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

and defining descriptions of each SNOMED CT concept, as well as, synonymy based on US English. Files were obtained from the US National Library of Medicine's (NLM) knowledge sources website [16]. The open source version of the graph database, Neo4j, was used as the platform for this study. (Neo Technology, Inc., San Mateo, CA) Neo4j is a Java-based, graph database platform that supports transactions with ACID properties (i.e., Atomicity, Consistency, Isolation, Durability) [17] which is a requirement for transactional databases such as EHR and clinical data repositories.

A graph DB consists of data elements represented by nodes. Nodes are connected by edges which are used to represent established relationships between pairs of elements. In the Neo4j environment, both nodes and edges may contain property definitions which provide additional informational artifacts relevant for the node or edge. SNOMED CT content was extracted from the RF2 files and loaded into Neo4j using a series of scripts written in Python, Groovy and Cypher (Neo4j's declarative query language) programming languages. Nodes were created for all SNOMED CT core concepts, and edges were created for all IS_A relationships and defining attributes. All nodes and edges contained properties for each aspect of a SNOMED CT concept such as, the concept's fully specified name, SNOMED CT concept ID, definition status of primitive or fully defined, module ID, active status and effective date (Fig. 1).

2.2. Validation of SNOMED CT logical model within the graph DB

To validate the integrity of the graph DB and the correct representation of the SNOMED CT concept model, a TC table of the graph DB model was created and compared to TC tables created directly from the RF2 relationship text file using the MySQL (Oracle Corporation, Redwood Shores, CA) method described in the SNOMED CT Technical Implementation Guide [18], as well as, Perl and Python scripting methods using the algorithm presented by Ionnidis et al. [9]. (Perl script developed and graciously provided by Kent Spackman, MD.) To ensure robustness of the methodology to create a SNOMED CT graph DB, graph DB instances were created using the 2014-07-31 International edition and for the 2013-09-01 and 2014-09-01 US national editions of SNOMED CT. TC tables were created and validated for each.

2.3. Test of graph DB queries based on SNOMED CT

To test the graph DB as a viable method to instantiate a patient data repository built with the SNOMED CT ontology in its core, a version of the graph DB containing the 2014-09-01 US national release files was populated with 465,171 patient records. Patient records were obtained from the University of Nebraska Medical Center (UNMC) de-identified clinical research database (IRB# 132-14-EP: OneTeam Enterprise Data Warehouse and Analysis System). The patient data set contained problem lists for each patient represented in SNOMED CT and exceeded 2.1 million clinical findings. A distinct node was created for each individual patient, and edges were created between each patient node and the SNOMED CT concept node(s) enumerated in the patient's problem list. Each edge was labeled 'HAS_DX' and contained properties representing the status of the problem entry as active, inactive, or deleted and the date when the finding status was made.

After populating the graph DB with patient information, subsumption queries were executed to identify patients with specific SNOMED CT clinical findings and all subtypes. To validate the query results, identical queries were run against the UNMC de-identified clinical data repository which is stored and queried within an instance of i2b2 (Informatics for Integrating Biology to the Bedside, Partners Healthcare, Boston, MA) operating in an enterprise Oracle environment. i2b2 is a well-established, mature clinical data repository architecture built upon RDBMS technology. The i2b2 clinical data repository managed SNOMED CT concept logic and hierarchies using a TC table. Query results and execution times were recorded and compared to the graph DB results.

Queries of clinical finding concepts involving negation and disjunction were developed and run against the graph DB. However, queries of negation and disjunction executed directly on the i2b2 RDBMS were not possible. The representation of the SNOMED CT concept model in the i2b2 data repository consisted of the TC table solely representing the structural IS_A hierarchy of SNOMED CT. Negation and disjunctive queries required the ability to query defining attributes of SNOMED CT concepts, and defining attribute–value pairs are not enumerated in the TC table.

A sample disjunctive query to identify all SNOMED CT concepts for pneumonia due to an influenza or parainfluenza virus is

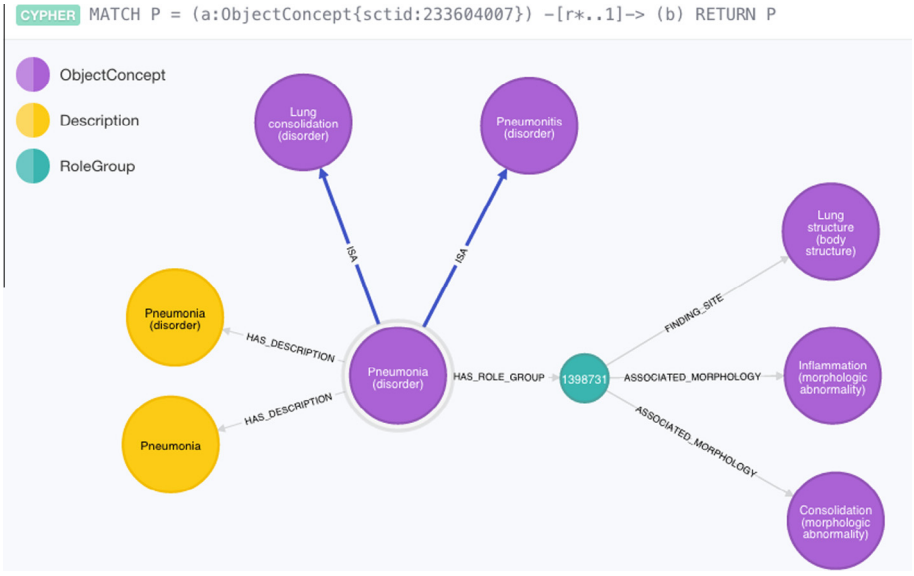

Fig. 1. Example of SNOMED CT concept representation. The target concept Pneumonia (disorder) is shown. Represented are the two proximal supertype concepts of Pneumonia, Lung consolidation and Pneumonitis, connected by directed IS_A edges. Defining attributes are shown using the HAS_ROLE_GROUP, FINDING_SITE, and ASSOCIATED_MORPHOLOGY directed edges. Finally, synonyms for Pneumonia are shown in yellow and connect to Pneumonia via HAS_DESCRIPTION edges. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

graphically shown in Fig. 2. The query identifies all possible pneumonia concepts using the IS-A hierarchy and then identifies all pneumonia concepts with a causative agent (defining attribute) of human parainfluenza virus or any influenza virus including subtyped influenza viruses. To further test the graph DB to search in depth and beyond the IS-A structure of SNOMED CT, a query was developed to identify all patients with a diagnosis of pneumonia caused by any influenza virus or human parainfluenza virus that subsequently consolidated into a particular region of the lung.

2.4. Test of graph DB queries based on SNOMED CT and LOINC

To extend the graphDB to additional, non-SNOMED CT encoded data, all LOINC version 2.50 concepts (74,600 concept codes) were obtained from the Regenstreif Institute website and added to the graphDB as nodes with each defining LOINC part instantiated as properties of the node, and patient laboratory data results encoded using LOINC were imported into the graphDB. Furthermore, a subset of medication data for anti-neoplastic agents encoded using RxNorm were imported into the graphDB. Complex queries involving SNOMED CT, LOINC encoded laboratory data and RxNorm encoded medication data were then conducted. For example, list the serial CA125 serum cancer levels and administered anti-neoplastic drugs by date for all patients with a recorded positive BRCA1 and/or BRCA2 gene mutation who have developed breast cancer.

3. Results

3.1. Transitive closure computation results

The graph DB containing the full SNOMED CT content model represented within the RF2 Snapshot for the 2014-01-31 International release was successfully instantiated in a Neo4j graph DB platform. Repeatability of the database creation process was supported by successful creation of separate graph DBs based using the 2014-07-31 International SNOMED CT edition and two separate release dates of the US edition of SNOMED CT. The size of the database containing the SNOMED CT concept model and definitional content (<870 MB) was similar to the RF2 text files used to construct the database (<690 MB). TC tables created from the graph DB to those TC tables created using current and validated methods were identical. Compute times to create the TC tables using the graph DB were consistent with "in memory" TC computational methods and superior to RDMBS calculation times.

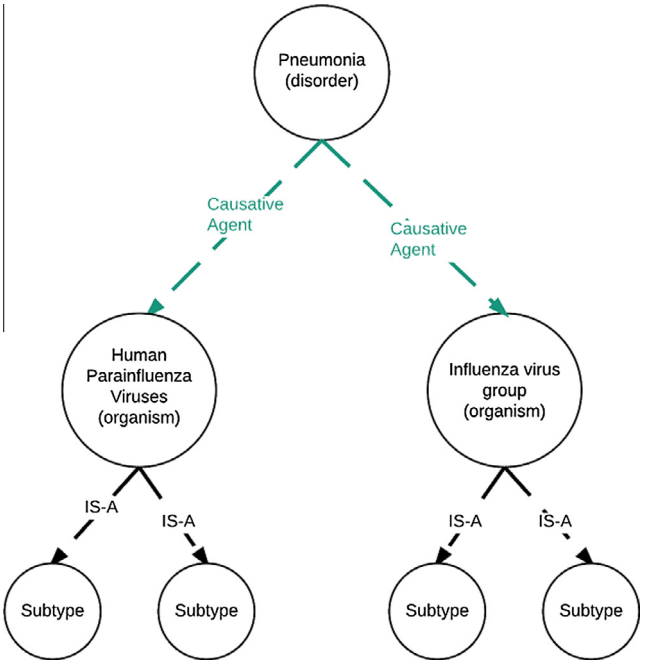

Fig. 2. Queries using defining relationships. The defining attribute, Causative agent (green arrows), is used to identify all pneumonia concepts due to human parainfluenza virus or any type of influenza virus. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Queries on patient data

Queries against the graph DB containing patient diagnostic data were compared with identical queries performed against the UNMC i2b2 clinical data repository. Query results were identical between the graph DB and the i2b2 Oracle SQL queries. Comparison of query runtimes between database designs showed that the graph DB performed in similar times as the SQL query times for most queries. For patient count queries, the graph DB was instantiated on a local workstation (Dell Optiplex 990, 16 GB RAM, 8 CPU cores). The clinical data repository queries were performed on an enterprise Oracle SQL database (Oracle v. 11.2) operating on a virtual machine provisioned for 8 CPU (Intel® Xeon® E5-4620 2.2 GHz) cores, 96 GB RAM. Specific queries and results are listed in Table 2.

Query time comparisons were not possible for negation or disjunction queries that required explicit use of defining attributes of a SNOMED CT concept. The graph query was done in a single step, but a multistep process between multiple, distinct databases was required to perform the same query of the i2b2 data.

To satisfy the queries of negation and disjunction using a RDBMS, a separate RDBMS external to i2b2 was required that consisted of three tables: (a) the SNOMED CT RF2 relationship table; (b) the SNOMED CT RF2 concept table; and (c) the TC table. For example, to perform a query to identify all patients with Pneumonia due to any influenza virus or human parainfluenza virus, four separate and distinct queries of the external RDBMS were required to identify candidate SNOMED CT concepts meeting the query definition. Query 1used the TC table to identify all pneumonia concepts and subtypes. Queries 2 and 3 used the TC table to identify all influenza virus and human parainfluenza virus concepts and subtypes, respectively. With results from the first three TC table queries, a query of the SNOMED CT relationships table, Query 4, identified all pneumonia concepts listing in Query 1 that had a causative agent relation with any virus concept returned in Queries 2 and 3. Finally, Query 5 identified all patients in i2b2 with a clinical finding concept contained in Query 4's output (Fig. 3).

Extension of the query to further identify patients with pneumonia due to influenza or human parainfluenza viruses that progressed to a localization of the pneumonia to the left, lower zone of the lung returned a single patient. The patient was diagnosed with pneumonia due to influenza and subsequently was given a new diagnosis of left lower zone pneumonia in the same episode of care.

An unexpected result was observed during negation and disjunction query testing. Specifically, the graph DB identified concepts that were not consistently modeled and were subsequently not classified by the DL classifier as intended. The concept for |Congenital group A hemolytic streptococcal pneumonia (disorder)| was not classified as a subtype of infective pneumonia in the IHTSDO release files, and patients with this condition were not returned using the TC approach. However, they were identified when employing definitional attributes of pneumonia in the query logic as done in the graph DB (Fig. 4). This finding was a result of the two pneumonia concepts being modeled with different

Table 2
Query results and times of i2b2 RDBMS and Neo4j graph DB. Queries based on SNOMED CT encoded patient diagnoses. Note: Query times for i2b2 the disjoint query represented in the last row could not be computed based on the requirement of multiple queries of external data necessary to complete prior to interrogation of i2b2.

Query	i2b2/Oracle		Neo4j	
	Query time (seconds)	Query result (#patients)	Query time (seconds)	Query result (#patients)
<<27624003|Chronic disease (disorder)	2.1	55512	3.2	55512
<<233604007|Pneumonia (disorder)	1	4833	0.1	4833
<<233604007 Pneumonia (disorder) and NOT <<34020007|Pneumonia due to Streptococcus (disorder)	1 s	4789	0.1 (7.3)	4810[a] (4789)
<<233604007 Pneumonia (disorder) and 'HAS_CAUSITIVE_AGENT' =<<58800005|Genus Streptococcus (organism) OR <<65119003|Genus Staphylococcus (organism)	N/A	61	1.6	61

[a] Graph algorithm accounted for patients with multiple forms of pneumonia. I2b2 logic does not. Results of graphDB replication of i2b2 query logic shown in parentheses.

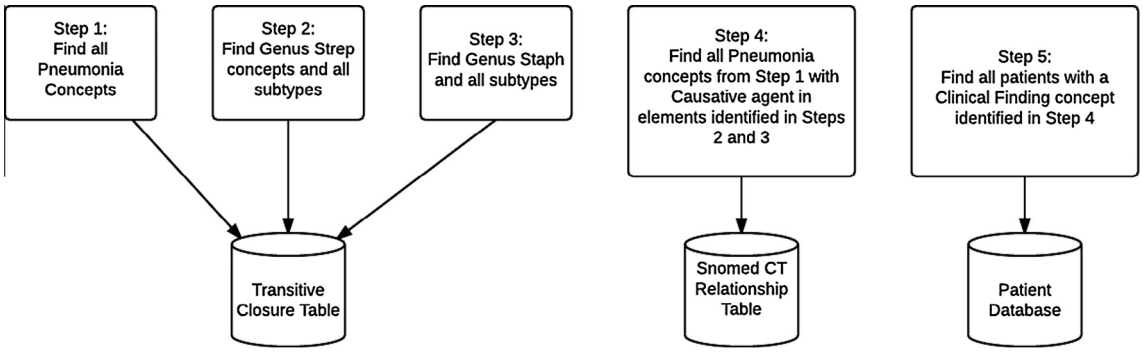

Fig. 3. Five step process required to identify patients with Pneumonia due to all influenza viruses and subtypes or human parainfluenza virus when using RDBMS and TC tables. The graphDB returned the results in a single query against the patient database.

primitive supertypes in the SNOMED CT release files and represented a concept modeling inconsistency error. (Note: the error was reported to IHTSDO's quality improvement system and will be corrected in the upcoming release.)

Complex queries of negation and disjunction of SNOMED CT concepts in conjunction with discrete data based upon separate medical terminologies were successful. Query results showing the CA125 serum levels by date and chemotherapeutic medications delivery by date for all patients with a positive BRCA1 or BRCA2 gene mutation who have developed some form of breast cancer (Table 3).

4. Discussion

A graph DB was a logical tool for serialization of the SNOMED CT concept model because SNOMED CT is, by definition, a directed, acyclic graph which is a well understood mathematical model [19]. A Neo4j graph model instantiating a RF1 version of SNOMED CT has previously been developed by the National Health Services [20], and the results of this study extended the NHS work to include the construction of a graph DB containing the full definitional content and logical model of SNOMED CT using the RF2 format, which includes concept module identification and additional revision data not contained in RF1 files. This distinction was important because historicity of terminology is important for terminology maintenance, medical-legal considerations and connection between past medical knowledge and new or developing medical knowledge.

By using the RF2 release which contained SNOMED CT concept history data from the creation of SNOMED CT to the current release, the authors have created a graph DB containing the entire, fully classified SNOMED CT content for every dated international release. The model is capable of producing TC tables on demand for any given date, as well as, TC table differences between any two release dates. *When associated with historical patient data, the database is capable of querying patient information with historical perspectives. For example, how many patients with disease X in year 2003 would have their diagnoses changed to disease Y in the current year based on new scientific evidence?* To the authors' knowledge, the methods developed in this study are the first to use the RF2 files to create a graph DB of the SNOMED CT concept model, the first to incorporate patient level data within the same model, and the first to instantiate a fully classified version of all published SNOMED CT content from the first release to the present in a single database.

The novelty of the research presented was the design of a patient database model using the fully classified semantic core of SNOMED CT as the basis for the data model. This approach is a departure from the more typical patient database design where patient data is modeled using paradigms of episodes of care, disease processes or treatment to which terminologies are subsequently bound. Based on the methods presented in this research, data queries using the entirety of the SNOMED CT ontology and not solely the structural IS_A hierarchy could be executed within the patient database. Current state-of-the-art clinical information systems interact with the SNOMED CT terminology through the use of limited value sets, specific concept hierarchies, by employing a pre-calculated TC table within the information model or interacting with an external terminology server (service).

The approach presented in this research did not place limits upon the amount of SNOMED CT content to be included within the database. An external terminology server was not required nor was a pre-calculated TC table necessary. In addition, the model enabled defining attribute–values of SNOMED CT concepts to be used as query parameters. This feature permitted queries beyond those of subsumption to include queries of negation and disjunction. In fact, queries based solely on defining attribute–value pairs

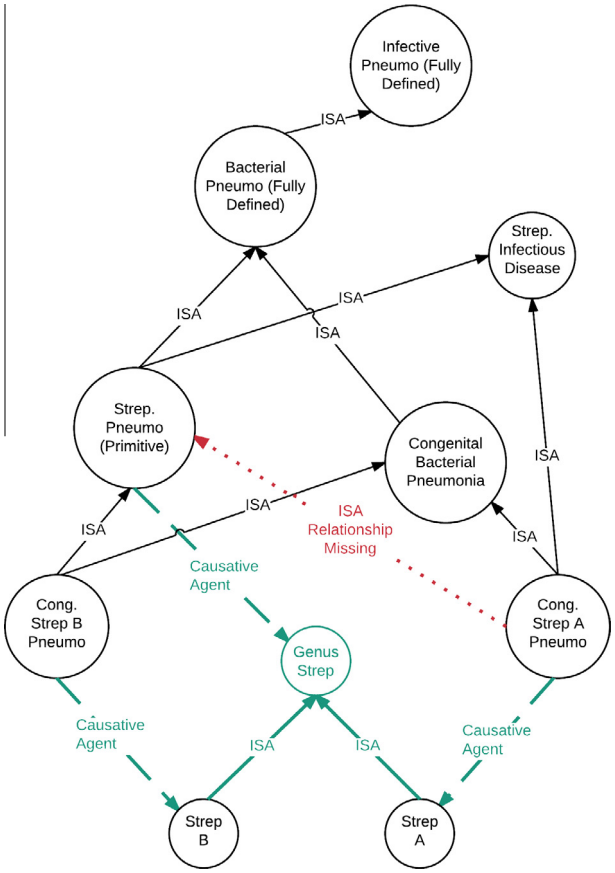

Fig. 4. Hierarchical view of Infective pneumonia and queries by defining attributes. A query to find all pneumonia concepts caused by streptococcus. The concept for congenital strep A would not be returned by simple subsumption. The 'ISA' relationship is missing due to a concept modeling inconsistency (red arrow). However, the concept is returned correctly when the defining attribute value pair of Causative agent = Genus streptococcus. (Green arrows). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

in any combination were possible (e.g., patients with pneumonia caused by some pathogen and disease progression to a particular anatomic location.) Queries involving defining relationships and attributes could not be performed directly against the i2b2 database in which the SNOMED CT concept model representation was limited to the IS_A hierarchy. Negation and disjunction queries of patient data sets that are relevant to clinical situations for individual patient management, population management and quality improvement require searches involving defining attributes of SNOMED CT concepts and exceed the scope of TC tables [21].

Because SNOMED CT is based on the EL [22] version of DL which does not support negation and disjunction [23], a query of SNOMED CT for the identification of concepts meeting the negated or disjoint requirements using ontology development and management tools such as Protégé was not possible regardless of any specific DL classifier invoked. To perform searches based on defining relationships using a RDBMS, multiple preparatory subsumption queries against the TC table were necessary, as well as, additional queries of the relationship release files to identify all concepts satisfying the non-ISA relationship parameters of the query prior to executing the desired patient DB query (Fig. 2). The result of this approach was a non-trivial complication to query workflow and a requirement to develop and maintain additional terminology database.

Extending the dataset to include non-SNOMED CT encoded data, such as LOINC encoded laboratory test results and RxNorm encoded medication data, enabled searches of greater complexity and importance. The example shown in Table 3 provided an example of the types of queries that can be performed within the graphDB model that take into account semantically structured data supporting inference, as well as, data encoded using standard terminologies with limited inference support (e.g., LOINC).

A graph DB supports data queries of "undefined depth" and "undefined connections" [24]. As a result, the graph DB developed in this research supported queries to find relationships between concepts without explicitly stating the particular length or structure of the search. For example, queries to find SNOMED CT concepts with common defining values and subtypes of those values entailed traversing the graph to find connections between concepts

Table 3
Subset of query results listing CA125 serum levels by date, chemotherapy drugs administered by date for patients with positive BRCA1 or BRCA2 gene mutations and breast cancer. Result shows declining CA125 levels during chemotherapy. Query execution time = 1.2 s.

Patient	Date of CA125	CA125 result	Drug name (RxNorm Name)	RxNorm RXCUI code	Medication date
23248924	20-Aug-11	28			
			Doxorubicin injectable solution	374376	4-Mar-13
			Cyclophosphamide injectable solution	376666	4-Mar-13
			Doxorubicin injectable solution	374376	18-Mar-13
			Cyclophosphamide injectable solution	376666	18-Mar-13
			Doxorubicin injectable solution	374376	1-Apr-13
			Cyclophosphamide injectable solution	376666	1-Apr-13
			Doxorubicin injectable solution	374376	15-Apr-13
			Cyclophosphamide injectable solution	376666	15-Apr-13
			Paclitaxel injectable solution	377111	29-Apr-13
			Paclitaxel injectable solution	377111	13-May-13
			Paclitaxel injectable solution	377111	27-May-13
			Paclitaxel injectable solution	377111	10-Jun-13
	19-Feb-14	8			
			Docetaxel injectable solution	376888	1-Sep-14
			Carboplatin injectable solution	378363	1-Sep-14
			Docetaxel injectable solution	376888	22-Sep-14
			Carboplatin injectable solution	378363	22-Sep-14
			Docetaxel injectable solution	376888	13-Oct-14
			Carboplatin injectable solution	378363	13-Oct-14
			Docetaxel injectable solution	376888	3-Nov-14
			Carboplatin injectable solution	378363	3-Nov-14

without a need to explicitly define the types of relationships that bound them nor the degrees of separation between concepts (i.e., in RDBMS parlance, the number of join operations.)

RDBMS architectures excel in the realm of queries involving well characterized, structured data. However, as the structure of the data becomes more complex and diverse, RDBMS query times begin to grow due to the number of computationally expensive and recursive JOIN operations necessary to search the data as demonstrated in the compute times to create TC tables for SNOMED CT. Graph database architectures accommodate data where aspects of the data relationships are not explicitly defined which is useful when analyzing data when attempting to identify relationships between concepts that may not be known or well-described.

Although not the focus of this research, the graph DB demonstrated utility as a quality improvement tool for SNOMED CT. Modeling consistency and completeness within a large, complex terminology such as SNOMED CT is a difficult process, and modeling issues have been cited for the unwary [25,26]. The complex task of SNOMED CT model quality assurance is further increased based on the number of primitively defined concepts. This phenomenon was observed when identifying all pneumonia concepts caused by streptococcus. In the 2014-09-01 US edition, there were 301,390 concepts of which 231,067 (76.7%) had a definition status of primitive. Many concepts were intermediate primitives which can confound the DL classifiers and result in the omission of inferences between otherwise linked SNOMED CT concepts. As the SNOMED CT user community demands additional content, the task of quality assurance will become more complex and susceptible to inconsistencies. A graph representation of the terminology may provide an additional tool to SNOMED CT modelers to quality assure future releases such as identifying all intermediate primitive concepts within any given hierarchy.

Because a graph DB emphasizes the connectedness of data elements using edges to link concepts, each SNOMED CT concept was represented simultaneously within the graph DB in several forms. First, each concept was represented as a single node consisting of the pre-coordinated SNOMED CT concept. Second, each individual SNOMED CT concept was bound to additional SNOMED CT concepts by edges indicating the relationships between the connected nodes, namely the types of defining attribute and the IS_A relationships. This was equivalent to the close-to-user form [18]. Finally, each defining concept(s) and supertype concept(s) were related by IS_A paths to their individual proximal primitive (s), that is, its normal form.

As the adoption of SNOMED CT increases, users will demand increased levels of expressivity from the terminology [27,28]. The use of post-coordinated expressions is likely to expand as it is impractical for the IHTSDO to add concept definitions for all requested terms [26,29]. Therefore, this database design may prove useful in management of post-coordinated expressions or equivalence and subsumption testing between concepts.

While Neo4j and other graph DB platforms support ACID properties for transactions and are capable of high volume data I/O, this research project did not test the SNOMED CT graph DB in the context of a fully operational transactional setting. This research involved clinical finding data encoded in SNOMED CT. Other SNOMED CT hierarchies were not included in query examples solely because there was limited data contained in UNMC i2b2 database represented by SNOMED CT outside of the clinical findings hierarchy. The inclusion of LOINC encoded laboratory test data and RxNorm encoded medication data into the model demonstrated that the graph DB architecture was capable of managing non-SNOMED CT based data, as well. The results of this study do not imply, however, that all aspects of a formal terminology server [30,31] could be replaced by the graph model (e.g., lexical equivalence).

5. Conclusion

Population of a graph DB with the SNOMED CT concept model and substantial amounts of patient data with problem lists, laboratory results and medications represented small subset of an electronic health record. Although not at the scale of an enterprise EHR, query accuracy and speeds using the graph DB compared favorably to an enterprise-class RDBMS instantiation of same patient data. The findings supported the hypothesis that an EHR or clinical data repository could be established using a graph DB that included complex terminology management without loss of performance. These findings provide a basis to further investigate the benefits of alternative database platforms for clinical systems that incorporate ontologic terminologies and models within the core function of the database.

Acknowledgements

Praveen Rao was supported by the National Science Foundation under Grant No. 1115871.

References

[1] B.H. Forman, J.J. Cimino, S.B. Johnson, S. Sengupta, R. Sideli, P. Clayton, Applying a controlled medical terminology to a distributed, production clinical information system, Proc. Annu. Symp. Comput. Appl. Med. Care (1995) 421–425.
[2] R.A. Rocha, S.M. Huff, P.J. Haug, H.R. Warner, Designing a controlled medical vocabulary server: the VOSER project, Comput. Biomed. Res. 27 (6) (1994) 472–507.
[3] J.J. Cimino, Desiderata for controlled medical vocabularies in the twenty-first century, Methods Inf. Med. 37 (4–5) (1998) 394–403.
[4] J.J. Cimino, In defense of the Desiderata, J. Biomed. Inform. 39 (3) (2006) 299–306.
[5] G. Schadow, M.R. Barnes, C.J. McDonald, Representing and querying conceptual graphs with relational database management systems is possible, in: Proc AMIA Symp., 2001, pp. 598–602.
[6] D. Lee, R. Cornet, F. Lau, N. de Keizer, A survey of SNOMED CT implementations, J. Biomed. Inform. 46 (1) (2013) 87–96.
[7] D.H. Lee, F.Y. Lau, H. Quan, A method for encoding clinical datasets with SNOMED CT, BMC Med. Inform. Decis. Making 10 (2010) 53. 6947-10-53.
[8] M. Krotzsch, F. Simancik, I. Horrocks, A description logics primer, eprint arXiv:1201 4089 [Internet]. 2013 June 3, 2013 [cited April 24, 2015]. <http://arxiv.org/abs/1201.4089v3>.
[9] Y. Ionnidis, R. Ramakrishnan, L. Winger, Transitive closure algorithms based on graph traversal, ACM Trans. Database Syst. 18 (3) (1993) 512–576.
[10] P. Hendler, R. Piro, M. Rossman, Clinical modeling and description logics, a collaboration between Oxford and Kaiser Permanente, in: SNOMED CT Implementation Showcase; October 31, 2014; Amsterdam, The Netherlands, International Health Terminology Standards Development Organization, Copenhagen, Denmark, 2014.
[11] R. Piro, RDFox, a heavily optimized RDF triple store and parallel Datalog reasoner, in: SNOMED CT Implementation Showcase; October 31, 2014; Amsterdam, The Netherlands, International Health Terminology Development Organization, Copenhagen, Denmark, 2014.
[12] Clinical Models and SNOMED Kaiser Perspective [Internet], CIMI, Rochester, Minnesota, 2012, <http://informatics.mayo.edu/CIMI/index.php/Clinical_Models_and_SNOMED_Kaiser_Perspective> [updated 31.07.12; cited 24.04.15].
[13] R. Sumbaly, J. Kreps, L. Gao, A. Feinberg, C. Soman, S. Shah, Serving large-scale batch computed data with project Voldemort, in: Proceedings of the 10th USENIX Conference on File an Storage Technologies (FAST '12), 2012; Berkeley, CA, USENIX Association, Berkeley, CA, 2012.
[14] F. Chang, J. Dean, S. Ghemawat, W. Hsieh, D. Wallach, M. Burrows, et al., Bigtable: a distributed storage system for structured data, ACM Trans. Comput. Syst. 26 (2) (2008). Article 4, 26p.
[15] S. Sivasubramanian, Amazon dynamoDB: a seamlessless scalable non-relational database service, in: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (SIGMOD '12), 2012; New York, NY, ACM, New York, NY, 2012.
[16] SNOMED CT International Release Files [Internet], US National Library of Medicine, Washington, DC, 2015, <http://www.nlm.nih.gov/research/umls/licensedcontent/snomedctfiles.html> [updated 09.02.15; cited 24.04.15].
[17] T. Haerder, A. Reuter, Principles of transaction-oriented database recovery, ACM Comput. Surv. 15 (4) (1983) 287–317.
[18] International Health Terminology Standards Development Organization, SNOMED-CT Technical Implementation Guide, July 2012 International Release (US English) ed., in: D. Markwell (Ed.), International Health Terminology Standards Development Organization, Copenhagen, 2012.

[19] J. Gross, J. Yellen, Handbook of Graph Theory, in: J. Gross, J. Yellen (Eds.), first ed., CRC Press, LLC, New York, NY, 2004.
[20] SNOGRAPH [Internet]. United Kingdon, 2013, <https://github.com/ysgao/SnoGraph> [updated 25.10.13; cited 2014].
[21] P. Hendler, Application of SNOMED and DL in Kaiser's EHR. Semantic Web Health Care and Life Sciences Special Interest Group Face-to-face Meeting; November 2, 2009; Santa Clara, CA. W3C HCLS, 2009.
[22] EL [Internet]. W3C, 2008, <https://www.w3.org/2007/OWL/wiki/EL> [updated 18.01.08; cited 15.04.15].
[23] A.L. Rector, S. Brandt, Why do it the hard way? The case for an expressive description logic for SNOMED, J. Am. Med. Inform. Assoc. 15 (6) (2008) 744–751.
[24] C. Vicknair, M. Macias, Z. Zhao, X. Nan, Y. Chen, D. Wilkins, A comparison of a graph database and a relational database, in: Association for Computing Machinery Southeast; April 15, 2010; Oxford, MS, ACM, Oxford, MS, 2010.
[25] A.L. Rector, S. Brandt, T. Schneider, Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications, J. Am. Med. Inform. Assoc. 18 (4) (2011) 432–440.
[26] A. Rector, L. Iannone, Lexically suggest, logically define: quality assurance of the use of qualifiers and expected results of post-coordination in SNOMED CT, J. Biomed. Inform. 45 (2) (2012) 199–209.
[27] W.S. Campbell, J.R. Campbell, W.W. West, J.C. McClay, S.H. Hinrichs, Semantic analysis of SNOMED CT for a post-coordinated database of histopathology findings, J. Am. Med. Inform. Assoc. 21 (5) (2014) 885–892.
[28] J.R. Campbell, J. Xu, K.W. Fung, Can SNOMED CT fulfill the vision of a compositional terminology? Analyzing the use case for problem list, in: AMIA Annu Symp Proc., 2011, pp. 181–188.
[29] R. Cornet, M. Nystrom, D. Karlsson, User-directed coordination in SNOMED CT, Stud. Health Technol. Inform. 192 (2013) 72–76.
[30] C.G. Chute, P.L. Elkin, D.D. Sherertz, M.S. Tuttle, Desiderata for a clinical terminology server, in: Proc AMIA Symp., 1999, pp. 42–46.
[31] FHIR: Terminology services [Internet]. Ann Arbor, MI: HL7, 2015, <http://hl7.org/implement/standards/fhir/2015Jan/terminology-service.html> [updated 23.02.15; cited 24.04.15].