doi:10.3233/978-1-61499-921-8-65

Ontologies in Big Health Data Analytics: Application to Routine Clinical Data

Harshana LIYANAGE[a] , John WILLIAMS[a], Rachel BYFORD[a],
Lampros STERGIOULAS[b] and Simon de LUSIGNAN[a,1]
[a]*Department of Clinical & Experimental Medicine, University of Surrey, UK*
[b]*Surrey Business School, University of Surrey, UK*

Abstract. Ontologies are an important big-data analytics tool. Historically code lists were created by domain experts and mapped between different coding systems. Ontologies allow us to develop better representations of clinical concepts, data and facilitate better data extracts from routine clinical data. It also makes the process of case identification and key outcome measures transparent. We describe a process we have operationalised in our research. We use ontologies to resolve the semantics of complex health care data. The use of the method is demonstrated through a pregnancy case identification method. Pregnancy data are recorded in different coding systems and stored in different general practice systems; and pregnancy has its own complexities in that not all pregnancies proceed to term, they have different lengths and involve multiple providers of health care.

Keywords. Medical Record systems, computerized; Biomedical Ontologies, Information Storage and Retrieval; Controlled Vocabulary

1. Introduction

Routine clinical data are commonly recorded using clinical codes (e.g. International Classification of Diseases-10 (ICD-10), Read Clinical Terms version 3 (CTv3) [1] etc.) The heterogeneity between these coding systems introduces complexity when extracting data for research. Historically, experts have produced coding lists to select cases or define study outcomes. These and their associated conclusions can be open to challenge [2].

Biomedical ontologies provide the opportunity to make these processes transparent, with scope to publish the ontology online for others to use and modify. We advocate their use for key variables: case identification and main outcome measures because it makes the process more transparent and reproducible. By working from concepts it avoids the potential hazards associated with one-to-one mapping of codes, which has the inherent weakness where different coding systems have varying scope; e.g. ICD-10 disease only; CTv3 contains drug codes that might imply a diagnosis (e.g. prescribed insulin, and “Attended annual diabetes review” both imply a diagnosis of diabetes).

Our three-step ontological process includes the formal use of Web Ontology Language (OWL) and publication online on a standard biomedical ontology repository such as BioPortal;[2] we first used it to identify diabetes cases (Table 1) [3].

[1] Corresponding Author, Simon de Lusignan, Department of Clinical & Experimental Medicine, University of Surrey, GUILDFORD, Surrey, GU2 7XH, UK; Email: s.lusignan@surrey.ac.uk.

[2] https://bioportal.bioontology.org/

2. Methods

2.1. Three-step ontological process

Our ontological process is typically initiated by clinicians who identify the essential concepts associated with case definition. In the second step this ontology is mapped to one or more coding systems to be used in the study. In the third step, the ontology is annotated with the clinical codes is used to extract the required data.

Table 1. A three-step ontological process identifying a case from routine data

Steps in the ontological process	Objective	Examples	Actors involved with the process
Ontological layer	Organisation of concepts within the domain and their relationships	Diagnostic criteria Symptom & examination findings Pathology and other test results Therapies and other treatments	Clinicians, Epidemiologists
Coding layer	Mapping to coding systems	Created in relevant coding systems	Clinicians, Researchers
Logical query layer	Implementing data extraction based on nuances of the system	Test extract Results feedback into ontological layer	Researchers, Database developers

The ontological process is particularly useful in situations where routine data analysis will involve extracting semantically similar data from heterogeneous data sources. Figure 1 illustrates this scenario where the ontology (describing the data of interest) and the logical query (algorithmic specification of how data should be extracted) is applied to two realities. For example, a study of diabetes in two countries would start with a common ontology describing a diabetes case and a common algorithm to identify various diabetes types. This would be annotated with their coding systems and the algorithm to extract the data would then be implemented.

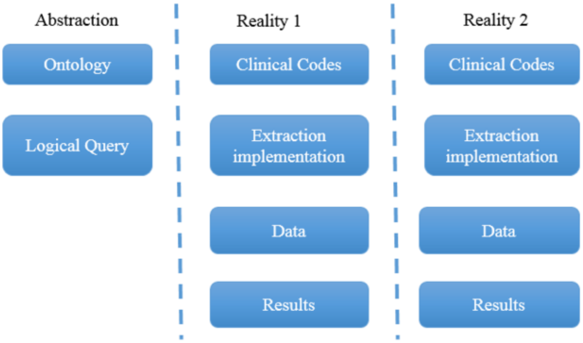

Figure 1. Abstraction of data at different levels of granularity being applied within two different realities (e.g. primary care/ secondary care setting). The ontology could be mapped to different coding systems in the two realities. The extraction process for the two settings maybe different, but the results would be consistent with the ontology.

2.1.1. Ontological layer

An ontology captures concepts and relationships within the domain of interest. This step is carried out independent of the clinical coding system(s). Our ontologies are implemented using Web Ontology Language (OWL) in the Protégé ontology development environment and shared through the BioPortal web repostiory. We recommended alignment with the upper level ontology Basic Formal Ontology.

When specifying an ontology for identifying cases, we include multiple case definitions that specify the certainty with a case can be defined [4].

- **Definite:** case ascertainment with a high degree of certainty (e.g. using concepts related to diagnosis)
- **Probable:** case ascertainment with a moderate degree of certainty (e.g. using concepts related to a pattern of symptoms and signs)
- **Possible:** case ascertainment with a low degree of certainty (e.g. using concepts related to lab tests without clear indication of result)

2.1.2. Coding layer

In the second step, the concepts in the ontology are mapped to corresponding codes in coding systems used by the data sources participating in the study. In the UK, different coding and classifications systems are implemented by the various computerized medical records (CMR) system vendors. These include Read Version 2, CTv3 and SNOMED CT classifications. The codes chosen may not be semantically be equivalent to the ontological concept. Hence mapping is classified using one of three levels of equivalence [5].

- **Directly mapping:** concept can be directly mapped to specific code(s)
- **Partial mapping:** concept can be mapped to a code in the coding system which is incompletely or partially representative
- **No clear mapping:** concept cannot be mapped to any code(s)

A single concept in the ontology can have one or more annotations.

2.1.3. Logical query layer

The final step is involves modifying the ontology to accommodate coding variations of the target data source. The existence of code in a coding system does not guarantee that code will be recorded in a clinical database. In addition there can be errors in recording clinical data in CMRs. The following three errors are frequently observed [6]:

- **Miscoding:** code selected which lack specificity for concept. Other data may allow more specific definition (e.g. pregnancy test done).
- **Misclassification:** coding indicates the right concept but had incorrect detail (e.g. singleton pregnancy when the patient has twins).
- **Misdiagnosis:** coding indicates an indirect
- **No coding** (i.e. false negative)**:** could be due to: a) information not being known to anyone anywhere b) information not known to the provider but known elsewhere in the health system c) information known to provider but not coded entry recorded (e.g. free text entry or information in hospital letter)

2.2. Implementing the process for extracting routine data

This process was used to identify pregnancy in routine data. We used the Royal College of General Practitioners Research (RCGP) and Surveillance Centre (RSC) database [7].

The database contains >2 million patient records. The general practices in the network use four different CMR systems. Three use Read Version 2 and one CTV3 (The Phoenix Partnership – TPP). Pregnancy data might be recorded by general practitioners (GPs), midwives, or coded from hospital attendance; hospital data is generally ICD-10.

3. Results

We developed a "Pregnancy Ontology" consisting of concepts associated to pregnancy and pregnancy related to care. We validated a random sample of the results by examining their complete medical record. The challenges in pregnancy is that not all pregnancies run to term (codes need to be identified that mark the start (e.g. positive pregnancy test) and end of pregnancy (e.g. Caesarian section), many are involved in care, and data are heterogeneous. The concepts were identified from the literature and from experienced GPs (ontological layer). We identified codes and annotated the concepts in the ontology. The pregnancy ontology represents the abstraction, the two sets of codes represent two possible realities of the coding layer. We implemented a pregnancy case identification algorithm using Structured Query Language (SQL). The code sets were optimised by examining the frequency of code usage in the database (logical query layer). The pregnancy case identification algorithm was executed on the RCGP RSC database and results were analysed for the realities based on coding system (Table 2, Figure 2).

Table 2. No of pregnancies identified (in 2015) by applying the ontology to two different coding systems

	Read v2	CTV3
No of women in RCGP database	719436	48347
All pregnancies (% pregnant)	29123 (4.0%)	1554 (3.2%)
Delivered pregnancies (% delivered)	24663 (3.4%)	1487 (3.1%)

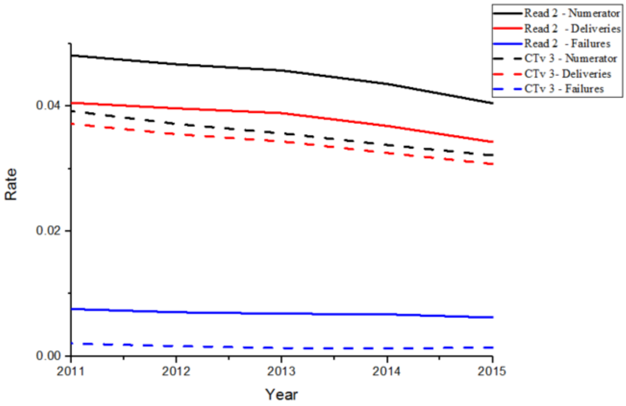

Figure 2. Rate of pregnancy outcomes (in relation to the denominator population) for the GP systems using the two coding systems during the period 2012 - 2015

4. Conclusion

Our ontological approach is a transparent process, running from the conceptual level. The key conceptual elements were placed in the public domain (online). Changes in criteria can be clearly recorded. Certainty about cases is also made clear: definite, probably and possible; and where cases are reclassified we have a clear taxonomy: miscoding (too vague a code – e.g. Caesarian section); misclassification (right illness or condition, wrong type - e.g. single pregnancy should be a twin); and misdiagnosis (where the diagnosis/outcome is incorrect – e.g. Emergency Caesarian section recorded when actually it was an elective section). Certainty about coding mapping to concepts is also classified by degree of certainty, directly or partially mapping or no clear mapping.

Pregnancy is an excellent and challenging example for developing an ontology because of the need to include a wide range of concepts and codes that might be associated with the start and end of pregnancy. They must exist within a valid interval, as "booking-in" may not occur until around 10 weeks with a 27-35 week window until delivery. There are also many concepts and codes that are associated with an early end of pregnancy (miscarriage and termination), as well as at the end of pregnancy (vaginal delivery by a midwife at home, or birth including by Caesarian section in hospital).

The ontological approach has decreased the semantic gap in research concepts used for extracting data from different representations of health data while increasing the traceability of ontological concepts from research question to outcome data. We advocate those working with big health data adopt an ontological approach, publishing their ontologies and using this methodological approach.

References

[1] S. de Lusignan, Codes, classifications, terminologies and nomenclatures: definition, development and application in practice, *Inform Prim Care* **13** (1) (2005), 65-70.

[2] S. de Lusignan, B. Sun, C. Pearce, C. Farmer, P. Steven, S. Jones, Coding errors in an analysis of the impact of pay-for-performance on the care for long-term cardiovascular disease: a case study, *Inform Prim Care* **21** (2) (2014), 92-101.

[3] S. de Lusignan, In this issue: Ontologies a key concept in informatics and key for open definitions of cases, exposures, and outcome measures, *J Innov Health Inform* **22** (2) (2015), 170.

[4] A.R. Sadek, J. van Vlymen, K. Khunti, S. de Lusignan, Automated identification of miscoded and misclassified cases of diabetes from computer records, *Diabet Med.* **29** (3) (2012), 410-414.

[5] S. de Lusignan, S. Shinneman, I. Yonova, J. van Vlymen, A.J. Elliot, F. Bolton, G.E. Smith, S. O'Brien, An Ontology to Improve Transparency in Case Definition and Increase Case Finding of Infectious Intestinal Disease: Database Study in English General Practice, *JMIR Med Inform* **5** (3) (2017), e34.

[6] S. de Lusignan, N. Sadek, H. Mulnier, A. Tahir, D. Russell-Jones, K. Khunti, Miscoding, misclassification and misdiagnosis of diabetes in primary care, *Diabet Med* **29** (2) (2012), 181-189.

[7] A. Correa, W. Hinton, A. McGovern, J. van Vlymen, I. Yonova, S. Jones, S. de Lusignan, Royal College of General Practitioners Research and Surveillance Centre (RCGP RSC) sentinel network: a cohort profile, *BMJ Open* **6** (4) (2016), e011092.