e-Health – For Continuity of Care
C. Lovis et al. (Eds.)

doi:10.3233/978-1-61499-432-9-171

EHR-based disease registries to support integrated care in a health neighbourhood: an ontology-based methodology

Siaw-Teng LIAW[a,b,c], Jane TAGGART[a], Hairong YU[a]
[a]*UNSW Medicine Australia;* [b]*SW Sydney Local Health District, Australia,* [c]*Ingham Institute of Applied Medical Research*

Abstract. Disease registries derived from Electronic Health Records (EHRs) are widely used for chronic disease management. We approached registries from the perspective of integrated care in a health neighbourhood, considering data quality issues such as semantic interoperability (consistency), accuracy, completeness and duplication. Our proposition is that a realist ontological approach is required to accurately identify patients in an EHR or data repository, assess data quality and fitness for use by the multidisciplinary integrated care team. We report on this approach with routinely collected data in a practice based research network in Australia.

Keywords. EHR, patient registries, data quality, routinely collected data, data repository, health neighbourhood, integrated care

Introduction

Disease registries derived from Electronic Health Records (EHR) are widely used for chronic disease management (CDM). However, not enough is known about the quality of EHR-based registers in the UK [1] and Australia [2]; even less is known about whether they improved CDM, patient safety or quality outcomes. Usually created through "blackbox" extraction tools, increasing use of registries for clinical care can increase the likelihood and scope of data errors and adverse events [2]. The design and development of EHR-based disease registries is not transparent [3] and aspects of their quality have been examined in the UK [4] and Australia [5]. Our proposition [6] is that a realist [7] and ontological [8] approach is required to systematically and accurately identify patients in an EHR or data repository [9], and assess/manage data quality and fitness for use by the multidisciplinary care or research team [5].

The realist approach to this evolving yet complex domain includes: **1. Context (**i.e. CDM, integrated care, evidence based practice); **2. Mechanisms** (i.e. methods to assess/manage data quality (DQ) and support data, knowledge, clinical and interdisciplinary integration); and **3. Impacts/outcomes** (i.e. DQ and fitness for use of disease registries, and, over the long term, safety and quality of integrated care).

The ontological approach includes the collection of formal, machine-processable and human-interpretable representations of the entities and their relations within a defined domain [10]. A formal ontological model of the domain data and metadata can specify a unified context, enabling intelligent software agents to act in spite of

differences in concepts and terminology from different EHRs. By incorporating defined rules, ontologies can generate logical inferences and control the inclusion/exclusion of relevant objects, such as the patient with a diagnosis of diabetes mellitus (DM) Reason For Visit (RFV), pathology (Path) tests, medication (Rx), or cycle of care service payments [10]. We have summarized this realist ontological approach to automated assessment/management of the quality of routinely collected data, relevant concepts and their relationships in integrated care [11].

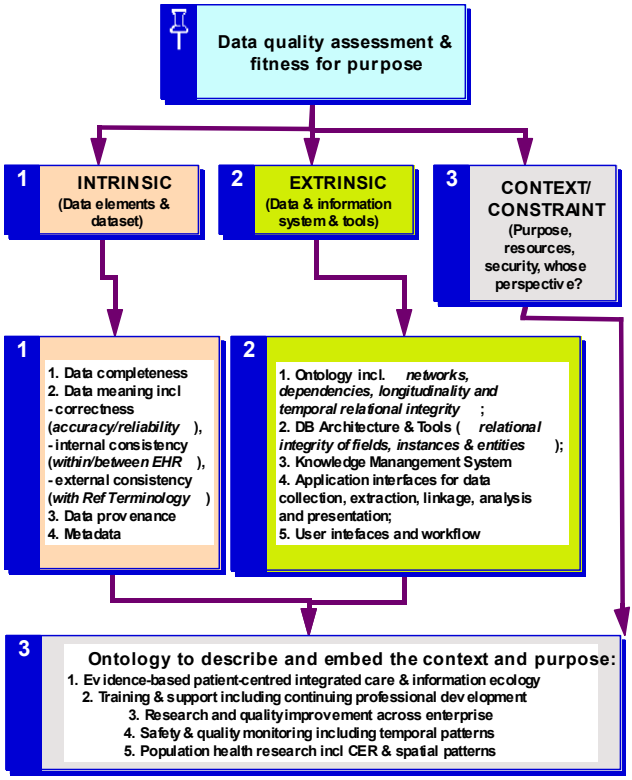

Figure 1. Data quality & fitness for purpose framework.

The International Standards Organisation (ISO) defined data quality (DQ) as: "*the totality of features & characteristics of an entity that bears on its ability to satisfy stated and implied needs*" (ISO 8402-1986, Quality Vocabulary). Fitness for purpose (FFP) is a multidimensional concept with intrinsic components and extrinsic associations to meet benchmark [12]. The literature [11], guided the development of a conceptual framework for DQ and FFP (Figure 1), with intrinsic, extrinsic and contextual dimensions. Intrinsic concepts cover the data elements and dataset, including the metadata, semantics (data meaning), provenance (who authored, where, when?) and constraints to the data meanings. Extrinsic concepts cover the information system, including concept representation, ontology, temporal relationships system architecture and user interface. Contextual determinants include the objectives of stakeholders such as the integrated care practitioner/team, resource constraints, security requirements and legislation.

Data elements are assessed intrinsically in terms of consistency, correctness; data sets in terms of completeness and duplicate records [6]. We are developing ontology-based tools to assess the information required to support integrated care in terms of relational, historical and temporal integrity between concepts. Contextual determinants constrain and guide clinical and organizational strategies to improve DQ as well as unify different concepts and terminology from different EHRs. The formal process of ontology development includes knowledge acquisition, conceptualisation, semantic modelling, knowledge representation and validation [13]. A layered approach [14] is used to incorporate clinical guidelines and rule-based modules. This paper discusses

the realist ontological approach to developing automated, valid and reliable methods to define *T2DM cases*, manage DQ and determine fitness for purpose.

1. Methods

Setting: The UNSW electronic Practice Based Research Network (ePBRN) pilot group of 4 general practices (N=64,770 patients) has validated tools, data, and management and governance protocols. Internal validation included regular data and metadata checks, including probabilistic matching to assess the extent of duplicate patients and patients shared within the local health neighbourhood of hospital, community health, general practice and other primary care services. External validation involved comparisons with other tools [4]. The overall aim is to transition from traditional SQL and schematic relational database management of "big data" to use of ontologies and semantic tools with minimally-relational databases using.

DM ontology for case-finding: Defined rules were used to generate logical inferences and control the inclusion/exclusion of patients with a DM RFV; DM Pathology (Path) such as HbA1C, glucose tolerance test; DM medication (Rx) or glucose testing scripts, or a DM cycle of care item in the Medicare Benefit Schedule (MBS). Duplicate records/patients within and across EHRs were excluded.

DQ ontology: The conceptualization and specification of the DQ ontology (Figure 1) included core dimensions such as accuracy, completeness, correctness, consistency and timeliness [6], duplicate records (to account for aggregating multiple EHRs), temporal pattern (to account for the constantly changing clinical "big data") and timeliness which is important in integrated care. Validation of the conceptualization included discussions with practitioners and consumers of health care.

Formalisation: Tools used included a popular open source ontology editor and knowledgebase framework (Protégé, http://protege.stanford.edu/); reference terminology (SNOMED-CT-Au); representation languages (Web Ontology Language (OWL) http://www.w3.org/TR/owl-features/), XML and RDF (Resource Description Framework)); query languages (SPARQL Protocol and RDF Query Language); rules languages (Semantic Web Rule Language (SWRL)); logic ontology reasoners to provide automated support for reasoning tasks in ontology and instance checking through -ontopPro- (http://ontop.inf.unibz.it/), an ontology based data access (OBDA) application [15]. Patient data, associated with instances of ontology classes or properties, were populated through -ontopPro-. The knowledge component of the infrastructure, relating to defined conceptual terminologies, was built using SNOMED CT-AU and OWL. The RDF schema is mapped to logics to support formal semantics and reasoning [16], which describes precisely the meaning of specific knowledge to minimise subjective intuitions and different interpretations by different actors or machines [14].

Implementation used Microsoft SQL Server and Transact-SQL™ to link server objects with heterogeneous datasets from multiple EHRs [16]. A manual validation of the results of the SQL queries was conducted with the smallest participating ePBRN practice (Practice 1).

2. Results

Ontological approach to find cases for a diabetes registry: A lower than expected T2DM prevalence rate of 2.8% was found. Table 1 shows the effect of data completeness of relevant indicators (RFV, Rx, Path) on the accuracy of identification. The ontological (1.1-5.7% across practices) was more sensitive than the single factor (0.2-4.8%) approach, compensating for individual data incompleteness and inconsistency. Data quality is therefore a significant factor. The denominator was also important as patients may not be accurately flagged as active and inactive in the EHRs.

Table 1. T2DM cases by Reason for Visit (RFV), Medication (Rx), Pathology (Path) and ontology approach

Attributes studied	**Practice1 (N=3 863)**	**Practice2 (N=7 028)**	**Practice3 (N=2 3,162)**	**Practice4 (N=3 0,717)**	**All practices (N=6 4,770)**
Data Completeness of :	**All (DM only)**	**All (DM only)**	**All (DM only)**	**All (DM only)**	**All (DM only)**
All RFV (All DM RFV)	95% (4.3%)	87% (5.7%)	92% (4.9%)	99% (6.5%)	95% (5.8%)
All Rx (All DM Rx)	80% (2.4%)	94% (8.4%)	96% (5.4%)	96% (6.6%)	95% (6.4%)
All Path (All DM Path)	16% (0.8%)	61% (8.0%)	63% (1.3%)	66% (1.5%)	62% 2.4%)
All 3 (RFV+Rx+Path)	82%	90%	90%	92%	90%
T2DM identified by:	**N (%)**	**N (%)**	**N (%)**	**N (%)**	**N (%)**
T2DM RFV	37 (0.9)	231 (3.3)	387 (1.4)	787 (2.6)	1,442 (2.2)
T2DM Rx	19 (0.5)	332 (4.7)	446 (1.9)	803 (2.6)	1,600 (2.5)
T2DM Path	8 (0.2)	334 (4.8)	468 (2.0)	809 (2.6)	1,619 (2.5)
T2DM ontology	43 (1.1)	403 (5.7)	602 (2.5)	1,042 (3.4)	2,090 (3.2)

Duplication and other dimensions of data quality: Up to 13% patient records matched across the participating EHRs, suggesting that DQ assessment should include the extent of duplication of data across the health neighbourhood and within practices, where there can be up to 3% duplication. This has implications for clinical use of EHR data in integrated and shared care as well as secondary uses for research, population health and policy guidance.

Specifying and formalising the ontological approach: The formal specification of the ontologies developed is available as Protégé files. The ontology was validated in Practice 1, using –ontopPro- to map to the relational database and implement the built-in reasoners. This technical aspect is the subject of another paper, which will also compare the utility and validity of SQL-based schematic and inductive versus ontology-based semantic approaches and tools.

3. Discussion

Ontologies deal with reality (*being*) and the transformation (*becoming*) of concepts as they interact with one another over time. The ePBRN research confirmed the need for a realist and ontological approach to the quality of routinely collected data in EHRs and EHR-based disease registries - to understand what is being done in what context and with what impact. This is important, given that the processes and knowledge base are continually evolving and require ongoing monitoring, evaluation and reflection. The research and development must be grounded in the real world of health practice, where data is noisy and continually changing and the DQ is variable. DQ management of

information from multiple EHRs to support integrated care and population health must exclude duplicated records.

An ontological approach to the creation of patient registries from EHRs is essential to optimise accuracy (10). The quality of the disease registry is only as good as the EHR from which it is created. Improved DQ require integrated and ecological approaches to the governance and provenance of DQ across the data cycle from collection to management to display and secondary use in other applications such as electronic decision support [17]. The quality of electronic data collected as part of routine clinical practice is determined by more than just the GIGO – garbage in garbage out - principle. Data models are influenced by the database management system, security and access management software, processes for data collection and management, and the people who enter and use data [4].

The validated ontologies and software tools will support automated methods to extract, link and manage data as well as assess/manage the data quality and semantic interoperability challenges in various semantic contexts in collaborative "big clinical data" environments. The challenges are surmountable and strategies sustainable.

References

[1] de Lusignan S, Sadek N, Mulnier H, Tahir A, Russell-Jones D, Khunti K. Miscoding, misclassification and misdiagnosis of diabetes in primary care. Diabet Med. 2012 Feb; 29(2): 181-9.

[2] Liaw S, Taggart J, Yu H, de Lusignan S. Data extraction from electronic health records – existing tools may be unreliable and potentially unsafe. Aust Fam Physician. 2013; 42(11): 820-3.

[3] Mehta A. The how (and why) of disease registers. Early Human Development. 2010;86(11):723-8.

[4] Martin D, Wright J. Disease prevalence in the English population: a comparison of primary care registers and prevalence models. Soc Sci & Med 2009; 68(2): 266-74.

[5] Liaw S, Taggart J, Dennis S, Yeo A. Data quality and fitness for purpose of routinely collected data – a case study from an ePBRN. AMIA Annual Symposium 2011; Washington DC: Springer Verlag; 2011.

[6] Liyanage H, Liaw S, Kuziemsky C, de Lusignan S. Ontologies to improve chronic disease management research and quality improvement studies – a conceptual framework. In: Aronsky D, Leong S, editors. Medinfo 2013; Copenhagen: Elsevier Press; 2013.

[7] Pawson R, Greenhalgh T, Harvey G, Walshe K. Realist review - a new method of systematic review designed for complex policy interventions. J Health Serv Res Policy. 2005;10(Suppl 1):21-34.

[8] Gruber TR. Toward principles for the design of ontologies used for knowledge sharing. Int J Human-Comput Stud. 1995;43(5-6).

[9] de Lusignan S, Liaw S, Michalakidis G, Jones S. Defining data sets and creating data dictionaries for quality improvement and research in chronic disease using routinely collected data: an ontology driven approach BCS Informatics in Primary Care. 2011;19(3):127-34(8).

[10] Rubin D, Lewis S, Mungall C, et al. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. OMICS (Summer). 2006; 10(2): 185-98.

[11] Liaw S, Rahimi A, Ray P, et al. Towards an ontology for data quality in integrated chronic disease: a realist review of the literature. Int J Med Inform 2013; 82(1): 10–24.

[12] Wang RY. A product perspective on total data quality management. Communications of the ACM. 1998; 41(2):58-65.

[13] Kuziemsky C, Lau F. A four stage approach for ontology-based health information system design. Artificial Intelligence in Medicine 2010; 50: 18.

[14] Chalortham N, Buranarach M, Supnithi T. Ontology Development for T2DM Clinical Support System 2009 3 March 2011. URL: http://text.hlt.nectec.or.th/ontology/sites/default/files/CRdm2css_0.pdf.

[15] Rodríguez-Muro M, Calvanese D. Quest, a System for Ontology Based Data Access. KRDB Research Centre for Knowledge and Data, Free University of Bozen-Bolzano, 2012.

[16] Yu H, Liaw S, Taggart J, Rahimi A. Using Ontologies to Identify Patients with Diabetes in EHRs. Int Semantic Web Conference 2013; Sydney Australia: Springer-Verlag Berlin Heidelberg.

[17] de Lusignan S, Liaw S, Krause P, et al. Key concepts to assess the readiness of data for International research: Data quality, lineage and provenance, extraction and processing errors, traceability, and curation. IMIA Yearbook of Medical Informatics. 2011: 112-21