Journal of Biomedical Informatics 57 (2015) 350–357

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

An alternative database approach for management of SNOMED CT and improved patient data queries

W. Scott Campbell [a,*], Jay Pedersen [b], James C. McClay [c], Praveen Rao [d], Dhundy Bastola [b], James R. Campbell [e]

[a] University of Nebraska Medical Center, Department of Pathology and Microbiology, 985900 Nebraska Medical Center, Omaha, NE 68198-5900, United States
[b] University of Nebraska at Omaha, College of IS&T, PKI 170, 6001 Dodge Street, Omaha, NE 68182-0116, United States
[c] University of Nebraska Medical Center, Department of Emergency Medicine, 985900 Nebraska Medical Center, Omaha, NE 68198-5900, United States
[d] University of Missouri – Kansas City, Department of Computer Science and Electrical Engineering, 550H Robert Flarsheim Hall, 5100 Rockhill Road, Kansas City, MO 64110, United States
[e] University of Nebraska Medical Center, Department of Internal Medicine, 985900 Nebraska Medical Center, Omaha, NE 68198-5900, United States

ARTICLE INFO

Article history:
Received 29 April 2015
Revised 11 August 2015
Accepted 12 August 2015
Available online 21 August 2015

Keywords:
SNOMED CT
Medical terminology
Ontology
Databases

ABSTRACT

Objective: SNOMED CT is the international *lingua franca* of terminologies for human health. Based in Description Logics (DL), the terminology enables data queries that incorporate inferences between data elements, as well as, those relationships that are explicitly stated. However, the ontologic and polyhierarchical nature of the SNOMED CT concept model make it difficult to implement in its entirety within electronic health record systems that largely employ object oriented or relational database architectures. The result is a reduction of data richness, limitations of query capability and increased systems overhead. The hypothesis of this research was that a graph database (graph DB) architecture using SNOMED CT as the basis for the data model and subsequently modeling patient data upon the semantic core of SNOMED CT could exploit the full value of the terminology to enrich and support advanced data querying capability of patient data sets.

Methods: The hypothesis was tested by instantiating a graph DB with the fully classified SNOMED CT concept model. The graph DB instance was tested for integrity by calculating the transitive closure table for the SNOMED CT hierarchy and comparing the results with transitive closure tables created using current, validated methods. The graph DB was then populated with 461,171 anonymized patient record fragments and over 2.1 million associated SNOMED CT clinical findings. Queries, including concept negation and disjunction, were then run against the graph database and an enterprise Oracle relational database (RDBMS) of the same patient data sets. The graph DB was then populated with laboratory data encoded using LOINC, as well as, medication data encoded with RxNorm and complex queries performed using LOINC, RxNorm and SNOMED CT to identify uniquely described patient populations.

Results: A graph database instance was successfully created for two international releases of SNOMED CT and two US SNOMED CT editions. Transitive closure tables and descriptive statistics generated using the graph database were identical to those using validated methods. Patient queries produced identical patient count results to the Oracle RDBMS with comparable times. Database queries involving defining attributes of SNOMED CT concepts were possible with the graph DB. The same queries could not be directly performed with the Oracle RDBMS representation of the patient data and required the creation and use of external terminology services. Further, queries of undefined depth were successful in identifying unknown relationships between patient cohorts.

Conclusion: The results of this study supported the hypothesis that a patient database built upon and around the semantic model of SNOMED CT was possible. The model supported queries that leveraged all aspects of the SNOMED CT logical model to produce clinically relevant query results. Logical disjunction and negation queries were possible using the data model, as well as, queries that extended beyond the structural IS_A hierarchy of SNOMED CT to include queries that employed defining attribute–values of SNOMED CT concepts as search parameters. As medical terminologies, such as SNOMED CT, continue to

* Corresponding author at: DRC2, Room 8064, 985900 Nebraska Medical Center, Omaha, NE 68198-5900, United States. Tel.: +1 402 559 9593 (O).
E-mail address: wcampbel@unmc.edu (W.S. Campbell).

http://dx.doi.org/10.1016/j.jbi.2015.08.016

expand, they will become more complex and model consistency will be more difficult to assure. Simultaneously, consumers of data will increasingly demand improvements to query functionality to accommodate additional granularity of clinical concepts without sacrificing speed. This new line of research provides an alternative approach to instantiating and querying patient data represented using advanced computable clinical terminologies.

1. Background

Based on knowledge developed in controlled medical terminology development [1,2], Cimino stated [3] the fundamentals of a controlled medical terminology entail the capability of consistent, unambiguous recording and communication of medical concepts for information use and reuse. As medical terminologies have adopted these recommendations and evolved, they have become more complex, taking on polyhierarchical architectures and ontologic features [4]. SNOMED CT is one such terminology. Incorporating the full concept model of SNOMED CT directly into a relational database system (RDBMS) or object-oriented database system (OODBMS), technologies commonly used for electronic health records (EHR) applications, has rarely been implemented due to the complexity and size of the terminology [5,6] and/or the scope of the EHR use case [7]. As a result, data queries and clinical decision support functionality based upon the SNOMED CT terminology cannot leverage the full semantic richness of the terminology. This research investigated the feasibility and implications of modeling a patient indexed, transactional DB around the full logical model of SNOMED CT as opposed to modeling a patient indexed DB and subsequently binding the terminology to the information model. It was hypothesized that modeling data around a semantic core would improve data queries apart from the use of external terminology services.

SNOMED CT is based in Description Logics (DL) [8] and functions under the open world assumption. Terminologies established upon the open world assumption and DL facilitate the creation of inferred relationships that exist beyond those that are explicitly stated as determined by the logical axioms of the terminology. Semantic inferences enable data queries to identify data elements beyond those that are specifically enumerated in the query to also include those data elements that are logically linked to the stated query elements. Therefore, clinically important queries can be performed that utilize this logic to identify patient populations of interest for purposes of research, quality assurance, or population management.

To perform queries that leverage the conceptual inferences contain within SNOMED CT, access to the full logical model, or some representation thereof, is necessary. When SNOMED CT, or other polyhierarchical ontologies, are deployed within the context of RDBMS and OODBMS, the logical model is often instantiated in the form of transitive closure (TC) tables. A TC table represents all subsumptive relationships within a concept model in a table containing all ancestor–descendant concept pairs [9] and enable rapid queries of subsumption (see Table 1). TC tables require extensive recursive calculations when created using RDBMS or OODBMS frameworks [5], and are therefore, typically precalculated and incorporated into the database versus calculating at run-time. (Note: The TC table for SNOMED CT exceeds 5 million rows.)

The TC approach works well for queries of subsumption, such as finding all patients with diagnoses of any form of diabetes or all patients who have had some form of operative procedure on the knee. However, queries beyond subsumption including those of negation and disjunction are of clinical interest. For example, find all patients with pneumonia caused by streptococcus or staphylococcus (disjunction) but not klebsiella (negation), or identify all patients assessed for BRCA1 and/or BRCA2 gene mutations (disjunction) who have developed cancer without metastases (negation). These types of queries cannot be performed using the ISA (subsumptive) hierarchy exclusively and require more robust representations of the SNOMED CT concept model within EHR and clinical data repositories [10–12].

Table 1
Sample section of a SNOMED CT transitive closure table.

Supertype (Ancestor)	Subtype (Descendant)
95436008|Lung consolidation (disorder)	233604007|Pneumonia (disorder)
205237003|Pneumonitis (disorder)	233604007|Pneumonia (disorder)
233604007|Pneumonia (disorder)	312342009|Infective Pneumonia (disorder)
233604007|Pneumonia (disorder)	105977003|Non-infectious pneumonia (disorder)
312342009|Infective Pneumonia (disorder)	53084003|Bacterial pneumonia (disorder)
312342009|Infective Pneumonia (disorder)	75570004|Viral pneumonia (disorder)
...	...

While not broadly employed in clinical systems, Not Only SQL (i.e., NoSQL) databases, including graph databases (e.g., RDF (Research Description Framework) triple stores, Neo4j), document stores (e.g., MongoDB), column stores (e.g., HBase), and key-value/tuple stores (e.g., Voldemort [13]) represent new methods of managing and querying large amounts of complex data. These technologies have been successfully employed by corporations to interrogate and explore Big Data to achieve business objectives. Google uses column stores with distributed architecture (i.e., BigTable [14]), Amazon uses the tuple-store DynamoDB [15], and Facebook has developed a form of graph database to manage social networks. The use of these proven database technologies in clinical systems or clinical data repositories may improve the types and extent of clinical queries and increase the usefulness of clinical information. This research investigated the feasibility and implications of using a graph DB incorporating the full logical model of SNOMED CT with patient data encoded with SNOMED CT. Additional controlled medical terminologies, specifically LOINC and RxNorm, were also incorporated into the graph DB in order to perform sophisticated, multi-terminology based data queries.

2. Methods and materials

2.1. Population of the graph DB with SNOMED CT

To test the hypothesis, it was necessary to create a graph DB with the SNOMED CT concept model. The 2014-01-31 International edition RF2 (release format 2) Snapshot release files were used for this portion of the study. The RF2 files consisted of a series of tab delimited text files defining each SNOMED CT concept, enumerating classified (stated and inferred) relationships between concepts