vocabularies and optimizing semantic interoperability; curation of data for relevant 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, Ludäscher, Grethe, & Martone, 2003; Jara, Blaya, Zamora, Skarmeta, & Ieee, 2009). We can also scale across to other domains such as other chronic diseases to examine the generalizability and flexibility of this method. This ontology-based experiment shows that it is possible to automate the interpretation process and build a reusable conceptual infrastructure over various databases to cope with other research purposes. This approach is interoperable with applications of ontologies and the real-world databases as well as a reasoning-based query solution to identify patients with chronic disease from EHRs.

This work is in progress with our primary and integrated care “big data” repository of more than 100,000 patients (S. Liaw, et al., 2011) from general practices and hospital based services. We are developing automated approaches to support semantic reasoning with clinical data from EHRs; assessment and management of the

quality of clinical data such as reason for visit, chronic conditions, pathology tests and prescriptions; and representation of the meaning of the data and knowledge as they interpreted by clinicians (S. Liaw, et al., 2011). This will address the gap identified in the literature: insufficient practical research on the development and validation of ontology approaches in clinical scenarios for the assessment and management of large patients' datasets (S. T. Liaw, et al., 2013) and insufficient studies on the development and testing of information models based on clinical scenarios on systematically test quality of information in chronic diseases (Rahimi, et al., 2014).

The ontology-based query to identify T2DM patients is modular, enabling the development of intelligent software agents to act in various semantic contexts and identify patients with other chronic diseases, support decision making about health care, conduct audit and evaluate research (S. Liaw, et al., 2011). This formal ontology based approach also guide the development of an application to enable automated assessment of the quality of data of patients with other chronic diseases such as COPD or hypertension and identify these patients at various levels of clinical course of the disease to guide clinical care, quality improvement and research (Buranarach, et al., 2009; Chalortham, et al., 2009; Colombo, et al., 2010; Coltell, et al., 2004). This experiment has reinforced the significant theoretical advantages of the ontological approach to guide and support the development of automated methods to manage "big data" routinely collected in EHRs cost-effectively (S. T. Liaw, et al., 2013; Rahimi, et al., 2014; Taggart et al., 2012).

5. Limitations of the research

This ontology based approach was tested in a relatively small general practice in a very specific domain area. While it enabled manual validation to confirm the accuracy of the ontology-based approach, the generalisability may be limited. The data had a range of quality issues from a range of reasons, but they were not significant problems. Rather it highlighted the effect of completeness and correctness of data on the accuracy, or fitness for purpose, of the approach used. Other sociotechnical sources of data errors such as problems with data extraction, encryption and management are difficult to isolate and assess. We also did not allow for the possibility that some cases of T2DM are either undetected because of reasons other than poor DQ, such as the patient data being recorded outside the time frame established for this study or are cared for entirely

in specialty settings with no records of T2DM captured within this dataset. However, the relevance of all these issues are being clarified as part of our ongoing research program (S. Liaw, et al., 2011).

6. Conclusion

This study validated an automated ontology-based semantic query of routinely collected data from EHRs to identify and assess patients with T2DM. The accuracy was established and a direct relationship to DQ metrics demonstrated. The ontology-based approach reduced the impact of the incompleteness of the data set. However, more subtle questions about the impact of quality of patients' data on the accuracy measures remain and needs further research. Fitness for purpose includes DQ but may not require perfect DQ metrics if an ontology-based multi-attribute approach is adopted. The ontology to assess and manage DQ could be extended to include the sources of errors, both human and technical, at all points in the data cycle. Further research is needed to examine the application of this method in other technical, health and social domains.

Authors' contributions

AR, STL and JT developed the theoretical framework and templates for the validation of study. AR managed the study, audited all patients' information from MD3 and also queried data. AR and STL prepared this paper iteratively with input from all co-authors prior to submission. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank Dr Andrew Knight and Dr Sanjyot Vagholkar for their useful consultations during patients' data manual audit and Dr Sarah Dennis for previous and ongoing contributions in this study.

Statement on conflicts of interest

The authors declare that they have no competing interests.

Summary table

What was already known on the topic?	What this study added to our knowledge?
<ul style="list-style-type: none"> • Lack of valid DQ assurance activities to ensure fitness for a range of uses 	<ul style="list-style-type: none"> • Ontology can automate purposeful extraction of data from EHRs so that they are fit for purpose
<ul style="list-style-type: none"> • Lack of validation studies for semantic approaches to DQ in CDM 	<ul style="list-style-type: none"> • The ontological approach is sufficiently accurate to define patients with T2DM from EHRs
<ul style="list-style-type: none"> • Routinely collected data raises issues of poor DQ that can affect accuracy of data models 	<ul style="list-style-type: none"> • Accuracy of case finding is affected by incomplete DQ but an ontological approach can offset this problem
<ul style="list-style-type: none"> • The challenges to evaluation of ontologies include methodological immaturity, an immature knowledge base, and a lack of tools to support ontological approaches 	<ul style="list-style-type: none"> • An ontological approach has practical advantages in developing reusable and flexible tools to access and assess DQ in the large datasets

References:

- Abidi, S. R. (2011). *Ontology-based knowledge modeling to provide decision support for comorbid diseases*. Paper presented at the 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>
- Armstrong, D. G., Lavery, L. A., Vela, S. A., Quebedeaux, T. L., & Fleischli, J. G. (1998). Choosing a practical screening instrument to identify patients at risk for diabetic foot ulceration. *Arch Intern Med*, 158(3), 289-292.
- Arts, D., De Keizer, N., & Scheffer, G. J. (2002). Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework. *Journal of the American Medical Informatics Association*, 9(6).
- Arts, D., de Keizer, N., Scheffer, G. J., & de Jonge, E. (2002). 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, D. G., Bosman, R. J., de Jonge, E., Joore, J. C., & de Keizer, N. F. (2003). Training in data definitions improves quality of intensive care data. *Crit 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.
- Berg, M. (2003). The search for synergy: interrelating medical work and patient care information systems. *Methods Inf Med*, 42(4).
- Borst, W. N. (1997). *Construction of Engineering Ontologies*. University of Twente, Enschede, NL.
- Brown, P., Warmington, V., Laurence, M., & Prevost, A. (2003). Randomised crossover trial comparing the performance of Clinical Terms Version 3 and Read Codes 5 byte set coding schemes in general practice. *BMJ*, 326(7399), 1127.
- 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.
- Calvanese, D., Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M., et al. (2009). *Ontologies and Databases: The DL-Lite Approach, Reasoning Web. Semantic Technologies for Information Systems: 5th International Summer School 2009, Brixen-Bressanone, Italy, August 30-September 4, 2009, Tutorial Lectures*: Springer-Verlag, Berlin, Heidelberg.
- Chalortham, N., Buranarach, M., & Supnithi, T. (2009). Ontology Development for Type II Diabetes Mellitus Clinical Support System.
- Chen, X. H., Lu, J., & Liu, Z. Y. (2007). Assistance ontology of quality control for enterprise model using data mining. In M. Helander, M. Xie, M. Jaio & K. C. Tan (Eds.), *2007 Ieee International Conference on Industrial Engineering and Engineering Management, Vols 1-4* (pp. 602-606).
- Choquet, R., Qouiya, 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.
- Colombo, G., Merico, D., Boncoraglio, G., De Paoli, F., Ellul, J., Frisoni, G., et al. (2010). An ontological modeling approach to cerebrovascular disease studies: The NEUROWEB case. *Journal of Biomedical Informatics*, 43(4), 469-484.
- Coltell, O., Arregui, M., Perez, C., Domenech, M. A., Corella, D., & Chalmers, R. (2004). *Building an ontology on genomic epidemiology of cardiovascular diseases*. Orlando: Int Inst Informatics & Systemics.
- Corcho, O., Fernandez, M., & Gomez, A. (2003). Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data & Knowledge Engineering*, 46, 41-64.
- Cowell, L. G., & Smith, B. (2010). The Infectious Disease Ontology. In S. V. (Ed.), *Infectious Disease Informatics* (Vol. Chapter 19, pp. P373-395). New York: Springer
- Cummings, E., Showell, C., Roehrer, E., Churchill, B., Yee, K., Wong, M., et al. (2010). *Discharge, Referral and Admission: A Structured Evidence-based Literature Review*.
- Cur, O. (2012). Improving the Data Quality of Drug Databases using Conditional Dependencies and Ontologies. *J. Data and Information Quality*, 4(1), 1-21.
- Cur, O., & #233. (2012). Improving the Data Quality of Drug Databases using Conditional Dependencies and Ontologies. *J. Data and Information Quality*, 4(1), 1-21.
- 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. *Diabet Med*, 27, 203-209.
- Dentler, K., Cornet, R., Teije, A., & de Keizer, N. (2011). Comparison of Reasoners for large Ontologies in the OWL 2 EL Profile. In B. C. Grau (Ed.), *Semantic Web* (Vol. 1, pp. 1-51). Oxford University, UK: IOS Press.
- 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.
- . Diabetes Management in General Practice Guidelines for Type 2 Diabetes (2012). In D. A. a. R. A. C. o. G. Practitioners (Ed.), (Seventeenth edition 2011/12 ed., Vol. 2011/12): Diabetes Australia.
- Dixon, B., McGowan, J., & Grannis, G. (2011). *Electronic Laboratory Data Quality and the Value of a Health Information Exchange to Support Public Health Reporting Processes*. Paper presented at the AMIA 2011 Annual Symposium Improving Health: Informatics and IT Changing the World.

- Eccher, C., Purin, B., Pisanelli, D. M., Battaglia, M., Apolloni, I., & Forti, S. (2006). Ontologies supporting continuity of care: the case of heart failure. *Comput Biol Med*, 36(7-8), 789-801.
- 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.
- Fernandez, M. (1999). *Overview Of Methodologies For Building Ontologies*. Paper presented at the IJCAI-99 workshop on Ontologies and Problem-Solving Methods (KRR5), Stockholm, Sweden.
- Gamper, J., Nejd, W., & Wolpers, M. (1999). *Combining Ontologies and Terminologies in Information Systems*. Paper presented at the Proceedings of the 5th International Congress on Terminology and Knowledge Engineering, Innsbruck, Austria.
- 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.
- Ganguly, P., Ray, P., & Parameswaran, N. (2005). Semantic Interoperability in Telemedicine through Ontology-Driven Services. *Telemedicine & e-Health*, 11(3), 8.
- Gedzelman, S., Simonet, M., Bernhard, D., Diallo, G., Palmer, P., & Ieee. (2005). Building an ontology of cardio-vascular diseases for concept-based information retrieval *Computers in Cardiology 2005, Vol 32* (Vol. 32, pp. 255-258). New York: Ieee.
- Gennari, J. H., Musen, M. A., Ferguson, R. W., Grosso, W. E., Eriksson, H., Noy, N. F., et al. (2003). The evolution of Protege: an environment for knowledge-based systems development. *Int. J. Hum.-Comput. Stud.*, 58(1), 89-123.
- Ghapanchi, A. H., & Aurum, A. (2011, 4-7 Jan. 2011). *Measuring the Effectiveness of the Defect-Fixing Process in Open Source Software Projects*. Paper presented at the System Sciences (HICSS), 2011 44th Hawaii International Conference on.
- Ghapanchi, A. H., & Aurum, A. (2012). The impact of project capabilities on project performance: Case of open source software projects. *International Journal of Project Management*, 30(4), 407-417.
- Gilbert, M., & Ddembe, W. (2008). A Flexible Approach for User Evaluation of Biomedical Ontologies. *International Journal of Computing and ICT Research*, 2(2), 62-74.
- Gillies, A. (2000). Assessing and improving the quality of information for health evaluation and promotion. *Methods Inf Med*, 39(3), 208-212.
- Grenon, P., Smith, B., & Goldberg, L. (2004). Biodynamic ontology: Applying BFO in the biomedical domain. In D. M. Pisanelli (Ed.), *Ontologies in Medicine* (Vol. 102, pp. 20-38).
- Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Human-Comput. Stud.*, 43(5-6), 907-928.
- Gupta, A., Ludäscher, B., Grethe, J. S., & Martone, M. E. (2003). Towards a formalization of disease-specific ontologies for neuroinformatics. *Neural Networks*, 16(9), 1277-1292.
- Hadzic, M., & Chang, E. (2004). Role of the ontologies in the context of grid computing and application for the human disease studies. *Semantics of a Networked World: Semantics for Grid Databases*, 3226, 316-318.
- Hadzic, M., Dillon, D. S., & Dillon, T. S. (2009). *Use and Modeling of Multi-agent Systems in Medicine*.
- Hamilton, W. T., Round, A. P., Sharp, D., & Peters, T. J. (2003). The quality of record keeping in primary care: a comparison of computerised, paper and hybrid systems. *Br J Gen Pract*, 53(497), 929-933; discussion 933.
- Hevner, A. R., March, J., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 31.

- Hoorn, H. F., & Wijngaarden, T. D. (2010). Web Intelligence for the Assessment of Information Quality: Credibility, Correctness, and Readability In Z.-U.-H. Usmani (Ed.), *Web Intelligence for the Assessment of Information Quality* (pp. 305): InTech.
- Huaman, M. A., Araujo-Castillo, R. V., Soto, G., Neyra, J. M., Quispe, J. A., Fernandez, M. F., 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 Medical Informatics and Decision Making*, 9.
- Huang, T., Li, W., & Yang, C. (2008). *Comparison of Ontology Reasoners: Racer, Pellet, Fact++* Paper presented at the American Geophysical Union, Fall Meeting 2008.
- Jara, A. J., Blaya, F. J., Zamora, M. A., Skarmeta, A. F. G., & Ieee. (2009). *An Ontology and Rule Based Intelligent Information System to Detect and Predict Myocardial Diseases*. New York: Ieee.
- Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information quality benchmarks: product and service performance. *Communications of the ACM*, 45(4), 8.
- Kahn, M. G., Batson, D., & Schilling, L. M. (2012). Data model considerations for clinical effectiveness researchers. *Med Care*, 50 Suppl, S60-67.
- Kerr, K., Norris, T., & Stockdale, R. (2007). *Data Quality Information and Decision Making: A Healthcare Case Study*. Paper presented at the 18th Australasian Conference on Information Systems.
- Kuziemsky, C., & Lau, F. (2010). A four stage approach for ontology-based health information system design. *Artificial Intelligence in Medicine* 50, 133-148.
- Lain, S. J., Roberts, C. L., Hadfield, R. M., Bell, J. C., & Morris, J. M. (2008). How accurate is the reporting of obstetric haemorrhage in hospital discharge data? A validation study. *Australian & New Zealand Journal of Obstetrics & Gynaecology*, 48(5), 481-484.
- Lee, C. S., Wang, M. H., Acampora, G., Loia, V., Hsu, C. Y., & Ieee. (2009). *Ontology-based Intelligent Fuzzy Agent for Diabetes Application*. New York: Ieee.
- Lezcano, L., Sicilia, M.-A., & Rodríguez-Solano, C. (2011). Integrating reasoning and clinical archetypes using OWL ontologies and SWRL rules. *Journal of Biomedical Informatics*, 44(2), 343-353.
- Liaw, S., Chen, H., Maneze, D., Dennis, S., & Vagholkar, S. (2011). Use of the "principal diagnosis" in emergency department databases to identify patients with chronic diseases (in press). *Electronic Health Informatics Journal*.
- 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)*. Paper presented at the AMIA 2011 Annual Symposium Improving Health: Informatics and IT Changing the World.
- Liaw, S. T., 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. *Int J Med Inform*, 82(1), 10-24.
- Lin, Y., Xiang, Z., & He, Y. (2011). Brucellosis Ontology (IDOBRO) as an extension of the Infectious Disease Ontology. *J Biomed Semantics*, 2(1), 9.
- 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.
- Maragoudakis, M., Lymberopoulos, D., Fakotakis, N., Spiropoulos, K., & Ieee. (2008). A Hierarchical, Ontology-Driven Bayesian Concept for Ubiquitous Medical Environments- A Case Study for Pulmonary Diseases 2008 30th Annual International Conference of the Ieee Engineering in Medicine and Biology Society, Vols 1-8 (pp. 3807-3810). New York: Ieee.
- March, S., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, 15, 15.
- McBride, S. J., Lawley, M. J., Leroux, H., & Gibson, S. (2012). Using Australian medicines terminology (AMT) and SNOMED CT-AU to better support clinical research. *Stud Health Technol Inform*, 178, 144-149.

- 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).
- Michalakidis, G., Kumarapeli, P., Ring, A., van Vlymen, J., Krause, P., & de Lusignan, S. (2010). A system for solution-orientated reporting of errors associated with the extraction of routinely collected clinical data for research and quality improvement. *Stud Health Technol Inform*, 160(Pt 1)), 724-728.
- MIE Conference Proceedings: *Quality of Life through Quality of Information*. (2012, 29 August 2012). Paper presented at the 24th International Conference of the European Federation for Medical Informatics (MIE), Pisa, Italy.
- Min, H., Manion, F. J., Goralczyk, E., Wong, Y. N., Ross, E., & Beck, J. R. (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? *Emerg Med J*, 25(11), 784.
- Moro, M. L., & Morsillo, F. (2004). Can hospital discharge diagnoses be used for surveillance of surgical-site infections? *Journal of Hospital Infection*, 56(3), 239-241.
- Nimmagadda, S. L., Nimmagadda, S. K., Dreher, H., & Ieee. (2008). Ontology based data warehouse modeling and managing ecology of human body for disease and drug prescription management 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).
- O-Hoon, C., Jung-Eun, L., Hong-Seok, N., & Doo-Kwon, B. (2008). *An Efficient Method of Data Quality using Quality Evaluation Ontology*. Paper presented at the Third 2008 International Conference on Convergence and Hybrid Information Technology.
- Pan Du, P., Feng, G., Flatow, J., Song, J., Holko, M., Kibbe, W., et al. (2009). From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations *Bioinformatics* 25,(12), i63-i68.
- 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, R. C., Bielinski, S. J., & Chute, C. G. (2012a). Applying semantic web technologies for phenome-wide scan using an electronic health record linked Biobank. *J Biomed Semantics*, 3(1), 10.
- Pathak, J., Kiefer, R. C., Bielinski, S. J., & Chute, C. G. (2012b). Mining the human phenome using semantic web technologies: a case study for Type 2 Diabetes. *AMIA Annu Symp Proc*, 2012, 699-708.
- Pathak, J., Kiefer, R. C., & Chute, C. G. (2012). Using semantic web technologies for cohort identification from electronic health records for clinical research. *AMIA Summits Transl Sci Proc*, 2012, 10-19.
- Peleg, M., Keren, S., & Denekamp, Y. (2008). Mapping computerized clinical guidelines to electronic medical records: knowledge-data ontological mapper (KDOM). *J Biomed Inform*, 41(1), 180-201.
- 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. *Comput Biol Med*, 36(7-8), 712-730.
- Pinto, H. S. (2004). Ontologies: How can They be Built? *Knowledge and Information Systems*, 6(4), 441-464.

- Pipino, L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Association for Computing Machinery. *Communications of the ACM* 45(4), 211-218.
- Poulsen, M. K., Henriksen, J. E., Vach, W., Dahl, J., Møller, J. E., Johansen, A., et al. (2010). Identification of asymptomatic type 2 diabetes mellitus patients with a low, intermediate and high risk of ischaemic heart disease: is there an algorithm. *Diabetologia*, 53(4), 659-667.
- 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 & Experience*, 20(3), 253-264.
- Quan, H., Li, B., Saunders, L. D., Parsons, G. A., Nilsson, C. I., 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.
- Rahimi, A., Liaw, S., Ray, P., & Taggart, J. (2012). *Developing an ontology for data quality in chronic disease management*. Paper presented at the the 24th European Medical Informatics Conference.
- Rahimi, A., Liaw, S., Ray, P., Taggart, J., & Yu, H. (2014). Ontological specification of quality of chronic disease data in EHRs to support decision analytics: a realist review. *Decision Analytics*, 1(5), 31.
- Rahimi, A., Liaw, S. T., Taggart, J., Ray, P., & Yu, H. (in-press). Validating an ontology-based algorithm to identify patients with Type 2 Diabetes Mellitus in electronic health records.
- Rector, A., & Rogers, J. (2005). Ontological & practical issues in using a description logic to represent medical concepts: experience from GALEN *Tech rep CS* (Vol. 35, pp. 1-35). Manchester, England: School of Computer Science, University of Manchester.
- Redman, T. (2005). Measuring data accuracy. In R. e. a. Wang (Ed.), *Information Quality* (Vol. 1, pp. 21). Armonk NY: ME Sharpe Inc.
- Richesson, R. L., & Krischer, J. (2007). Data standards in clinical research: Gaps, overlaps, challenges and future directions. *J Am Med Inform Assoc*, 14(6), 687-696.
- Rodriguez-Muro, M., & Calvanese, D. (2012). *Quest, a system for ontology based data access*: OWLED.
- Rodriguez-Muro, M., Kontchakov, R., & Zakharyashev, M. (2013). OBDA with Ontop. *Proc. of the OWL Reasoner Evaluation Workshop*.
- Rubin, D. L., Lewis, S. E., Mungall, C. J., Misra, S., Westerfield, M., Ashburner, M., et al. (2006). National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *OMICS*, 10(2), 185-198.
- Spasic, I., & Ananiadou, S. (2005). A flexible measure of contextual similarity for biomedical terms. *Pac Symp Biocomput*, 197-208.
- Taggart, J., Liaw, S. T., Dennis, S., Yu, H., Rahimi, A., Jalaludin, B., et al. (2012). The University of NSW electronic practice based research network: disease registers, data quality and utility. *Stud Health Technol Inform*, 178, 219-227.
- Talaei-Khoei, A., Solvoll, T., Ray, P., & Parameshwaran, N. (2011). Policy-based Awareness Management (PAM): Case study of a wireless communication system at a hospital. *Journal of Systems and Software*, 84(10), 1791-1805.
- Talaei-Khoei, A., Solvoll, T., Ray, P., & Parameshwaran, N. (2012). Maintaining awareness using policies; Enabling agents to identify relevance of information. *Journal of Computer and System Sciences*, 78(1), 370-391.
- Terzi, E., Vakali, A., & Hacid, M.-S. (2003). Knowledge Representation, Ontologies, and the Semantic Web. In X. Zhou, M. Orłowska & Y. Zhang (Eds.), *Web Technologies and Applications* (Vol. 2642, pp. 382-387): Springer Berlin Heidelberg.
- 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.

- Tu, S., Tennakoon, L., O'Connor, M., Shankar, R., & Das, A. (2008). *Using an integrated ontology and information model for querying and reasoning about phenotypes: The case of autism*. Paper presented at the AMIA Annu Symp Proc.
- Uschold, M. (2005). An ontology research pipeline. *Applied Ontolog*, 1(1).
- Uschold, M., King, M., Moralee, S., & Zorgios, Y. (1998). The enterprise ontology. *Knowledge Eng Rev* 13(1), 31-89.
- Valencia-Garcia, R., Fernandez-Breis, J. T., Ruiz-Martinez, J. M., 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.
- Valle, R., Bagolin, E., Canali, C., Giovinazzo, P., Barro, S., Aspromonte, N., et al. (2006). The BNP assay does not identify mild left ventricular diastolic dysfunction in asymptomatic diabetic patients. *Eur J Echocardiogr*, 7(1), 40-44.
- 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 C. S. Leung, M. Lee & J. H. Chan (Eds.), *ICONIP 2009, Part II* (pp. 360–366). Berlin Springer-Verlag
- Verma, A., Kasabov, N., Rush, A., & Song, Q. (2008, 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
- Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11 (Nov)), 86-95.
- Wand, Y., & Wang, Y. (1996). Anchoring Data Quality Dimensions in Ontological Foundations. *COMMUNICATIONS OF THE ACM*, 36(11), 86-95.
- Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, 41(2 (Feb)), 58-65.
- Wang, R. Y., Strong, D. M., & Guarascio, L. M. (1996). Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 5-33.
- Weingarten, S. R., Henning, J. M., Badamgarav, E., Knight, K., Hasselblad, V., & Gano, A. (2002). Interventions used in disease management programmes for patients with chronic illness which ones work? Meta-analysis of published reports. *BMJ* 325(7370), 925-928.
- Ying, W., Wimalasiri, J., Ray, P., Chattopadhyay, s., & Wilson, C. (2010). An Ontology Driven Multi-Agent Approach to Integrated e-Health Systems *International Journal of E-Health and Medical Communications (IJEHMC)*, 1(1), 28-40.
- Yu, H., Liaw, S., Taggart, J., & Rahimi, A. (2013). *Using Ontologies to Identify Patients with Diabetes in Electronic Health Records*. Paper presented at the Proceedings of the 12th International Semantic Web Conference and the 1st Australasian Semantic Web Conference, Sydney, Australia.

CHAPTER 5

DISCUSSION

5.1 Methodology to Develop an Ontology

This research started with a motivation to develop a methodology that would enable better utilisation of routinely collected data that had quality problems. We carried out a systematic survey of the literature that demonstrated two significant findings. Firstly, there is a growing research and development in ontology-based approaches to assess and manage DQ within a CDM context across various functions in health and biomedical informatics areas to support the semantic interoperability, flexibility and quality of decision support systems for diagnosis and management. Secondly, “fitness for use” could be the better approach to specify and assess DQ.

Most Australian CISs have functionalities to manage the quality of patients’ registers. An important differentiation between the approach described herein and typical managing capabilities is to include attributes to identify diabetes, a key component of diabetes management based on Australian Diabetes Guidelines (2013 ed). By comparison, in British CISs, the United Kingdom Quality and Outcomes Framework (QOF) aims to improve UK general practice by encouraging GPs to use evidence-based interventions, particularly in the management of chronic diseases (such as diabetes) (*Quality and Outcomes Framework guidance for GMS Contract 2008/09: delivering investment in general practice*, 2008). The QOF relates specifically to documenting and reporting quality of patients’ registers developed from general practice EHRs; it includes many elements related to CDM generally and diabetes management specifically.

The QOF consists of many quality indicators across five general areas: clinical, patient experience, organisational, additional services and holistic care. The clinical domain includes attributes for chronic conditions, including diabetes. Linking the QOF directly to GP systems and clinical outcomes has had mixed reviews, with some reporting improvement in certain conditions (Campbell et al., 2007; Steel, Maissey, Clark, Fleetcroft, & Howe, 2007) while others reported more skeptical attitudes

(O'Dowd, 2008; Heath, Hippisley-Cox, & Smeeth, 2007; Fleetcroft & Cookson, 2006). Despite these varied findings, introduction of the QOF was associated with marked improvements in the management of diabetes in primary care in the UK (Millett et al., 2007).

We believe it is important to give individual authorities the ability to create their own regional/local level indicators. For example, the Aboriginal and Torres Strait Islander populations in Australia have inherently higher diabetes risk, all other factors being equal, than their European counterparts (Australian Government, 2013); thus, defining correct range for 'HbA1C' based on the Australian National Guidelines for T2DM, is more appropriate than using only the "HbA1C=<7%" as used in the DM12 indicator of the QOF. We expect our approach to be used for similar purposes as the QOF, but with somewhat different details depending on the context. A key strength of our approach is its modularity and extensibility – the relevant components of the unified ontology can be easily extended to include other contexts and clinical concepts, as well as extend current query capabilities. For example, UMLS and MeSH could be included, over and above the SNOMED-CT-AU currently being used in Australia.

The literature survey (Chapter 2) revealed that medical informatics has suffered from a lack of a comprehensive approach to develop a semantic model for the assessment and management of quality of patient data. Chapter 2 also assessed the importance of the methodology for the ontology-based approaches and demonstrated its significance in the use of ontology-based approaches for DQ based on the 'fitness for purpose' of EHR data (page 30, paragraph 2).

Chapter 2 also highlighted gaps in the methodological approaches to develop ontologies. The first important gap is that the specification of DQ for implementation is incomplete and there is no comprehensive methodological approach for this purpose (Rahimi, et al., 2014). Only a few papers used DQ with the definition of "fitness for purpose" and examined ontology-based approaches to support DQ. No data was found on the basic research into the association between methodologies to develop ontology-based models for DQ and 'fitness for purpose' in various contexts, specifically CDM (page 45, paragraph 2).

Hence, we designed the methodology MDQO (Chapter 3) that helped develop an ontology-based approach for metadata for DQ specification and assessment based on

the context of the domain (page 60, paragraph 1). In MDQO, DQ can be expressed by constraints and axioms (presented by object and data properties) for DQ specification. For example, specific range, format and value of HbA1c value (as a data property) were defined to capture correctness and consistency of patients' data for all instances allocated in the HbA1C class. Also, Protégé as a knowledge management tool allows specifications of properties of classes, such as disjoint, so that an individual (or object) cannot be an instance of more than one of the specified classes (Appendix 2, page 144). The ontology classes and data properties guide research teams to ensure that fields, records, tables and relationships in the database are appropriately presented (Appendix 4, page 156). MDQO can also enable the development of intelligent software agents to act in various semantic contexts to identify patients with a range of diseases, support decision making in health care, and conduct audit, evaluation and research on patients with other chronic diseases.

We validated the MDQO through the DMO to corroborate the belief that an ontology-based model would have theoretical and practical advantages in developing automated methods for identifying patients with chronic diseases. For example, the conceptualisation and semantic modelling stages of this approach showed that the ontology-based model is particularly useful to enable quick development and testing, because feedback can be cycled back into the development process. Also, the results of Chapters 2 (page 40, paragraph 1), 3 (page 76, paragraph 3) and 4 (page 105, paragraph 2) reveal that the ontology-based approach can access and manage the DQ in a way that is generalisable and reusable, to examine the issues and challenges in data extraction, linkage and semantic interoperability in other domains (Appendix 1, page 137).

Although our ontology-based approach focused on identifying T2DM and the assessment of the quality of T2DM registers, we believe the principles underlying our work are generalisable to other domains and application ontologies. Because our approach is interoperable with applications of ontologies and the real-world databases as well as a reasoning-based query solution to identify various patients with other chronic disease from EHRs, our methodology also can provide a basis for creating or enhancing application ontologies.

Chapter 3 showed that the significant strength of this methodology is to assess DQ using the three core dimensions of DQ - completeness, correctness and

consistency - but also to potentially add other dimensions of DQ, such as timeliness, to better assess the quality of patients' data (page 80, paragraph 4).

The interaction of DQ and patient safety has been explored substantially. For example, there are similarities between the categories of factors on quality of incident reporting described by Magrabi and colleagues (2010) and the factors influencing the correctness and completeness of patient registers described in Chapter 4. Magrabi et al. (2010) aimed to identify categories in a classification of IT problems that will provide a clinically useful, comprehensive means of eliciting information about, and collating and classifying computer-related patient safety incident reports (Magrabi, Ong, Runciman, & Coiera, 2010). They found common human errors - such as knowledge deficit, erroneous computer data entry, use of ambiguous abbreviations, and faulty dose calculations - to be leading causes of incidents.

Other contributing factors were inexperienced staff, heavy workloads and computer system failure. These findings are consistent with the correctness and consistency dimensions of DQ described in this thesis. Both are focused on the safe entry and retrieval of clinical information and the support of users to detect and correct errors and malfunctions. These findings corroborated the manual EHR audit described in Chapter 4 that showed the accuracy of the T2DM identification algorithm was influenced by DQ, such as incorrect data due to mistaken units of measurement, and unavailable data such as non-documentation, incorrect documentation, data extraction problems, encryption and data management errors.

Although the methodology to develop ontology was a useful approach for identifying T2DM patients, there were some limitations associated with its use in this project and the framework should be further investigated. One limitation is that we only focused on T2DM in the knowledge acquisition stage. Further, in the conceptualisation stage we only focused on the three knowledge acquisition resources (ePBRN dataset, literature review and general practitioner and nurse meeting) because of time limitations. These resources may not be comprehensive and could be complemented by other resources, such as specialists, patients and clinical data managers.

In the knowledge representation stage, we used only two ontology reasoners to check internal consistency of the model - RacerPro and Pellet - that both support SPARQL queries. RacerPro allowed us to check the model for inconsistent

(unsatisfiable) concepts and Pellet justified any inference that it can compute. However, other reasoners, like Fact++, may perform those acts faster than others (Dentler, Cornet, Teije, & de Keizer, 2011). Also, in the validation stage, some SWRL rules could be helpful to support SPARQL queries to show accuracy of the result of the ontology-based model in the identification of T2DM patients. Our approach was not able to show the differences when comparing such an ontological approach with other non-semantic techniques. The functionality of this methodology could be enhanced by creating new ontology-based models and improving the semantic rules to deal with DQ issues.

Even though our methodological approach has been fully implemented and integrated within a specific context, further empirical studies are necessary to validate its applicability in different domains. Finally, we note that these kinds of improvements to the methodology are important for the evolution of a model engineered to cope with the implementation of such an approach in the practical world, thus fostering a new generation of methods, tools and strategies to assess DQ.

The MDQO represents a methodology for data quality ontology and produced a semantic knowledge management approach to identify T2DM and assess the accuracy of ontology. In contrast, Kuziemsky and Lau (2010) only applied a four-stage methodological approach to capture user knowledge in severe pain management, and used that knowledge to design an ontology without focusing on DQ (Kuziemsky & Lau, 2010). The MDQO was developed based on the observation of an advanced medical practice's need to incorporate DQ from EHRs into practice, coupled with recognition of a dearth of a specific methodology to develop an ontology-based approach in CDM. The methodology was designed to develop an ontology-based approach in artefacts result, in this case the DMO. However, Kuziemsky and Lau (2010) used the ontology and problem-solving approaches to design and implement a CIS that tested favourably in usability testing.

Our approach can extend existing broader ontologies in biomedicine and bioinformatics by providing an empirical basis to expand concepts and their relations in the context of CDM, specifically in diabetes management. For example, several specific medical ontologies exist: SNOMED CT (Stearns, Price, Spackman, & Wang, 2001), while nominally a terminology system, has been examined critically as an ontology; and the Gene Ontology (GO) (Harris et al., 2004; Smith, Williams, & Schulze-Kremer,

2003) has become the standard terminology for describing the function of genes and gene products across species. However, those kind of high-level ontologies are intended for use across biomedical domains. Our approach and methodology contributes to the development and support of ontology-based clinical and health service conceptual frameworks in the future.

The experiences associated with developing, validating and using the artefacts of an ontology-based approach add to the body of knowledge about the methodology to develop ontology in health and medical informatics. The project was also fuelled by curiosity about the phenomenon of an ontology-based approach, specifically how it develops DQ and how it could be validated. For instance, the process of mapping and querying in the knowledge representation stage differed from the more familiar ontology-based approach to map and query patients' information using OntopPro 1.8 as a plugin for Protégé. The goal of applying this open source model is to recognise and use the new tools for mapping, querying and inferring of patients' data at the same time in the unique semantic environment of Protégé.

5.2 Validation of the MDQO

The MDQO was validated for a major case of CDM (T2DM) using the clinical data of the ePBRN. The validation involved two stages: 1) the construction of a DMO using MDQO, and 2) the validation of the DMO with regard to ePBRN data for T2DM. The main significant finding of the DMO validation presented in Chapter 4 was that this approach is able to automate the assessment of patients' data using a formulised model (presented in Chapter 3) with axioms and constraints to the concepts based on the specific purposes of DQ assessment and management (page 98, paragraph 1).

The ability to extend the DMO by adding new types of patient attributes also enables the ontology to query and maintain relevance if diversity in patient data continues to increase. Furthermore, the ability to add new types of patient attributes to the ontology may enable it to be used to identify various types of diabetes as well as other chronic diseases. Therefore, this formal ontology-based approach guides the development of an application to enable automated assessment of DQ for patients (Appendix 4, page 155) with other chronic diseases, such as COPD or hypertension and identify these patients at various levels of clinical course of the disease, quality improvement and research.

The findings of Chapter 4 are important because the validation of DMO results particularly reveals that the ontology-based approach can contain more explicit semantic information compared with non-semantic and non-ontological approaches. Ontologies enable the modelling of the domain and representation of information requirements to specify the context in collaborative environments (Appendix 1, page 136). Also, the ontology-based approach mapped only a small part - a unique general practice with 908 active patients - from the larger data repository (i.e., ePBRN). This verified the scalability of the ontology-based approach.

Furthermore, the DMO validation results confirmed the significant theoretical advantages of the ontology query-based approach to support the development of automated methods to manage “big data” routinely collected in EHRs (page 107, paragraph 2). The DMO helped with: the testing of automated approaches to support semantic reasoning with clinical data from EHRs (Appendix 4, page 156); assessment and management of the quality of clinical data such as reason for visit, chronic conditions, pathology tests and prescriptions; and representation of the meaning of the data and knowledge (S. Liaw, et al., 2011; S. T. Liaw, et al., 2013; Rahimi, et al., 2014; Taggart, et al., 2012).

The validation of the DMO as an ontology-based query approach presented in Chapter 4 (page 92, paragraph 2) was a useful automated semantic technique for identifying patients with T2DM and the quality of their registers. This was because SPARQL queries automatically infer medical relationships (based on the defined object and data properties) and quality of the registers (completeness and correctness), vital steps for representing this logical knowledge in a computable format and developing a formalised diabetes knowledge base (Appendix 4, page 157).

5.2.1 Strengths and weaknesses of the semantic queries

The result of the validation of DMO demonstrated in Chapter 4 showed that the ontology-based query approach can play a major role in the specification of ‘fitness for purpose’ because the DMO could express DQ. The manual validation of the model using RFV, Rx and Path revealed that the DMO could accurately recognise T2DM cases in a general practice EHR (page 100, paragraph 2). Accuracy of the identification of T2DM patients is nearly 100% when we combined all patient attributes by queries. Also this revealed that accuracy of the patients’ identification from EHRs can be

impacted by poor DQ, specifically incomplete data. The validation of DMO partially substantiated that completeness and correctness of patient data for some attributes are poor (e.g., incorrect HbA1C ranges cited percentages versus SI units, as well as incomplete or non-documented pathology test results).

Also, it has been demonstrated that there are some specific reasons why a patient may not be identified using this model; these included human errors in data entry; organisational problems like coping with changes in units of measurements for HbA1C; system idiosyncrasies like not importing pathology tests and results, or poorly designed system upgrades that lose data; technical problems with data management; and, potentially, technical problems at the data repository level. However, the validation of DMO demonstrated how correct and complete data in patient attributes lead to high sensitivity and specificity of the model and it is fit for patient identification.

Also, queries presented in the validation of DMO can include other concepts and properties related to patients with T2DM, such as Referrals to a T2DM-related service, and risk factors such as BMI (obesity), family history of T2DM and ethnicity, for our purpose. Using more complex ontology queries to identify diabetes can elevate the accuracy of this semantic query approach. Hence, it could conceivably be hypothesised that the ontological approach can be accurate and flexible to model the real world of clinical practice where the patient is usually someone who requires multidisciplinary integrated care for multiple health issues.

However, despite the validation of DMO showing an accurate ontology-based approach to identify patients, there were some weaknesses with its application. One limitation was some technical errors, such as problems with data extraction, encryption and management, which are difficult to isolate and unavoidable. Hence, they can effect execution of semantic queries and rules to facilitate inference of knowledge. The validation of the DMO query approach was tested in a relatively small general practice in a very specific domain area. While it enabled manual validation to confirm the accuracy of the ontology-based query approach, the generalisability may be limited. The patients' data also had a range of quality issues due to a range of reasons, but they were not significant problems in the validation of DMO. Rather these issues highlighted the effect of completeness and correctness of data on the accuracy, or 'fitness for purpose', of this query approach. Other sociotechnical sources of data errors - such as problems

with data extraction, encryption and management - are difficult to isolate and assess. This study has corroborated previous work and added new knowledge to explain non-identification of cases from EHRs (Armstrong, et al., 1998).

5.2.2 Contribution of validation to the knowledge base

Chapter 2 presented the gaps in the validation of ontology-based approaches in the current literature. The valuable result of the review demonstrated that there is insufficient research to address ontology evaluation metrics comprehensively (page 43, paragraph 1). The development and deployment of ontology validation techniques must include evaluation metrics. The review also has shown that the ontological approach to develop DQ is poorly validated (S. T. Liaw, et al., 2013; Rahimi, et al., 2014). The review in Chapter 2 found the most common criteria to assess the validity of ontologies and data models are flexibility, reusability and scalability versus non-ontology-based models for big data (Appendix 4, page 158). Also, there were a small number of evaluative studies on the cost-effectiveness of ontological approaches in DQ and quality of care (Rahimi, et al., 2014).

The approach presented in Chapter 4 fills the current gap in the application and validation of ontological models to assess and manage quality of information in EHRs. The validation results (Chapter 4, page 104, paragraph 3) highlight that the model is accurate when all three attributes (RFV, Rx and Path) are used by the ontological definition of the patient with T2DM as someone with a relevant reason for visit (RFV), is prescribed a relevant medication (Rx) and has a relevant and correct pathology test (Path). It addresses that the ‘fitness for purpose’ definition is more than just the DQ metrics of each attribute (Appendix 1, page 132). In this case, the ontology-based query to identify patients with T2DM was fit for purpose even when the completeness and correctness of pathology data, and to a certain extent prescribing data, was not perfect. This finding confirms previous research that demonstrated the role of ontology-based approaches to assess DQ based on ‘fitness for purpose’ in the health context (Rahimi, et al., 2014).

Importantly, the validation part of this research in Chapter 4 demonstrated that the model can address the lack of valid and reliable DQ assurance to ensure fitness for a range of uses by health providers and professionals (page 106, paragraph 2). The findings represented the relationship between DQ, as measured by metrics, and fitness

for purpose, that is, finding cases of T2DM in the EHR using RFV, Rx and Path. By using this ontology-based query approach, the DMO can support the following tasks for DQ (Rahimi, et al., 2014):

1. The automation of data extraction from EHRs into clinical data warehouses
2. Assessment and management of the intrinsic and extrinsic DQ so that they are fit for purposes such as research, quality improvement and health information exchange and sharing
3. Management of controlled vocabularies and optimising semantic interoperability
4. Curation of data for use by users and applications such as electronic decision support systems
5. Data mining to discover relationships between concepts
6. Discovery of new knowledge
7. Reuse of knowledge in the management of chronic diseases.

This study emphasised the role of ontologies in two aspects: firstly to identify cases of T2DM in a dataset and secondly to assess the DQ required to identify T2DM cases accurately.

Chapter 1 addressed the notion of ontology and DQ as important research topics, and showed that an ontology-based approach can support DQ research. This is particularly because of their inherent (and potential) ability to address semantic interoperability.

Chapter 2 expanded on these notions with a detailed specification of DQ and the role of ontology-based approaches to develop DQ based on ‘fitness for purpose’ within the health context. The lack of comprehensive ontological approaches for DQ based on ‘fitness for purpose’ specifically or in health generally is an important research gap to be addressed. Compared with non-hierarchical data models, there may be more advantages and benefits in the use of ontologies to solve clinical DQ issues semantically and improve the validity and reliability of data retrieval, collection, storage, extraction and linkage algorithms and tools. Chapter 2 addressed the first research question, in particular sub question 1.1. It guided the ontology-based approach to manage various

patients with chronic diseases because of their intrinsic capability to relate various patients' clinical data semantically. Chapter 2 also introduced popular ontology development methods such as METHONTOLOGY. However, there are few and limited studies on the identification of cases of diabetes. Theoretically, ontology-based applications could support automated processes to address DQ and semantic interoperability in the health area. The current evidence also supports moving to the ontology-based design of information systems to enable more flexible use of clinical data. Chapter 2 guided the development of a DQ ontology "fitness for specific purpose" in CDM. The published paper summarises the ontological specification of the quality of data in EHRs to support decision analytics. In addressing research question 1, Chapter 2 defined what is needed in a comprehensive ontology-based approach to DQ and semantic interoperability issues.

Chapter 3 discussed a step-by-step process of developing a methodology of data quality ontology (MDQO) to support a semantic knowledge management approach to identify T2DM and assess the accuracy of the ontology-based identification algorithm. Chapter 3 presented the intuitions as well as the formalism for a semantically-accurate mechanism for capturing DMO-related data from EHRs. DQ was assessed using three core dimensions, namely completeness, correctness and consistency. The longer-term objective is to develop a flexible, generalisable and reusable semantic approach and mechanism that can be used to design intelligent software agents to identify patient cohorts and the quality of data.

Chapter 4 addressed research question 2 and reported on the validation of the DMO developed in Chapter 3. This used real-world EHR data from the ePBRN in South Western Sydney. The sensitivity and specificity (accuracy) of the algorithm to identify patients with T2DM were bench-marked by a manual EHR audit. Accuracy was determined using Reason for Visit (RFV), Medication (Rx) and Pathology (Path), singly and in combination. The combination was based on the DMO. Chapter 4 addressed the gap identified in Chapter 2: insufficient practical research on the development and validation of ontology-based approaches in the assessment and management of large patient datasets and insufficient studies on the development and testing of information models based on clinical scenarios to systematically test quality of data in chronic diseases.

However, our ontology-based approach was tested only in a relatively small general practice in a very specific domain area. While it enabled manual validation to confirm the accuracy of the ontology-based approach, the generalisability may be limited. Further research is required in other clinical domains and with larger repositories of more EHR-derived datasets.

The validation of DMO to identify T2DM patients is accurate and modular, enabling the development of intelligent software to act in various semantic contexts and identify patients with other chronic diseases, support decision making about health care, conduct audits and evaluate research (S. Liaw, et al., 2011). This also addresses the gap identified in the literature around insufficient practical research on the development and validation of ontology approaches in clinical scenarios for the assessment and management of large patient datasets (S. T. Liaw, et al., 2013) and lack of studies on the development and deployment of information models based on clinical scenarios to systematically test quality of information in chronic diseases (Rahimi, et al., 2014).

5.3 Summary

The ontology-based approach to represent knowledge about patient DQ is a priority in medical informatics. We added two main notions to our understanding of ontology-based models. First, an ontology approach can improve DQ so it is useful for various purposes such as clinical research, teaching, audit and evaluation (e.g., quality assurance and clinical decision making). Second, compared with non-hierarchical data models, ontological approaches may have more theoretical and practical advantages in developing automated methods to address DQ, solve semantic clinical DQ issues, enable reuse of knowledge and discovery of new knowledge in CDM, and improve the validity and reliability of data retrieval, collection, storage, extraction and linkage algorithms and tools.

This project's success in creating and validating a T2DM ontology based on 'fitness for purpose' demonstrated that ontology is an appropriate information structure for formalising the depth and breadth of medical knowledge about complex phenomena, and that such representations may be useful in supporting DQ assessment from EHRs. The methodology applied to develop, formalise and validate the DMO may provide a methodology for future medical ontology-based projects. Successful application and validation of the DMO demonstrated that formalised representation of T2DM patients'

data may support integrating CDM into population health, which is dependent on underlying DQ from EHRs. In this manner, medical informatics research on ontology-based approaches may help create truly personalised health care, ultimately having a profound and positive effect on the health care of individuals. These thesis studies led to a novel MDQO and a DMO to identify patients with T2DM. The accuracy of the ontology was tested on a real-world EHR and validated with a manual audit of the same EHR.

5.4 References

- Armstrong, D. G., Lavery, L. A., Vela, S. A., Quebedeaux, T. L., & Fleischli, J. G. (1998). Choosing a practical screening instrument to identify patients at risk for diabetic foot ulceration. *Arch Intern Med*, 158(3), 289-292.
- Dentler, K., Cornet, R., Teije, A., & de Keizer, N. (2011). Comparison of Reasoners for large Ontologies in the OWL 2 EL Profile. In B. C. Grau (Ed.), *Semantic Web* (Vol. 1, pp. 1-51). Oxford University, UK: IOS Press.
- Kuziemy, C., & Lau, F. (2010). A four stage approach for ontology-based health information system design. *Artificial Intelligence in Medicine* 50, 133-148.
- 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)*. Paper presented at the AMIA 2011 Annual Symposium Improving Health: Informatics and IT Changing the World.
- Liaw, S. T., 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. *Int J Med Inform*, 82(1), 10-24.
- Rahimi, A., Liaw, S., Ray, P., Taggart, J., & Yu, H. (2014). Ontological specification of quality of chronic disease data in EHRs to support decision analytics: a realist review. *Decision Analytics*, 1(5), 31.
- Taggart, J., Liaw, S. T., Dennis, S., Yu, H., Rahimi, A., Jalaludin, B., et al. (2012). The University of NSW electronic practice-based research network: disease registers, data quality and utility. *Stud Health Technol Inform*, 178, 219-227.

CHAPTER 6

CONCLUSION

This thesis has presented the design of a methodology for data quality ontology (MDQO) specification and assessment that can be used in any domain based on the principle of ‘fitness for purpose’ DQ. The thesis has validated the MDQO in the context of DQ routinely collected data for CDM (T2DM). This ontological approach to collecting, annotating, analysing, and presenting clinical and scientific data is possibly the only practical and sustainable solution to the clinical information and data explosion. It is important to optimise the availability of good quality and relevant information to facilitate the safety and quality of integrated care as well as accurate and valid research. The DMO approach can support that ‘fitness for purpose’ includes DQ and decrease the potential side effects of poor quality big data.

This methodology also is important because it can be applied in future research to represent other complex phenomena of importance to assess DQ in CDM. For instance, the process of conceptualising a phenomenon for ontology development is a useful approach that can be demonstrated by applying this MDQO to develop ontologies for DQ in other contexts. We highlighted technical challenges associated with ontology development and also validation. Further research is needed on the development of this approach in other technical, health and social domains.

Practically, this research can also support general practitioners in diagnosing patients with T2DM in their EHRs and those at risk of developing T2DM. The findings also suggest that a semantic ontology-based approach can be used to help clinicians manage patients with T2DM specifically, and any other chronic illness generally, through the development of high-quality patient/disease registers.

6.1 Future Work

Future work based on this research could include expanding the ontology to other aspects of diabetes management (such as other types of diabetes and their

attributes) as well as applying the ontology-based approach methodology to other chronic diseases and other areas of healthcare, such as primary health care. Further, the contextualisation of concepts identified in Chapter 3 currently only serve as ontology annotations and properties for identifying T2DM and their information retrieval. It would be interesting to implement formal semantics into the ontology by mapping the ontology concepts and relationships to formal medical terminologies including UMLS and SNOMED-CT-AU. Considerably more work is required to test the congruence of the ontology, using formal ontology modeling principles, to other more established ontologies. Finally, further possibilities include implementing the ontology in OWL or RDF, or using SWRL rules for patient information storage and retrieval.

Continuing research will focus mostly on the methodology (MDQO), and the extension of the DMO to address study limitations. Major areas are as follows.

1. *Expanding case studies into various types of diabetes with more complex DQ problems and big patient data sets.* This will allow us to test extendibility and scalability of the DMO and whether the DMO can be used in other domains to identify patients as well as assess DQ problems.
2. *Testing reusability and implementation performance.* In this thesis, we used the DMO in the diabetes context. Reusability studies need to be conducted for real users in other domains (i.e., patients with different chronic diseases like COPD, etc.). Other implementations also need to be performance tested.
3. *Support for more generation of axioms and SPARQL queries and SWRL rules into code.* By having a tool that can generate more axioms, we can identify patients with different chronic diseases and assess more dimensions of DQ from EHRs in the development process.
4. *Assess different emerging technologies.* The agent development component for the DMO was treated as a “black box”. Although the ontology-based approaches were intended for consumption by agents, it would be interesting to see if the DMO would work with different emerging technologies other than agents, e.g., semantic web services.
5. *Comparing ontological and non-ontological approaches to identify patients and assess DQ from multiple EHRs in big data sets.* Comparing ontological (e.g.,

MDQO and DMO) versus non-ontological (e.g., database schemas and SQL techniques) will allow us to compare the methodologies in terms of flexibility, reusability and scalability as well as accuracy of the results in both approaches.

APPENDIX 1

INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 82 (2013) 10–24



journal homepage: www.ijmijournal.com



Review

Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature

S.T. Liaw^{a,b,c,*}, A. Rahimi^{a,d,e}, P. Ray^{a,d}, J. Taggart^b, S. Dennis^b, S. de Lusignan^f,
B. Jalaludin^{a,g}, A.E.T. Yeo^h, A. Talaei-Khoei^d

^a University of NSW School of Public Health & Community Medicine, Sydney, Australia

^b University of NSW Centre for Primary Health Care & Equity, Sydney, Australia

^c General Practice Unit, South West Sydney Local Health District, Australia

^d Asia Pacific ubiquitous Healthcare research Centre (APuHC), University of NSW, Sydney, Australia

^e Isfahan University of Medical Sciences, Faculty of Management and Medical Information Sciences, Iran

^f Department of Health Care Management and Policy, University of Surrey, Guildford, UK

^g Population Health Unit, South West Sydney Local Health District, Australia

^h Ingham Institute of Applied Medical Research, Australia

ARTICLE INFO

Article history:

Received 14 April 2012

Received in revised form

3 October 2012

Accepted 5 October 2012

Keywords:

Realist

Research design

Chronic disease

Information system

Data quality

Ontology

ABSTRACT

Purpose: Effective use of routine data to support integrated chronic disease management (CDM) and population health is dependent on underlying data quality (DQ) and, for cross system use of data, semantic interoperability. An ontological approach to DQ is a potential solution but research in this area is limited and fragmented.

Objective: Identify mechanisms, including ontologies, to manage DQ in integrated CDM and whether improved DQ will better measure health outcomes.

Methods: A realist review of English language studies (January 2001–March 2011) which addressed data quality, used ontology-based approaches and is relevant to CDM.

Results: We screened 245 papers, excluded 26 duplicates, 135 on abstract review and 31 on full-text review; leaving 61 papers for critical appraisal. Of the 33 papers that examined ontologies in chronic disease management, 13 defined data quality and 15 used ontologies for DQ. Most saw DQ as a multidimensional construct, the most used dimensions being completeness, accuracy, correctness, consistency and timeliness. The majority of studies reported tool design and development (80%), implementation (23%), and descriptive evaluations (15%). Ontological approaches were used to address semantic interoperability, decision support, flexibility of information management and integration/linkage, and complexity of information models.

* Corresponding author at: PO Box 5, General Practice Unit, Fairfield Hospital, Fairfield, NSW 1860, Australia. Tel.: +61 2 96168520; fax: +61 2 96168400.

E-mail address: siaw@unsw.edu.au (S.T. Liaw).

1386-5056/\$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
<http://dx.doi.org/10.1016/j.ijmedinf.2012.10.001>

Conclusion: DQ lacks a consensus conceptual framework and definition. DQ and ontological research is relatively immature with little rigorous evaluation studies published. Ontology-based applications could support automated processes to address DQ and semantic interoperability in repositories of routinely collected data to deliver integrated CDM. We advocate moving to ontology-based design of information systems to enable more reliable use of routine data to measure health mechanisms and impacts.

© 2012 Elsevier Ireland Ltd. All rights reserved.

Contents

1. Introduction	11
2. Objective	12
3. Methodology	12
4. Findings	14
4.1. General and methodological	14
4.2. Definitions of DQ and its operationalisation and measurement in various studies	15
4.3. Documented uses of ontologies for DQ	15
4.4. Documented uses of ontology in CDM	17
5. Discussion	18
6. Conclusions	20
Conflict of interest statement	20
Authors' contributions	20
Acknowledgments	20
References	20

1. Introduction

The increasing global burden of chronic disease due to the ageing population, scarcity of resources and costs of health care delivery has led to the WHO's prediction that, by the year 2020, chronic disease will be responsible for three-quarters of the world's deaths [1]. Globally, integrated care [2–5] has the potential to improve the quality and efficiency of chronic disease management (CDM) [6], but depends on the sharing of good quality patient information, including results of investigations or referrals. A definition of integrated care is “a coherent set of methods and models on the funding, administrative, organisational, service delivery and clinical levels designed to create connectivity, alignment and collaboration within and between the care and care sectors” [7]. This is consistent with the dimensions of the chronic care model [8,9]: *health care organisation, delivery system design, decision support, clinical information systems (CIS), self-management support and community resources/policies*. Systematic reviews have found that, despite methodological shortcomings, inconsistent definitions and considerable heterogeneity in interventions, patient populations, processes and outcomes of care [10], integrated care programmes can improve the quality of patient care [11]. Good quality data collected as part of routine clinical care is required to address this evidence gap cost-effectively. Routinely collected electronic health care data, aggregated into large clinical data warehouses (CDW), are increasingly being mined, linked and used for audit, continuous quality improvement in clinical care, health service planning, epidemiological study and evaluation research. Managing the increasing amount of routinely collected data is a priority.

However, data quality (DQ) is poor in about 5% of records in health organisations [12*,13*,14]. Many studies regularly report a range of deficiencies in the routinely collected electronic information for clinical [15–18] or health promotion [12*,19] purposes in hospital [20] and general practice [21] settings. The evidence was more encouraging for data for administrative purposes [22,23]. Hybrid record keeping systems in primary care were believed to be more complete than computer-only or paper-only systems [24]. Prescribing data are generally more complete than diagnostic or lifestyle data [21,25].

Improving the quality of routinely collected data can improve the quality of care. Every year, 10% of hospital admissions and >1 million general practice encounters in Australia experience an adverse event, and evidence-based care is delivered only about half the time [26–29]. Linkages between primary and secondary care information systems are important to improve the quality of information exchange to support optimum clinical handover between the levels of care. Information-enhanced integrated care can benefit health care providers and consumers through more accurate and timely information exchange, improve work efficiency by avoiding repetitive work, and improve decision-making [30,31]. Complete and accurate information sharing such as in clinical handover is vital to maintain continuous and safe patient care across primary and acute services [32]. In response, Australian governments [33–36] have emphasized the need for effective use of clinical information systems (CIS) and electronic decision support tools to collect, share and use information to guide ongoing health reform, policy development and strategic work plans to implement safe, effective and coordinated care over the life cycle and across the “patient journey” in the health system [27–29,37].

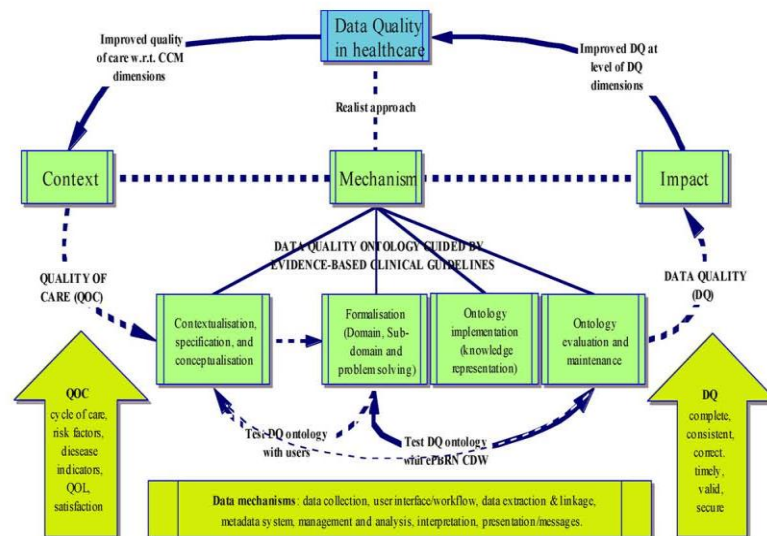


Fig. 1 – Conceptual framework for data quality (DQ) research program and literature review.

Since 2001 there has been an increasing use of ontological approaches to health, particularly chronic disease management. Historically, ontologies are rooted in philosophy as the study of being or reality, including their basic categories and relations. The biomedical and health informatics definition of an ontology is “collections of formal, machine-processable and human interpretable representation of the entities, and the relations among those entities, within a definition of the application domain” [38]. Explicit concepts and the relationships and constraints are clearly defined and understood by the user. A formal ontology is computer-readable, allowing the computer to ‘understand’ the relationships – the ‘formal semantics’ – of the ontology. By incorporating defined rules, ontologies may also generate logical inferences and control the inclusion/exclusion of relevant objects [39].

This is the background for this literature review on ontological approaches to data quality and quality of care, with a specific focus on integrated chronic disease management. The scope was guided by the knowledge and experience of this multidisciplinary group of authors.

2. Objective

To conduct a literature review to address the following questions:

- (1) How is data quality (DQ) currently defined/described, assessed and managed in health care?
- (2) How are ontologies being used to assess and manage DQ?
- (3) What is/are role(s) of ontologies in the assessment and management of DQ to support better decision making and measurement of health outcomes in integrated chronic disease management (CDM)?

3. Methodology

A realist literature review [40] was adopted, as this was an evolving and complex domain. The conceptual framework developed for the literature review included (Fig. 1):

- **Context:** Integrated CDM, care based on evidence based practice;
- **Mechanisms:** Methods to achieve data quality, including ontology-based approaches;
- **Impacts/outcomes:** Measurable health outcomes based on improved data quality.

The following databases (January 2001–March 2011) were searched: MEDLINE, the Cochrane Library, ISI Web of Knowledge, Science Direct, Scopus, IEEE Xplore and Springer (Table 1).

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

1. **Context:** Diseases (chronic diseases, chronic illnesses, chronic disease management, chronic illness management);
2. **Mechanisms:** Ontology (ontology based models, ontological approaches, ontology based multi agent systems (OBMAS), and ontological framework);
3. **Impacts:** Data quality (data quality, information quality, data quality management, data quality assessment, data and information).

The search was repeated three times with the following phrases:

- (data quality OR information quality) AND (chronic diseases OR chronic illnesses) in Title, Abstract or Keywords, Subject or MESH

Table 1 – Scope of literature review - online databases and research fields.

Database	Subjects	# papers
PubMed	Medicine, Health Science, Medical Informatics and Bioinformatics	57
Cochrane Central Databases	Medicine and Health Science	8
ISI Web of Sciences	Computer Science, Information Technology, Medical Informatics, Bioinformatics and Health Science	25
ScienceDirect	Computer Science, Medical Informatics, Engineering, Decision Science, Engineering, Mathematics, Psychology, Social Sciences, and Medicine	60
Scopus	Computer Science, Health Science, Medical Informatics, Bioinformatics, Information Technology, Psychology, Social and Behavioural Sciences	61
IEEE Xplore	Computing and Processing, Medical Informatics, Bioinformatics, Communication Networking and Cybernetics	20
SpringerLink	Computer Science, Medical Informatics, Bioinformatics, information science and Engineering	14
	Total	245

- ontology in Title, Abstract or Keywords, Subject or MESH (data quality or information quality) in Title, Abstract or Keywords, Subject or MESH
- ontology in Title, Abstract or Keywords, Subject or MESH AND chronic diseases in Title, Abstract or Keywords, Subject or MESH.

All English language papers published from January 2001 to March 2011 were included if they met the following eligibility criteria: (a) examined data and information quality in chronic diseases; (b) involved some form of ontology to improve DQ; (c) used data models and ontology-based approaches in CDM.

These papers were screened by title and abstract content for inclusion by AR and STL. The references of the included papers were hand-searched for other eligible papers. Following this comprehensive process, the included papers were distributed for review among all the authors according to their expertise and experience. All papers were reviewed by AR, STL and one of the co-authors. Authors used a

data extraction template (Fig. 2), with a realist “context-mechanism-impacts/outcomes” overlay. The template kept the extracted information consistent: study types, methods, tools, outputs and impacts. The quality appraisal included: validity (internal and external), reliability, generalisability and relevance of the research methods, tools and measurements, and interpretations.

AR and STL collated all appraised papers, using a specific template (Fig. 3) which summarised 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 collated appraisals were then distributed among the reviewers, and two workshops were arranged to discuss and achieve final consensus and synthesis of the findings. Further iterative feedback was obtained on specific areas of ambiguity prior to this final report on the literature review.

Critical Appraisal Template

- Research questions:
1. How is DQ currently defined/described, assessed and managed in health?
 2. How are ontologies being used to assess and manage DQ?
 3. What is/are the role(s) of ontologies in the assessment and management of DQ in CDM?

Author / title / reference	Study type ¹	Context & Population studied	Aims of project being reported on	Details of terminology, DQ models & ontology	Methods / Tools used in project	Results / Outputs of project	Critical Appraisal: 1. quality ² of methods & tools 2. relevance to review questions

¹ This classification of study types has been developed to cover the pattern of R&D in this multidisciplinary field. There are 5 types broadly based on the stage in the development lifecycle:

1. Requirements analysis e.g. literature reviews, qualitative research, etc
2. Design and tools development: data/information models and ontologies
3. Implementation, deployment and testing of information system
4. Evaluation: descriptive evaluation
5. Evaluation: comparative (e.g. pre and post, time series, etc) with/without contemporary control (e.g. RCT)

This matrix (study types X current column headings) will focus the analysis and synthesis of the literature review e.g. by study types, methods, tools, outputs and impacts.

² The quality appraisal will include traditional methods of critical appraisal: validity (internal and external), reliability, generalisability, relevance, etc of the research methods, tools and measurements

Fig. 2 – Template for critical appraisal of allocated papers.

Summary table: Systematic review of ontology-based approaches to data quality in chronic disease management

Research questions: 1. How is DQ currently defined/described, assessed and managed in health? 2. How are ontologies being used to assess and manage DQ? 3. What is/are the role(s) of ontologies in the assessment and management of DQ in CDM?

Paper reviewed: Smith J et al. *The Idiot's guide to Ontology for DQ. JAMIA 2011*

	Stage of development (Study Type ^a)	Reviewer 1 appraisal	Reviewer 2 appraisal	Consensus validity & relevance	Notes
1.	Requirements analysis				
2.	Design and tools development				
3.	Implementation, deployment and testing				
4.	Evaluation: descriptive evaluation				
5.	Evaluation: comparative and/or contemporary control				

^a This classification of study types has been developed to cover the pattern of R&D in this multidisciplinary field. There are 5 types broadly based on the stage in the development lifecycle:

1. Requirements analysis e.g. literature reviews, qualitative research, etc
2. Design and tools development: data/information models and ontologies
3. Implementation, deployment and testing of information system
4. Evaluation: descriptive evaluation
5. Evaluation: comparative (e.g. pre and post, time series, etc) with/without contemporary control (e.g. RCT)

This matrix (study types X current column headings) will focus the analysis and synthesis of the literature review e.g. by study types, methods, tools, outputs and impacts.

Fig. 3 – Summary template for collating critical appraisal differences between two reviewers.

4. Findings

4.1. General and methodological

We identified 245 articles, of which 135 were excluded on abstract review because they did not meet inclusion criteria and 26 articles were duplicates. After full text review 23 papers were excluded because they did not meet inclusion criteria: (a) examined data and information quality in chronic diseases; (b) involved some form of ontology to improve DQ; (c) used data models and ontology-based approaches in CDM. This left 61 papers: of these 33 implemented ontology in CDM, 13 used a defined process for DQ generally and 15 used ontology to improve DQ in various contexts. While the focus was on chronic disease, we also included general health domains (24.6%), non-health (9.8%) and non-specific

(4.9%) domains where the methodology appeared relevant and appropriate. The chronic diseases most frequently studied were diabetes mellitus (18%), cardiovascular diseases (8.2%), respiratory diseases (8.2%) and communicable diseases (8.2%). Other conditions included nervous system diseases (6.6%), neoplasms (4.9%), autism (3.3%), urologic diseases and obesity (1.6%).

The majority of studies (80.4%) examined the design and development of tools for DQ and/or ontologies. This was followed by system implementation, deployment and testing of information systems (23%), formal requirements analysis (16%) and descriptive evaluation (15%). There was little comparative evaluation of outcomes; the one paper found was focused on DQ (Table 2). While most of the studies designed, developed, assessed and evaluated information models and ontologies in the chronic diseases context, there were no comprehensive ontological approaches for the development

Table 2 – Distribution of papers by study types and research questions.

Study type	Study type		Research questions					
	n	%	Q1		Q2		Q3	
			n	%	n	%	n	%
1. Formal requirements analysis, e.g. literature reviews, qualitative research	10	16	4	6	2	3	3	4
2. Design & tools development: including data/information models and ontologies	49	80	13	21	11	18	35	57
3. Implementation, deployment and testing of information systems	14	23	3	4	3	4	10	16
4. Evaluation: descriptive evaluation of DQ or ontology in health area	9	15	6	9	2	3	4	6
5. Evaluation: comparative with/without contemporary control (e.g. RCT)	1	2	1	2	0	0	0	0

Table 3 – Ontology development tools.

Ontology functions	Tools
Ontology environment editors	Protégé, HOZO, Web ODE, JAVA ontology editor (JOE)
Reference terminology, metathesaurus, thesaurus	SNOMED CT, MESH and UMLS
Ontology development methods	METHONTOLOGY, Enterprise Ontology and TOVE
Representation languages	OWL, SWRL, XML and RDF
Ontology logic reasoners	Pellet, Fact++, Jena and Racer
Improve semantic interoperability	Ontology based multi agent systems (OBMAS)

of DQ in CDM described. In Table 2 the total number of study types or research questions is greater than the number of included papers ($n=61$) because each paper may be classified as two or more study types, or may address two or more review questions.

A number of tools (Table 3) used to develop ontological based models were documented, including: ontology environment editors such as Protégé [41]; reference terminology such as SNOMED CT [42], metathesaurus such as UMLS [43] and thesaurus such as MESH [44]; ontology development methods such as METHONTOLOGY [45]. Enterprise Ontology [46] and TOronto Virtual Enterprise (TOVE) [47]; representation languages such as OWL, SWRL, XML and RDF; logic ontology reasoners [48] to provide automated support for reasoning tasks in ontology and instance checking [46] such as Pellet, Hermit, Fact++, Cyc; and layered ontology methodology and tools such as ontology-based multi-agent systems (OBMAS) [49,50]. The tasks involved in the development of a DQ ontology [51,52] include the: review of concepts required for ontological views of DQ, capture of terms to produce ontologies for DQ [52], identification of errors in DQ and DQ ontologies, integration of data from heterogeneous clinical databases [39], and evaluation of DQ and DQ ontology [53]. Ontology tools are currently the subject of a more detailed literature review.

4.2. Definitions of DQ and its operationalisation and measurement in various studies

DQ is consistently defined in terms of its “fitness for purpose/use” [54], in this case, to describe and assess the safety and quality of care. This functional and product approach is consistent with the International Standards Organisation (ISO) definition of quality as “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). To be fit for purpose, some authorities have asserted that data must possess three attributes: utility, objectivity and integrity [55]. The Canadian Institute for Health Information (CIHI) information quality framework is based on a DQ work cycle, a DQ assessment tool and documentation about DQ. It comprises 5 quality dimensions (accuracy, timeliness, comparability, usability, and relevance), further subdivided into 24 quality characteristics and 58 quality criteria [55].

There was agreement that DQ is a multidimensional construct, with a number of dimensions such as “accuracy, perfection, freshness and uniformity” [56] and “completeness, ‘unambiguity’, meaningfulness and correctness” [57] and “currency” [58] across a number of application domains

(Table 4). There was no general consensus on the definitions of the DQ dimensions; however, the five most frequently reported dimensions were “accuracy”, “completeness”, “consistency”, “correctness” and “timeliness”. Various quantitative and statistical methods were used to assess timeliness (currency), accuracy (precision), reliability, representativeness and completeness (Table 4). Usability, privacy, comparability and relevance were evaluated with qualitative methods like interviews and reports analysis, usually interpreted using grounded theory. Consistency of clinical data has been assessed with concept mapping in non-health contexts.

The review process confirmed that points of ambiguity in the data model were potential sources of data errors. A comparison of “persons consulting prevalence rates of musculoskeletal disease” among four databases in the UK found considerable variation and suggested that the prevalence rates were determined by the database used to generate them and methods used to calculate the rates [59]. A popular Australian CIS did not allow the recording of BP in different positions during the same consultation or the changing of smoking status over time, contributing to poor DQ for these data elements [21]. Data are stable whereas data models are influenced by the database management system, security and access management software, organisational processes for data collection and management, and the people in the organisation who enter and use data. A conceptual framework has been proposed to assess the quality of data models using a combination of metrics and subjective assessments, which included correctness, implementability, completeness, understandability, integration, flexibility and simplicity [60].

The strategic use of related data fields relevant to research questions to improve the accuracy and “fitness for use” of the dataset [61] highlighted the need for people working with large data sets to understand fully the complexity of the context within which data collection and management takes place. Metadata are important to guide users about how to find relevant data, select appropriate research methods and ensure that the correct inferences are drawn [62,63,64] as well as to explain the source, context of recording, validity check and processing method of any routinely collected data used in research [65]. There were very few studies that examined the comprehensiveness, efficiency or effectiveness of DQ management in health care.

4.3. Documented uses of ontologies for DQ

A number of definitions of ontology were found [38,66–68], with most revolving around Gruber’s “an explicit, formal specification of a shared conceptualisation” [69] and providing a vocabulary of terms, their meanings and relationships to be

Table 4 – DQ dimensions – definitions and measures.

Definitions of DQ dimensions	Measures of DQ dimensions
1. Completeness The extent to which information is not missing and is of sufficient breadth and depth for the task at hand (121) The ability of an information system to represent every meaningful state of the represented real world system (57) Degree to which information is sufficient to depict every possible state of the task (122) All values for a variable are recorded (121) Availability of defined minimum number of records/patient	Completeness Ratio of total number of records with data to the total number of records. Include an assessment of missing values Set a threshold value for acceptable completeness within an appropriate time frame in context
2. Consistency Representation of data values is same in all cases (121). Includes values and physical representation of data (57). The extent to which information is easy to manipulate and apply to different tasks (121) The equivalence, and process to achieve, equivalence of information stored or used in applications, and systems (111) The extent of use of a uniform data type and format (e.g. integer, string, date) with a uniform data label (internal consistency) and codes/terms that can be mapped to a reference terminology (external consistency)	Consistency 1 – (ratio of violations of a specific consistency type to the total number of consistency checks) Ratio: The most commonly-used data type, format or label divided by total number of data type, formats or labels used (<i>internal consistency</i>) Proportion of data labels that can be mapped to a relevant reference terminology or data dictionary (<i>external consistency</i>) Distance to reference terminology
3. Correctness The free-of-error dimension (104, 123) Credibility of source and user's level of expertise (121) Data values, format and types are valid and appropriate; an example is height is in metres and within range for age Data correctness includes accuracy and completeness (124)	Correctness 1 – (ratio of number of data units in error to the total number of data units) Correctness is indicated by accuracy, completeness and depth (125, 126)
3.1 Accuracy (⇒ correctness) Recorded value is in conformity with actual value (121) Refers to values and representation (127) of output data (98)	Ratio: The number of correct (accurate) values divided by the overall number of values (99)
3.2 Reliability (⇒ correctness) Extent to which a data can be expected to perform its intended function with required/defined accuracy (121, 122) How data conforms with user requirements or reality (57) Data can be counted on to convey the right information (57)	Descriptive statistics with comparison to validated population surveys to ensure representativeness, i.e. no statistically significant differences in data values
4 Timeliness Data is not out of date; availability of output is on time (57) Extent to which information is up to date for task (121, 123) The delay between a change of the real-world state and the resulting modification of the information system state (57)	Ratio: number of reports sent on time divided by total reports Ratio: number of data values within a defined time frame divided by the total records in the same time frame
5. Relevance The extent to which information is applicable and helpful for the task at hand (121)	Descriptive qualitative measures with group interviews and interpreted with grounded theory
6. Usability The degree to which data can be accessed, used, updated, maintained, managed (57) to enable effective decisions (121)	Descriptive qualitative measures with semi structured interview and interpreted with grounded theory
7. Security Personal data is not corrupted and access suitably controlled to ensure privacy and confidentiality (57, 121)	Analyses of access reports Data corruption can be measured by DQ measures

used in various application contexts. The ontological view described the closed semantic loop of observation and action, linking the reality and information realms.

There were numerous uses of ontologies for DQ in health and general contexts (Table 5). The major categories of use were in semantic data interoperability [51*,70*,71*,72*];

Table 5 – Documented use of ontologies for DQ.

Ontology in DQ	Findings	Context
Ontology-based description of DQ: <i>Based on a paper describing a DQ ontology and 27 papers describing healthcare ontologies</i>	Represent DQ factors, terms and terminology standard	CDM
	Describe concepts (and relationships) DQ ontology	General
	Describe logic processes and semantics in ontology	General
	Represent how good DQ facilitate accurate decisions	General
	Describe variations in meaning of terms and coding	COPD
	Represent/model terminological and semantic relationships among concepts in a disease map	CNS+ vector-borne diseases
	Represent/model a DQ evaluation framework	Health care
	Methodology to develop, assess, interpret, manage DQ	Severe Pain Management
	A sharable/extensible analysis tool to identify patient data, semantic interoperability and terms	CVD
	Guide the use of NLP to convert text to coded data	COPD
Ontology-based assessment of DQ: <i>Based on 6 papers describing ontology for assessment of DQ and 14 papers describing ontology for assessment of healthcare.</i>	Approach to collect/retrieve information intelligently and address semantic interoperability of data from multiple information sources, e.g. OBMAS	CDM
	Approach to efficiently share, integrate and manage scientific data in a timely manner, e.g. OBMAS	Prostate cancer
	Guide the development and use of metrics to measure the complexity and cohesion of ontologies	CDM
	Facilitate the ability of researchers to analyse data	Genetic
	Augment data repositories with rule-based abstractions	Autism
	Systematic approach to DQ assessment, e.g. OBMAS and 5-step methodology	Autism
	Enhance inter-professional collaboration	CDM
	Automated approach to identify data errors/variations	CDM
	Consistency checking, duplicate detection, metadata mx	Heart diseases
	Capture correct terms in ontology production and relationships between concepts in ontology	General
Ontology-based management of DQ or health care: <i>Management of DQ is done at the levels of the DQ dimensions and in context.</i> <i>Based on 9 papers describing ontology for management of DQ and 14 papers describing ontology for management of healthcare.</i>	Intelligent agents to integrate data from many sources	CNS+ vector-borne diseases
	Facilitate semantic interoperability in CDM, e.g. OBMAS	CVD
	Represent new methods for fuzzy medical relationship using taxonomical knowledge	T2DM
	Reduce uncertainty for decision making	Diabetes
	Facilitate data integration and re-use	Diabetes
	Guide integration of OWL and RDF with SWRL for better expressiveness of data	Immunology
	A tool for intelligent data integration from remote biomedical resources	Brain abnormalities
	Facilitate semantic interoperability through domain terminologies	CVD
		CDM

information retrieval, DQ management [73], data collection, data sharing and data integration [39,74*,75*,76*] in clinical information systems (CIS) for CDM; DQ in geographical information systems (GIS) and other non-health areas [77*,78*]; and regular validation of key data items in clinical data warehouses (CDW) [39,79*]. In the case of CDWs, a formal ontological model of the domain and representation of data and metadata can specify a unified context which allows intelligent software agents to act in spite of differences in concepts and terminology. This is the potential of layered ontologies and ontology-based multi-agent systems to enable the systematic development of automated, valid and reliable methods to extract, link and manage data as well as assess the DQ and semantic interoperability issues [46,76*,77*,79*].

4.4. Documented uses of ontology in CDM

Documented uses of ontology in CDM included clinical decision support systems [49,80*,81*,82*,83*] for diagnosis [84*,85*] and management [51*], clinical data analysis, information management [86*,87*,88*] and retrieval [71*,89*], diagnostic support in telecare services and remote patient monitoring, enhanced flexibility in database architecture and configuration, and reducing complexity of Bayesian networks [90*] and inferences [91*,92*] (Table 6). A few studies examined ontology-based approaches to support data consistency [72*] and accuracy. However, we found no reports on a systematic and comprehensive ontological approach to DQ issues or evaluation in CDM.

Table 6 – Documented uses of ontology in CDM.

Ontology in CDM	Findings	Context
Description or definition	Identify relevant entities to successfully integrate and represent heterogeneous data and knowledge	T2DM
	A method to generate more intuitive concepts, properties, relations and restrictions	Diabetes
	A layered approach to provide guidance and constraints based on domain knowledge	CVD
Management	Embedding clinical guidelines and rule based approaches in ontology development	CVD
	A method to formalise genomic data inclusion	CVD
	A tool to improve retrieval information for users	CVD
	Enhance and facilitate temporal querying requirements in general practice medicine	CVD (hypertension)
	Detect and predict diseases in patients with CD in telecare services	CVD
	Support decision making for physicians	COPD
	Predict risk analysis semantically	T2DM
	Ontology-based data warehouse modelling and data mining tools to manage large data sets	COPD
	An approach to support semantic decision making	Diabetes
	A method to classify patients with CD	Diabetes
	Simplify fuzzy medical relationships through ontology guided taxonomical knowledge	Diabetes
	Facilitate semantic interoperability in diseases treatment	Human diseases
	A basis for diet care knowledge management	T2DM
Assessment	An approach for terms extractors, concordance checking, and a terminology server	CVD
	Guide the development of a flexible data architecture with multiple configuration options, allowing users to define their own solutions.	CVD
	A tool to reduce the complexity of Bayesian Networks (BNs), BN-based inference and clinical information systems, including diagnostic systems.	Obesity
	Use for the retrieval and the assessment of data	Obesity
	Represent multiple semantic relationships among concepts with UMLS ancestors through MESH descriptors to develop retrieval information	Breast cancer
	Present a language independent approach for extracting knowledge from natural language documents, to improve retrieval information	Breast cancer
	An automated approach to detect errors and abnormalities due to diseases	Heart diseases

There were some significant technical trends for ontology development in CDM, with most methodologies comprising knowledge acquisition, conceptualisation, semantic modelling, knowledge representation and validation [50,51]. Most used clinical guidelines and rule based approaches [93] to guide ontology development, including a layered approach [94]. An example of the layered ontology framework and methodology was the use of ontology-based multi-agent system (OBMAS) to address semantic interoperability problems of terminology and/or structure amongst e-health systems [49,82,83,95,96]. This approach enables intelligent software agents to act in various semantic contexts in multilingual [97] and collaborative environments [94,95,96].

5. Discussion

The DQ domain is fragmented. While there was general agreement that DQ is a multidimensional concept, there was no apparent consensus on what the dimensions are and how they should be defined and operationalised. Preferences for the dimensions were often based on intuitive understanding,

industry experience or literature review [98]. This variation is probably inherent in the contextual definition of DQ in terms of “fitness for purpose/use” [54]. Specific operational definitions of the dimensions of DQ have been proposed [57] but they tended to add to the confusing variety and variability. Fairly sophisticated measures of the most frequently used DQ dimensions (accuracy, completeness, consistency, correctness and timeliness) have been developed [63,70,99,100]. These are likely core dimensions on which to build a consensus DQ ontology [20,21] to enable the consistent measurement of DQ across all contexts and domains.

The quality of routinely collected electronic information and their fitness for purpose is determined by more than just the GIGO – *garbage in garbage out* – principle. Determinants of poor DQ include the lack of coding rules, leading to much of the data being incomplete or in relatively inaccessible text format; wrong diagnoses; incomplete or inaccurate data entry; errors in spelling or coding; corruption of the database architecture or management system; mal-compliance to the organisational data protocols and errors in data extraction [101]. The large and increasing amount of potentially relevant health and health services data collected as part of routine practice compounds the DQ challenge. However, apart from

a computerised solution and a need to filter and sort data in terms of their quality characteristics [58], there was no agreement on whether and how these data should be curated and preserved [46].

The consistency dimension of DQ is concerned with semantic interoperability, a significant issue in CDWs, where the different data sources often used different models, schemas and vocabularies [102]. The semantic interoperability problem increases with the growing secondary use of the data in CIS, in both primary and secondary care settings, for health care, public health and epidemiological research [65]. This is compounded in international multicentre studies by the logistic difficulties and different levels of commitment to DQ [103]. A standard terminology such as SNOMED-CT [42] is part of a comprehensive ontology-based solution to provide a unified semantic framework to harmonise the contribution of different data sources to the specific purpose. This includes data dictionaries with accurately specified metadata and production rules are needed to standardise the assessment of DQ [104] for the fitness for purpose in different clinical domains and contexts. This ontology-driven integration of local architectures with flexible network infrastructures for unified data access will enable automated assessment, management and monitoring of the DQ of large datasets of routinely collected data [79,105,106,107,108], through intelligent software agents with or without guidance from human users.

The increasing research and development in ontologically rich approaches to data quality (DQ) and chronic disease management (CDM) across a range of tasks in a range of health and biomedical informatics domains is promising. These tasks included the addressing of semantic interoperability, data quality to underpin the safety and quality of electronic decision support in diagnoses and management, improving flexibility of information management and linkage in clinical information systems, and reducing complexity of data analyses in complex information models and networks such as Bayesian networks. However, research to date has mainly focused on the design and development of tools, with little substantial research into the use of ontologies to assess and/or manage DQ in CDM [46,77,109]. There were few evaluative studies on the cost-effectiveness of ontological approaches in DQ and quality of care. This was due to a number of scope, methodological, contextual and ethical-legal issues and challenges identified for this relatively immature field. Nevertheless, some guidance on the directions to examine DQ at both data and ontology levels was demonstrated by the Information Quality Triangle project [99], which benchmarked technical standards for information quality at the model/ontology (Health Level 7) reference information model and the data (WHO-ATC terminology for drugs) levels.

Reference or domain ontologies had been shown to improve DQ by influencing data collection and analysis [110]. A reference DQ ontology can potentially act as a benchmark for assessing DQ. The DQ ontology can be constructed from dimensions such as completeness, correctness, consistency and timeliness; all of which can be measured using a ratio scale. Other time-related dimensions can be defined and measured in terms of system currency, storage time and volatility [57]. However, we need a greater understanding of the

relationships and overlaps between the dimensions, which requires significant quantitative and qualitative research. DQ ontologies can be complex and may have to be defined in different layers, such as application and domain ontologies. However, meaningful relationships to real world situations must not be lost with increasing levels of abstraction and reduction with formalisation and implementation of the conceptual models.

DQ management (DQM) is important because poor DQ is a substantial economic and social burden: it consumes up to 10% of an organisation's revenues [111]; leads to poor planning and delivery of health services, takes longer to make poorer decisions; lowers consumer satisfaction; and increases difficulty in reengineering work and information flows to improve service delivery [56]. DQM will and must address these challenges through the establishment and deployment of roles, responsibilities, policies, and procedures concerning the acquisition, maintenance, dissemination, and disposition of data [73] within and across organisations. DQM of CDWs include optimising data extraction, cleansing and/or transformation, periodical updates and data federation [46,51,81,85,92,96,112].

Judicious presentation of good quality information can improve decision-making in health organizations [49,76,105,111,113,114,115,116] as it enables more efficient and effective use of data in health care [113,114,117]. In addition to the content e.g. measures of DQ [64], the way the information is presented can also affect health care and health literacy. This is where the relationships between the real world and the information model, which are often weakened with the level of abstraction and modelling, can be revisited and the messages made more relevant through ontology-based approaches.

However, this literature review was limited by the immaturity of the field. Most of the papers reported on studies that designed, developed, assessed and evaluated information models and ontologies in the chronic diseases context. The lack of systematic and comprehensive ontological approaches for the development of DQ in CDM is compounded by a lack of studies that evaluated the efficacy of the ontological approach or the relationship to DQ or improved integrated CDM.

In summary, this review suggests that ontologically rich approaches to DQ may be more cost-effective than the traditional data/information modelling [52,76,77,79]. The mapping between the real world and information systems, using a design-oriented method with ontological foundations [98], is logically and intuitively advantageous to ensure a well-grounded approach to the design and development of a practical and useful DQ ontology. Ontologically rich approaches that are well contextualised in the professional, legal and social environments, can inform policy development, planning and implementation; quality monitoring [118]; control of costs of external data failure and complementary costs of data-quality assurance [119]; and improvement of the accuracy, validity and reliability of data collection, storage, extraction and linkage algorithms and tools [120]. It is also applicable to information retrieval and analysis, intelligent data mining (seeking concepts and relationships), discover new knowledge, and reuse knowledge for decision support systems and patient decision aids [121].

6. Conclusions

DQ is a multidimensional concept, but lacks a consensus framework and definitions, partly because DQ is defined in terms of “fitness for use”. The key barriers to the optimal use of routinely collected data are increasing data quantity, poor data quality, and lack of semantic interoperability. Poor DQ and data not fit for purpose have significant economic costs, both in terms of direct costs and indirect costs in terms of poor decisions and planning by organisations and individuals, and poor quality and safety of care.

DQ must be measurable consistently across all domains and contexts. The most frequently reported DQ dimensions – completeness, correctness, consistency and timeliness – can be a starting point for a DQ ontology. An ontology-based approach to DQ would be flexible and modular, enabling intelligent software agents to act in various semantic contexts to specify metadata and assess/manage DQ accurately within specified constraints and contexts. The formalisation and implementation of the DQ ontology as an application will enable automated and cost-effective assessment of DQ.

The challenges to the development and validation of a DQ ontology in CDM include methodological immaturity, an immature knowledge base, and a lack of tools to support ontology-based database design for CIS and CDW, evaluation of ontological approaches, and engagement of users in design and implementations. A systematic data/information quality R&D program focused on routinely collected clinical data in information systems in primary and secondary care settings, focussed on how to measure the quality of CDM, would improve the quality of our health datasets, understanding of the health of our communities, and the quality of care provided.

Conflict of interest statement

The authors declare that they have no competing interests.

Authors' contributions

STL developed the conceptual framework and templates for the literature review and guided AR in the management of the review. AR appraised all included papers as part of his PhD studies. The same papers were also distributed equally among all the co-authors for independent appraisal. All authors discussed their appraisals with AR and STL to achieve consensus; all participated in the consensus and synthesis workshops. STL prepared this paper iteratively with input from all co-authors prior to submission.

Acknowledgments

The authors would like to thank A/Prof Elizabeth Comino, Prof Jim Warren and Dr Hairong Yu for comments on drafts.

Summary points

What was already known on the topic?

- DQ is a multidimensional concept, but lacks a consensus framework and definitions.
- Aggregating increasingly large datasets raises issues of semantic interoperability and a need for automated methods to assess and manage DQ.
- Lack of certainty about ontological approaches to DQ in chronic disease management (CDM)

What this study added to our knowledge?

- The literature suggests that the core dimensions of the DQ conceptual framework are completeness, consistency, correctness and timeliness.
- An ontological approach has theoretical and practical advantages in developing cost-effective automated methods to address DQ and semantic interoperability
- There is an increasing amount of work on ontology of chronic disease, but little on ontological approaches to DQ in CDM specifically or in health generally. This gap needs to be addressed.

REFERENCES¹

- [1] WHO, 2008–2013 Action Plan For The Global Strategy For The Prevention and Control of Noncommunicable Diseases: Prevent and Control Cardiovascular Diseases, Cancers, Chronic Respiratory Diseases and Diabetes, World Health Organization, Geneva, 2008 (Report No.: 978 92 4 159741 8).
- [2] G. Esselens, R. Westhovens, P. Verschueren, Effectiveness of an integrated outpatient care programme compared with present-day standard care in early rheumatoid arthritis, *Musculoskelet. Care* 7 (March (1)) (2009) 1–16.
- [3] K. Grimmer-Somers, W. Dolejs, J. Atkinson, A. Worley, Integrated GP and allied health care for patients with type 2 diabetes, *Aust. Fam. Physician* 37 (September (9)) (2008) 774–775.
- [4] T. Hammar, P. Rissanen, M.L. Perala, The cost-effectiveness of integrated home care and discharge practice for home care patients, *Health Policy March* (2009).
- [5] L.E. Olsson, E. Hansson, I. Ekman, J. Karlsson, A cost-effectiveness study of a patient-centred integrated care pathway, *J. Adv. Nurs.* 65 (August (8)) (2009) 1626–1635.
- [6] N. Zwar, M. Harris, R. Griffiths, M. Roland, S. Dennis, G.P. Davies, I. Hasan, APHRI Stream Four: A Systematic Review of Chronic Disease Management, Australian Primary Health Care Research Institute, Sydney, 2006.
- [7] D.L. Kodner, C. Spreeuwenberg, Integrated care: meaning, logic, applications, and implications – a discussion paper, *Int. J. Integr. Care* 2 (2002) e12.
- [8] T. Bodenheimer, E. Wagner, K. Grumbach, Improving primary care for patients with chronic illness, *J. Am. Med. Assoc.* 288 (October (14)) (2002) 1775–1779.

¹ Note: Papers included for the literature review are marked with an *. Four of the 61 included papers were not referenced in this paper.

- [9] T. Bodenheimer, E.H. Wagner, K. Grumbach, Improving primary care for patients with chronic illness: the chronic care model, part 2, *J. Am. Med. Assoc.* 288 (October (15)) (2002) 1909–1914.
- [10] S. Smith, S. Allwright, T. O'Dowd, Effectiveness of shared care across the interface between primary and specialty care in chronic disease management, *Cochrane Datab. Syst. Rev.* 3 (July) (2007) CD004910.
- [11] M. Ouwens, H. Wollersheim, R. Hermens, M. Hulscher, R. Grob, Integrated care programmes for chronically ill patients: a review of systematic reviews, *Int. J. Qual. Health Care* 17 (April (2)) (2005) 141–146.
- [12] A. Gillies, Assessing and improving the quality of information for health evaluation and promotion, *Methods Inf. Med.* 39 (3) (2000) 4.
- [13] M.A. Huaman, R.V. Araujo-Castillo, G. Soto, J.M. Neyra, J.A. Quispe, M.F. Fernandez, C.C. Mundaca, D.L. Blazes, 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. Making* 9 (March) (2009).
- [14] A.N. Kiragga, B. Castelnuovo, P. Schaefer, T. Muwonge, P.J. Easterbrook, Quality of data collection in a large HIV observational clinic database in sub-Saharan Africa: implications for clinical research and audit of care, *J. Int. AIDS Soc.* 14 (1) (2011).
- [15] A. Azaouagh, J. Stausberg, Frequency of hospital-acquired pneumonia – comparison between electronic and paper-based patient records, *Pneumologie* 62 (May (5)) (2008) 273–278.
- [16] J. Mitchell, F. Westerduin, Emergency department information system diagnosis: how accurate is it? *Emerg. Med. J.* 25 (November (11)) (2008) 784.
- [17] M.L. Moro, F. Morsillo, Can hospital discharge diagnoses be used for surveillance of surgical-site infections? *J. Hosp. Infect.* 56 (March (3)) (2004) 239–241.
- [18] S. de Lusignan, K. Khunti, J. Belsey, A. Hattersley, J. van Vlymen, H. Gallagher, C. Millett, N. Hague, C. Tomson, K. Harris, A. Majeed, A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: a pilot and validation study of routinely collected data, *Diabet. Med.* 27 (2010) 203–209.
- [19] C. Soto, K. Kleinman, S. Simon, Quality and correlates of medical record documentation in the ambulatory care setting, *BMC Health Serv. Res.* 2 (December (1)) (2002).
- [20] S. Liaw, H. Chen, D. Maneze, J. Taggart, S. Dennis, S. Vagholkar, J. Bunker, Health reform: is current electronic information fit for purpose? *Emerg. Med. Australasia* September (2011).
- [21] S. Liaw, J. Taggart, S. Dennis, A. Yeo, Data quality and fitness for purpose of routinely collected data – a case study from an electronic Practice-Based Research Network (ePBRN), in: *American Medical Informatics Association Annual Symposium 2011*, Springer Verlag, Washington DC, 2011.
- [22] S.J. Lain, C.L. Roberts, R.M. Hadfield, J.C. Bell, J.M. Morris, How accurate is the reporting of obstetric haemorrhage in hospital discharge data? A validation study, *Aust. N. Z. J. Obstet. Gynaecol.* 48 (October (5)) (2008) 481–484.
- [23] H. Quan, B. Li, L.D. Saunders, G.A. Parsons, C.I. Nilsson, A. Alibhai, W.A. Ghali, I. Investigators, Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database, *Health Serv. Res.* 43 (August (4)) (2008) 1424–1441.
- [24] W.T. Hamilton, A.P. Round, D. Sharp, T.J. Peters, The quality of record keeping in primary care: a comparison of computerised, paper and hybrid systems, *Br. J. Gen. Pract.* 53 (December (497)) (2003) 929–933 (Discussion 33).
- [25] K. Thiru, A. Hassey, F. Sullivan, Systematic review of scope and quality of electronic patient record data in primary care, *Br. Med. J.* 326 (May (398)) (2003) 1070.
- [26] P. Davis, R. Lay-Yee, S. Schug, R. Briant, A. Scott, S. Johnson, et al., Adverse events regional feasibility study: indicative findings, *N. Z. Med. J.* 114 (1131) (2001) 203–205.
- [27] W. Runciman, R. Webb, S. Helps, E. Thomas, B. Sexton, D. Studdert, et al., A comparison of iatrogenic injury studies in Australia and the USA. II. Reviewer behaviour and quality of care, *Int. J. Qual. Health Care* 12 (5) (2000) 379–388.
- [28] E. Thomas, D. Studdert, W. Runciman, R. Webb, E. Sexton, R. Wilson, et al., A comparison of iatrogenic injury studies in Australia and the USA. I. Context, methods, casemix, population, patient and hospital characteristics, *Int. J. Qual. Health Care* 12 (5) (2000) 371–378.
- [29] C. Vincent, G. Neale, M. Woloshynowych, Adverse events in British hospitals: preliminary retrospective record review, *Br. Med. J.* 322 (March) (2001) 517–519.
- [30] A. Adaji, P. Schattner, K. Jones, The use of information technology to enhance diabetes management in primary care: a literature review, *Inform. Prim. Care* 16 (3) (2008) 229–237.
- [31] S. Liaw, D. Boyle, Primary care informatics and integrated care of chronic disease, in: E. Hovenga, M. Kidd, S. Garde, C.H.L. Cossio (Eds.), *Health Informatics: An Overview*, vol. 151, Studies in Health Technology and Informatics, IOS Press, 2010 (Chapter 20, ISBN:978-1-60750-092-6).
- [32] E. Cummings, C. Showell, E. Roehrer, B. Churchill, K. Yee, M. Wong, P. Turner, Discharge, Referral and Admission: A Structured Evidence-based Literature Review, eHealth Services Research Group, University of Tasmania (on behalf of the Australian Commission on Safety and Quality in Health Care, and the NSW Department of Health), Australia, 2010.
- [33] Commonwealth of Australia, Primary Health Care Reform in Australia, Report to Support Australia's First National Primary Health Care Strategy, Australian Government, Canberra, 2009.
- [34] National Health & Hospital Reform Commission, in: Ageing DoHa (Ed.), *A Healthier Future For All Australians – Final Report of the National Health and Hospitals Reform Commission – June 2009*, Commonwealth of Australia, Canberra, 2009.
- [35] National Preventative Health Taskforce, Australia: The Healthiest Country by 2020 – National Preventative Health Strategy – Overview, Commonwealth of Australia Department of Health and Ageing, Canberra, June 20, 2009, Report No.: Publications Approval Number P3-5457 Contract No.: ISBN:1-74186-925-0.
- [36] P. Garling, Final Report of the Special Commission of Inquiry: Acute Care in NSW Public Hospitals, 2008 – Overview, November 27, 2008 ed. NSW Government, Sydney, 2008.
- [37] D. Bates, A. Gawande, Improving safety with information technology, *N. Engl. J. Med.* 348 (2003) 2526–2534.
- [38] D.L. Rubin, S.E. Lewis, C.J. Mungall, S. Misra, M. Westerfield, M. Ashburner, I. Sim, C.G. Chute, H. Solbrig, M.A. Storey, B. Smith, J. Day-Richter, N.F. Noy, M.A. Musen, National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge, *OMICS* 10 (Summer (2)) (2006) 185–198.
- [39] D. Perez-Rey, V. Maojo, M. Garcia-Remesal, R. Alonso-Calvo, H. Billhardt, F. Martin-Sanchez, A. Sousa, ONTOFUSION: ontology-based integration of genomic and clinical databases, *Comput. Biol. Med.* 36 (July–August (7–8)) (2006) 712–730.
- [40] R. Pawson, T. Greenhalgh, G. Harvey, K. Walshe, Realist review – a new method of systematic review designed for

- complex policy interventions, *J. Health Serv. Res. Policy* 10 (Suppl. 1) (2005) 21–34.
- [41] Biomedical Informatics Unit, Protege User Documentation, Stanford University, Palo Alto, 2012, Available from: <http://protege.stanford.edu/doc/users.html> (cited 12.04.12).
- [42] International Health Terminology Standard Development Organisation (IHTSDO), SNOMED Clinical Terms (SNOMED CT), 2012, Available from: http://ihtsdo.org/fileadmin/user_upload/doc/ (cited 12.04.12).
- [43] U.S. National Library of Medicine, Unified Medical Language System® (UMLS®), US National Library of Medicine, Bethesda, 2012, Available from: <http://www.nlm.nih.gov/research/umls/> (cited 12.04.12).
- [44] U.S. National Library of Medicine, Medical Subject Headings (MESH), U.S. National Library of Medicine, Bethesda, 2012, Available from: <http://www.nlm.nih.gov/mesh/> (cited 12.04.12).
- [45] H.S. Pinto, Ontologies: how can they be built? *Knowl. Inform. Syst.* 6 (4) (2004) 441–464.
- [46] A. Preece, P. Missier, S. Ernbury, B. Jin, M. Greenwood, An ontology-based approach to handling information quality in e-science, *Concurr. Comput. Pract. Exp.* 20 (March (3)) (2008) 253–264.
- [47] M. Fox, Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, in: F.J. Belli FaR (Ed.), *Lecture Notes in Artificial Intelligence* #604, Springer-Verlag, Berlin, 1992, pp. 25–34.
- [48] F. Baader, I. Horrocks, U. Sattler, Chapter 3. Description logics, in: V.L. Frank van Harmelen, Bruce Porter (Eds.), *Handbook of Knowledge Representation*, Elsevier, Berlin, 2007.
- [49] Ontology-based multi-agent systems support human disease study and control, in: M. Hadzic, E. Chang (Eds.), *International Conference on Self Organization and Adaptation of Multi-Agent and Grid Systems (SOAS)*, 2005 Dec 11 2005, IOS Press, Glasgow, UK/Amsterdam, The Netherlands, 2005.
- [50] W. Ying, J. Wimalasiri, P. Ray, S. Chattopadhyay, C. Wilson, An ontology driven multi-agent approach to integrated e-health systems, *Int. J. E-Health Med. Commun.* 1 (1) (2010) 12.
- [51] C. Kuziemsky, F. Lau, A four stage approach for ontology-based health information system design, *Artif. Intell. Med.* 50 (2010) 18.
- [52] R. Valencia-Garcia, J.T. Fernandez-Breis, J.M. Ruiz-Martinez, F. Garcia-Sanchez, R. Martinez-Bejar, A knowledge acquisition methodology to ontology construction for information retrieval from medical documents, *Expert Syst.* 25 (July (3)) (2008) 314–334.
- [53] C. Van Buggenhout, W. Ceusters, A novel view on information content of concepts in a large ontology and a view on the structure and the quality of the ontology, *Int. J. Med. Inf.* 74 (March (2–4)) (2005) 125–132.
- [54] R.Y. Wang, A product perspective on total data quality management, *CACM* 41 (February (2)) (1998) 58–65.
- [55] Canadian Institute for Health Information, The CIHI Data Quality Framework, CIHI, Ottawa, Ontario, 2009.
- [56] T. Redman, Measuring data accuracy, in: R. Wang, e. Rea (Eds.), *Information Quality*, ME Sharpe, Inc., Armonk NY, 2005, p. 21.
- [57] Y. Wand, R.Y. Wang, Anchoring data quality dimensions in ontological foundations, *CACM* 39 (November (11)) (1996) 86–95.
- [58] R. Wang, D. Strong, L. Guarascio, Beyond accuracy: what data quality means to data consumers, *J. Manage. Inform. Syst.* 12 (4) (1996) 5–33.
- [59] K. Jordan, A. Clarke, D. Symmons, D. Fleming, M. Porcheret, U. Kadam, P. Croft, Measuring disease prevalence: a comparison of musculoskeletal disease using four general practice consultation databases, *Br. J. Gen. Pract.* 57 (2007) 7–14.
- [60] Measuring the quality of data models: an evaluation of the use of quality metrics in practice, in: D. Moody (Ed.), *11th European Conf on Information Systems*, 2003.
- [61] S. de Lusignan, N. Hague, J. van Vlymen, P. Kumarapeli, Routinely-collected general practice data are complex, but with systematic processing can be used for quality improvement and research, *Inform. Prim. Care* 14 (1) (2006) 59–66.
- [62] S. de Lusignan, C. van Weel, The use of routinely collected computer data for research in primary care: opportunities and challenges, *Fam. Pract.* 23 (April (2)) (2006) 253–263.
- [63] D. Arts, N. de Keizer, G.J. Scheffer, E. de Jonge, Quality of data collected for severity of illness scores in the Dutch National Intensive Care Evaluation (NICE) registry, *Intens. Care Med.* 28 (May (5)) (2002) 656–659.
- [64] H. Britt, G. Miller, C. Bayrarn, The quality of data on general practice – a discussion of BEACH reliability and validity, *Aust. Fam. Physician* 36 (January–February (1–2)) (2007) 36–40.
- [65] S. de Lusignan, J. Metsemakers, P. Houwink, V. Gunnarsdottir, J. van der Lei, Routinely collected general practice data: goldmines for research? A report of the European Federation for Medical Informatics Primary Care Informatics Working Group (EFMI PCIWG) from MIE2006, Maastricht, The Netherlands, *Inform. Prim. Care* 14 (3) (2006) 203–209.
- [66] L. Will, Glossary of Terms Relating to Thesauri and Other Forms of Structured Vocabulary for Information Retrieval, 2007, Available from: <http://www.willpowerinfo.co.uk/glossary.htm>
- [67] Jernst, What are the Differences Between a Vocabulary, a Taxonomy, a Thesaurus, an Ontology, and a Meta-model? 2003, Available from: <http://www.metamodel.com/article.php?story=2003011223271>
- [68] K. Vanopstal, J. Buyschaert, R. Vander Stichele, G. Laureys, Vocabularies and retrieval tools in biomedicine: disentangling the terminological knot, *J. Med. Syst.* 12 (November) (2009).
- [69] T.R. Gruber, Toward principles for the design of ontologies used for knowledge sharing, *Int. J. Hum.-Comput. Stud.* 43 (5–6.) (1995).
- [70] C. Jacquelinet, A. Burgun, D. Delamarre, N. Strang, S. Djabbour, B. Boutin, P. Le Beux, Developing the ontological foundations of a terminological system for end-stage diseases, organ failure, dialysis and transplantation, *Int. J. Med. Inf.* 70 (2–3) (2003) 317–328, [http://dx.doi.org/10.1016/S1386-5056\(03\)00046-7](http://dx.doi.org/10.1016/S1386-5056(03)00046-7).
- [71] T. Mabotuwana, J. Warren, An ontology-based approach to enhance querying capabilities of general practice medicine for better management of hypertension, *Artif. Intell. Med.* 47 (2) (2009) 87–103.
- [72] P. Topalis, E. Dialynas, E. Mittra, E. Deligianni, I. Siden-Kiamos, C. Louis, A set of ontologies to drive tools for the control of vector-borne diseases, *J. Biomed. Inform.* 44 (February (1)) (2011) 42–47.
- [73] S. Brüggemann, F. Grüning, Using ontologies providing domain knowledge for data quality management, *Stud. Comput. Intel.* 221 (2009) 187–203.
- [74] H. Min, F.J. Manion, E. Goralczyk, Y.N. Wong, E. Ross, J.R. Beck, Integration of prostate cancer clinical data using an ontology, *J. Biomed. Inform.* 42 (December (6)) (2009) 1035–1045.

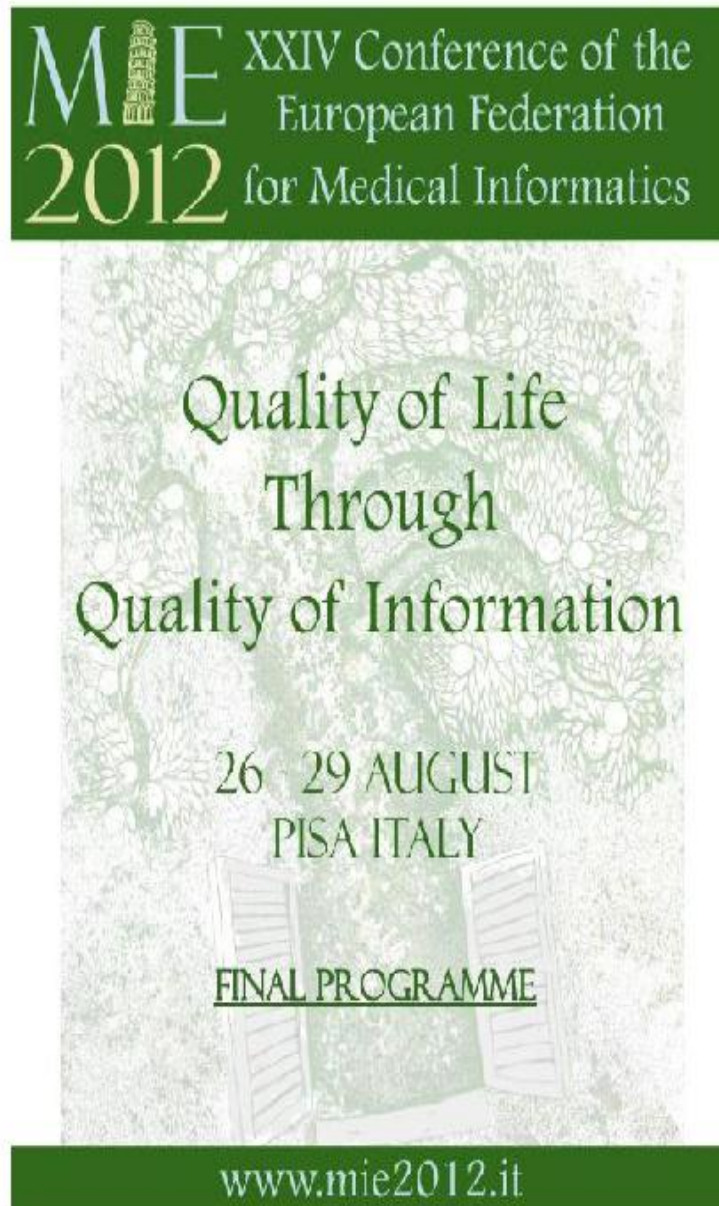
- [75] T. Young, S.W. Tu, L. Tennakoon, D. Vismer, V. Astakhov, A. Gupta, J.S. Grethe, M.E. Martone, A.K. Das, M.J. McAuliffe, IEEE, Ontology driven data integration for autism research, in: 22nd IEEE International Symposium on Computer-Based Medical Systems, New York, 2009, pp. 54–60.
- [76] J. O'Donoghue, J. Herbert, P. O'Reilly, D. Sammon, 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, 2009, pp. 299–302.
- [77] A.M. Orme, H. Yao, L.H. Etzkorn, Indicating ontology data quality, stability, and completeness throughout ontology evolution, *J. Softw. Maint. Evol. Res. Pract.* 19 (January–February (1)) (2007) 49–75.
- [78] O. Vorochek, Y. Biletskiy, Toward assessing data quality of ontology matching on the web CNSR, in: 2007 Proceedings of the Fifth Annual Conference on Communication Networks and Services Research, 2007, Available from Go to ISI://000246988200045 (serial on the Internet).
- [79] S.L. Nimmagadda, S.K. Nimmagadda, H. Dreher, IEEE, 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, 2008, pp. 465–473.
- [80] S. Abidi, Ontology-based knowledge modeling to provide decision support for comorbid diseases, in: The 19th European Conference in Artificial Intelligence, Lisbon, 2011, pp. 27–39.
- [81] M. Buranarach, N. Chalortham, P. Chatvorawit, Y. Thein, T. Supnithi, An Ontology-based Framework for Development of Clinical Reminder System to Support Chronic Disease Healthcare, 2009, Available from <http://text.hlt.nectec.or.th/ontology/sites/default/files/reminder.jsbme09.cr.0.pdf>
- [82] N. Chalortham, M. Buranarach, T. Supnithi, Ontology Development for Type II Diabetes Mellitus Clinical Support System, 2009, Available from: <http://text.hlt.nectec.or.th/ontology/sites/default/files/CRdm2css.0.pdf>
- [83] M. Hadzic, D.S. Dillon, T.S. Dillon, Use and modeling of multi-agent systems in medicine, in: A.M. Tjoa, R.R. Wagner (Eds.), *Proceedings of the 20th International Workshop on Database and Expert Systems Application*, 2009.
- [84] A.J. Jara, E.J. Blaya, M.A. Zamora, A.F.G. Skarmeta, IEEE, An Ontology and Rule Based Intelligent Information System to Detect and Predict Myocardial Diseases, IEEE, New York, 2009.
- [85] Ontology based personalized modeling for chronic disease risk analysis: an integrated approach, in: A. Verma, N. Kasabov, A. Rush, Q. Song (Eds.), *The 15th International Conference on Advances in Neuro-Information Processing 2008*, Springer-Verlag, Berlin/Heidelberg, 2009.
- [86] A. Baneyx, J. Charlet, M.C. Jaulent, Building an ontology of pulmonary diseases with natural language processing tools using textual corpora, *Int. J. Med. Inf.* 76 (February–March (2–3)) (2007) 208–215 (Proceedings Paper).
- [87] O. Coltell, M. Arregui, C. Perez, M.A. Domenech, D. Corella, R. Chalmers, Building an ontology on genomic epidemiology of cardiovascular diseases, in: N. Callaas, M. Sanchez, J.M. Pineda (Eds.), *8th World Multi-Conference on Systemics, Cybernetics, and Informatics*, Vol. XVI, Proceedings, Int Inst Informatics & Systemic, Orlando, 2004.
- [88] A. Gupta, B. Ludäscher, J.S. Grethe, M.E. Martone, Towards a formalization of disease-specific ontologies for neuroinformatics, *Neural Netw.* 16 (9) (2003) 1277–1292.
- [89] S. Gedzelman, M. Simonet, D. Bernhard, G. Diallo, P. Palmer, IEEE, Building an ontology of cardio-vascular diseases for concept-based information retrieval, in: *Computers in Cardiology 2005*, vol. 32, IEEE, New York, 2005, pp. 255–258.
- [90] K. McGarry, S. Garfield, S. Wermter, Auto-extraction, representation and integration of a diabetes ontology using Bayesian networks, in: P. Kokol, V. Podgorelec, D. Micetic-Turk, M. Zorman, M. Verlic (Eds.), *Twentieth IEEE International Symposium on Computer-Based Medical Systems*, Proceedings, 2007, pp. 612–617.
- [91] B.J. Jeon, I.Y. Ko, in: D. Howard, P.K. Rhee, S. Halgamuge, S.J. Yoo (Eds.), *Ontology-based Semi-automatic Construction of Bayesian Network Models for Diagnosing Diseases in e-Health Applications*, IEEE Computer Soc., Los Alamitos, 2007.
- [92] M. Maragoudakis, D. Lymberopoulos, N. Fakotakis, K. Spiropoulos, Ieee, 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, IEEE, New York, 2008, pp. 3807–3810.
- [93] S. Tu, L. Tennakoon, M. O'Connor, R. Shankar, A. Das, Using an integrated ontology and information model for querying and reasoning about phenotypes: the case of autism, in: *AMIA Annu Symp Proc.*, AMIA, 2008, pp. 727–731.
- [94] G. Colombo, D. Merico, G. Boncoraglio, F. De Paoli, J. Ellul, G. Frisoni, Z. Nagy, A. van der Lugt, I. Vassányi, M. Antoniotti, An ontological modeling approach to cerebrovascular disease studies: the NEUROWEB case, *J. Biomed. Inform.* 43 (4) (2010) 469–484.
- [95] G. Ganendran, Q. Tran, P. Ganguly, P. Ray, G. Low, An ontology-driven multi-agent approach for healthcare, *HIC* (2002) 464–469.
- [96] P. Ganguly, P. Ray, N. Parameswaran, Semantic interoperability in telemedicine through ontology-driven services, *Telemed. e-Health* 11 (3) (2005) 8.
- [97] N. Collier, A. Kawazoe, L. Jin, M. Shigematsu, D. Dien, R.A. Barrero, K. Takeuchi, A. Kawtrakul, A multilingual ontology for infectious disease surveillance: rationale, design and challenges, *Lang. Res. Eval.* 40 (3–4) (2006) 405–413.
- [98] Y. Wand, Y. Wang, Anchoring data quality dimensions in ontological foundations, *CACM* 36 (11) (1996) 10.
- [99] R. Choquet, S. Qouiya, D. Ouagne, E. Pasche, C. Daniel, O. Boussaid, M. Jaulent, The Information Quality Triangle: a methodology to assess clinical information quality, *Stud. Health Technol. Inform.* 160 (Pt 1) (2010) 699–703.
- [100] W. Hogan, M. Wagner, Accuracy of data in computer-based patient records, *J. Am. Med. Inform. Assoc.* 4 (5) (1997) 342–355.
- [101] G. Michalakidis, P. Kumarapeli, A. Ring, J. van Vlymen, P. Krause, S. de Lusignan, A system for solution-oriented reporting of errors associated with the extraction of routinely collected clinical data for research and quality improvement, *Stud. Health Technol. Inform.* 160 (Pt 1) (2010) 724–728.
- [102] J.-Y. Han, L.-Z. Xu, Y.-S. Dong, An overview of data quality research, *Comput. Sci.* 35 (2) (2008).
- [103] S. de Lusignan, S. Liaw, P. Krause, V. Curcin, M. Vicente, G. Michalakidis, L. Agreus, P. Leysen, N. Shaw, K. Mendis, Key concepts to assess the readiness of data for International research: data quality, lineage and provenance, extraction and processing errors, traceability, and curation, in: *IMIA Yearbook of Medical Informatics*, 2011, pp. 112–121.
- [104] C. O-Hoon, L. Jung-Eun, N. Hong-Seok, B. Doo-Kwon, An efficient method of data quality using quality evaluation ontology, in: *Third 2008 International Conference on Convergence and Hybrid Information Technology*, IEEE CS, Busan, Korea, 2008, pp. 1058–1061.

- [105] O.H. Choi, J.E. Lim, H.S. Na, D.K. Baik, Ieee Computer SOC, An efficient method of data quality using quality evaluation ontology, in: Third 2008 International Conference on Convergence and Hybrid Information Technology, 2008.
- [106] W.L. Chen, S.D. Zhang, X. Gao, Ieee Computer SOC, Anchoring the consistency dimension of data quality using ontology in data integration, in: 2009 Sixth Web Information Systems and Applications Conference, Proceedings, 2009.
- [107] A.U. Frank, Data quality ontology: an ontology for imperfect knowledge, in: S. Winter, M. Duckham, L. Kulik, B. Kuipers (Eds.), Spatial Information Theory, Proceedings, 2007, pp. 406–420.
- [108] B. Stvilia, L. Mon, Y.J. Yi, A model for online consumer health information quality, *J. Am. Soc. Inform. Sci. Technol.* 60 (September (9)) (2009) 1781–1791.
- [109] L. Rao, H. Reichgelt, K.M. Osei-Bryson, An approach for ontology development and assessment using a quality framework, *Knowl. Manage. Res. Pract.* 7 (September (3)) (2009) 260–276.
- [110] P.J. Brown, V. Warmington, M. Laurence, A.T. Prevost, Randomised crossover trial comparing the performance of Clinical Terms Version 3 and Read Codes 5 byte set coding schemes in general practice, *Br. Med. J.* 326 (May (7399)) (2003) 1127.
- [111] Data quality, information and decision making a healthcare case study, in: K. Kerr, A. Norris, R. Stockdale (Eds.), 18th Australasian Conference on Information Systems, Tbowoomba, Australia, 2007.
- [112] C. Cunningham-Myrie, M. Reid, T.E. Forrester, 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 Med. J.* 57 (4) (2008) 383–392.
- [113] T.H. Chen, Modeling the effect of information quality on risk behavior change and the transmission of infectious diseases, *Math. Biosci.* 217 (2) (2009) 125–133.
- [114] C.S. Lee, M.H. Wang, G. Acampora, V. Loia, C.Y. Hsu, IEEE, Ontology-based Intelligent Fuzzy Agent for Diabetes Application, IEEE, New York, 2009.
- [115] L. Lima, P. Novais, R. Costa, J. Cruz, J. Neves, Decision making based on quality-of-information a clinical guideline for chronic obstructive pulmonary disease scenario, in: F. de Leon, A. de Carvalho, S. Rodríguez-González, J. De Paz Santana, J. Rodríguez (Eds.), Distributed Computing and Artificial Intelligence, Springer, Berlin/Heidelberg, 2010, pp. 417–424.
- [116] L. Lima, P. Novais, R. Costa, J.B. Cruz, J. Neves, Group decision making and quality-of-information in e-health systems, *Logic J. IGPL* 19 (April (2)) (2011) 315–332.
- [117] D.G. Arts, R.J. Bosman, E. de Jonge, J.C. Joore, N.F. de Keizer, Training in data definitions improves quality of intensive care data, *Crit. Care* 7 (2) (2003 Apr) 179–184.
- [118] R.Y. Wang, E.M. Pierce, S.E. Madnick, C.W. Fisher (Eds.), Information Quality, ME Sharpe, Inc., Armonk, NY, 2005.
- [119] R.Y. Wang, V. Storey, C. Firth, A framework for analysis of data quality research, *IEEE Trans. Knowl. Data Eng.* 7 (August (4)) (1995) 623–640.
- [120] D. Arts, N. De Keizer, G.J. Scheffer, Defining and improving data quality in medical registries: a literature review, case study, and generic framework, *JAMIA* 9 (6) (2002) 600–611.
- [121] M. Wang, C. Lee, H. Li, W. Ko, Ontology-based fuzzy inference agent for diabetes classification, in: Annual Meeting of the North American Fuzzy Information Processing Society, NAFIPS'07, San Diego, CA, June 24–27, 2007.

INCLUDE PAPERS NOT QUOTED IN THIS PAPER

- [R1] B.K. Kahn, D.M. Strong, R.Y. Wang, Information quality benchmarks: product and service performance, *Commun. ACM* 45 (4) (2002) 8.
- [R2] M. Esposito, Congenital heart disease: an ontology-based approach for the examination of the cardiovascular system, in: I. Lovrek, R. Howlett, L. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Springer, Berlin/Heidelberg, 2008, pp. 509–516.
- [R3] M. Esposito, An ontological and non-monotonic rule-based approach to label medical images, in: Third International IEEE Conference on Signal-Image Technologies and Internet-Based System, 2007 SITIS'07, Shanghai, December 16–18, 2007.
- [R4] H. Li, W. Ko, Automated food ontology construction mechanism for diabetes diet care, in: 2007 International Conference on Machine Learning and Cybernetics, Hong Kong, August 19–22, 2007.

APPENDIX 2



Developing an ontology for data quality in chronic disease management

Rahimi, A, MSc ^{a,b}, Liaw, ST, PhD, FRACGP, FACHI ^{a,1}, Taggart, J, MPH ^a, Ray, P ^a, PhD, Yu, H ^a, PhD

^a *University of New South Wales, Sydney, Australia,*

^b *Isfahan University of Medical Sciences, Isfahan, Iran*

Abstract. The growing use of electronic health records raises the question of quality control of routinely-collected data for clinical care and research. Improving data quality (DQ) can improve the quality of decisions, evidence-based care and patient outcomes. This paper describes the methodology and progress of the development of an ontology to assess DQ in diabetes. The specification and conceptualization phase is largely completed. The implementation of the DQ ontology will be tested on the electronic Practice Based Research Network (ePBRN) dataset.

Keywords: Data quality, Chronic Disease Management, Ontology, Clinical information systems

Introduction

There is a growing recognition of the use of Clinical Information Systems (CIS) for chronic disease management (CDM), public health services and epidemiological research ¹. Improving data quality (DQ) can improve the quality of decisions and lead to better policy, evidence-based care and patient outcomes.

In the biomedical informatics literature, ontologies have been described as “collections of formal, machine-process able and human interpretable representation of the entities, and the relations among those entities” ². Our literature review indicates that there are few studies on the application of ontology for DQ in health care of CDM.

Objective: This study will develop a conceptual framework and methodology to guide the development and validation of DQ ontology in CDM. It draws on our current work into completeness, correctness, and consistency (the 3Cs) of DQ of routinely collected data from general practice in diabetes

1. Methods

We conducted a literature review to provide input into the ontology development. The literature review addressed the role(s) of ontologies in the assessment, collation and management of DQ in health care. Key researchers informed the literature review

¹ Corresponding Author: siaw@unsw.edu.au

process and provided input into the findings. There are 5 stages in our ontology development:

Specification: The purpose and scope of this ontology is to improve the 3C of DQ within the context of diabetes. **Conceptualization:** There are 2 parts to this stage. i) Auditing the electronic Practice Based Research Network (ePBRN) data for the 3C of DQ and collecting qualitative information from ePBRN participants about how they use the clinical systems for the care of diabetes. Frequencies of routinely collected data from the ePBRN practices are determined for the 3C as an indication of DQ for diabetes (5). (ii) Semi-structured interviews with GPs, specialists and nurses seeking specific information needs to scope the ontology domain (concepts, vocabularies and their relationships). A hybrid grounded theory-participatory design (GT-PD) methodology³ will guide the data collection and analysis.

Formalization: The first step in formalization is developing the domain ontology (diabetes disease register ontology). This ontology and the defined rules can generate logical inferences and control the relevant objects, such as the patient with a diagnosis of diabetes mellitus (DM), and other related properties. The second step is developing the DQ ontology. This ontology uses the definitions of the 3Cs of DQ. The outputs are a domain ontology and sub-ontologies, which represent the ontological structure and concepts relationships.

Implementation: The upper ontology has been implemented first to represent the domain broadly. It is firstly used for describing necessary general concepts from CDM point of view and secondly, in our case, particularly adds the constraints for lower datasets from database in order to meet our DQ goals. **Evaluation:** Differences in the quality of data will be compared using the final DQ ontology model with not using the model for the 3C of diabetic patient records. Health care data consumers' evaluation can assess. **Maintenance** updates and corrects the implemented ontology e.g. DQ assessment and management in the ePBRN.

2. Results

2.1 Conceptual Stage: We identified 61 papers for synthesis in the literature review. The 3C of DQ were the most frequently reported dimensions of DQ and CDM were poorly evaluated. Ontological approaches to improving DQ for decision making have been effective in both primary and secondary care settings⁴. Analysis of diabetes management found that 200+ concepts were required for a comprehensive conceptual model.

Useful ontology tools for developing the ontology for CDM were identified including: Protégé^{5, 6} which has a number of advantages: (i) it is an open source standard; (ii) it provides additional semantics; (iii) it enables simultaneous editing of the same database; (iv) it uses OWL, SWRL and SPARQL³, as standard ontology representation languages; and (v) it can manage large-scale domain ontologies³; SNOMED CT and UMLS are comprehensive medical vocabularies in widespread use³; METHONTOLOGY is the most mature^{3, 7} of ontology methodologies⁷; Ontology-based multi-agent systems (OBMAS)⁵ that use a layered ontology-driven methodology and tools to enable their application in different domains and layers. This can resolve the semantic interoperability problem.

2.2. Formalization stage: Our research team had created the rules of 3C for ePBRN data using Australian National Guidelines, Australian National Data Dictionary (ANDD), ICD-10, ICPC and SNOMED CT to define data properties, use uniform data types and formats for each variable.

2.3. Results of the implementation and evaluation stages are still to be completed. We identified and made decisions to select the most frequently used ontology reasoners (Pellet 2.3.0 and HermiT 1.3.6)⁸. Both reasoners found no logical inconsistencies (e.g., loops) in our ontology. These reasoners were used to infer logical consequences from a set of asserted facts or axioms, and provide automated support for the logical reasoning tasks such as classification, debugging and querying in the system. Expected results are to identify the individuals in ontology or records from datasets of databases which follow or violate the rules (represented in DQ ontology) of the 3C of data.

3. Discussion

Work to date supports an ontological approach to develop the 3C of DQ for diabetes management in the ePBRN. However, we need to consider the actual and potential challenges as we transition between the conceptualization and formalization and implementation stages. So, will this ontology differ with different chronic diseases and care settings? Will it differ with different provider and patient socio-demographics? Will the limitations of the tools and development environment used for formalization cause the loss of too much contextual richness and render the exercise irrelevant? The loss of reality with increasing abstraction is a challenge inherent in all informatics and ontology work. Multiple views and taxonomies, often with conflicting semantics, present another challenge.

References

1. Choquet R, Qouiyy S, Ouagne D, Pasche E, Daniel C, Boussaid O, et al. The Information Quality Triangle: a methodology to assess clinical information quality. *Stud Health Technol Inform.* 2010;160(Pt 1):699-703.
2. Rubin DL, Lewis SE, Mungall CJ, Misra S, Westerfield M, Ashburner M, et al. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *OMICS.* 2006 Summer;10(2):185-98.
3. Kuziemy C, Lau F. A four stage approach for ontology-based health information system design. *Artificial Intelligence in Medicine* 2010 2010;50:18.
4. Liaw S, Taggart J, Dennis S, Yeo A. Data quality and fitness for purpose of routinely collected data – a general practice case study from an electronic Practice-Based Research Network (ePBRN). *AMIA 2011 Annual Symposium Improving Health: Informatics and IT Changing the World*; 2011 October 22-26, 2011; Washington DC, US. Washington DC: AMIA; 2011. p. 785-94.
5. Ying W, Wimalasiri J, Ray P, Chattopadhyay s, Wilson C. An Ontology Driven Multi-Agent Approach to Integrated e-Health Systems *International Journal of E-Health and Medical Communications (IJEHMC).* 2010 2010;1(1):12.
6. Esposito M. Congenital Heart Disease: An ontology-based approach for the examination of the cardiovascular system. In: Lovrek I, editor. *Knowledge - Based Intelligent Information and Engineering Systems, Pt 1, Proceedings*; 2008. p. 509-16.
7. Corcho O, Fernandez M, Gomez A. Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data & Knowledge Engineering.* 2003;46:41-64.
8. Huang T, Li W, Yang C, editors. *Comparison of Ontology Reasoners: Racer, Pellet, Fact++* American Geophysical Union, Fall Meeting 2008; 2008. American Geophysical Union.

APPENDIX 3

Health Informatics: Building a Healthcare Future Through Trusted Information
A.J. Maeder and F.J. Martin-Sanchez (Eds.)
IOS Press, 2012
© 2012 The authors and IOS Press. All rights reserved.
doi:10.3233/978-1-61499-078-9-219

219

The University of NSW electronic Practice Based Research Network: Disease registers, data quality and utility

J. TAGGART^a, S.T. LIAW^{a,b,c}, S. DENNIS^a, H. YU^{a,d},
A. RAHIMI^{a,d,e}, B. JALALUDIN^{b,c}, M. HARRIS^a

^a University of NSW Centre for Primary Health Care and Equity,

^b University of NSW School of Public Health and Community Medicine

^c South West Sydney Local Health District

^d Asia-Pacific Ubiquitous Healthcare Research Centre (APuHC), School of Information Systems & Technology Management, Australian School of Business, UNSW.

^e Isfahan University of Medical Sciences, Faculty of Management and Medical Information Sciences, Iran

Abstract. Introduction: Accurate well-maintained registers are a prerequisite to co-ordinated care of patients with chronic diseases. Their effectiveness in enabling improved management is dependent on the quality of the information captured. This paper provides an overview into the methodology and data quality of the electronic Practice Based Research Network. **Methods:** Clinical records with no identifying information are routinely extracted from four general practices. The data are linked in the data warehouse. Data quality is assessed for completeness, correctness and consistency. Reports on data quality are given back to practices and semi-structured interviews provide information to interpret the results and discuss how data quality could be improved. **Findings:** Data quality is mostly complete for sex and date of birth but indigenous status, smoking and weight were incomplete. There are generally high levels of correctness and internal consistency. Completeness of records in assisting the management of diabetes patients using the annual cycle of care was poor. GPs often use the progress notes to enter information during the consultation and coding diagnoses was considered onerous. **Discussion:** The routine capture of electronic clinical health records from primary health care and health services can be used to monitor performance and improve the quality of clinical records. There is a need for accurate and comprehensive clinical records to ensure the safety and quality of clinical practice. Understanding the true reasons for poor data quality is complex. Having a community-based research network may assist in answering some of these questions. **Conclusion:** Electronic health records are increasingly being used for secondary research and evaluation, beyond the primary purpose of supporting clinical care. The data must be of sufficient quality to support these purposes.

Keywords. electronic health information, quality improvement

Introduction

Electronic Health Records (EHR) are needed to support quality health care for the increasing prevalence and burden of chronic diseases: 77% of Australians report one or more long-term health problems and more than half of those aged 65 years and older

have five or more chronic conditions [1]. Australian governments [2-5] have emphasised the need for the effective use of EHR and electronic decision support tools to collect, share and use information to implement safe, effective and coordinated care [6-9]. The EHR may have a role in improving management through enabling the ready identification of cases for inclusion in disease registers and measuring and monitoring of safety and quality of care [10]. Accurate well-maintained registers have long been recognised as a prerequisite to co-ordinated care of patients with chronic diseases [11-19]. Their effectiveness in enabling improved management is dependent on the quality of the information it holds.

Data quality (DQ) is defined by the International Standards Organisation as "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). The Canadian Institute for Health Information quality framework comprises of six quality dimensions: accuracy, timeliness, comparability, usability, relevance and privacy & security [20] while research has mostly focused on accuracy, currency and completeness [21] or completeness, correctness, consistency and timeliness [22, 23].

A range of deficiencies in the routinely collected electronic information for clinical [24-27] purposes has been reported in hospital [28] and general practice [29] information systems, where the lack of coding rules meant that much of the data are often incomplete or in relatively inaccessible text format as opposed to a structured data field. Prescribing data are generally of better quality than diagnostic or lifestyle data [29, 30]. Obvious sources of inaccuracies include the wrong diagnoses, incomplete or inaccurate data entry, errors in spelling or coding, corruption of the database architecture or management system, mal-compliance to the organisational protocols and errors in data extraction [31].

The lack of a common terminology among different EHR and disease registers is another barrier to the effective use of EHR-based disease registers in both research and quality improvement. Routinely collected electronic health care data, aggregated into large data repositories, are increasingly being mined, linked and used for audit, continuous quality improvement in clinical care, health service planning, epidemiological study and evaluation research.

We have identified these DQ issues in the electronic Practice Based Research Network (ePBRN) of general practices and health services in south western Sydney. The network is being established to support clinical audit, quality improvement and research into integrated health services with an initial focus on diabetes. The ePBRN pilot study is examining the issues and challenges for data extraction, linkage and utility.

Patient clinical information is extracted from participating general practices and health service information systems, captured in a clinical data warehouse, linked and used for research purposes. Data quality of clinical records and assurance that terminologies are consistent are paramount in achieving the aim of the network to support integrated care and research.

An approach to improving data quality using ontologies is being pursued for the ePBRN. An ontology is an explicit, formal specification of a shared conceptualisation [32] that provides a vocabulary of terms, their meanings and relationships to be used in various application contexts.

An ontological approach has theoretical and practical advantages in developing cost-effective automated methods to address data quality, including semantic interoperability in large repositories of routinely collected data such as those from

practice-based research networks. This approach enables the systematic development of automated, valid and reliable methods to extract, link and manage data as well as assess the data quality and semantic interoperability issues.

The aims of this paper are to: provide an overview into the methodology of the electronic Practice Based Research Network; and to report some of the findings on the ePBRN data, data quality and linkage.

1. Methods

1.1. Recruitment

General practices involved in the Primary Health Care Research Network (PHReNet) or involved in integrated care with the South West Sydney Local Health District diabetes services are invited to participate in the ePBRN. GPs must consent to the extraction of their patient data.

The first three recruited practices participated in a pilot study to test the system and processes of extracting the required data, sending and saving it in the repository and managing the data. A fourth practice has been recruited since and further recruitment is underway.

1.2. Data Processes

Electronic health records (EHR) of all patients are extracted routinely from the general practices (3 using MD3, one originally used MD2 and one using Practix) using GRHANITE™ and sent encrypted to a secure data warehouse at the University of NSW. No identifying information is extracted from the practice. The GRHANITE™ tool creates a unique patient identifier using components of the patients name, address, date of birth and Medicare number. This unique identifier can be matched probabilistically to enable linkage with other data extracted by GRHANITE™ from the EHRs of different participating services.

The data arrives at the warehouse encrypted. The GRHANITE™ data manager decrypts and links the data which is viewed in SQL Server Manager 2005. Data is captured in a number of tables (e.g. patient, pathology, diagnoses). Patients are identified within each table by the pseudonymised patient identifier. The patient table has one record for each patient while other tables may have multiple records for patients (e.g. medication).

Once data is in the warehouse all records are linked to identify patients who have a record at more than one of the practices or duplicate records with the practice.

1.3. Data Mining

Pathology results from Practix and Medical Director 2 are extracted in text or HTML format. A data mining tool we developed will extract the pathology results for HbA1c, lipids, microalbumin and kidney function at the data warehouse.

1.4. Patient Privacy

A number of strategies are in place to protect patient privacy. These include the use of GRHANITE™ to extract pseudonymised data containing no patient identifying information; providing an opt out option for patients using the GRHANITE™ consent management function (patients are informed about the study via a poster at the practices and are able to have their records excluded from the data extraction by informing the practice staff); storing the data in a password protected and secure electronic format on a physical web server hosted by the University of NSW.

1.5. Literature Review on Data Quality

A literature review was conducted to identify how data quality is defined, assessed and managed in health care, how ontologies are being used and the role(s) of ontologies in chronic disease management. The results of this review are being used to inform the development and implementation of an ontological approach to improve identification of patients and data quality.

1.6. Data Quality

A range of data quality dimensions were identified based on a conceptual framework developed from the literature review. We assessed these and decided to address completeness, correctness and consistency as the first priorities. The following definitions are being used:

- **Completeness** – Two levels of completeness are defined. The first is the availability of at least one record per patient for some social determinants (sex, age, indigenous status) and risk factors (BMI, blood pressure, HbA1c, total cholesterol and smoking). The second level is the availability of information required to make a clinical decision. For the information required to manage diabetes, we started with some of the components of the diabetes annual cycle of care: HbA1c recorded at least once a year, blood pressure twice a year, lipids once a year and BMI twice a year.
- **Correctness** – A valid and appropriate clinical record with correct unit of measurements and within acceptable clinical range. For example, the unit for weight is expressed as kilograms, is appropriate for age and within an acceptable weight range. Outlying results were identified and assessed by two researchers as being within or outside an acceptable range based.
- **Consistency** – A uniform data type, format (e.g. string, numeric) and standard terminology and coding within the practice or externally (e.g. Australian National Data Dictionary, SNOMED or ICPC).

1.7. Feedback to Practices

Reports on data quality and diabetes related care is provided back to practices. Meetings and/or semi-structured interviews between practice staff and the researchers are organised to interpret results and discuss how data quality can be improved.

1.8. Validation

Internal and external validation of the GRHANITE™ extraction tool is being conducted. For internal validation the extraction specifications are being checked against the XML extraction file and the data in the ePBRN data repository.

The external validation involves extracting diabetes data from four practices using the GRHANITE™ extraction tool, the PEN Computer Systems Clinical Audit Tool (CAT) and the Canning Division of General Practice Tool consecutively. The GRHANITE™ extraction specification is being updated to extract data from the same tables as the comparison tools after an initial test showed differences. The data being examined in the validation study include some demographic and diabetes related items. All the data extracted are compared and checked for completeness, correctness and consistency within each practice where possible. Differences within practices will be compared across the practices. The research questions for the external validation are: What is the data quality at each practice using the different extraction methods?; What is the comparative data quality between practices for the extraction methods?; and What are the underlying reasons for the results?

2. Findings

Four practices in south western Sydney are currently providing data routinely to the ePBRN. Table 1 show some information about the practices involved and Table 2 presents the number of records captured in one extraction from each of the practices.

Table 1. ePBRN practices

	Practices			
	Practice 1	Practice 2	Practice 3	Practice 4
Clinical software	MD2 then MD3	MD3	Practix	MD3
GPs	5	4	>7	7
Practice nurse/s	No	No	Yes	Yes

Note: MD=Medical Director

Table 2. Frequencies of some records captured in the data warehouse in December 2011 for ePBRN practices

Records for:	Practice 1	Practice 2	Practice 3	Practice 4	TOTALS
Patients	15,215	3,188	26,621	29,637	74,661
Consultations	310,978	14,657	641,787	488,204	1,455,626
Prescriptions	220,167	7,398	93,071	272,515	593,151
Pathology	233,020	2,871	360,390	423,061	958,798
Measures*	172,476	13,356	21,245*	361,647	547,479

*All measures such as blood pressure, temperature, weight, height, BMI and some pathology (e.g. lipids, HbA1c, kidney function); *Blood pressure, height and weight only included

2.1. Data Quality

Data quality was mostly complete for sex and date of birth but indigenous status, smoking and weight were incomplete. The data generally had high levels of correctness and was internally consistent. Data quality for diabetes patients was similar to the data quality of all patients. However, completeness of records to assist with the management

of diabetes patients and for the annual cycle of care was poor. Table 3 shows the completeness, correctness and consistency of records for all patients in the practices, Table 4 shows the completeness of records for diabetes patients and Table 5 shows the completeness of records for some variables included in the diabetes annual cycle of care for one practice in one year.

Table 3. Data quality of all patient records as at December 2011* (results are percentages)

	Practice 1 (15,215)			Practice 2 (3,188)			Practice 3 (n=26,621)		
	Comp	Corr.	Consist.	Comp.	Corr.	Consist.	Comp.	Corr.	Consist.
Sex	99.9	100	100	99.5	99.5	100	100	99.2	100
Date of birth	99.9	100	100	100	99.9	100	100	99.8	100
Indigenous status	8.3	97.4	NA	3.8	100	100	7.5	100	100
Smoking status	31.9	100	100	48.7	100	100	67.3	100	100
Weight [‡]	0	NA	NA	8.2	100	100	13.57	99.9	100

Table 3 Note: Comp.=complete; corr.=correct; consist.=internal consistency *denominator is the number of patients except for weight correctness and weight consistency that has the denominator as the number of records; NA not available; [‡]Includes multiple records/patient

Table 4. Completeness of records - Percentage of all diabetes patients with at least 1 record for some social determinants and risk factors at three of the ePBRN practices since 2000

	Practice 2 %	Practice 3 %	Practice 4 %
Sex	100	100	100
Date of birth	100	100	100
Height results	27	40	48
Weight results [†]	39	43	52
BMI results	19	39	42
BP results [†]	66	52	56
HbA1c records	10	86	21
HbA1c results [‡]	75	NA	21
TC records	X	65	52

[†] denominator is number with a record; NA Not available; X not extracted

Table 5. Completeness of records for components of the diabetes annual cycle of care for one practice in 2010

	Practice 3(n=453)
HbA1c (1/yr)	156 (34.4)
BP (2/yr)	72 (15.9)
Lipids (1/yr)	301 (66.4)
BMI (2/yr)	62 (13.6)

More details on the data quality findings are published elsewhere [33].

Three semi-structured meetings (one with each of the pilot practices) were held with 7 GPs. GPs said they often used the progress notes rather than use the structured fields in the clinical software.

"I think a lot of the time we just write it in the progress notes and it won't extract."

One practice had tried coding diagnoses for a period of time but found it onerous and not useful for patient care as well as detracting from the consultation.

"I mean we can code for the purposes of extracting data and that sort of thing.it is an added 2 minutes on to a consultation which is difficult to do and that is difficult to actually code because it takes a different style of thinking instead of actually doing the consultation."

2.2. Data Linkage and Probabilistic Matching

5,293 patient records from the four practices created 2,575 patient linkages. These include linkages within each practice (duplicate patient records) and between the practices (the same patient at different practices). Despite significant distances between practices (up to 40 km), there are a significant number of shared patients.

3. Discussion

The routine capture of electronic clinical health records from primary health care and health services can be used to monitor performance and improve the quality of clinical records. The aggregation and linkage of these records can be used for translational, clinical and health services research and quality assurance.

The aim of our ePBRN is to improve integrated care with an initial focus on diabetes. At the heart of this is the need for accurate and comprehensive clinical records to ensure the safety and quality of clinical practice.

Practice based research networks are valuable resources for recruiting clinicians and patients into research studies. It is therefore important that the ePBRN is representative of primary health care services in the area. To date we have medium and large sized practices participating. Solo practices in the area are difficult to recruit. Our long term plan is to recruit 50 general practices and health services in south west Sydney.

The infrastructure required to support a larger network is being established to cope with the large quantity of data received from each extraction. Automated processes to manage the data are gradually being put in place.

Improvements in data quality are being tackled at two levels: in the data warehouse and at the practice. At the data warehouse level we are taking an ontological approach by defining the relevant concepts and relationships. Initially, the concepts and relationships are for diabetes. This will assist in identifying all information in the warehouse related to a particular concept. For example, to identify diabetes patients may involve looking for information in diagnoses, pathology (HbA1c), prescriptions for diabetes related medications and referrals to diabetes educators.

Reports back to the practices aim to improve data entry where it is incomplete or not consistent or correct. For instance, to improve the identification of patients in the registry the practice may decide to use consistent terminology or coding within the practice and ensure diagnoses are entered into the structured fields rather than free text.

The poor completion rates of diabetes management records that we found may be due to a number of factors. The qualitative findings suggest that GPs don't always enter patient information into structured fields. Clinical information may be entered as text in clinical notes which are not picked up in the extraction. GPs are also not the only providers of care. Patients may be referred for diabetes related tests to specialists rather than being organised by the GP. Understanding the true reasons for poor data quality is

complex. Having a community-based research network may assist in answering some of these questions.

The ability to link patient records between practices and health services provides a more comprehensive picture of patient care and use of health services. We found that despite being up to 40km apart, patients were found to have attended 2 or more of the 4 practices used in this pilot study. Further study and examination of the data extracted from practices in spatio-geographical network is required.

4. Conclusion

Electronic health records are increasingly being used for secondary research and evaluation purposes, beyond the primary purpose of supporting clinical care. The data in these EHRs must be of sufficient quality to support these purposes. However, there is little systematic research into the quality of routinely collected data in EHRs and the disease registers created from them. This ePBRN research program aims to address this gap systematically.

References

- [1] Glasgow N, Zwar N, Harris M, Hasan I, Jowsey Y. In: Nolte E, Knai C, McKee M, editors. Managing chronic conditions : experience in eight countries. Copenhagen: World Health Organization on behalf of the European Observatory on Health Systems and Policies; 2008. p. 60 -131.
- [2] Commonwealth of Australia. Primary Health Care Reform in Australia. Report to Support Australia's First National Primary Health Care Strategy. Canberra: Australian Government; 2009.
- [3] National Health & Hospital Reform Commission. A Healthier Future For All Australians – Final Report of the National Health and Hospitals Reform Commission – June 2009. In: Ageing DoHa, editor. Canberra: Commonwealth of Australia; 2009.
- [4] National Preventative Health Taskforce. Australia: The Healthiest Country by 2020 – National Preventative Health Strategy – Overview. Canberra: Commonwealth of Australia Department of Health and Ageing 2009 20 June 2009. Report No.: Publications Approval Number- P3-5457 Contract No.: ISBN: 1-74186-925-0.
- [5] Garling P. Final Report of the Special Commission of Inquiry: Acute Care in NSW Public Hospitals, 2008 - Overview. 27 November 2008 ed. Sydney: NSW Government; 2008.
- [6] Runciman W, Webb R, Helps S, Thomas E, Sexton B, Studdert D, et al. A comparison of iatrogenic injury studies in Australia and the USA. II: Reviewer behaviour and quality of care. *Int J Quality in Health Care*. 2000;12(5):379-88.
- [7] Thomas E, Studdert D, Runciman W, Webb R, Sexton E, Wilson R, et al. A comparison of iatrogenic injury studies in Australia and the USA. I: Context, methods, casemix, population, patient and hospital characteristics. *Int J Quality in Health Care*. 2000;12(5):371-8.
- [8] Vincent C, Neale G, Woloshynowych M. Adverse events in British hospitals: preliminary retrospective record review. *British Medical Journal* 2001;322(3 March):517-9.
- [9] Bates D, Gawande A. Improving safety with information technology. *NEJM*. 2003;348:2526-34.
- [10] Muttitt S, Alvarez R. Chronic disease management: it's time for transformational change! . *Healthc Paper*. 2007;7(4):43-7; discussion 68-70.
- [11] Bodenheimer T, Wagner E, Grumbach K. Improving Primary Care for Patients with Chronic Illness. *JAMA*. 2002 Oct 9;288(14):1775-9.
- [12] Bodenheimer T, Wagner EH, Grumbach K. Improving Primary Care for Patients With Chronic Illness: The Chronic Care Model, Part 2. *JAMA*. 2002 October 16;288(15):1909-14.
- [13] Hyrich K, Symmons D, Watson K, Silman A, Consortium BCC. Baseline comorbidity levels in biologic and standard DMARD treated patients with rheumatoid arthritis: results from a national patient register. *Annals of the Rheumatic Diseases*. 2006 July 1, 2006;65(7):895-8.
- [14] Cheales N, Howitt A. Personal experience with a district diabetic register located in general practice. London: HMSO 1996.

- [15] Jacobsen AF, Skjeldestad FE, Sandset PM. Incidence and risk patterns of venous thromboembolism in pregnancy and puerperium—a register-based case-control study. *Am J Obstet & Gynecology*. 2008;198(2):233.e1-e7.
- [16] Fink P. Physical disorders associated with mental illness. A register investigation. *Psychological Medicine*. 1990;20(04):829-34.
- [17] Espehaug B, Havelin LI, Engesaeter LB, Langeland N, Vollset SE. Patient-related risk factors for early revision of total hip replacements: A population register-based case-control study of 674 revised hips. *Acta Orthopaedica*. 1997;68(3):207-15.
- [18] Mors O, Mortensen PB, Ewald H. A population-based register study of the association between schizophrenia and rheumatoid arthritis. *Schizophrenia research*. 1999;40(1):67-74.
- [19] Carstensen B, Kristensen J, Ottosen P, Borch-Johnsen K, Register obotsgotND. The Danish National Diabetes Register: trends in incidence, prevalence and mortality. *Diabetologia*. 2008;51(12):2187-96.
- [20] Devantier A, Kjer JJ. The national patient register—a research tool? *Ugeskrift for laeger*. 1991;153(7):516-7.
- [21] Wang R, Strong D, Guarascio L. Beyond accuracy: what data quality means to data consumers. *J Management Information Systems*. 1996;12(4):5-33.
- [22] Liaw S, Chen H, Maneze D, Taggart J, Dennis S, Vagholkar S, et al. Health reform: is current electronic information fit for purpose? *Emergency Medicine Australasia*. 2011 2011 (Sep).
- [23] Liaw S, Taggart J, Dennis S, Yeo A. Data quality and fitness for purpose of routinely collected data – a case study from an electronic Practice-Based Research Network (ePBRN). American Medical Informatics Association Annual Symposium 2011; Washington DC: Springer Verlag; 2011.
- [24] Azaouagh A, Stausberg J. [Frequency of hospital-acquired pneumonia—comparison between electronic and paper-based patient records]. *Pneumologie*. 2008 May;62(5):273-8.
- [25] Mitchell J, Westerduin F. Emergency department information system diagnosis: how accurate is it? *Emerg Med J*. 2008 November;25(11):784.
- [26] Moro ML, Morsillo F. Can hospital discharge diagnoses be used for surveillance of surgical-site infections? *J Hosp Infect*. 2004 Mar;56(3):239-41.
- [27] de Lusignan S, Khunti K, Belsey J, Hattersley A, van Vlymen J, Gallagher H, et al. A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: a pilot and validation study of routinely collected data. *Diabet Med*. 2010;27:203-9.
- [28] Liaw S, Chen H, Maneze D, Taggart J, Dennis S, Vagholkar S, et al. Health reform: is current electronic information fit for purpose? *Emergency Medicine Australasia*. 2011 2011 (Sep).
- [29] Liaw S, Taggart J, Dennis S, Yeo A. Data quality and fitness for purpose of routinely collected data – a case study from an electronic Practice-Based Research Network (ePBRN). American Medical Informatics Association Annual Symposium 2011; Washington DC: Springer Verlag; 2011.
- [30] Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. *BMJ*. 2003 May 17;326(7398):1070.
- [31] Michalakidis G, Kumarapeli P, Ring A, van Vlymen J, Krause P, de Lusignan S. A system for solution-orientated reporting of errors associated with the extraction of routinely collected clinical data for research and quality improvement. *Stud Health Technol Inform*. 2010;160(Pt 1):724-8.
- [32] Ganguly P, Ray P, Parameswaran N. Semantic Interoperability in Telemedicine through Ontology-Driven Services. *Telemedicine & e-Health*. 2005;11(3):405-412.
- [33] Liaw ST, Taggart J, Dennis S, Yeo AET. Data quality and fitness for purpose of routinely collected data – a case study from an electronic Practice-Based Research Network (ePBRN). Proceedings of the AMIA Annual Symposium 2011, Washington DC 2011:785-94. Epub 2011 Oct 22.

APPENDIX 4

Using Ontologies to Identify Patients with Diabetes in Electronic Health Records

Hairong Yu, Siaw-Teng Liaw, Jane Taggart, and Alireza Rahimi Khorzoughi

School of Public Health & Community Medicine and Research Centre for Primary Health Care & Equity, Faculty of Medicine, University of New South Wales, Sydney, NSW 2052, Australia

{hairong.yu, siaw, j.taggart}@unsw.edu.au
alireza.rahimikhorzoughi@student.unsw.edu.au

Abstract. This paper describes a work in progress that explores the applicability of ontologies to solve problems in the medical domain. We investigate whether it is feasible to use ontologies and ontology-based data access (OBDA) to automate common clinical tasks faced by general practitioners (GPs), which are labor-intensive and error prone in terms of relevant information retrieved from electronic health records (EHRs). Our study aims to improve the selection of diabetes patients for clinical trials or medical research. The biggest impediment to automating such clinical tasks is the essential bridging of the semantic gaps between existing patient data in EHRs, such as reasons for visit, chronic conditions and diagnoses, pathology tests and prescriptions stored in general practice EHRs (GPEHR), and the ways which medical researchers or GPs interpret those records. Our current understanding is that automated identification of diabetes patients can be specified systematically as a solution supported by semantic retrieval. We detail the challenges to building a realistic case study, which consists of solving issues related to conceptualization of data and domain context, integration of different datasets, ontology creation based on the SNOMED CT-AU® standard, mapping between existing data and ontology, and the challenge of data fitness for research use. Our prototype is based on data which scale to thirteen years of approximately 100,000 anonymous patient records from four general practices in south western Sydney.

Keywords: Ontology, Diabetes Mellitus, Electronic Health Records, eHealth, Knowledgebase Management, Ontology-Based Database Access

1 Introduction

This paper reports on work that explores the applicability of ontologies for solutions in the health domain. In Australia, the main health applications of ontologies appear to be the SNOMED terminology services. We investigated the feasibility of the use of ontologies and OBDA to automate clinical tasks, such as identifying patients with specific diabetes mellitus (DM) phenotypes in EHRs contributing to the data repository of the electronic Practice Based Research Network (ePBRN). The ePBRN is used

ePBRN, p. 1, 2011.

© Springer-Verlag Berlin Heidelberg 2011

to conduct translational research on primary and integrated care, including tracking patients, managing chronic disease, and providing quality evidence-based care. The biggest barrier to automating the clinical task of identifying a patient with DM is the semantic gaps between patient data in EHRs, such as reasons for visit, diagnoses, pathology tests and prescriptions, and how these EHR data are interpreted. For example, in addition to a diagnostic label, DM can be implied by a blood glucose test with suggestive levels of diabetes, certain medications such as oral hypoglycaemics or insulin, or the use of DM supplies such as glucose diagnostic strips. By using ontologies, our experiments show that it is possible to automate this interpretation process and build a reusable conceptual infrastructure over diverse standards or experience or datasets. Currently most efforts at automation is only limited within individual clinics or in a physician-driven process or at data levels.

The SNOMED CT-AU®, the Australian extension to SNOMED CT® (Systematized Nomenclature Of Medicine Clinical Terms), is an ontology which formally defines classes of medical procedure, pharmaceutical or biologic product, and body structure and so on. The SNOMED CT-AU® Ontology (SCAO) is the reference terminology for EHRs in Australia. SCAO is available in Web Ontology Language (OWL) format from the Australian National E-Health Transition Authority (NEHTA). Our experiments showed that the integration of SNOMED CT-AU and the Diabetes Identification Ontology (DIO) based on ePBRN data to select patients with DM is well suited for our case study. Our key approach is that the automation of the process of identifying DM patients is an issue of semantic retrieval, i.e. selection criteria can be expressed as semantic queries, which are processed by a reasoner to retrieve explicit information on eligible patients from datasets and infer implicit knowledge from ontologies simultaneously.

The objective of this study is to assess the practicality and utility of ontologies in a real world environment. The technical challenges of conceptualization of data and domain context, ontology integration of different datasets or ontologies, mapping between existing datasets and ontologies, and finding solutions to ensure data fitness for clinical or research use will be described and discussed in the following sections.

2 Methodology

The architecture for this study comprises six parts separated by dashed lines as shown in Figure 1. Patient data were extracted from individual GPEHRs, e.g. Medical Director™¹ at each clinic by GRHANITE™². The software provides a data repository over server called GRHANITE™ Databank, in our case the ePBRN repository operated by MS SQL Server™. The ABox, associated with instances of ontology classes or properties, is populated through ontopPro (formerly known as an OBDA plugin for Protégé³). Another primary component in our knowledgebase, the TBox, related to concep-

¹ <http://www.hcn.com.au/Products/Medical+Director>

² <http://www.grhanite.com/>

³ <http://protege.stanford.edu/>

tual terminologies defined in ontologies, is built through Protégé, a popular open source ontology editor and knowledgebase framework.

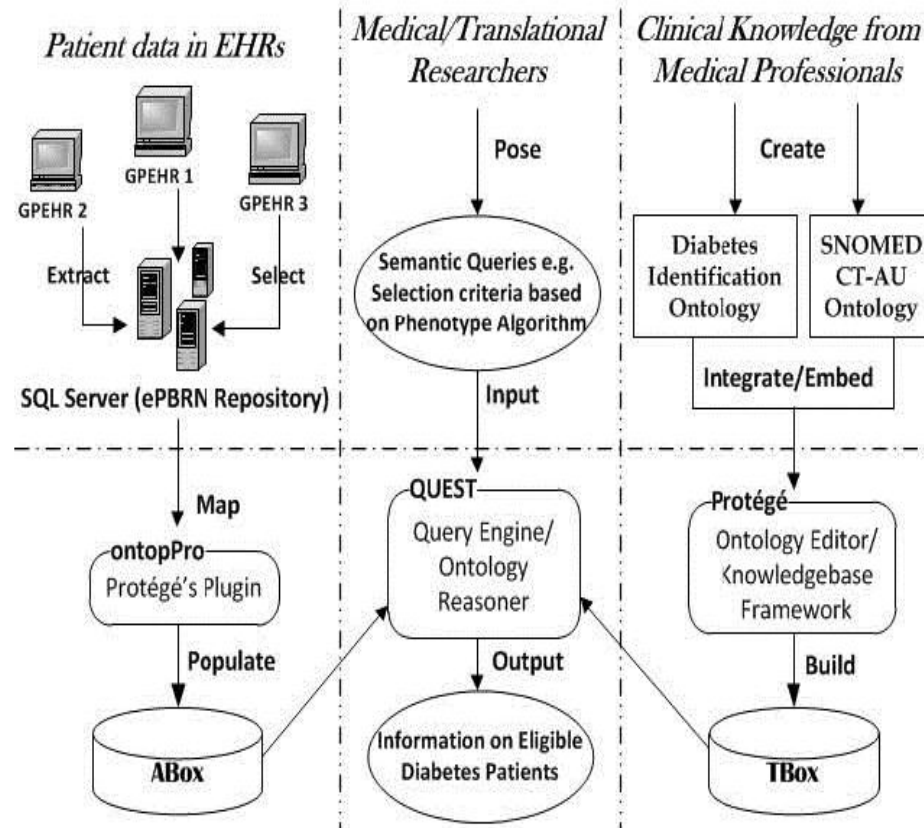


Fig. 1. ePBRN Diabetes Identification Case Study Solution Architecture

Clinical selection criteria are formulated as semantic queries in SPARQL Protocol and RDF Query Language (SPARQL). The SPARQL query engine QUEST⁴ that comes with ontopPro⁵ checks the queries against the knowledgebase to retrieve matched patients. A demonstration is given for each of six parts in Figure 1.

The first step in building this solution is creating the specific DIO in hierarchical conceptual modeling, based on the Australian National Guidelines for Type 2 Diabetes Mellitus (T2DM) and discussions with the research team and GPs participating in the ePBRN. The output of this first task is a formalized ontology which consists of 4 main classes Actor, Content, Mechanism and Impact and 68 subclasses with object/data properties. Some of them can be mapped to the SNOMED CT-AU Ontology (SCAO), which has more than 300,000 concepts.

Due to the small number of concepts captured by the DIO, the mapping can be operated manually. For example, T2DM is a Disease under the subclass of Problem which has a superclass Context in DIO. In the SCAO, T2DM is a disorder of glucose metabolism which is a subclass of Disease under the highest level concept of Clinical

⁴ <http://semanticweb.org/wiki/Quest/>

⁵ <http://ontop.inf.unibz.it/>

finding. Similarly, Actor class in DIO corresponds to Environment or Geographical location in SCAO. However the automation of integration of two ontologies can be complex for large terminologies.

Next we linked the server objects in SQL Server to integrate other heterogeneous datasets by T-SQL™. The SQL query results are mapped by ontopPro for ABox associated with relevant classes in ontologies. This meant that the schematic or semantic heterogeneity challenges faced were solved at either data or ontology level. The mapping mechanism supplied by ontopPro theoretically based on OBDA [1], provided a big advantage on populating class members, assigning property values, and incorporating schematic data in the ePBRN repository with semantic concepts in ontologies. The raw data in EHRs that contribute to the ePBRN repository are incomplete, incorrect and inconsistent (against external standards or internal logic perspectives). We used definitions of properties in DIO or mappings created in ontopPro to solve core data quality issues before preparation of semantic queries.

We then wrote semantic queries in SPARQL according to requirements from domain experts, and ran them through QUEST, the query engine and OWL reasoner. The query results are expected to identify DM patients and help clinicians to manage the cycle of care for the cohort. The SPARQL queries were validated using SQL over an artificial dataset of 100 patients schematically similar to the ePBRN dataset. The approach that we developed and tested on the artificial dataset will be scalable to the ePBRN repository of more than 100,000 patient records. Other use case scenarios, for example assisting researchers to conduct association and/or controlled studies will contribute to the validation of the architecture.

3 Discussion and Conclusion

We have briefly presented a feasibility study of the use of ontologies to detect patients with DM in real world EHRs. Using real patient datasets, we solved some engineering challenges around ontology creation and integration, bridging between ontologies and datasets, and data quality [2]. Apart from usability, interoperability and scalability aforementioned, other quality attributers are assessed closely for architecture evaluation for instance, modifiability with many facades/locations where data/data types are transferred in our solution, integrability and extensibility which are especially critical as several open source software components are used in our design.

References

1. M. Lenzerini, Data Integration: A Theoretical Perspective, Proc. of the 21st ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS'02), pp. 233 – 246.
2. S.-T. Liaw, et al. Data quality and fitness for purpose of routinely collected data – a general practice case study from an electronic Practice-Based Research Network, in: AMIA Annual Symposium Proceedings, 2011:785–794.