INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 83 (2014) 768–778

journal homepage: www.ijmijournal.com

Validating an ontology-based algorithm to identify patients with Type 2 Diabetes Mellitus in Electronic Health Records

Alireza Rahimi[a,b,d], *Siaw-Teng Liaw*[a,c,e,*], *Jane Taggart*[c], *Pradeep Ray*[d], *Hairong Yu*[c]

[a] *UNSW, School of Public Health & Community Medicine, Sydney, Australia*
[b] *Isfahan University of Medical Sciences, Health Information Research Centre, Isfahan, Iran*
[c] *UNSW, Centre for Primary Health Care & Equity, Sydney, Australia*
[d] *UNSW, Asia-Pacific Ubiquitous Healthcare Research Centre, Sydney, Australia*
[e] *General Practice Unit, South Western Sydney Local Health District*

ARTICLE INFO

Article history:
Received in revised form 2 June 2014
Accepted 2 June 2014

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

ABSTRACT

Background: Improving healthcare for people with chronic conditions requires clinical information systems that support integrated care and information exchange, emphasizing 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 were 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 measured, 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 or correct 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 incorrect data due to mistaken units of measurement and unavailable data due to non-documentation or documented in the wrong place or progress notes, problems with data extraction, encryption and data management errors.

* *Corresponding author at*: General Practice Unit, Fairfield Hospital, P.O. Box 5, Fairfield, NSW 1860, Australia. Tel.: +61 2 96168520; mobile: +61 407818281.
E-mail address: siaw@unsw.edu.au (S.-T. Liaw).
http://dx.doi.org/10.1016/j.ijmedinf.2014.06.002

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 incorrect data. The extent of compromise requires further study, using ontology-based and other approaches.

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 [1]. It also can lead to better policy that actually meet needs, strategies, evidence-based care and patient outcomes in chronic disease management (CDM) [2].

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" [3], drawing on the general definition by Gruber: "an explicit, formal specification of a shared conceptualization" [4]. 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 [5,6]; information retrieval, DQ management [7], data collection, data sharing and data integration [8,9] in clinical information systems (CIS) for CDM; and regular validation of key data items in clinical data warehouses (CDW) [10,11]. 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 [12]. 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 [13].

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 [14,15]. 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 [11,16,17].

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 [18]. Ontology-based semantic methodologies, formalized 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 [19], leading to more accurate decisions [20–22].

1.1. *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 [23]. 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 [23].

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