ihtsdo
Delivering
SNOMED CT

Data Analytics with SNOMED CT

MAY 2015

Date: 2015-05-21

Version: 1.00
Status: FINAL

TABLE OF CONTENTS

Copyright and Acknowledgements

The document is a draft publication of the International Health Terminology Standards Development Organisation (IHTSDO), the association that owns and maintains SNOMED Clinical Terms. IHTSDO acknowledges the work undertaken by Malcolm Duncan in co-authoring this document during his participation in the SNOMED CT Implementation Advisor (SIA) scheme.

1 Executive Summary

SNOMED CT is a clinically validated, semantically rich, controlled terminology designed to enable effective representation of clinical information. SNOMED CT is widely recognized as the leading global clinical terminology for use in Electronic Health Records (EHRs). SNOMED CT enables the full benefits of EHRs to be achieved by supporting both clinical data capture, and the effective retrieval and reuse of clinical information.

The term 'analytics' is used to describe the discovery of meaningful information from healthcare data. Analytics may be used to describe, predict or improve clinical and business performance, and to recommend action or guide decision making [3].

Using SNOMED CT to support analytics services can enable a range of benefits, including:

- Enhancing the care of individual patients by supporting:
    - Retrieval of appropriate information for clinical care
    - Guideline and decision support integration
    - Retrospective searches for patterns requiring follow-up
- Enhancing the care of populations by supporting:
    - Epidemiology monitoring and reporting
    - Research into the causes and management of diseases
    - Identification of patient groups for clinical research or specialized healthcare programs
- Providing cost-effective delivery of care by supporting:
    - Guidelines to minimize risk of costly errors
    - Reducing duplication of investigations and interventions
    - Auditing the delivery of clinical services
    - Planning service delivery based on emerging health trends

SNOMED CT has a number of features, which makes it uniquely capable of supporting a range of powerful analytics functions. These features enable clinical records to be queried by:

- Grouping detailed clinical concepts together into broader categories (at various levels of detail);
- Using the formal meaning of the clinical information recorded;
- Testing for membership of predefined subsets of clinical concepts; and
- Using terms from the clinician's local dialect.

SNOMED CT also enables:

- Clinical queries over heterogeneous data (using SNOMED CT as a common reference terminology to which different code systems can be mapped);
- Analysis of patient records containing no original SNOMED CT content (e.g. free text);
- Powerful logic-based inferencing using Description Logic reasoners;
- Linking clinical concepts recorded in a health record to clinical guidelines and rules for clinical decision support; and
- Mapping to classifications, such as ICD-9 or ICD-10, to utilize the additional features that these provide.

Analytics tasks, which may be enabled or enhanced by the use of SNOMED CT techniques, can be considered in three broad categories:

1. Point-of-care analytics, which benefits individual patients and clinicians. This includes historical summaries, decision support and reporting.
2. Population-based analytics, which benefits populations. This includes trend analysis, public health surveillance, pharmacovigilance, care delivery audits and healthcare service planning, and
3. Clinical research, which is used to improve clinical assessment and treatment guidelines. This includes identification of clinical trial candidates, predictive medicine and semantic searching of clinical knowledge.

While the use of SNOMED CT for analytics does not dictate a particular data architecture, there are a few key options to consider, including:

- Analytics directly over patient records;
- Analytics over data exported to a data warehouse;
- Analytics over a Virtual Health Record (VHR);
- Analytics using distributed storage and processing; and
- A combination of the above approaches.

Practically all analytical processes are driven by database queries. To get the most benefit from using SNOMED CT in patient records, record-based queries and terminology-based queries must work together to perform integrated queries over SNOMED CT enabled data. To this end, IHTSDO is developing a consistent family of languages to support a variety of ways in which SNOMED CT is used. Clinical user interfaces can also be designed to harness the capabilities of SNOMED CT, and to make powerful clinical querying more accessible. Innovative data visualization and analysis tools are becoming more widespread as the capabilities of SNOMED CT content are increasingly utilized.

A number of challenges exist when performing analytics over clinical data, irrespective of the code system used. These include the reliability of patient data, terminology/information model boundary issues, concept definition issues and versioning. Many of these challenges, however, are able to be mitigated using the unique features of SNOMED CT.

A number of software vendors are now realizing the competitive advantage that using SNOMED CT can provide to unlock the analytics potential of clinical data. Several commercial tools are now available that support analytics using SNOMED CT, while others are following a roadmap of increasing functionality driven by SNOMED CT.

As the SNOMED CT encoding of healthcare data increases, so too have the benefits being realized from analytics processes performed over this data.

2 Introduction

2.1 BACKGROUND

SNOMED CT is a clinically validated, semantically rich, controlled terminology. SNOMED CT is comprised of meaning-based concepts, human-readable descriptions and machine-readable definitions. SNOMED CT is used within electronic health records to support data capture, retrieval, and subsequent reuse for a wide range of purposes. SNOMED CT is also used to enable or enhance analysis of patient records and other clinical documents containing no original SNOMED CT content.

SNOMED CT hierarchies and formal concept definitions allow selective information retrieval to support analysis – from patient-based queries to operational reporting, public health reporting, strategic planning, predictive medicine and clinical research. As the SNOMED CT encoding of healthcare data increases, so too have the benefits being realized from analytics processes performed over this data.

2.2 PURPOSE

The purpose of this document is to review current approaches, tools and techniques for performing data analytics using SNOMED CT and to share developing practice in this area. It is anticipated that this report will benefit members, vendors and users of SNOMED CT by promoting a greater awareness of both what has been achieved, and what can be achieved by using SNOMED CT to enhance analytics services.

2.3 SCOPE

This document presents different data approaches, tools, terminology techniques, query languages, data architectures and user interfaces that may be used in performing analytics using SNOMED CT. Analytics services considered include patient-based queries, operational reporting, the application and audit of evidence-based medical practice, strategic planning, predictive medicine, public health reporting and clinical research. The benefits and challenges of these approaches are also presented. The case study summaries describe a selection of SNOMED CT analytics projects and tools.

This document does not provide an exhaustive list of analytics projects and tools, and does not mandate a specific approach. The development of clinical case definitions [1] is also outside of the scope of this document.

2.4 AUDIENCE

The target audience of this document includes:

- IHTSDO members who wish to learn about current analytics activities in other jurisdictions and inform future directions;
- Clinicians, informatics specialists and technical staff involved in the planning, management, design or implementation of clinical record applications or healthcare analytics tools;

- Software vendors, data analysts, epidemiologists and others designing SNOMED CT based solutions.

This document assumes a basic level of understanding of SNOMED CT. For background information it is recommended that the reader refers to the SNOMED CT Starter Guide [2].

2.5 DOCUMENT OVERVIEW

This document presents an introduction to analytics over data with SNOMED CT content.

Section 1 (Executive Summary) provides a concise summary of the document.

Section 2 (Introduction) introduces the document by explaining the background, purpose, scope, audience and overview of the document.

Section 3 (Analytics Overview) introduces the topic by presenting a definition of analytics and describing the scope, purpose and substrates of analytics services which use SNOMED CT.

Section 4 (SNOMED CT Overview) describes the main features of SNOMED CT which may be used to support analytics over health data, and the specific benefits that using SNOMED CT enables.

Section 5 (Preparing Data for Analytics) describes some approaches used to prepare clinical data for analytics using SNOMED CT, including mapping and natural language processing.

Section 6 (SNOMED CT Analytics Techniques) presents a range of techniques for using SNOMED CT to perform data analytics, including using value sets, subsumption, defining relationships and description logic.

Section 7 (Task-Oriented Analytics) looks at how these SNOMED CT based techniques can be used to assist with specific analytics tasks for point of care analytics, population health monitoring and reporting, and clinical research.

Section 8 (Data Architectures) presents a number of approaches for architecting analytics services, including querying directly over patient data, using a data warehouse, querying a virtual medical record and using distributed storage and processes.

Section 9 (Database queries) considers the query languages that are needed to perform analytics over the combination of the patient record and terminology content.

Section 10 (User Interface Design) presents a selection of user interface styles that may be used with SNOMED CT to support querying and results visualization.

Section 11 (Challenges) discusses some of the challenges which are faced when performing analytics over SNOMED CT enabled data, including the reliability of patient data, information model/terminology boundary issues, concept definition issues, versioning and inactive content.

Section 12 (References) contains a list of documents that were referenced in the creation of this report.

Two appendices to this report present a variety of project case studies and vendor tooling case studies respectively. These appendices, which are referenced extensively throughout this document, can be found at http://snomed.org/analyticscasestudies.pdf.

3 ANALYTICS OVERVIEW

3.1 DEFINITION

The term '*analytics*' is used broadly in this document to describe the process of extracting useful information from healthcare data.

> "Analytics is the discovery and communication of meaningful patterns in data… Analytics may be applied to business data to describe, predict and improve business performance. The insights from data are used to recommend action or to guide decision making." [3]

Most analytical processes are driven by database queries. A 'query' is a means for retrieving information from a database consisting of a machine readable question presented to the database in a predefined format. Queries are used to inform or contribute to a human-readable report or produce a machine-actionable response. A human-readable report may be a list of patients, a graph, historical or projected resource utilization figures, or a summary dashboard display. Machine-actionable responses may include populating an order for a new laboratory test, based on the results of a previous test, or placing an order to restock medical devices on a hospital ward.

3.2 SCOPE AND PURPOSE

Full benefits of electronic health records only accrue with the implementation of effective retrieval and reuse of clinical information. The scope of analysis of health record data may cover:

- An individual patient, across time and/or care providers;
- An individual healthcare worker;
- Patient groups or cohorts, based on demographics, diagnoses, treatments or interventions;
- Enterprise groups, based on teams, wards, clinics, institutions or providers;
- Geographical groups, based on a local area, town, region or country.

Figure 3-1 illustrates the three main purposes of analytics with SNOMED CT. These are:

1. Clinical assessment and treatment;
2. Population monitoring; and
3. Research.

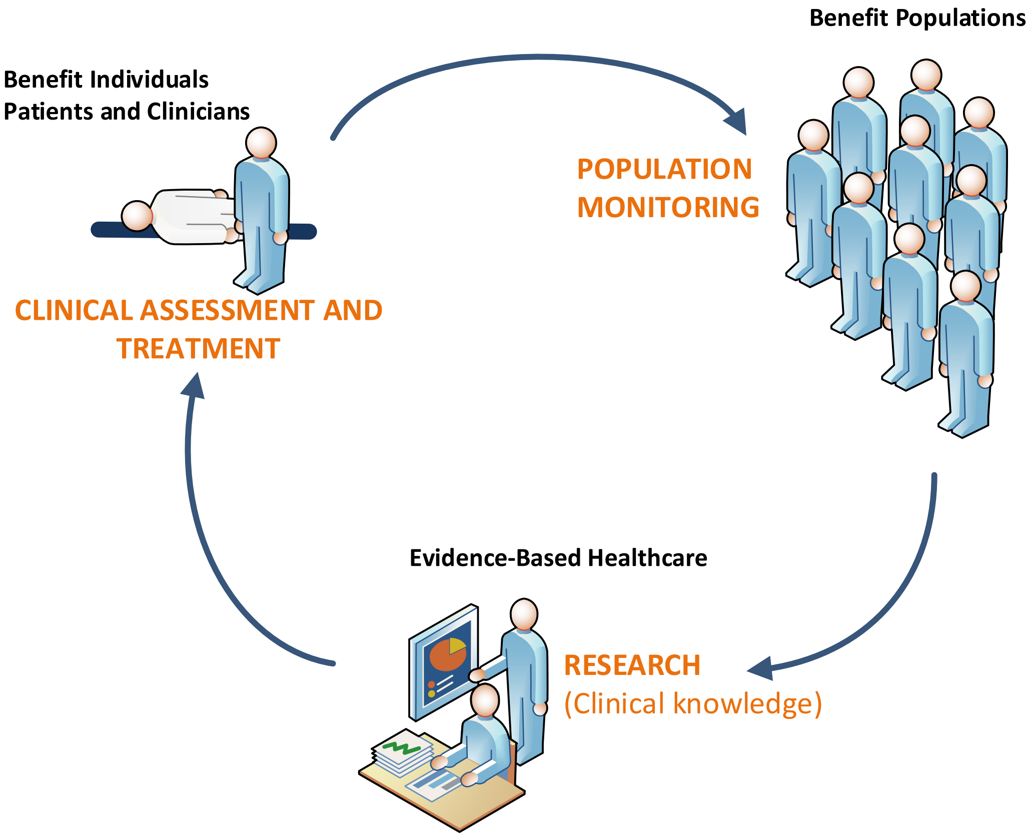


Figure 3-1: Purposes of analytics with SNOMED CT

SNOMED CT may be used to support analytics that:

- Improves the care of individual patients by enabling:
    - Retrieval of relevant information that better supports clinicians in assessing the condition and needs of a patient
    - Clinical records to be integrated with decision support tools to guide safe, appropriate and effective patient care – for example, allergy checking and potential drug contraindications identified at the point of prescribing
    - Reduction in the duplication of investigations and interventions through the effective retrieval of shared information about the patient
    - Meaning-based sharing of clinical information that is collected by different members of the health care team at different times and places (and potentially in different languages)
    - Identification of patients requiring follow-up or changes to treatment based on updated guidelines
    - Wellness management, for example, using genetic and behavioral risk profiles.
    - Context-sensitive presentation of guidelines and care pathways within the user interface
    - Labor-saving decision support systems for clinicians

  - Adaptive pick lists in clinical user interfaces
  - Professional logs and performance tracking for clinicians
  - Work list generation, for example, patients requiring follow-up based on specific criteria
  - Workload profiling and monitoring.
- Improves the care of populations by enabling:
  - Epidemiological monitoring and reporting, for example, monitoring of epidemic outbreaks, or hypothesis generation for the causes of diseases
  - Audit of clinical care and service delivery
  - Systems that measure and maximize the delivery of cost-effective treatments and minimize the risk of costly errors
- Supports evidence-based healthcare and clinical knowledge research by enabling:
  - Identification of clinical trial candidates
  - Research into the effectiveness of different approaches to disease management
  - Clinical care delivery planning, for example, determining optimum discharge time
  - Planning for future service delivery provision based on emerging health trends, perceived priorities and changes in clinical understanding.

3.3 SUBSTRATES FOR ANALYTICS

Analytics with SNOMED CT may be deployed on a wide range of data sources as summarized in the table below. These data sources are also known as the 'substrate' of the analytics. Please note that data which is not natively coded using SNOMED CT may be transformed using one of the techniques described in Section 5. These techniques may be used to transform heterogeneous data recorded using free text or a variety of code systems into SNOMED CT, which can serve as a common reference terminology for analysis.

Table 3-1: Direct and indirect substrates for SNOMED CT based analytics

Analytics Substrate	Examples	Coding	Information Model
Unstructured free text document	Dictated clinical letter Typed discharge summary letter	Natural language	None or informal headings
Structured documents with free text fields	Assessment form Discharge summary form	Natural language	Standardized headings and fields
Structured documents with free text and post-coded classification (i.e. added by clinical coders after the clinical event	Discharge summary form with post-coded classification	Classifications (e.g. ICD)	Formal information model (typically simple)
Structured documents with non-SNOMED CT coding (e.g. proprietary, local or other coding system)	Standalone clinical application using departmental codes Enterprise-wide healthcare system using local	Local code system, controlled vocabulary or legacy clinical terminology	Formal information model

	dictionaries and pick-lists Electronic patient record using regional coding system (such as UK Primary Care systems)		
Structured documents with SNOMED CT content	Cardiology report GP event summary	SNOMED CT	Formal information model
‘Big data’ data store	Data warehouse Data store containing a mixture of substrates	Various coding systems	Mixture of both structured and unstructured data

3.4 EXAMPLES OF APPROACHES

There are a number of ways in which SNOMED CT can be used in systems to support analytics, including:

- Analyzing free text with clinical Natural Language Processing (NLP) techniques, which use SNOMED CT as a resource;
- Mapping coded clinical data from SNOMED CT to a classification, to enable analysis using the features of the classification;
- Querying clinical data using the machine-processable definitions of clinical concepts defined in SNOMED CT;
- Mapping clinical data captured using a variety of code systems into SNOMED CT, to enable analysis over heterogeneous data using a common reference terminology.

These approaches (and others) are described in more detail in the following chapters.

4 SNOMED CT OVERVIEW

4.1 OVERVIEW

SNOMED CT is a clinical terminology containing concepts, with unique meanings and formal logic-based definitions, organized into hierarchies. The clinical content of SNOMED CT includes diagnoses and other clinical findings, clinical observations, drug products, organisms, specimen types, body structures, and surgical and non-surgical procedures.

SNOMED CT enables clinical information to be consistently represented at an appropriate level of detail within electronic health records. The relationships within SNOMED CT then facilitate meaning-based retrieval of this information at the preferred level of detail for the given query. This provides significant flexibility and facilitates the integration of data from divergent models of use, such as different user interfaces or databases, into convergent models of meaning, such as for the representation of data for reporting or statistical analysis purposes. Clinical systems can thereby query and analyze electronic health record data recorded in different settings, at varying levels of granularity and across multiple axes. This enables SNOMED CT to support a variety of clinical processes, which may require either detailed or high-level information - from investigation, to diagnosis and clinical research.

SNOMED CT content is represented using three main types of component:

- Concepts - unique clinical meanings
- Descriptions - human readable terms used to refer to a concept
- Relationships - links between concepts that help to define the meaning of each concept

In addition to these three types of components, SNOMED CT also supports:

- Expressions – a structured combination of one or more concept identifiers used to represent a new clinical meaning
- Reference sets – a mechanism for representing references to SNOMED CT components for a variety of purposes, including subsets, aggregation hierarchies, maps and language preferences

In this section we introduce these SNOMED CT features and explain how they may be used to support analytics over health data. For more detailed information about SNOMED CT features, please refer to the ‘SNOMED CT Starter Guide’ [2] and the ‘SNOMED CT Technical Implementation Guide’ [4].

We also discuss the specific benefits enabled by using SNOMED CT. For more details about the benefits of SNOMED CT please refer to ‘Building the Business Case for SNOMED CT’ [5].

4.2 CONCEPTS

SNOMED CT concepts represent clinical meanings. Each concept has a permanent concept identifier, which uniquely identifies the clinical meaning. For example:

- 22298006 |myocardial infarction|
- 160341008 |family history: epilepsy|
- 399208008 |plain chest X-ray|
- 319996000 |simvastatin 10mg tablet|

SNOMED CT's concepts, and their logic-based definitions, allow analytics services to perform meaning-based queries, rather than purely lexical (or string-matching) searching.

4.3 DESCRIPTIONS

SNOMED CT descriptions link appropriate human readable terms to concepts. Each concept can have many descriptions, which represent different synonymous ways of referring to the same clinical meaning. Each description is written in a specific language, and new descriptions can be created to support a variety of languages. Like concepts, descriptions also have a permanent unique identifier.

The richness of description content assists the process of searching and finding concepts using user interfaces or database queries. It may also be used to enhance string-matching in natural language processing applications, including analytics over multi-lingual data.

4.4 RELATIONSHIPS

SNOMED CT relationships represent an association between two concepts. Relationships are used to logically define the meaning of concept in a way that can be processed by a computer. A third concept, called a relationship type, is used to represent the meaning of the association between the source and destination concepts. There are different types of relationships available within SNOMED CT.

Subtype relationships, which use the |is a| relationship type, are the most widely used type of relationship. The SNOMED CT concept hierarchy is constructed from |is a| relationships. For example, the concept 128276007 |cellulitis of foot| has an |is a| relationship to both the concept 118932009 |disorder of foot| and the concept 128045006 |cellulitis|. Subtype relationships are used in many analytics scenarios to aggregate groups of concepts together, or to perform queries using more abstract (less detailed) concepts that match more specific (or more detailed) concepts stored in health records.

Attribute relationships contribute to the definition of the source concept by associating it with the value of a defining characteristic. For example, the concept |viral pneumonia| has a |causative agent| relationship to the concept |Virus| and a |finding site| relationship to the concept |lung|. Attribute relationships are used in analytics scenarios in which the meaning of a concept is needed to determine whether a record matches the query criteria.

4.5 CONCEPT MODEL

The rules which define how SNOMED CT concepts may be defined are called the SNOMED CT concept model. The SNOMED CT concept model defines the permitted attributes and values that may be applied to each kind of concept. For example, concepts in the |clinical finding| hierarchy are permitted to have a |finding site| relationship, and the valid values of these relationships must belong to the |anatomical or acquired body structure| hierarchy. The SNOMED CT concept model provides the foundation for processing the clinical meanings recorded in clinical records and enables the appropriate use of clinical information for decision support and other analytics services.

4.6 EXPRESSIONS

An expression is a structured combination of one or more concept identifiers used to represent a clinical meaning. SNOMED CT expressions enable clinical meanings to be represented, which cannot be represented using a single SNOMED CT concept. For example, the following expression represents 'pain in the left thumb:

```
53057004 |hand pain| :
    363698007 |finding site| = (76505004 |thumb structure| :
        272741003 |laterality| = 7771000 |left| )
```

SNOMED CT expressions allow analytics services to perform meaning-based queries over a more extensive set of clinical meanings than just individual concepts.

4.7 REFERENCE SETS

A reference set (or 'refset') is a mechanism used to refer to a set of SNOMED CT components and to add customized information to these components. Reference sets can be used for many different purposes, including representing subsets of concepts, descriptions or relationships, language and dialect preferences, maps to and from other code systems, ordered lists, navigation hierarchies and aggregation hierarchies. For more information about the different types of reference sets, please refer to section 5.6 of the Technical Implementation Guide [4].

Reference sets are used for a range of analytics purposes, including:

- Representing subsets of SNOMED CT concepts with which query criteria are defined and clinical records are matched;
- To represent non-standard aggregations of concepts for specific use cases;
- To define maps from other code systems to SNOMED CT so that clinical data can be prepared for analytics to be performed using SNOMED CT;
- To define language or dialect specific sets of descriptions over which lexical searches can be performed.

4.8 DESCRIPTION LOGIC FEATURES

SNOMED CT concepts are modelled in such a way that their meaning can be represented using a formal family of logics called Description Logic (DL). Description logic enables computers to make inferences about the concepts in SNOMED CT and their meanings, and to classify SNOMED CT using a DL reasoner. Description logic also allows the formal computation of:

- Subsumption – Testing pairs of expressions to see whether one is a subtype of the other
- Equivalence – Testing pairs of expressions to see whether they have the same logic-based meaning

Subsumption and equivalence are both extremely useful functions when retrieving or querying clinical information. For example, when retrieving all clinical records related to 73211009 |diabetes mellitus|, it would usually be necessary to retrieve records referring to any subtype of this concept, such as 23045005 |insulin dependent diabetes mellitus type 1A|.

4.9 BENEFITS OF USING SNOMED CT FOR ANALYTICS

In addition to providing the features already described in this section, SNOMED CT also offers a number of additional benefits for the provision of analytics including:

- SNOMED CT allows clinical data to be recorded at an appropriate level of detail, and then queried at either the same level or a less detailed level of detail;
- SNOMED CT's broad coverage can enable queries across data captured within different disciplines, specialties and domain areas;
- SNOMED CT provides a robust versioning mechanism, which helps to manage queries over longitudinal health records;
- SNOMED CT is international, which enables queries, decision support rules and code system maps to be shared and reused between countries;
- SNOMED CT includes localization mechanisms, which allow the same query to be applied to data from different countries, dialects, regions and applications;
- IHTSDO provides maps between SNOMED CT and other international coding systems and classifications, including LOINC (Logical Observation Identifiers Names and Codes) and ICD (International Classification of Diseases, both ICD-10 and ICD-9-CM). This enables the additional benefits of these other specialized standards to be integrated with the use of SNOMED CT.

Using SNOMED CT to support analytics services can also enable the following benefits:

- Enhancing the care of individual patients by supporting:
  - Retrieval of appropriate information for clinical care – e.g. for a clinical dashboard
  - Guideline and decision support integration
  - Retrospective searches for patterns requiring follow-up
- Enhancing the care of populations by supporting:
  - Epidemiology monitoring and reporting
  - Research into the causes and management of diseases
  - Identification of patient groups for clinical research or specialized healthcare programs
- Providing cost-effective delivery of care by supporting:
  - Guidelines to minimize risk of costly errors
  - Reducing duplication of investigations and interventions
  - Auditing the delivery of clinical services
  - Planning service delivery based on emerging health trends

5 PREPARING DATA FOR ANALYTICS

5.1 OVERVIEW

As discussed in Section 3.3, not all electronic health records represent clinical data using SNOMED CT. However, even when health records use free text or other code systems, it is still possible to use SNOMED CT for analytics over this data if the data is prepared appropriately. For example, Natural Language Processing can be used to encode free text data in SNOMED CT, subsequently enabling more sophisticated analytics to be performed. Similarly, clinical data using other code systems can be mapped into SNOMED CT to provide similar benefits.

In this section we discuss these alternative ways of preparing clinical data for analytics using SNOMED CT.

5.2 NATURAL LANGUAGE PROCESSING

5.2.1 OVERVIEW

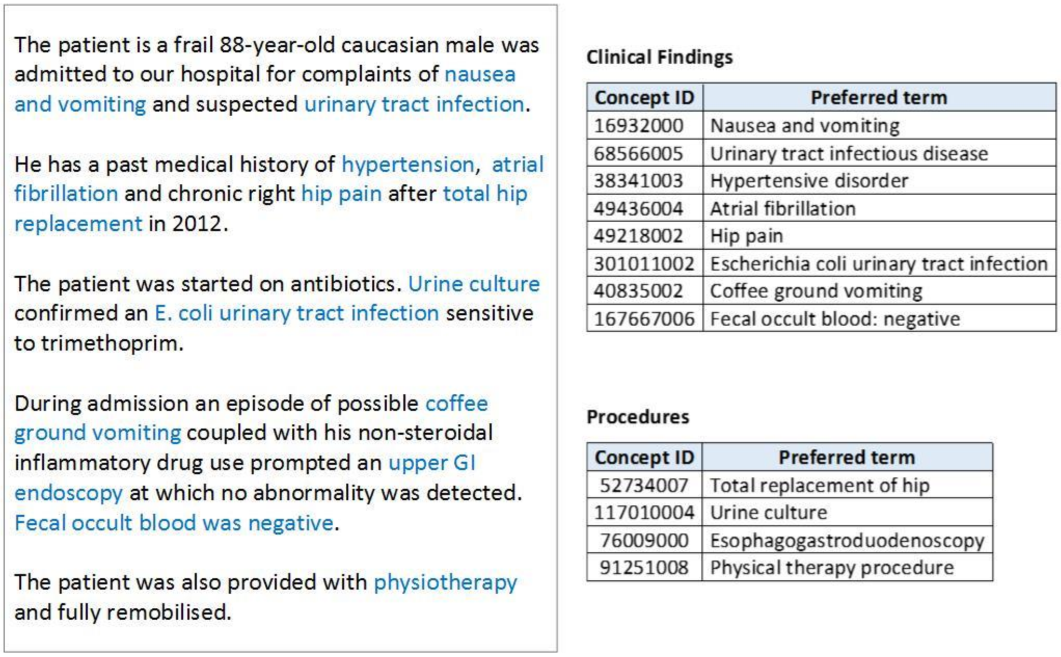
While there is a strong trend towards the direct coding of clinical data, the capture and retention of free text remains essential to record broader narratives about clinical history, physical examinations, clinical procedures and investigation results. Wider deployment of medical transcription technologies featuring speech recognition also add to the volume of free text in electronic format. Medical literature, clinical guidelines and published clinical research also remains largely in free text.

Natural Language Processing (NLP) is a linguistic technique that enables a computer program to analyze and extract meaning from human language. Clinical NLP, using SNOMED CT's concepts, descriptions and relationships, may be applied to repositories of clinical information to search, index, selectively retrieve and analyze free text. These techniques can be used to extract SNOMED CT encoded data from free-text patient records, and also support the retrieval of clinical knowledge documents.

It should be noted that while clinical NLP techniques have increased in sophistication over recent years, it is not possible to guarantee full accuracy or completeness using a computer-based algorithm. Spelling errors, grammatical errors, abbreviations, unexpected synonyms, unusual vernacular (i.e. local) phrases, and hidden contextual information continue to provide challenges that human intelligence is uniquely equipped to handle.

5.2.2 EXAMPLE

The example shown below in Figure 5-1 shows a free text section of a discharge summary that has been processed with clinical NLP to extract a set of potential SNOMED CT clinical findings and procedures. In order to ensure the correctness of this automatic encoding, the application should present this list of extracted codes to the user for confirmation, giving them the opportunity to refine, delete or append codes.



The patient is a frail 88-year-old caucasian male was admitted to our hospital for complaints of nausea and vomiting and suspected urinary tract infection.

He has a past medical history of hypertension, atrial fibrillation and chronic right hip pain after total hip replacement in 2012.

The patient was started on antibiotics. Urine culture confirmed an E. coli urinary tract infection sensitive to trimethoprim.

During admission an episode of possible coffee ground vomiting coupled with his non-steroidal inflammatory drug use prompted an upper GI endoscopy at which no abnormality was detected. Fecal occult blood was negative.

The patient was also provided with physiotherapy and fully remobilised.

Clinical Findings

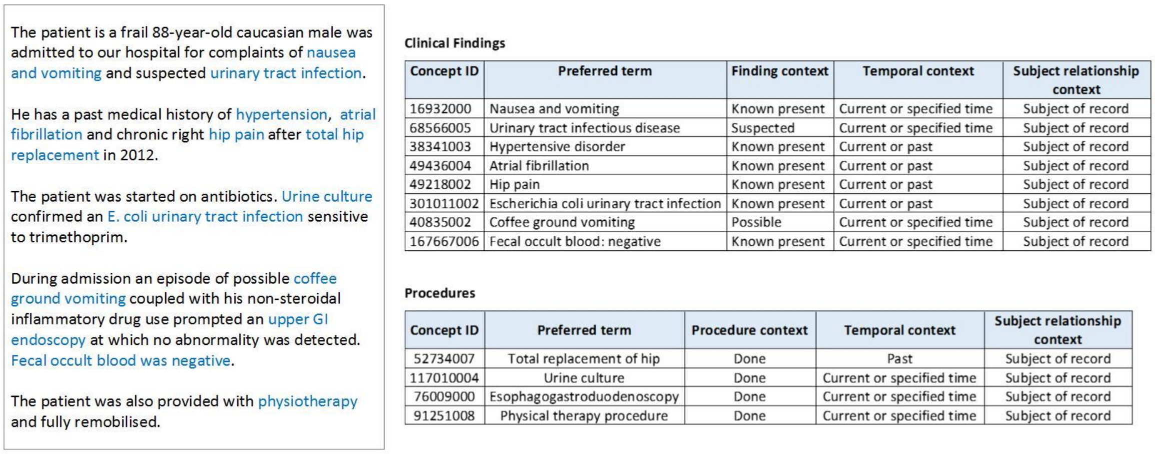
Concept ID	Preferred term
16932000	Nausea and vomiting
68566005	Urinary tract infectious disease
38341003	Hypertensive disorder
49436004	Atrial fibrillation
49218002	Hip pain
301011002	Escherichia coli urinary tract infection
40835002	Coffee ground vomiting
167667006	Fecal occult blood: negative

Procedures

Concept ID	Preferred term
52734007	Total replacement of hip
117010004	Urine culture
76009000	Esophagogastroduodenoscopy
91251008	Physical therapy procedure

Figure 5-1: Natural Language Processing encoding SNOMED CT

To improve the accuracy of clinical NLP and the value for analytics processes, it is important that the context of each statement expressed in natural language is clearly identified – for example, past history, suspected and negation/absence. Figure 5-2 shows the same discharge summary narrative as in Figure 5-1, but this time processed with clinical NLP that also extracts the explicit context of each clinical finding and procedure.



The patient is a frail 88-year-old caucasian male was admitted to our hospital for complaints of nausea and vomiting and suspected urinary tract infection.

He has a past medical history of hypertension, atrial fibrillation and chronic right hip pain after total hip replacement in 2012.

The patient was started on antibiotics. Urine culture confirmed an E. coli urinary tract infection sensitive to trimethoprim.

During admission an episode of possible coffee ground vomiting coupled with his non-steroidal inflammatory drug use prompted an upper GI endoscopy at which no abnormality was detected. Fecal occult blood was negative.

The patient was also provided with physiotherapy and fully remobilised.

Clinical Findings

Concept ID	Preferred term	Finding context	Temporal context	Subject relationship context
16932000	Nausea and vomiting	Known present	Current or specified time	Subject of record
68566005	Urinary tract infectious disease	Suspected	Current or specified time	Subject of record
38341003	Hypertensive disorder	Known present	Current or past	Subject of record
49436004	Atrial fibrillation	Known present	Current or past	Subject of record
49218002	Hip pain	Known present	Current or past	Subject of record
301011002	Escherichia coli urinary tract infection	Known present	Current or past	Subject of record
40835002	Coffee ground vomiting	Possible	Current or specified time	Subject of record
167667006	Fecal occult blood: negative	Known present	Current or specified time	Subject of record

Procedures

Concept ID	Preferred term	Procedure context	Temporal context	Subject relationship context
52734007	Total replacement of hip	Done	Past	Subject of record
117010004	Urine culture	Done	Current or specified time	Subject of record
76009000	Esophagogastroduodenoscopy	Done	Current or specified time	Subject of record
91251008	Physical therapy procedure	Done	Current or specified time	Subject of record

Figure 5-2: Natural Language Processing encoding SNOMED CT with context

When SNOMED CT codes with explicit context are extracted from free text narrative, the resulting clinical meanings may be captured using SNOMED CT postcoordinated expressions. For example, the following clinical statement:

> *Endoscopy revealed an acute gastric ulcer but no evidence of gastric bleeding or perforation of the stomach.*

can be encoded using the following SNOMED CT expressions with explicit context (case study 2.9):

- 243796009 |situation with explicit context| :
  {408731000 |temporal context| = 410512000 |current or specified time|,
  246090004|associated finding| = 95529005 |acute gastric ulcer|,
  408732007 |subject relationship context| = 410604004 |subject of record|,
  408729009 |finding context| = 410515003 |known present|
- 243796009 |situation with explicit context| :
  {408729009 |finding context| = 410516002 |known absent|,
  246090004 |associated finding| = 61401005 |gastric bleeding|,
  408731000 |temporal context| = 410512000 |current or specified|,
  408732007 |subject relationship context| = 410604004 |subject of record|}
- 243796009 |situation with explicit context| :
  {408729009 |finding context| = 410516002 |known absent|,
  246090004 |associated finding| = 235674005 |perforation of stomach|,
  408731000 |temporal context| = 410512000 |current or specified|,
  408732007 |subject relationship context| = 410604004 |subject of record|}

5.2.3 IMPLEMENTATION

5.2.3.1 NLP TECHNIQUES USING SNOMED CT

A clinical NLP engine can use SNOMED CT to encode free text narrative in patient records in a number of ways. Firstly, it can use SNOMED CT descriptions together with techniques such as:

- Stemming: The process of reducing a word to its stem, base or root form – for example "cardiology", "cardiac" and "cardiologist" may be reduced to the stem "cardi".
- Reordering: The process of reordering the words in a phrase – for example, reordering "hip fracture" to "fracture hip".
- Word substitution: The process of substituting a word or word phrase with an equivalent word or word phrase. The SNOMED CT Lexical Resources zip file [7], available from the SNOMED CT Document Library, includes an English Word Equivalents table that groups together equivalent words and phrases – for example, "Renal stone", "Kidney stone", "kidney calculus", "renal calculus" and "nephrolith" are grouped into the same word block group. This table can be modified or extended with additional word equivalent groups if required.
- Stop word removal: The process of removing words with limited semantic specificity – for example 'a', 'an', 'and', 'as', 'at', 'be', 'by', 'for', 'of', 'the'. The SNOMED CT Lexical Resources zip file [7], available from the SNOMED CT Document Library, includes an Excluded Words table, which suggests some common English stop words that may be used with SNOMED CT.

The SNOMED CT concept model can also be used to identify potential connections between related concepts – for example, the words "left", "hip" and "fracture" used in close proximity may indicate a |fracture| with finding site |hip| and a laterality of |left|. Similarly, the SNOMED CT concept model may help to identify context that is expressed within the text – for example, past history, certainty and absence.

Another commonly adopted NLP strategy is to use the location of the free text within the structure of a document to restrict the possible SNOMED CT code matches. For example, free text entered into a

'Diagnosis' field may restrict its SNOMED CT encoding to the |disorder| hierarchy, together with other concepts that may be linked to |clinical findings| via the SNOMED CT concept model.

When NLP techniques are applied to non-English (or dialect-specific) text, translations of relevant SNOMED CT descriptions may be required. The NLP methods themselves may also need to be adapted to reflect the structure and style of the language in which the text is written.

5.2.3.2 INDEXING

Another major application for Natural Language Processing technologies is indexing collections of free text transcripts or documents such that topic specific searches may be run on them, or relevant clinical knowledge sources may be identified and linked to a given patient's clinical data. The challenge is to return ranked matches which permit selection of texts with high sensitivity and high specificity (i.e. that relevant documents are rarely overlooked and that irrelevant documents are rarely returned).

SNOMED CT can be used to support these applications by enabling more powerful searching of free text data stores than using a purely lexical keyword-based approach. For example, the clinician may request "all documents which refer to cardiac rhythm disorders". Rather than relying purely on text matching, the search term may be matched with the concept 698247007 |cardiac arrhythmia (disorder)|, based on its synonym |disorder of heart rhythm|. The descendants of this concept (e.g. 276796006 |atrial tachycardia|, 49260003 |idioventricular rhythm|, 233917008 |atrioventricular block|) may then be used to search for any code which is a kind of cardiac arrhythmia. Non-|is a| attribute relationships may also be used in the retrieval process to find associations between the search term and the indexed concepts, and to calculate the relevance of each free text artefact to determine the order in which they should be presented to the user.

5.2.3.3 CASE STUDIES

Clinical NLP has been implemented for encoding free text narrative in health records by a number of vendors, including Caradigm (case study 2.7), Cerner (case study 2.8), Clinithink (case study 2.9) and Intelligent Medical Objects (case study 2.13).

NLP techniques for indexing and searching have also been implemented by Cerner (case study 2.8) and Dr Bevan Koopman (case study 1.5). Allscript's Sunrise InfoButton™ feature (case study 2.3) uses encoded patient problem lists and medication data elements, together with SNOMED CT-based indexes provided by third-party medical content providers, to present on-topic information to the clinician without manual searching.

5.3 MAPPING OTHER CODE SYSTEMS TO SNOMED CT

5.3.1 OVERVIEW

Mapping data from clinical records encoded using non-SNOMED CT code systems to SNOMED CT for analysis may be considered when there is a requirement to produce:

- Management information for care service audit or delivery planning
- Statistical information for epidemiology
- Links from clinical records to clinical knowledge resources
- Links between clinical records and decision support tools

- An integrated data warehouse for querying from multiple heterogeneous sources
- Other types of research, reports or surveillance that requires SNOMED CT

Two important characteristics of a map, which affect its ability to be used for a particular purpose, are the direction of the map, and the correlation between the source and target codes. Where the analytics use case requires SNOMED CT to be used, the direction of the map must be *from* the non-SNOMED CT codes *to* SNOMED CT codes. A map designed to move data from code system A to code system B will serve poorly (if at all) 'in reverse' if it is used to map from B to A, unless *all* the links are exact semantic matches.

For analytics purposes where patient safety or data accuracy is important (e.g. point of care clinical decision support or data integration), it is important that the correlation of the map is an 'exact match' (or equivalence). For other purposes (e.g. epidemiology or care service delivery planning) it may be acceptable for the SNOMED CT code to be broader than (or a supertype of) the non-SNOMED CT code. However, broad-to-narrow and narrow-to-broad maps need to be used with care.

When a non-SNOMED CT code is being mapped into SNOMED CT, and an equivalent precoordinated SNOMED CT concept does not exist, a number of options are possible, including:

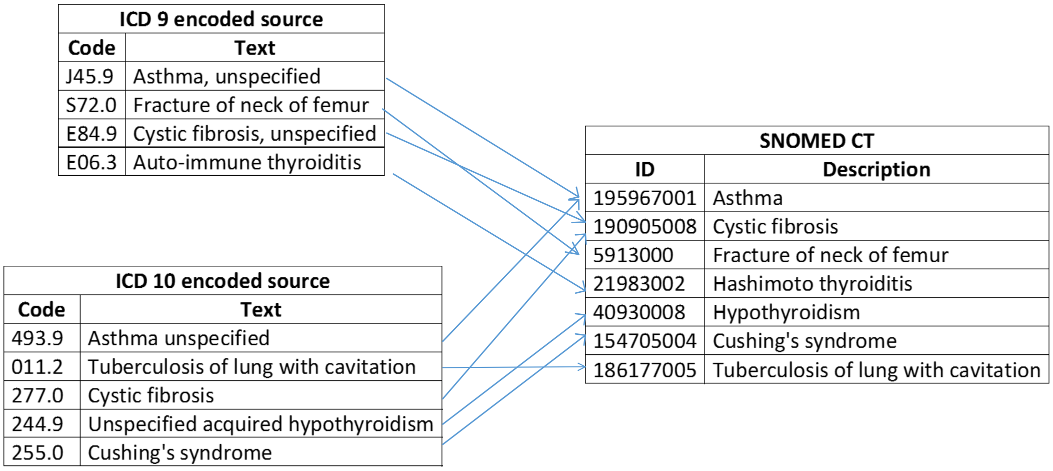
1. Map the code to a broader (supertype) SNOMED CT concept
    a. For example, map "DX0162: arthritis of left knee" to "371081002 |arthritis of knee|" with correlation 'broad to narrow'
2. Map the code to a SNOMED CT postcoordinated expression
    a. For example, map "DX0162: arthritis of left knee" to "371081002 |arthritis of knee| : 272741003 |laterality| = 7771000 |left|" with correlation 'exact match'
3. Create a new precoordinated SNOMED CT concept with the same meaning as the code, and map the code to this new concept
    a. For example, map "DX0162: suspected gastric ulcer" to a new extension concept 1111000000105 |arthritis of left knee|[1]

Designing and authoring maps requires expertise and appropriate resources. Large maps (e.g. tens of thousands of codes) are typically created and maintained by the IHTSDO, National Release Centers, large healthcare organizations, specialist data suppliers and large system vendors. However, smaller maps may be created and maintained by smaller system suppliers, hospitals or clinics. Maps must be maintained to ensure that both the SNOMED CT content and non-SNOMED CT content remains current whenever either code system is updated.

5.3.2 EXAMPLE

A typical scenario requiring mapping to SNOMED CT is shown below in Figure 5-3. In this example, two source systems (using ICD-9 and ICD-10 respectively) are being integrated into a data warehouse using SNOMED CT as the common 'reference terminology' for analysis. Once this mapping is done, the same analytic techniques as used on native SNOMED CT records may be applied (See Section 6).

[1] Please note that this concept does not exist in the international edition of SNOMED CT, but is shown here as a hypothetical example of a concept added in a SNOMED CT extension.



ICD 9 encoded source	
Code	**Text**
J45.9	Asthma, unspecified
S72.0	Fracture of neck of femur
E84.9	Cystic fibrosis, unspecified
E06.3	Auto-immune thyroiditis

ICD 10 encoded source	
Code	**Text**
493.9	Asthma unspecified
011.2	Tuberculosis of lung with cavitation
277.0	Cystic fibrosis
244.9	Unspecified acquired hypothyroidism
255.0	Cushing's syndrome

SNOMED CT	
ID	**Description**
195967001	Asthma
190905008	Cystic fibrosis
5913000	Fracture of neck of femur
21983002	Hashimoto thyroiditis
40930008	Hypothyroidism
154705004	Cushing's syndrome
186177005	Tuberculosis of lung with cavitation

Figure 5-3: Mapping from ICD classifications to SNOMED CT

5.3.3 IMPLEMENTATION

5.3.3.1 MAPPING USING SNOMED CT

Maps are represented in SNOMED CT's RF2 using a Simple map reference set, a Complex map reference set, or an Extended map reference set (depending on what additional information is required to support the implementation of the map). Code mappings are then performed by matching each non-SNOMED CT code in a patient's record with the 'mapTarget' field of the corresponding row of the map reference set, and using the SNOMED CT code found in the 'referencedComponentId'.

5.3.3.2 CASE STUDIES

The UK Terminology Centre's Data Migration Workbench demonstrates some advanced uses of data migration and mapping products published by the UKTC, including Read Code Version 2 and CTV3 maps to SNOMED CT (case study 1.2). A number of vendor products also map non-SNOMED CT codes to SNOMED CT for use in analytics, including Allscript's terminology service (case study 2.3), Apelon's Distributed Terminology System (case study 2.4), the Cerner Millennium Terminology (CMT) package (case study 2.8), and Epic's electronic patient record systems (case study 2.11).

6 SNOMED CT ANALYTIC TECHNIQUES

6.1 OVERVIEW

SNOMED CT offers a number of analytics techniques, which are not possible using other coding systems. SNOMED CT's hierarchical design improves upon the purely lexical query capabilities of free text lists or 'flat' controlled vocabularies. For example, a purely text based query for 'kidney disease' will not return the kidney disease 'glomerulonephritis'. Purely mono-hierarchies, however, limit querying to a single grouping of each code. For example, using a mono-hierarchy 'tuberculosis of the lung' must be assigned a code which makes it *either* a kind of 'lung disease' *or* a kind of 'tuberculosis' – however it cannot be both. Using SNOMED CT's polyhierarchy 'tuberculosis of the lung' can be represented as both a kind of 'lung disease' *and* a kind of 'tuberculosis'. The inclusion of other attribute-based defining relationships and the ability to represent SNOMED CT using OWL 2 EL enables additional Description Logic techniques for classifying and querying SNOMED CT. Extending these capabilities even further, it is possible to use Description Logic techniques across both the terminology and the structure of the patient records in which the codes are stored. Finally, in some specific use cases such as billing, reimbursement and statistics where double counting must be avoided, clinically recorded SNOMED CT codes can be used to map into more general statistical classifications, such as ICD (International Classification of Diseases).

In this section, we describe how the following analytics techniques can be used to support analytics over SNOMED CT enabled data. The techniques described include:

- Subsets – for example, find the patients with a diagnosis in the set of 'kidney disease codes'
- Subsumption – for example, find the patients with a diagnosis that is a subtype (or self) of 'kidney disease'
- Using defining relationships – for example, find the patients whose diagnosis has a finding site of 'kidney structure' (or a subtype of 'kidney structure')
- Description logic over terminology – for example, find the patients whose diagnosis is associated (directly or indirectly) with the 'Streptococcus pyogenes organism'
- Description logic over terminology and structure – for example, find the patients with a family history of heart disease (where this may either be recorded as 275120007 |family history: cardiac disorder| or recorded in a 'Family History' section on a form as 56265001 |heart disease|)
- Using statistical classifications – for example, to meet national reporting guidelines using ICD (International Classification of Diseases)

In practice, a query language may combine a number of these techniques in the same syntax. With the possible exception of the last two approaches, these SNOMED CT query techniques should then be embedded within an EHR query to ensure that the semantic context of the surrounding patient record is taken into account.

6.2 SUBSETS

6.2.1 OVERVIEW

One approach to analytics using SNOMED CT is to construct subsets of SNOMED CT identifiers, which are applicable to a specific clinical purpose, and to test the codes recorded in patient records to check for membership in the appropriate subset. Subsets of SNOMED CT identifiers may either be defined extensionally or intensionally.

Extensionally defined subsets are those in which each concept is individually enumerated. They are usually manually constructed and maintained, and can therefore be labor intensive and error prone. For example, one might construct a subset of kidney disease codes including 36171008 |glomerulonephritis|, 71110009 |hydrocalycosis| and 42399005 |renal failure|.

Intensionally defined subsets are those which are automatically populated (or expanded) based on a machine processable query. For example, one might construct a subset of kidney disease codes using the results of the query "<< 90708001 |kidney disease|" (i.e. descendants or self of 90708001 |kidney disease|). The query used to define an intensional subset may utilize SNOMED CT's hierarchical relationships, attribute values, descriptions, and membership in other intensionally or extensionally defined subsets. For more information about SNOMED CT query languages, which may be used to define intensional subsets, please refer to Section 9.

6.2.2 EXAMPLE

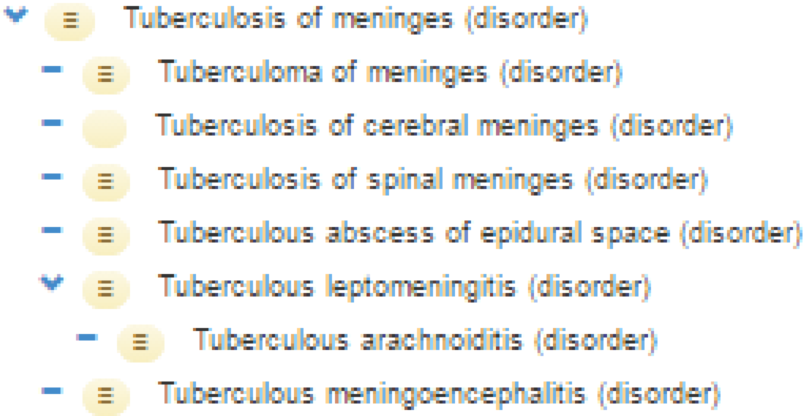
A subset containing types of 58437007 |tuberculosis of meninges| may be defined extensionally as follows:

Concept ID	Description
58437007	tuberculosis of meninges (disorder)
35786001	tuberculoma of meninges (disorder)
90302003	tuberculosis of cerebral meninges (disorder)
38115001	tuberculosis of spinal meninges (disorder)
447332005	tuberculous abscess of epidural space (disorder)
11676005	tuberculous leptomeningitis (disorder)
447253004	tuberculous arachnoiditis (disorder)
31112008	tuberculous meningoencephalitis (disorder)

With the help of the SNOMED CT hierarchy (as shown in Figure 6-1), this same subset can be defined intensionally as:

<< 58437007 |tuberculosis of meninges|

The expansion of an intensional subset defined using this query is the same as the extensionally defined subset shown above.



Tuberculosis of meninges (disorder)
Tuberculoma of meninges (disorder)
Tuberculosis of cerebral meninges (disorder)
Tuberculosis of spinal meninges (disorder)
Tuberculous abscess of epidural space (disorder)
Tuberculous leptomeningitis (disorder)
Tuberculous arachnoiditis (disorder)
Tuberculous meningoencephalitis (disorder)

Figure 6-1: Tuberculosis of meninges concept sub-hierarchy

Using a lexical query, it is also possible to intensionally define a subset of 'tuberculosis of meninges' findings. However, the results of purely lexical queries are not as reliable. For example, using the query:

<< 404684003 |clinical finding| {{ term = ".*tuberculosis.*meninges.*" }}

the following expansion can be calculated:

Concept ID	Description
58437007	tuberculosis of meninges (disorder)
90302003	tuberculosis of cerebral meninges (disorder)
38115001	tuberculosis of spinal meninges (disorder)

As can be seen, the results of this lexical query only includes 3 of the possible 8 values from the previous subset. In other cases, lexical queries may incorrectly find concepts which are not appropriate to the subset. It is therefore recommended that lexical queries are avoided in the definition of intentional subsets. However, they do serve a useful purpose in identifying candidates for an otherwise manually crafted subset.

6.2.3 IMPLEMENTATION

6.2.3.1 DEFINING SUBSETS IN SNOMED CT

Subsets of SNOMED CT may be defined locally as a flat list of concept identifiers, or as an independent query specification. However, where wider distribution and/or version control is required over these subsets, SNOMED CT reference sets offer the ideal solution.

Extensional subsets are commonly defined in SNOMED CT using a Simple reference set - however an Ordered reference set or Annotation reference set can be used if additional information needs to be recorded for each member of the subset. Intensional subsets are defined in SNOMED CT using a Query specification reference set. A Query specification reference set allows a serialized query to define the membership of a subset of SNOMED CT components. It also specifies the extensional reference set into which the results of executing the query are generated. Intensional reference sets are preferred in many circumstances as they enable their membership to be automatically recomputed over new versions of SNOMED CT. Version management of subsets is discussed further in Section 11.5.

Subsets can be created using the following methods, either alone or in any combination:

- Manual inclusion, using search and browse methods
- Existing subset, used as a starting point for further manual inclusion and update
- Lexical queries, to identify candidate members, followed by manual verification and update
- Hierarchical queries, to identify descendants of a given concept (e.g. descendants of <73211009 |diabetes mellitus|)
- Attribute queries, to identify concepts with a specific attribute value (e.g. disorders with a finding site of 80891009 |heart structure|)
- SNOMED CT queries, using the SNOMED CT Expression Constraint or Query languages, which offer additional query functionality. Please refer to Section 9 for more details.

6.2.3.2 CASE STUDIES

A number of vendor products, such as Apelon (case study 2.4) and B2i Healthcare (case study 2.5) allow users to create customized extensional and intensional subsets of SNOMED CT. Other vendor products, such as the Cambio COSMIC® Electronic Patient Record system (case study 2.6), Caradigm's population health solutions (case study 2.7), Cerner's data warehousing solution (case study 2.8) and Epic's decision support and reporting tools (case study 2.11) use subsets of SNOMED CT to support their analytics services.

6.3 SUBSUMPTION

6.3.1 OVERVIEW

Determining whether one concept (or expression) is a kind of another concept (or expression) is the fundamental capability enabled by SNOMED CT. For example, answering the question 'Which patients have an infectious disease?' involves finding all the patients with *any kind of* infectious disease (e.g. viral pneumonia, tuberculosis).

Subsumption occurs when one clinical meaning is a subtype of another clinical meaning, and testing for this is called 'subsumption testing'. If clinical meaning X is a subtype of clinical meaning Y, then Y is said to 'subsume' X and X is 'subsumed by' Y.

Subsumption testing between concepts is represented using a stated or implied |is a| relationship. For example, 75570004 |viral pneumonia| is a 40733004 |infectious disease| and therefore 40733004 |infectious disease| subsumes 75570004 |viral pneumonia|, and 75570004 |viral pneumonia| is subsumed by 40733004 |infectious disease|.

Subsumption testing between expressions tests to see if the *candidate expression* (often recorded in a patient record) is subsumed by a *predicate expression* (typically part of the query being run across the patient record). For example:

Candidate expression: 75570004 |viral pneumonia|

Predicate expression: 40733004 |infectious disease|:
363698007 |finding site| = 39607008 |lung structure|

In this case, the candidate expression *is* subsumed by the predicate expression.

Subsumption testing can be represented using the SNOMED CT Expression Constraint Language using the '<' (descendantOf) or '<<' (descendantOrSelfOf) operators. For example, the expression constraint:

<< 40733004 |infectious disease|

is satisfied by any expression that is subsumed by 40733004 |infectious disease|.

There are a variety of ways to implement subsumption testing. These are summarized in Section 0.

6.3.2 EXAMPLE

A typical example using subsumption would be an audit within a hospital, reviewing all patients with an infectious disease. In this scenario, the following simple query could be executed to find all the patients whose health record contains a diagnosis that is subsumed by the concept 40733004 |infectious disease|:

```
SELECT distinct patientID
FROM health_records
WHERE diagnosis = (<< 40733004 |infectious disease|)
```

If the health records contained the following data:

patientID	date	diagnosis
634711	16th January 2015	71620000 |fracture of femur|
634711	25th January 2015	415353009 |rotavirus food poisoning|
634711	3rd February 2015	66308002 |fracture of humerus|
158775	7th January 2015	40468003 |hepatitis A|
889125	7th January 2015	75570004 |viral pneumonia|
456872	15th January 2015	22298006 |myocardial infarction|
456872	15th January 2015	195967001 |asthma|

Then this query would return the following list of patients:

- 634711 (because 415353009 |rotavirus food poisoning| is a subtype of 40733004 |infectious disease|)
- 158775 (because 40468003 |hepatitis A| is a subtype of 40733004 |infectious disease|)
- 889125 (because 75570004 |viral pneumonia| is a subtype of 40733004 |infectious disease|)

Note that patient 456872 would not be returned by this query as neither 22298006 |myocardial infarction| or 195967001 |asthma| are subtypes of 40733004 |infectious disease|.

6.3.3 IMPLEMENTATION

6.3.3.1 TESTING SUBSUMPTION BETWEEN CONCEPTS

Rapid and efficient computation of whether a concept |is a| subtype descendant of another concept is essential for testing subsumption between expressions. A variety of approaches exist for testing subsumption. When the candidate and predicate expressions are both precoordinated concepts, subsumption testing can use the published relationships from the SNOMED CT release files. Approaches for testing subsumption between precoordinated concepts include:

- Exhaustive testing of subtype relationships
  In this approach, every possible sequence of |is a| relationships are recursively tested from the candidate concept until the predicate concept is reached or until all possible paths have been exhausted.
- Semantic type identifiers and hierarchy flags
  In this approach, flags are added to each concept to indicate the set of high-level concept nodes of which that concept is a subtype. A concept can only subsume concepts that include the same set of high-level concept flags. This reduces the number of tests that need to be performed to recursively test the subtype relationships.
- Use of proprietary database features
  In this approach, a database is used which supports the recursive testing of a chain of hierarchical relationships.
- Branch numbering
  In this approach, a depth first tree walk is performed that applies an incremental number to each concept. A second tree walk then allocates one or more branch number ranges to each concept, which contains the number of all of their descendants.
- Precomputed transitive closure table
  In this approach, a comprehensive list of all supertypes of each concept is created by recursively traversing all |is a| relationships and adding each stated and inferred subtype relationship to a table.
- Using a Description Logic Reasoner
  In this approach, a description logic reasoner (e.g. Snorocket, ELK, Fact++) is used to determine whether one concept is subsumed by another.

In most environments, the recommended approach is to either use a precomputed transitive closure table or a description logic reasoner. However, where disk capacity or distribution bandwidth are limiting factors, branch numbering provides an efficient alternative approach. For more information on these approaches, please refer to section 7.7.4 of the SNOMED CT Technical Implementation Guide [4].

6.3.3.2 TESTING SUBSUMPTION BETWEEN EXPRESSIONS

When either the candidate expression (in the patient data) or the predicate expression in the query are postcoordinated (or both), techniques based on description logic are needed to perform subsumption testing. Approaches for testing subsumption between postcoordinated expressions include:

- Comparing normal form expressions
  In this approach, the predicate expression is transformed to short normal form and the candidate expression is transformed to long normal form. The two normal form expressions are then tested for subsumption by checking that each focus concept in the predicate

expression subsumes at least one focus concept in the candidate expression, each attribute group in the predicate expression subsumes at least one attribute group in the candidate expression and each ungrouped attribute in the predicate expression subsumes at least one attribute in the candidate expression.

- Using a Description Logic Reasoner
  In this approach, a description logic reasoner (e.g. Snorocket, ELK, Fact++) is used to determine whether one expression is subsumed by another.

Where available, the recommended approach is to use a description logic reasoner to calculate subsumption between expressions. However, comparing normal form expressions provides an alternative approach when a reasoner is not available. For more information on these approaches, please refer to section 7.8.2.4 of the SNOMED CT Technical Implementation Guide [4].

6.3.3.3 CASE STUDIES

A number of vendor products use the SNOMED CT hierarchy to support subsumption testing in their analytics services, including the Cerner Millennium Terminology (CMT) package (case study 2.8) and Epic's decision support and reporting tools (case study 2.11). Terminology servers that provide the ability to perform subsumption testing include B2i Healthcare's Snow Owl® terminology server (case study 2.5). The UK Terminology Centre's Data Migration Workbench also uses subsumption testing in its query tool, and its case mix and caseload trends analysis tools (case study 1.2).

6.4 USING DEFINING RELATIONSHIPS

6.4.1 OVERVIEW

SNOMED CT attributes are used to represent a characteristic of the meaning of a concept. There are more than 50 attributes in SNOMED CT, which can each be used as the 'type' of a defining relationship, including:

- 363698007 |finding site|
- 116676008 |associated morphology|
- 246075003 |causative agent|
- 363704007 |procedure site|
- 260686004 |method|
- 272741003 |laterality|
- 127489000 |has active ingredient|

The SNOMED CT Concept Model provides rules about how these attributes can be used. Some database queries use the rules from the SNOMED CT Concept Model to match concepts based on the value of their defining relationships.

6.4.2 EXAMPLE

Figure 6-2 illustrates the execution of a query to retrieve a set of findings which have a benign tumor morphology. The query is executed by finding those concepts with an 'associated morphology' relationship with the value ‘benign neoplasm’. In this example, the concepts ‘benign tumor of kidney’, ‘benign neoplasm of bladder’ and ‘benign tumor of lung’ are found to have the required defining relationship value.

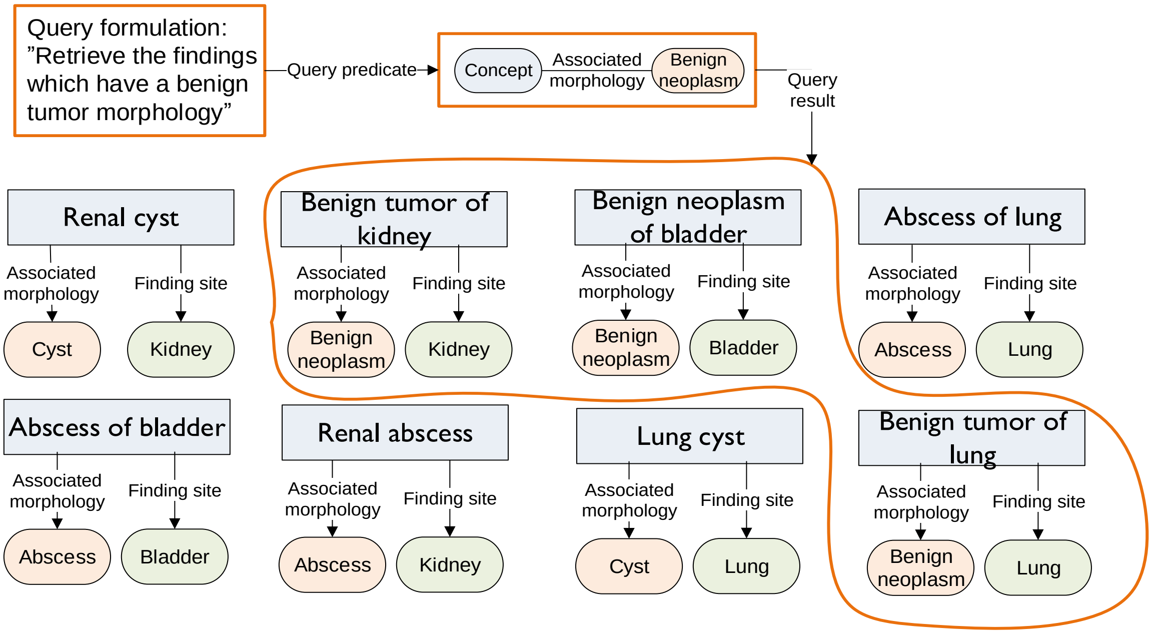


Figure 6-2: Query to retrieve benign neoplasm findings

In Figure 6-3 the same set of concepts are shown analyzed with the intention to identify those which have a finding site of kidney. In this example, the concepts ‘renal cyst’, ‘benign tumor of kidney’ and ‘renal abscess’ are found to have the required defining relationship value.

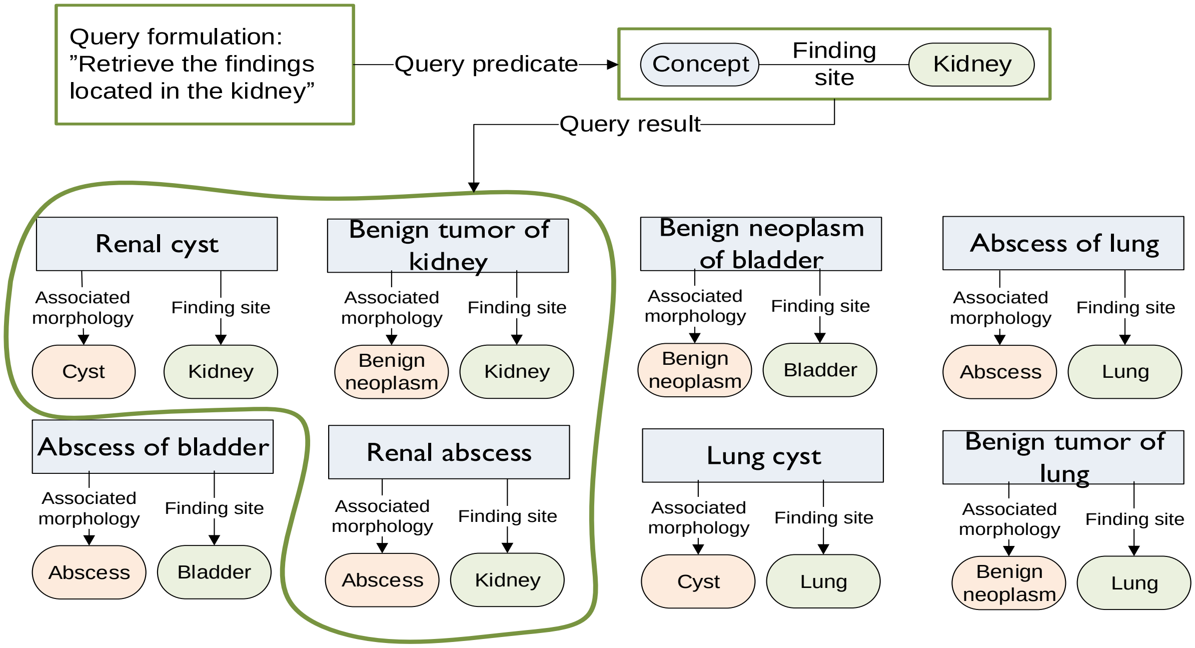


Figure 6-3: Query to retrieve findings in the kidney

If the queries from Figure 6-2 and Figure 6-3 are combined, then the query will return those concepts which are benign tumors of the kidney (see Figure 6-4). In this case, the concept ‘benign tumor of kidney’ is the only concept found to have the required defining relationship values.

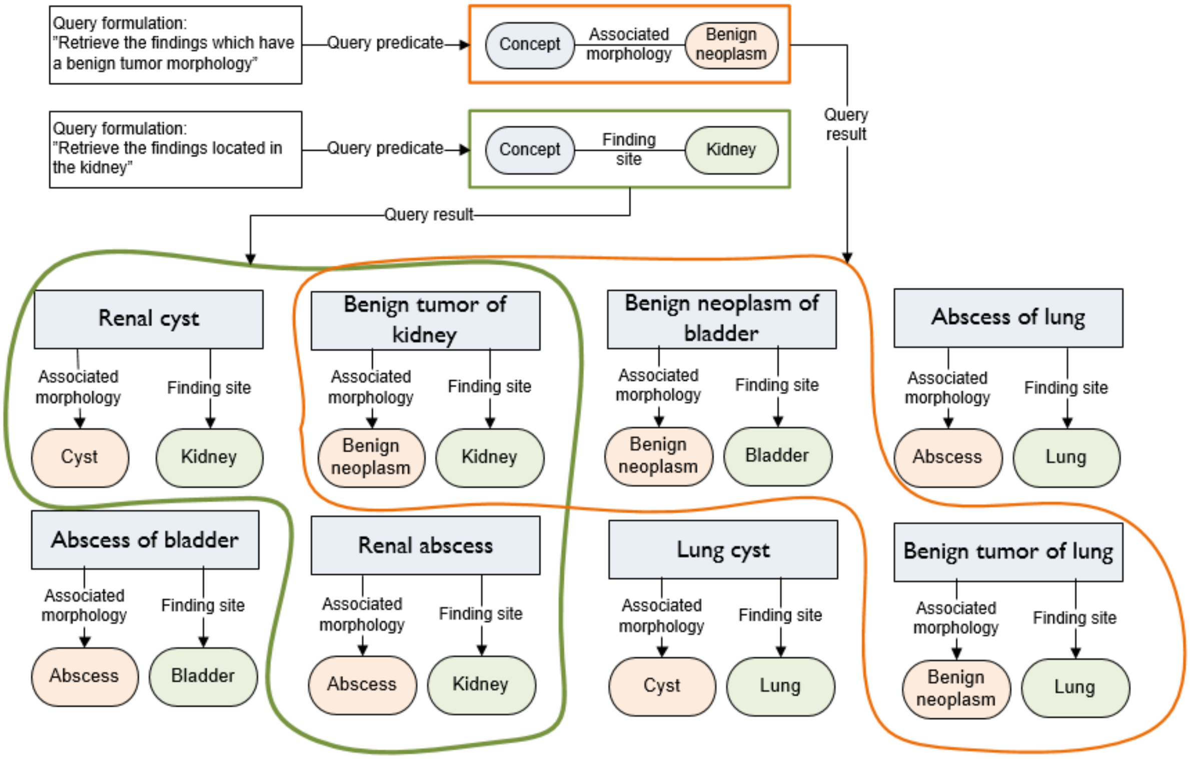


Figure 6-4: Query to retrieve benign neoplasms of the kidney

In most cases, these queries would be designed to return concepts with an associated morphology of ‘benign neoplasm’ *or any subtype of* ‘benign neoplasm (e.g. ‘angiomyolipoma’), and a finding site of ‘kidney’ *or any subtype of* ‘kidney’ (e.g. ‘papillary duct of kidney’, or ‘upper pole, left kidney’). This query could be expressed using the SNOMED CT Expression Constraint Language (see Section 9) as:

```
< 404684003 |clinical finding|:
    116676008 |associated morphology| = << 3898006 |benign neoplasm| AND
    363698007 |finding site| = << 64033007 |kidney structure|
```

When executed against the January 31st 2015 international edition of SNOMED CT, this query would return the following 12 concepts:

Concept ID	Preferred Term
254925008	Benign tumor of renal calyx
254919009	Cortical adenoma of kidney
269489006	Benign tumor of renal parenchyma
254920003	Cystadenoma of kidney

Concept ID	Preferred Term
254922006	Oncocytoma of kidney
276866009	Benign tumor of pelviureteric junction
254927000	Benign papilloma of renal pelvis
92319008	Benign neoplasm of renal pelvis
307618001	Juxtaglomerular tumor
254923001	Hemangiopericytoma of kidney
254921004	Angiomyolipoma of kidney
92165001	Benign neoplasm of kidney

6.4.3 IMPLEMENTATION

6.4.3.1 QUERIES OVER DEFINING RELATIONSHIPS

A query, which constrains the defining relationships of matching clinical meanings to specific values can either be represented informally using a set of attribute value pairs, or represented more formally using a machine processable language (e.g. the SNOMED CT Expression Constraint Language). Please see Section 9 for more information.

Approaches to implement such a query include:

- Using the distributed relationships
  In this approach, the distributed Relationship file is used directly to compare the target value of each defining relationship with the required attribute value in the query. This approach may be combined with a subsumption testing approach (e.g. transitive closure table) to enable subtypes of the required attribute value to also be matched.
- Comparing normal form expressions
  In this approach, the query is represented as a predicate expression containing the constrained attribute values, and the short normal form of this predicate expression is tested for subsumption against each candidate expression (as per the normal form subsumption test in section 6.3.3.2).
- Using a Description Logic Reasoner
  In this approach, a description logic reasoner (e.g. Snorocket, ELK, Fact++) is used to determine whether each candidate expression is subsumed by the query (represented by a predicate expression).

6.4.3.2 CASE STUDIES

Many organization-wide implementations of SNOMED CT, such as Kaiser Permanente's HealthConnect EHR (case study 1.3) and the Danish National Medication Decision Support System (case study 1.4), are taking advantage of SNOMED CT's definitional attributes to support advanced analytics.

A number of vendor products are also supporting analytics over SNOMED CT's defining relationships, including Apelon's Distributed Terminology System (case study 2.4), B2i Healthcare's SnowOwl terminology server (case study 2.5), and Cerner's Semantic Search tool (case study 2.8).

6.5 DESCRIPTION LOGIC OVER TERMINOLOGY

6.5.1 OVERVIEW

SNOMED CT's semantics are based on Description Logic (DL). This enables the automation of reasoning across SNOMED CT, and subsequently the implementation of more powerful analytics operations than is possible using most other approaches. In addition to the subsumption and defining relationship testing described in the previous approaches, DL reasoners and query engines are able to utilize a number of additional logic-based techniques including:

- Property chaining

  A property chain is a rule that allows you to infer the existence of a property from a chain of properties. For example, "x has parent y" and "y has parent z" implies "x has grandparent z" (which may be written as "|has parent|o|has parent|→|has grandparent|). The current release of SNOMED CT includes the property chain:

  363701004 |direct substance| o 127489000 |has active ingredient|
  → 363701004 |direct substance|

  However, more property chains may be added in local implementations if required.
- Reasoning over concrete values

  Some concepts in SNOMED CT (e.g. 374646004 |amoxicillin 500mg tablet|) require numbers or strings to fully define their meaning. By generating an OWL 2 representation of these concept definitions, Description Logic can be used to reason over their complete definition (including the concrete values)
- Testing equivalence and subsumption of postcoordinated expressions (without calculating normal forms)

  Description Logic enables equivalence and subsumption testing to be performed efficiently, without the need to manually calculate the normal form of each expression.
- Reasoning over minimum sufficient sets

  SNOMED CT definitions include the set of necessary and sufficient conditions that define the given concept. However, SNOMED CT does not currently distinguish the minimum sets which are sufficient to define these concepts. For example, the defining relationships of 154283005 |pulmonary tuberculosis| are:

  116680003 |is a| = 64572001 |disease|
  246075003 |causative agent| = 113858008 |mycobacterium tuberculosis complex|
  116676008 |associated morphology| = 6266001 |granulomatous inflammation|
  363698007 |finding site| = 39607008 |lung structure|

  However, while the associated morphology of 'granulomatous inflammation' is necessarily present, the following set of defining relationships are sufficient to infer 154283005 |pulmonary tuberculosis|:

  116680003 |is a| = 64572001 |disease|
  246075003 |causative agent| = 113858008 |mycobacterium tuberculosis complex|
  363698007 |finding site| = 39607008 |lung structure|

  Using Description Logic, it is possible to reason using multiple minimum sufficient sets for each concept.

6.5.2 EXAMPLE

For example, if we want to find all disorders that are associated with the organism 80166006 |streptococcus pyogenes|, we may discover (using the SNOMED CT Relationships file) that there is a direct 'causative agent' relationship from 302809008 |streptococcus pyogenes infection| to 80166006 |streptococcus pyogenes|. However, by introducing the following property chain rule:

47429007 |associated with| **o** 47429007 |associated with| → 47429007 |associated with|

and noting that 47429007 |associated with| has three subtypes:

255234002 |after|
42752001 |due to|
246075003 |causative agent|

it is possible to discover, using Description Logic, that 81077008 |acute rheumatic arthritis| and 58718002 |rheumatic fever| are also 'associated with' the concept 30209008 |streptococcus pyogenes infection|. Figure 6-5 illustrates these relationships that can discovered using property chaining.

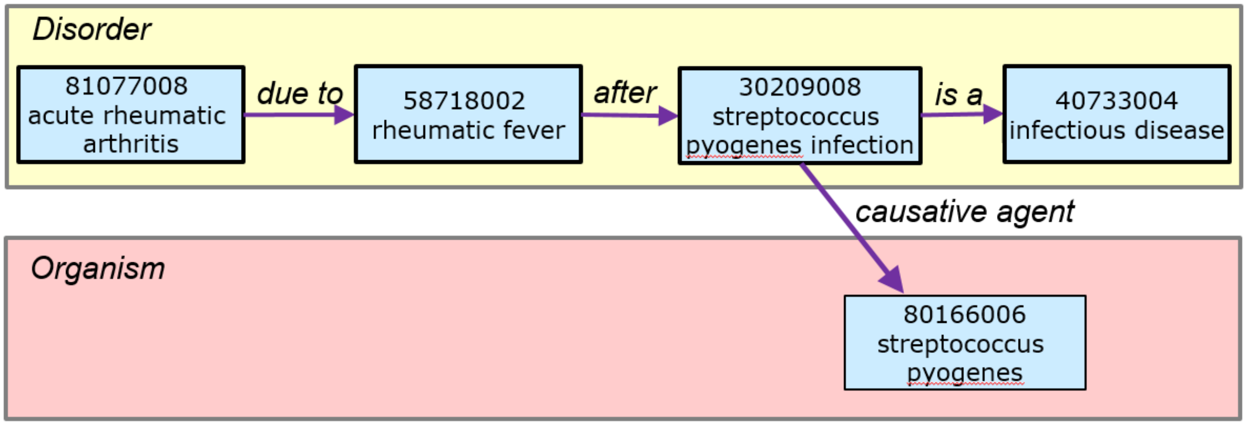

Disorder
81077008 acute rheumatic arthritis
due to
58718002 rheumatic fever
after
30209008 streptococcus pyogenes infection
is a
40733004 infectious disease
causative agent
Organism
80166006 streptococcus pyogenes

Figure 6-5: Property chaining

6.5.3 IMPLEMENTATION

6.5.3.1 OWL 2

Using Description Logic techniques to perform analytics over SNOMED CT involves first translating SNOMED CT into OWL 2 (Web Ontology Language). OWL 2 is an ontology language for the Semantic Web with formally defined meaning. The SNOMED CT international release comes with a Perl transform script that converts the RF2 files into OWL XML/RDF, Functional Syntax or KRSS files.

Once generated, the OWL files can then be loaded into a Description Logic Editor (such as Protégé) or used directly by a terminology service which offers description logic capabilities. The Description Logic Editor or terminology service then uses DL reasoners (also known as 'classifiers'), such as Snorocket, ELK and FACT++, to perform consistency checking and subsumption testing (also known as 'classification') over SNOMED CT. Subsumption testing can also be performed between two expressions. Semantic query languages, such as SPARQL, can be used to query over RDF representations of SNOMED CT.

6.5.3.2 CASE STUDIES

Some commercial terminology servers, such as B2i Healthcare's Snow Owl terminology server (case study 2.5), use Description Logic based techniques to support both classification and querying over SNOMED CT. Kaiser Permanente is collaborating with Oxford University (case study 1.3) to investigate ways of performing complex queries efficiently across extremely large numbers of patient records using scalable parallel processing and description logic reasoners.

6.6 DESCRIPTION LOGIC OVER TERMINOLOGY AND STRUCTURE

6.6.1 OVERVIEW

When performing analytics over patient data, an appreciation for the semantics represented in both the terminology and the information model is required. Different information models can use different amounts of precoordination in the terminology, and the same semantics can be represented using different information structures. By using description logic over both the terminology and the information structures, a consistent representation of the meaning of data can be achieved, irrespective of whether this meaning is captured in the data values or in the model itself.

6.6.2 EXAMPLE

Consider for example the two alternative ways of recording family history, as shown in Figure 6-6. The green rectangles represent the logical structure of the information model and the blue rectangles represent the concept identifiers that are used to populate this information model in the patient record.

The information model on the left uses a heading of 'Family history' to indicate that the named problem refers to a family history of that problem. The information model on the right uses the terminology value to indicate that the problem refers to a family history instance.

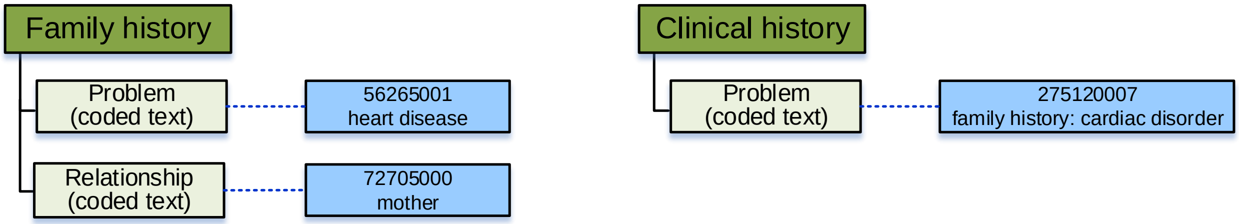


Figure 6-6: Two ways of recording family history

When querying over data, which may be collected in either format, both the semantics of the information model and the semantics of the data instances must be considered. One way of achieving this is to use an 'expression template' to convert all data instances into a Description Logic representation, and use this to reason over the data. Figure 6-7 shows an example of an expression template that could be used to create a SNOMED CT expression for each of the data instances shown in Figure 6-6. Please note that the orange parallelograms represent 'slots' which are subsequently populated with the value of the named data element (e.g. '$Problem').

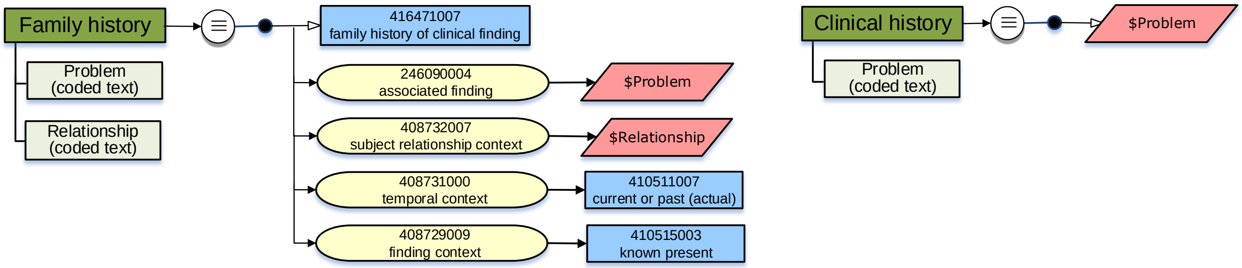


Figure 6-7: SNOMED CT expression representation of family history data

When the data instances from Figure 6-6 are used to populate the templates from Figure 6-7, the following two expressions are created:

416471007|family history of clinical finding|:
246090004 |associated finding| = 56265001 |heart disease|,
408732007 |subject relationship context| = 72705000 |mother|,
408731000 |temporal context| = 410511007 |current or past (actual)|,
408729009 |finding context| = 410515003 |known present|

275120007 |family history: cardiac disorder|

These expressions may then be compared using a DL reasoner to discover that the first expression is subsumed by the second, or queried using a semantic query language to allow the two data representations to be analyzed in a consistent way.

6.6.3 IMPLEMENTATION

6.6.3.1 OWL 2

Description Logic techniques, such as those described in section 6.5, can be used to reason over both the terminology and the information model. In addition to translating SNOMED CT to OWL 2, OWL 2 representations of the information model are also created using 'templates' that include 'slots' which are then filled with the patient record instance values. DL reasoners, such as Snorocket, ELK and FACT++, and semantic query languages, such as SPARQL, can then be used over both the terminology and the information model in a consistent way.

6.6.3.2 CASE STUDIES

Kaiser Permanente is collaborating with Oxford University (case study 1.3) to investigate ways of performing complex queries efficiently across extremely large numbers of patient records using scalable parallel processing and description logic reasoners. In this project, the analysis is being performed over an OWL-RL representation of the patient data, which incorporates both the terminology and the structure of the information.

6.7 USING STATISTICAL CLASSIFICATIONS

6.7.1 OVERVIEW

Clinical terminologies and classifications serve different but complementary purposes and both are an important part of the healthcare environment. There are therefore some situations in which it is necessary to map SNOMED CT codes to a classification, such as ICD-9 or ICD-10, for analytics or reporting purposes. The differences between the two holds the key to their distinct purpose.

A classification is a hierarchical organization of terms that allows aggregation into categories which can be counted and compared. A statistical classification is mono-hierarchical which means that each code in the hierarchy is classified underneath a single code in the level above. This avoids codes being counted twice because they are grouped into two distinct groupings (i.e. double counting), but means that arbitrary decisions must be made as to where codes are grouped. For example, the ICD-10 code J12 |viral pneumonia, not elsewhere classified| is classified under "Diseases of the respiratory system", but is not classified under "Certain infectious and parasitic diseases". Therefore, a query that asks "Is J12 a respiratory disease?" will return "Yes", while a query that asks "Is J12 an infectious disease?" will return "No".

Unlike clinical terminologies, classifications also explicitly enumerate 'known unknowns' (e.g. 'not otherwise specified (NOS)' and 'not elsewhere classified (NEC)'); they often use a single code to represent several closely related but clinically distinct entities (e.g. H65.9 represents 'allergic otitis media', 'exudative otitis media', 'mucoid otitis media' and others); and they are often presented in the form of coding manuals with rigid, well defined rules of use. Classifications emphasize coding discipline (rather than expressivity). This is helpful for driving formal billing and reimbursement. The lower number of codes also makes assigning prices to each code tractable (e.g. using either ICD-9, ICD-10, or one of the Diagnosis Related Group systems). Classifications are also deployable in low tech environments, including paper or simple spreadsheet based systems.

Classifications are primarily used for purposes in which terms must be grouped into categories, and double counting must be avoided. These purposes may include:

- Statistical reporting on major diagnoses, procedures or primary cause of morbidity
- Epidemiological reporting involving counting of disease categories
- Other administrative reporting based on specific WHO reporting requirements
- Billing and reimbursement

In contrast, SNOMED CT is a clinical terminology, in which each concept identifier represents a distinct clinical meaning. By providing a more detailed level of granularity than classifications, SNOMED CT enables clinicians to use SNOMED CT to record healthcare information at the clinically appropriate level of detail. Unlike statistical classifications, SNOMED CT uses a polyhierarchy, in which each concept may be grouped under more than one supertype, reflecting possible alternate ways of categorizing each clinical meaning. SNOMED CT also provides defining relationships between concepts, which further enhances its ability to support flexible and powerful analytics capabilities.

It is generally recommended that clinical data is recorded using a clinical terminology, such as SNOMED CT, and then mapped for reporting purposes to one or more classifications, such as ICD. IHTSDO publishes a map from SNOMED CT to both ICD-9 and ICD-10. This supports epidemiological, statistical and administrative reporting needs of the member countries and WHO Collaborating Centers. The collaborative work between IHTSDO and the WHO on the alignment of ICD-11 with SNOMED CT is in progress and promises tighter integration of the distinct use cases in the future.

6.7.2 EXAMPLE

The following example illustrates the rows of the Extended Map reference set that supports the mapping from the SNOMED CT concept 15296000 |sterility| to an appropriate ICD-10 code. The set of map rules associated with each SNOMED CT concept are grouped together into ‘map groups’ and then ordered within each map group by a ‘map priority’. The map rule provides a machine readable rule that indicates whether this map should be selected within its map group, and the map advice provides human readable advice. The correlation indicates the type of match between the source and the target (e.g. ‘exact match’ or ‘narrow to broad’) and the map category indicates the kind of map being represented.

Referenced component	Map target	Map group	Map priority	Map rule	Map advice	Correlation ID	Map Category Id
15296000 |sterility|	N97.9 |female infertility, unspecified|	1	1	IFA 10114008 |Female sterility (finding) |	IF FEMALE STERILITY CHOOSE N97.9	Not specified	Context dependent
15296000 |sterility|	N46 |male infertility|	1	2	IFA 49408009 |Male sterility (finding) |	IF MALE STERILITY CHOOSE N46	Not specified	Context dependent
15296000 |sterility|		1	3	OTHERWISE TRUE	MAP SOURCE CONCEPT CANNOT BE CLASSIFIED WITH AVAILABLE DATA	Not specified	Context dependent

6.7.3 IMPLEMENTATION

6.7.3.1 MAPPING TO CLASSIFICATIONS USING SNOMED CT

Maps from SNOMED CT to classifications are generally represented in SNOMED CT's RF2 using a Complex or Extended map reference set. Mappings are then performed by matching each SNOMED CT code in a patient's record with the corresponding row of the map reference set, and using the classification code found in the 'mapTarget' field.

6.7.3.2 CASE STUDIES

The UK Terminology Centre's Data Migration Workbench demonstrates the use of maps from SNOMED CT to ICD-10 International Edition (using the UK maps) and OPCS Classification of Interventions and Procedures (OPCS-4) (case study 1.2). The National Library of Medicine (NLM) has also developed a demonstration tool, which demonstrated the key principles of implementing map rules and advice. This tool, called I-MAGIC[2] (Interactive Map-Assisted Generation of ICD Codes) uses the SNOMED CT to ICD-10 map in a real-time, interactive manner to generate ICD-10 codes. It simulates a problem list interface in which the user enters problems with SNOMED CT terms, which are then used to derive ICD-10 codes using the map. A number of vendor products, such as Cerner Millenium (case study 2.8), also use maps from SNOMED CT to ICD-10 to enable statistical analysis.

[2] http://imagic.nlm.nih.gov/imagic2/code/map

7 TASK-ORIENTED ANALYTICS

7.1 OVERVIEW

The SNOMED CT analytics techniques described in the previous chapter only become useful when performing a specific analytics task intended to meet a business need. In this chapter, we consider a range of analytics tasks, which are either enabled or enhanced by using these SNOMED CT techniques.

The analytics tasks which can benefit from the use of SNOMED CT techniques can be considered in three broad categories, as shown in Figure 7-1:

- Point of care analytics, which benefits individual patients and clinicians. This includes historical summaries, decision support and reporting.
- Population-based analytics, which benefits populations. This includes trend analysis, public health surveillance, pharmacovigilance, care delivery audits and healthcare service planning.
- Clinical research, which is used to improve clinical assessment and treatment guidelines. This includes identification of clinical trial candidates, predictive medicine and semantic searching of clinical knowledge.

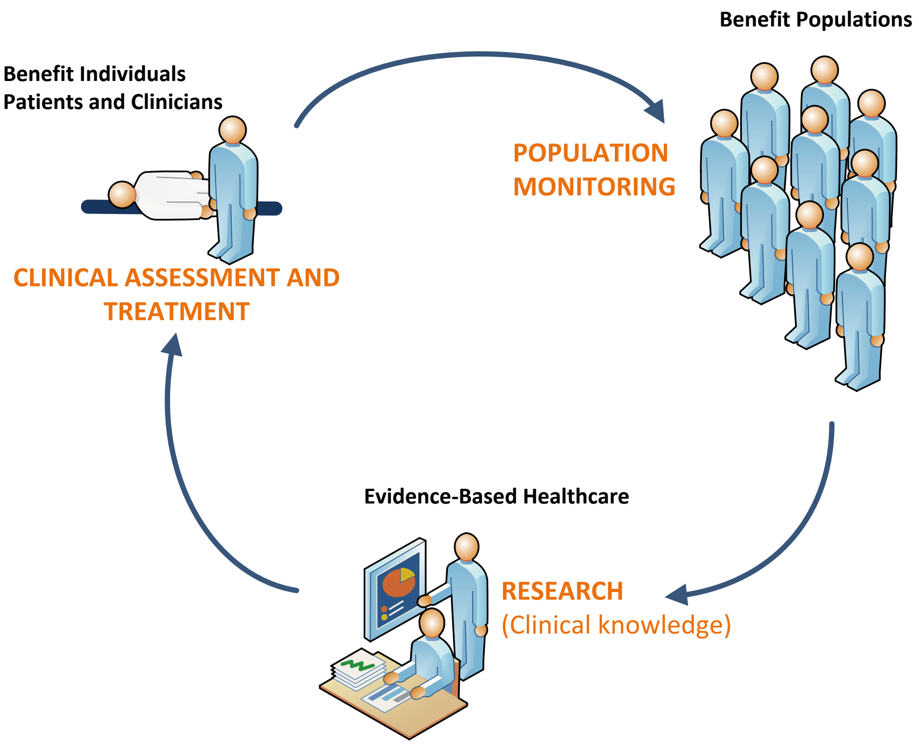


Figure 7-1: Analytics tasks that can benefit from SNOMED CT

Many of these tasks use business intelligence capabilities, similar to those used in other sectors, such as manufacturing, retail and transportation. *Business intelligence* is the provision of historic, current and predictive views of information. Such services include reporting, online analytical processing (OLAP), data mining, process mining, complex event processing, benchmarking, text mining, predictive analysis and prescriptive analytics. In many cases, a data warehouse is used as the platform on which these services are provided (see section 8.3).

The combination of these business intelligence techniques with the capabilities of SNOMED CT creates new opportunities to improve healthcare delivery.

7.2 POINT OF CARE ANALYTICS

7.2.1 OVERVIEW

Point of care analytics encompasses those analytics services that directly benefit individual patients and clinicians, including historical summaries, decision support and point of care reporting. These analytics tasks typically involve the summarization and mapping of patient data, and the linking of terminology with clinical knowledge artefacts.

7.2.2 HISTORICAL SUMMARIES

One major ambition of healthcare IT is to make effective summaries of a patient's clinical history available to healthcare providers (especially in emergency situations). Typically a patient's clinical data is scattered across a number of healthcare institutions using a variety of information models and coding systems. Even within a single institution patient data may be captured across many episodes of care, many devices, and often many software systems.

SNOMED CT can help to support the integration of this information by serving as a common reference terminology into which other code systems can be mapped (see section 5.3). It can also be used to unlock clinical data that was captured by source systems in free text narrative (see section 5.2), and to summarize large volumes of data by grouping codes together into more general categories (see section 6.3). SNOMED CT can also be used to enable clinicians to filter large volumes of data to select those records that are relevant to the current care episode – for example identifying all previous records of a heart attack.

One significant example of this is the UK NHS Summary Care Record (SCR) service [6], which uses SNOMED CT to represent a number of types of clinical information including medical history, medications, adverse reactions and allergies. This service uses a summary extracted from detailed patient care records held in a variety of disparate systems. Where the source data is not stored natively in SNOMED CT, they are mapped into SNOMED CT prior to transmission. Over 40 million people in England (80% of the population) now have a summary care record. This service now contributes to the safe and efficient assessment and treatment of these people, and has greatly improved the accuracy and timeliness of medicines reconciliation [9].

7.2.3 CLINICAL DECISION SUPPORT

Clinical decision support systems (CDSS) are designed to assist clinicians at the point of care on decision making tasks. Examples of applications of clinical decision support include:

- Checking conformance with clinical guidelines and protocols
- Guide clinicians through complex care pathways
- Protect against errors in prescribing (e.g. drug-drug and allergy-drug contraindication checking)

- Highlight critical laboratory results
- Display clinical knowledge resources upon request, that are relevant to the given patient's diagnosis, symptoms, procedures or medications

Most CDSSs consist of three parts:

1. The knowledge base, with rules and guidelines – for example:
    a. IF drug = << 48603004 |warfarin| AND 77386006 |pregnant| THEN alert user
    b. IF drug has active ingredient = << 387494007|codeine| AND past history of 292055008 |codeine adverse reaction| THEN alert user
    c. IF diagnosis = << 195967001 |asthma| THEN display Asthma Management Guidelines
2. The inference engine, which uses the data from the patient record to determine which rules from the knowledge base should be executed – for example:
    a. When a patient, with finding 77386006 |pregnant| is prescribed 375374009 |warfarin sodium 4mg tablet|, the inference engine triggers Rule a. above.
    b. When a patient, with past history of 292055008 |codeine adverse reaction| is prescribed 412575004 |aspirin 325mg/codeine 30mg tablet|, the inference engine triggers Rule b. above.
    c. When a patient's primary diagnosis is entered as "195949008 |chronic asthmatic bronchitis|" the inference engine triggers Rule c. above.
3. A mechanism to communicate, which allows the system to display alerts or clinical knowledge to the user

Using a combination of SNOMED CT techniques, including mapping (section 5.3), subsets (section 6.2), subsumption (section 6.3) and defining relationships (section 6.3), SNOMED CT helps to support the inference engine in determining the appropriate rules to execute.

For example, Kaiser Permanente's HealthConnect system uses SNOMED CT to support efficient translation of its business rules into decision support rules (case study 1.3). The National Board of E-Health in Denmark is developing a centralized decision support service based on the Danish SNOMED CT drug extension, which utilizes the hierarchical and defining relationships of SNOMED CT.

A number of commercial tools also use the capabilities of SNOMED CT to implement Clinical Decision Support. For example, Cambio's COSMIC tool (case study 2.6) binds GDL (Guideline Definition Language) rules to SNOMED CT concepts to support the triggering of appropriate rules. Allscript's Sunrise InfoButton™ feature (case study 2.3) provides relevant medical reference content to clinicians wherever patient care decisions are made, by using SNOMED CT encoded patient problem lists and medication data to query third-party medical content. The Epic system (case study 2.11) provides decision support alerts (called 'Best Practice Advisories'), which are able to use the SNOMED CT hierarchy to help define their criteria. And First DataBank (case study 2.12) delivers clinical decision support solutions linked to SNOMED CT, primarily to detect safety issues arising from certain combinations of medications, diagnoses and drug adverse reaction histories.

7.2.4 POINT OF CARE REPORTING

When it comes to reporting needs, the preference of most clinicians is to 'collect once and use many times'. SNOMED CT enables this goal to be achieved by allowing data to be captured at the appropriate level of detail and then queried at the same or less detailed level. SNOMED CT supports point of care reporting requirements using any (or all) of the SNOMED CT analytics techniques described in section 6, including subsets, subsumption, defining relationships and description logic. Examples of point of care reporting requirements may include:

- Helping clinicians remember preventative services (reminders)
- Identifying patients with care gaps and risk factors
- Monitoring patient compliance with prescribed treatments
- Reporting clinical data to registries, such as cancer, stroke, and infectious disease registries
- Billing and reimbursement[3]

When supporting a reporting requirement in which double counting must be avoided (such as statistical reporting, administrative reporting, billing, or reimbursement), SNOMED CT codes can be mapped to statistical classifications (such as ICD-9 and ICD-10) (see section 0).

When the source data uses a coding system without the same reporting capabilities as SNOMED CT, or when a variety of coding systems are used, coded data can be mapped into SNOMED CT to support the reporting requirements (see section 5.3).

7.3 POPULATION-BASED ANALYTICS

7.3.1 OVERVIEW

Population-based analytics encompasses those analytics services that benefit entire populations, including trend analysis, public health surveillance, pharmacovigilance, care delivery audits and healthcare service planning. Population-based analytics contributes to public health programs by helping to identify health threats, inform public policy and manage healthcare resources.

Efficient healthcare delivery and service planning depends on high quality clinical data. Clinical data is typically scattered between multiple different healthcare providers using different clinical systems. Collating this information for analysis requires both standardized terminologies and common information models. Identifying relevant and useful facts in large volumes of collated data also requires this data to be accurate, meaningful and machine processable.

SNOMED CT supports population-based analytics in a number of ways. Firstly, it enables more accurate capture of clinical data by allowing it to be represented at the appropriate level of clinical detail. Secondly, it supports the integration of disparate clinical data sources by serving as a reference terminology into which free text and other code systems can be mapped. And thirdly, it enables more meaningful and powerful queries to be performed over the data using the descriptions, hierarchies and logic-based definitions of each concept.

Vendor products, which provide population health solutions include Caradigm's Intelligence Platform (case study 2.7), Allscript's Clinical Quality Management and Clinical Performance Management tools (case study 2.3), Cerner's PowerInsight® Data Warehouse (case study 2.8) and Epic's analytics and reporting suite (case study 2.11).

In this section, we discuss three key types of population-based analytics: trend analysis, pharmacovigilance, and clinical audit.

[3] Note: In some healthcare environments this is a point of care activity, while in others it is not.

7.3.2 TREND ANALYSIS

Trend analysis is the practice of collecting information and attempting to spot a pattern, or trend, in the information. Trend analysis often refers to techniques for extracting an underlying pattern of behavior in a time series, which would otherwise be partly or nearly completely hidden by noise.

Detecting changes of either incidence or prevalence of a particular disease, treatment, procedure or intervention over time has major utility for population health monitoring, prediction of demand and effective resource allocation at enterprise and national levels. One challenge that is encountered when analyzing routinely collected patient data for trends, is distinguishing minor changes in coding style from real changes in disease incidence. Simply counting the use of individual concept identifiers may be highly misleading. For example, a fall in the use of the code 22298006 |myocardial infarction| might reflect a shift to using more specific codes (such as 314207007 |non-Q wave myocardial infarction| or 304914007 |acute Q wave myocardial infarction|), rather than a reduction in the incidence of myocardial infarctions. Use of subsumption testing on SNOMED CT encoded data (see section 6.3) can enable higher level trend analysis to be performed over more specific coded data.

SNOMED CT's polyhierarchy allows trends to be analyzed from multiple perspectives. However, deciding which level of aggregation to use for trend analysis can be arbitrary. Novel approaches to this task are emerging as the demand for trend analysis over SNOMED CT enabled data increases.

The UK Data Migration Workbench (case study 1.2), for example, includes a trend module which analyses the frequency with which individual SNOMED CT codes are used in the Electronic Patient Record (EPR) instance data, looking for those whose recording frequency has changed over the course of the data collection period. It also includes an Induce module, which performs a more sophisticated analysis of case mix and caseload trends within a clinical department. Instead of returning the most frequently used individual codes, the Induce module identifies the most frequently used types of codes. For example, an emergency department may use roughly 500 different SNOMED CT codes for a laceration in a particular anatomical location. While none of the site-specific codes may appear in a list of most common codes, the descendants of 312608009 |laceration| may collectively account for a significant part of the department's workload.

The algorithm used picks aggregation points at defined levels for analysis. The default setting finds roughly 100 sub-trees within the SNOMED CT hierarchy, where each sub-tree accounts for a more or less constant proportion of all coded episodes (around 1% of all coded events per sub-tree). The algorithm completes once the set of all codes within all identified sub-trees collectively accounts for the large majority of the dataset being analyzed. When applied to real emergency department attendance data, relatively low numbers of presentations (about 0.2%) were coded as occurring primarily as a result of endocrine disease. As a result, in order to get a big enough grouping of episodes, the algorithm chooses 362969004 |disorder of endocrine system| as the root of a single sub-tree covering these reasons for the patient's attendance. By contrast, a very high proportion (9.4%) of presentations relate to some subtype of 928000 |disorder of musculoskeletal system|. Therefore this part of the caseload is aggregated under multiple more granular sub-trees, including (separately) burns, abrasions, lacerations, blunt injury, crush injury and foreign body.

These code aggregations can then be tracked across time to reveal trends in demand, disease incidence or resource utilization.

7.3.3 PHARMACOVIGILANCE

Pharmacovigilance is the collection, detection, assessment, monitoring and prevention of adverse effects with pharmaceutical products. It is concerned with identifying the hazards associated with pharmaceutical products and minimizing the risk of any harm that may come to patients. An important part of pharmacovigilance is postmarketing surveillance, which monitors the safety of a pharmaceutical drug or medical device after it has been released on the market. Since drugs are approved on the basis of clinical trials, which involve relatively small numbers of people, postmarketing surveillance plays an important part in further refining, confirming or denying the safety of a drug in the general population.

Pharmacovigilance uses a number of data sources to assess and monitor the safety of licensed drugs, including clinical trial data, medical literature, spontaneous reporting databases, prescription events, electronic health records, and patient registries. Data mining of large volumes of clinical data can be used to highlight potential safety concerns. However, current mechanisms to analyze this data is often both costly and insensitive.

The availability of large datasets of richly encoded SNOMED CT data within longitudinal healthcare records can greatly assist pharmacovigilance. Where SNOMED CT is not used natively to capture clinical data, free text narrative and other code systems may be mapped to SNOMED CT (see sections 5.2 and 5.3) to support a homogeneous approach to querying across diseases, signs and symptoms, lab results, medications, devices, procedures, allergies, adverse reactions, body sites and substances. SNOMED CT's polyhierarchy and defining relationships, which provide links between these domains provide a rich source of meaning-based information across which queries can be performed.

Many drug regulatory authorities and pharmaceutical companies currently use the Medical Dictionary for Regulatory Activities (MedDRA) to classify adverse drug events. MedDRA is an international standard adverse event classification used from pre-marketing through to post-marketing activities. However, as MedDRA was not designed to support routine clinical data collection, its penetration into clinical systems is limited. Therefore mapping from SNOMED CT to MedDRA would enable both styles of analysis and reporting to be performed from the same clinical data. The UK Medicines and Healthcare products Regulatory Agency (MHRA) is working (with input from the MedDRA Maintenance and Support Services Organization) to develop a mapping from a subset of SNOMED CT to MedDRA for this purpose.

7.3.4 CLINICAL AUDIT

Clinical audit seeks to improve patient care and outcomes through systematic review of care against defined standards and the implementation of change. It informs care providers and patients where their healthcare services are doing well and where there could be improvements. The key component of clinical audit is that performance is reviewed (or audited) to ensure that what *should* be done *is being done*, and if not it provides a framework to enable improvements to be made.

Clinical audits can be performed in primary care facilities, individual clinics, hospitals, enterprises or jurisdictions. Audit can have major beneficial impacts in ensuring the consistent delivery of quality healthcare. The questions asked in audit are often chosen pragmatically according to local data collection practices. For example

- What proportion of patients invited to attend cervical screening did so?
- How many patents with ischemic heart disease are receiving appropriate drug treatments?
- Are all patients with diabetes mellitus reviewed within a stated time interval?

Current audit schemes use a combination of reporting against the classifications or questions specifically collected for audit purposes. SNOMED CT will facilitate an increase in such audits being able to collect some of the data by extracting data from the patient record thus reducing the additional burden of collection; it will also enable more a more accurate picture from say tertiary centers where some of their procedures may fall into the NOS or NEC classification and provide an unrepresentative comparison with other centers when the procedures are complex and innovative/new.

SNOMED CT is well suited to service the *ad hoc* requirements that emerge in clinical audit questions, using the techniques described in section 6. Using the SNOMED CT codes recorded during care delivery can reduce the additional burden of data collection specifically for audit purposes. SNOMED CT may also facilitate more accurate audit results than classifications, by distinguishing between distinct concepts (e.g. clinical findings or procedures) which may fall into the 'Not Otherwise Specified' or 'Not Elsewhere Classified' categories in these classifications.

A number of vendor products, such as Cerner's PowerInsight® Data Warehouse (case study 2.8), are able to support clinical audit using SNOMED CT enabled analytics tools.

7.4 CLINICAL RESEARCH

7.4.1 OVERVIEW

Clinical research is a branch of healthcare science that determines the safety and effectiveness of medications, devices, diagnostic products, and treatment regimens intended for human use. Clinical research may be used for the prevention, treatment, diagnosis or for relieving symptoms of a disease. In contrast to clinical practice, which applies established treatment regimes, clinical research collects evidence to extend knowledge and establish the value of novel treatments and other patient management practices.

Clinical research typically involves the analysis of data from well-defined and homogenous groups of patients with a specific disease, at a specific stage, receiving similar treatments and often without significant co-morbidities. The data may be captured prospectively or retrieved retrospectively.

SNOMED CT helps clinical research activities by assisting in the identification of clinical trial candidates, enabling the powerful analysis of trial data, supporting predictive medicine, and improving the effectiveness of semantic search over clinical knowledge.

In this section, we discuss three key aspects to clinical research that can benefit from the use of SNOMED CT: identification of clinical trial candidates, predictive medicine, and semantic search.

7.4.2 IDENTIFICATION OF CLINICAL TRIAL CANDIDATES

SNOMED CT can be used to assist the process of identifying clinical trial candidates for recruitment into formal clinical trials. Subsets of findings, procedures or medications (see section 6.2) can be used to filter trial candidates based on their clinical conditions or treatments. Subsumption techniques (see section 6.3) can be used to identify suitable candidates, irrespective of the level of granularity in which their clinical data is stored. SNOMED CT defining relationships (see section 6.4) can be used in a number of ways – for example, identifying patients with diseases of specific anatomical sites, with certain morphologies; patients who are taking medications with specific ingredients or dose forms; and patients who have had procedures on a specific body site.

Commercial tools, which can be used to support clinical research include Cerner's clinical research module (PowerTrials), which offers patient identification functionality (case study 2.8).

7.4.3 PREDICTIVE MEDICINE

Predictive medicine involves predicting the probability of disease and implementing measures to either prevent the disease altogether or significantly decrease its impact upon the patient. The outcomes of predictive medicine are often applied to the care of individual patients, but may also inform the deployment of resources to entire populations at high risk.

The goal of predictive medicine is to predict the probability of future disease so that healthcare professionals and the patient themselves can be proactive in implementing lifestyle modifications and increased physician surveillance, such as regular skin exams, mammograms, or colonoscopies. Predictive medicine changes the paradigm of medicine from being reactive to being proactive, and has the potential to significantly extend the duration of health and to decrease the incidence, prevalence and cost of diseases.

Much attention has been focused on the availability of genetic makers of vulnerability to specific illnesses. However the accurate capture of phenotypic (e.g. height and weight, blood pressure), environmental factors (e.g. smoking, alcohol consumption) and other lifestyle factors (e.g. exercise, nutrition, quality of life) is not to be overlooked. For example:

1. Patient is a smoker and has ischemic heart disease → predict excess risk of myocardial infarction
2. Patient has BRCA1 gene and is a 40 year old woman → predict (excess) risk of breast cancer

SNOMED CT can help to support predictive medicine by:

- Helping to identify clinical trial candidates (as described in section 7.4.2)
- Helping to analyze clinical data, such as family history, lifestyle and environmental findings, to improve predictive capabilities (using analytics techniques, as described in section 6)
- Providing a link between patient data and risk assessment rules, so that rules can be triggered based on subsumption of codes recorded in clinical data (see section 6.3). For example, matching against patient records could be improved by defined the above rules as:
    1. Criteria: 77176002 |smoker| AND 414545008 |ischemic heart disease|
       Risk: 22298006 |myocardial infarction|
    2. Criteria: 412734009 |BRCA1 gene mutation positive|
       Risk: 254837009 |breast cancer|

7.4.4 SEMANTIC SEARCH

With an ever increasing volume of medical literature and clinical reports, it is becoming increasingly important to be able to meaningfully search this information. A major application for Natural Language Processing technologies (see section 5.2) is to index collections of free text transcripts or documents such that topic specific searches may be run on them. The challenge is to move beyond the limitations of plain keyword searching strategies towards more advanced search techniques, which return ranked matches with high sensitivity and specificity. Clinical searches may be performed over documents within an electronic library, within medical records, or on the internet. Examples of searches include:

- "Show me articles on this website concerned with inflammatory bowel disease"
- "Does this patient have transcripts in their record suggesting a heart rhythm disturbance?"

SNOMED CT was used in techniques developed by Koopman (case study 1.5) to improve search performance by addressing vocabulary mismatch (using synonyms, e.g. hypertension vs high blood pressure), granularity mismatch (using hierarchical relationships, e.g. antipsychotic vs haloperidol), conceptual implication (using defining relationships, e.g. from renal cyst infer kidney) and inferences of similarity (e.g. using subset membership, e.g. comorbidities anxiety and depression). Koopman also assigned a measure of similarity to each SNOMED CT relationship type, and use this weighting to determine the relevance of each document.

Some commercial tools also provide semantic search, including Cerner’s semantic search tool (case study 2.8).

8 DATA ARCHITECTURES

8.1 OVERVIEW

While the use of SNOMED CT for analytics does not dictate a particular data architecture, there are a few key options to consider. In this section, we describe the major categories of data architecture that may be used to perform analytics over SNOMED CT enabled patient data, including:

1. Analytics directly over patient records;
2. Analytics over data exported to a data warehouse;
3. Analytics over a Virtual Health Record (VHR);
4. Analytics using distributed storage and processing.

Please note that some of these approaches may be used in combination. For example, data warehouses with large volumes of data may use distributed storage and processing for enhanced performance, and querying directly over disparate patient records could be performed using a Virtual Health Record.

8.2 PATIENT RECORDS FOR ANALYTICS

Electronic patient record systems typically require high performance, high reliability and no (or limited) downtime. Any operation that effects these key criteria need to be kept to an absolute minimum, so as not to disturb the clinical and documentation activities of busy clinicians.

Many analytics activities require large volumes of data to be processed, which may slow down or even ‘lock out’ clinical transactions that are being performed at the same time. For this reason, population-based analytics and clinical research is typically not performed directly on patient records in their native clinical system. Instead, analytics directly over ‘live’ patient records tends to be restricted to point of care analytics activities, such as historical summaries, clinical decision support and point of care reporting. These analytics activities tend to demand the most up to date data possible to ensure its accuracy. They also tend to only require data for a single patient, which can be efficiently accessed using a patient identifier index.

Figure 8-1 illustrates a simple architecture in which the data store for patient records is directly used for reporting and analytics purposes.

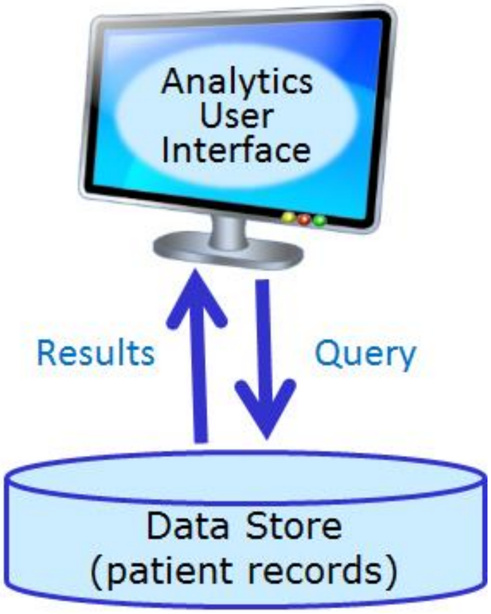


Figure 8-1: Analytics directly over patient records

8.3 DATA WAREHOUSE

A data warehouse is a central data store of integrated data from one or more disparate sources, used for reporting and data analysis. While operational systems are optimized for the preservation of data integrity and the speed of recording transactions, data warehouses are optimized for the high performance execution of queries.

The typical Extract-Transfer-Load (ETL) based data warehouse uses a staging layer to clean the extracted data and transform it into a homogeneous structure and standardized terminology. During this process, the techniques from section 5, such as mapping codes to SNOMED CT, can be used to prepare the data for analytics. The transformed data is then loaded into the data warehouse, and indexed, so that optimized analysis of the data can begin.

The benefits of using a data warehouse include:

- Data from multiple heterogeneous sources can be integrated to enable consistent querying over data from all sources
- The operational clinical system does not suffer performance degradation when running large analytics queries over historical data
- The data quality can be improved by cleaning the data, and mapping non-SNOMED CT codes to SNOMED CT
- The data can be restructured to optimize query performance

Figure 8-2 illustrates an architecture in which the patient record data is extracted from its operational data store and loaded into a data warehouse for reporting and other analytics purposes.

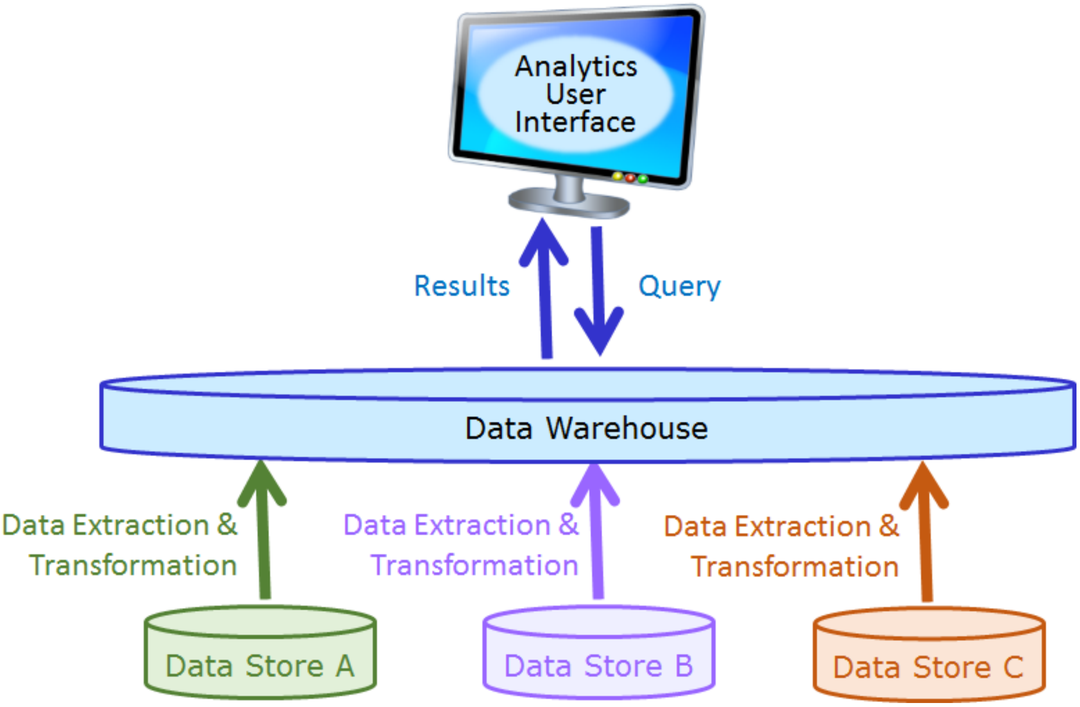

Analytics User Interface
Results
Query
Data Warehouse
Data Extraction & Transformation
Data Extraction & Transformation
Data Extraction & Transformation
Data Store A
Data Store B
Data Store C

Figure 8-2: Querying using a data warehouse

Commercial data warehousing solutions that support SNOMED CT include Cambio's COSMIC Intelligence (case study 2.6), Cerner's PowerInsight Data Warehouse (PIDW) and Cerner's Health Facts Data Warehouse (case study 2.8).

8.4 VIRTUAL HEALTH RECORD

A Virtual Health Record (VHR) provides a virtual view of heterogeneous data sources, using a common data model. In contrast to the data warehousing approach in which heterogeneous data is extracted, transformed and stored in a homogeneous form, the VHR approach does not require clinical data to be extracted from existing data stores. Instead, logical queries are defined in terms of a common data model and then transformed into a set of physical queries which can each be executed locally on an individual data store. Figure 8-3 illustrates an architecture which supports querying over a VHR.

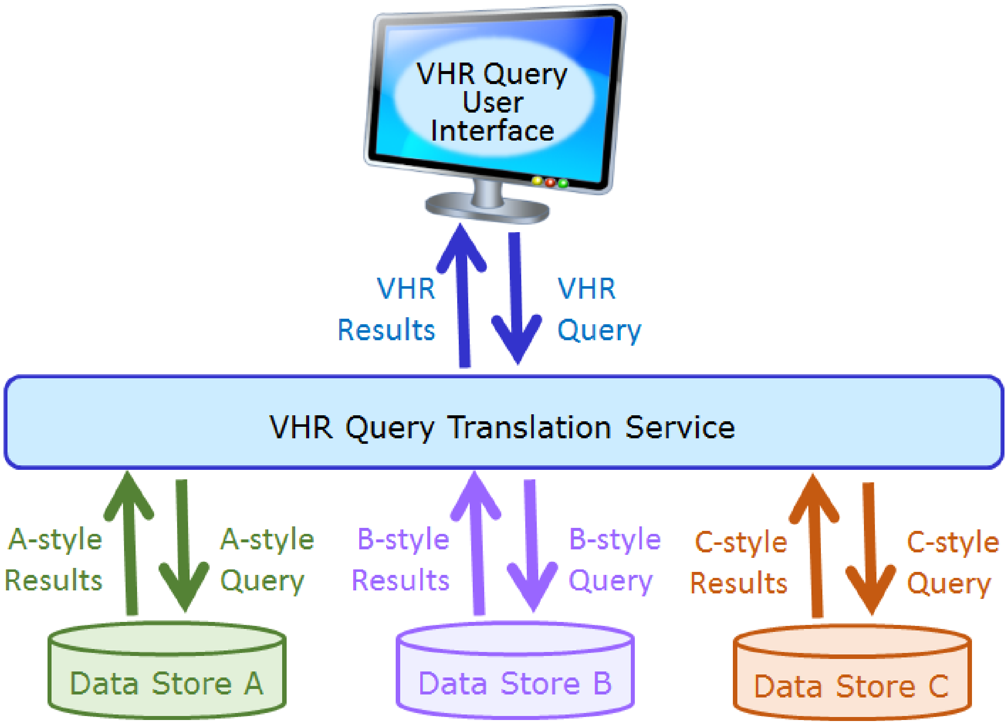


Figure 8-3: Querying using a Virtual Health Record

The process of transforming the logical query into separate physical queries may involve translating:

- *The Query Language* – from a common query language to the local data store's native query language
- *Data Model References* – from the common data model to the local data model
- *Terminology References* – from the standard terminology to the local code system

For example, if the user poses the following SQL query, written in terms of the VHR's common data model, to select those patients with a diagnosis that is a subtype of 40733004 |infectious disease|:

```
SELECT patient_id FROM Health_Records
WHERE diagnosis IN (<40733004 |infectious disease|)
```

This query may be translated into the following 3 queries for local execution on each data store:

Data Store A:

```
Patient_record/patient_id[@diagnosis=typeOf(INF)]
```

Data Store B:

```
SELECT id FROM EHR NATURAL JOIN DSummary
WHERE discharge_diagnosis IN (descendantsOf (40733004)
```

Data Store C:

```
SELECT patient FROM record
WHERE diag IN (<40733004)
```

Similarly, when the query results are returned by each data store, these need to be transformed and mapped into the common data model and then combined for presentation to the user.

The VHR approach provides an alternative architecture to a data warehouse for integrating heterogeneous systems. It is most commonly used when copying clinical data into a data warehouse is not possible (e.g. due to legislative requirements), or when the currency of the data is imperative. The challenges with this approach lie with the potential complexity of the transformations required. The implementation of this approach is considered to be a type of heterogeneous distributed database, as described in Section 8.5.

8.5 DISTRIBUTED STORAGE AND PROCESSES

The increasing volume and variety of data collected by healthcare enterprises is a challenge to traditional relational database management systems. This increase in data is due both to an increase in computerization of health records, and to an increase in the capture of data from other sources, such as medical instruments (e.g. biometric data from home monitoring equipment), imaging data, gene sequencing, administrative information, environmental data and medical knowledge. The proliferation of large volumes of both structured and unstructured data sets has led to the popularity of the term 'Big data' within the healthcare context. Big data refers to any collection of data sets that is so large and complex that it becomes difficult to process them using traditional data processing applications.

Accommodating and analyzing this expanding volume of diverse data (i.e. 'Big Data') requires distributed database technologies. A distributed database is a federation of loosely coupled data stores with separate processing units, which are controlled by a common distributed database management system. It may be stored in multiple computers located in the same physical location, or dispersed over a network of interconnected computers. Distributed databases may be categorized as either:

- *Homogeneous* – A distributed database with identical software and hardware running on all database instances.
- *Heterogeneous* – A distributed database supported by different hardware, operating system, database management systems and even data models (e.g. using the VHR strategy described in section 8.4).

In both cases, however, the database appears through a single interface as if it were a single database.

Distributed databases are used for Big Data analytics for a number of reasons, including:

- Transparency of querying over heterogeneous data stores (as described in section 8.4)
- Increase in the reliability, availability and protection of data due to data replication
- Local autonomy of data (e.g. each department or institution controls their own data)
- Distributed query processing can improve performance, as the load can be balanced among the servers

A number of tools are available for the distributed storage and processing of big data, including Apache Hadoop. Apache Hadoop is an open-source software framework, which splits files into large blocks and distributes these blocks amongst the nodes in the cluster. To process the data, Hadoop sends code to the nodes that have the required data, and the nodes then process the data in parallel. Hadoop supports horizontal scaling – that is, as data grows additional servers can be added to distribute the load across them.

Many distributed database solutions use NoSQL (Not Only SQL) systems. NoSQL systems are increasingly being used for big data, as they provide a mechanism for storage and retrieval of data in a variety of structures, including relational, key-value, graph or documents. The Oxford University, in collaboration with Kaiser Permanente (case study 1.3) are using a NoSQL database (RDFox) to investigate how to perform complex queries efficiently across extremely large numbers of patient records. RDFox is a highly scalable and performant NoSQL database that is readily distributed across parallel processing units.

9 DATABASE QUERIES

9.1 OVERVIEW

Practically all analytical processes are driven by database queries. A database query is a machine readable question presented to a database in a predefined language.

Unlike other code systems, which either have no hierarchy or a hierarchy that is fully represented within the code (e.g. H65.9), just retrieving the SNOMED CT codes recorded in a patient record does not fully utilize the analytics capabilities of SNOMED CT. To get the most benefit from using SNOMED CT in patient records, one must be able to not only query the records themselves, but also query SNOMED CT.

In this section, we describe how record and terminology queries can work together to perform powerful queries over SNOMED CT enabled data. In section 10, we will then consider how user interfaces can be designed to make these queries more accessible to non-technical users.

9.2 TERMINOLOGY QUERIES

9.2.1 SNOMED CT LANGUAGES

IHTSDO is developing a consistent family of languages to support a variety of use cases involving SNOMED CT, including querying and defining intensional subsets. The SNOMED CT family of languages will include:

- Compositional Grammar – for defining SNOMED CT expressions
- Expression Constraint Language – for constraining a set of possible expressions
- Query Language – for querying over SNOMED CT content
- Template Languages – using the other languages with slots that may be filled at a later time

Compositional grammar, which provides a common foundation for the functionality added to the other languages, was adopted as an IHTSDO standard in 2010. The other languages in the SNOMED CT family of languages are expected to be available to the community during 2015/2016. Both the SNOMED CT Expression Constraint Language and the SNOMED CT Query Language can be used to define queries against SNOMED CT content.

The SNOMED CT Expression Constraint Language is a formal language used to represent SNOMED CT Expression Constraints. A SNOMED CT Expression Constraint is a computable rule that can be used to define a bounded set of clinical meanings represented by either precoordinated or postcoordinated expressions. SNOMED CT Expression Constraints allows a set of clinical meanings to be defined using hierarchical relationships, attribute values, reference set membership, and other features such as cardinality, conjunction, disjunction and exclusion. For example, the following expression constraint represents the set of clinical findings, which have both a finding site of 'pulmonary valve structure' (or a subtype of 'pulmonary valve structure') and an associated morphology of 'stenosis' (or a subtype of 'stenosis').

```
   << 404684003 |clinical finding| :
        363698007 |finding site| = << 39057004 |pulmonary valve structure| ,
        116676008 |associated morphology| = << 415582006 |stenosis|
```

The SNOMED CT Query Language is a formal language used to represent SNOMED CT Queries. This language is based on the same features as the SNOMED CT Expression Constraint Language, with the addition of SNOMED CT specific filters. These filters allow the author of the query to restrict the results based on the version of SNOMED CT being used and the value of SNOMED CT's release file fields (e.g. definitionStatus, characteristicType, languageCode, term and typeId). Additional keywords are also provided (e.g. preferredTerm, fullySpecifiedName) to simplify the use of common filter combinations. For example, the following SNOMED CT query finds all fully defined diseases which have a preferredTerm (in the GB English language reference set) that contains the substring "heart".

```
<< 64572001 |disease| {{ definitionStatus = 900000000000073002 |defined|,
     preferredTerm = ".*heart.*", languageRefSet = 900000000000508004 |GB English| }}
```

B2i's Snow Owl Terminology Server (case study 2.5) supports the execution of SNOMED CT queries using a precursor to the SNOMED CT Expression Constraint Language (referred to as 'Extended SNOMED CT Compositional Grammar' or 'ESCG').

9.2.2 SNOMED CT APIs

An Application Programming Interface (API) for a SNOMED CT enabled terminology server can be used to request the execution of SNOMED CT searches and queries. Using a terminology server API, record management systems are able to effectively access terminology services without re-implementing their functionality in every system.

A number of commercial terminology servers offer proprietary APIs that enable SNOMED CT search and query, including Dataline's SnAPI solution and B2i's Snow Owl Terminology Server (case study 2.5). An example of a script which uses the B2i's Snow Owl API to execute a SNOMED CT query is shown below:

```
import com.b2international.snowowl.scripting.services.EscgEvaluatorService

def escgQuery = """

<<404684003|Clinical finding| :

  246454002|Occurrence| = 255399007|Congenital|,

  370135005|Pathological process|=<<263680009|Autoimmune|

"""                                                    //initialize a query

def escgEvaluator = new EscgEvaluatorService()         //initialize a service for evaluating a query

def concepts = escgEvaluator.evaluate(escgQuery)       //evaluate the query

concepts.each { println "ID: ${it.id}, ${it.label}" }  //prints the result to the console
```

Standardized APIs for terminology services are also available. In particular, HL7's Common Terminology Services 2 (CTS 2) provides a standardized API that supports access to terminology servers that may contain a variety of code systems, including SNOMED CT.

9.3 PATIENT RECORD QUERIES

The query language used to query a set of patient records is usually dependent on the type of database used to store the patient records. For example:

- Relational databases may be queried using SQL (Structured Query Language)
- Object-oriented databases may be queried using OQL (Object Query Language)
- RDF databases may be queried using SPARQL (SPARQL Protocol And RDF Query Language)
- XML databases may be queried using XQuery (XML Query Language)
- OLAP databases may be queried using MDX (Multidimensional Expressions)

However, some query languages support logical queries that are independent of the application, programming languages, system environment and storage models - for example, AQL (Archetype Query Language) and EQL (EHR Query Language). These languages instead focus on queries based on the relevant information models (called 'archetypes').

To get the most benefit from using SNOMED CT in patient records, however, one must be able to not only query the records themselves, but also query SNOMED CT.

One way of achieving this is to include a list of all possible SNOMED CT codes that are required within the query. For example, to find the patients with a Respiratory system disorder, one could include every individual code that is a descendant of 50043002 |disorder of respiratory system| (around 3000 codes) within the patient record query. Using SQL, this would look like:

```
SELECT DISTINCT PatientID FROM ProblemList
WHERE Code IN (140004, 181007, 222008, 490008, 517007, 599006, 652005, 663008, etc)
```

However, this creates a lengthy query that is difficult to both validate and maintain. In some cases, it may also be too long to be accepted by the query engine.

Another approach would be to use a subset of respiratory system disorders, and load these into a separate table – for example:

```
SELECT DISTINCT PatientID FROM ProblemList
WHERE Code IN (SELECT * FROM RespiratorySystemDisorders)
```

However, it may not be scalable to create a new table for each terminology query that is required.

A third approach would be to use a transitive closure table to test the hierarchical relationship between each SNOMED CT code and 50043002 |disorder of respiratory system|. For example,

```
SELECT DISTINCT PatientID FROM ProblemList PL
INNER JOIN SNOMEDTransitiveClosure TC ON TC.SourceId = PL.Code
WHERE TC.TargetId = 50043002
```

However, to support a more advanced style of query that utilizes the full capabilities of SNOMED CT, SNOMED CT query languages or API calls must be embedded within the patient record query languages. For example, the following queries use the SNOMED CT Expression Constraint Language embedded within a SQL query.

```
SELECT DISTINCT PatientID FROM ProblemList
WHERE Code IN (< 50043002 |disorder of respiratory system| )
```

```
SELECT DISTINCT PatientID FROM ProblemList
WHERE Code IN (<< 404684003 |clinical finding| :
        363698007 |finding site| = << 39057004 |pulmonary valve structure|,
        116676008 |associated morphology| = << 415582006 |stenosis|)
```

10 USER INTERFACE DESIGN

10.1 OVERVIEW

In this section, we consider how user interfaces can be designed to harness the capabilities of SNOMED CT, and to make clinical querying more accessible to non-technical users. We describe both user interfaces for authoring queries, as well as user interfaces for viewing query results.

10.2 QUERY INTERFACE

When querying patient records containing SNOMED CT-enabled data, a variety of interfaces may be adopted to support the user in authoring queries. In this section we first consider user interfaces for querying SNOMED CT, and then look at user interfaces for querying SNOMED CT enabled patient records.

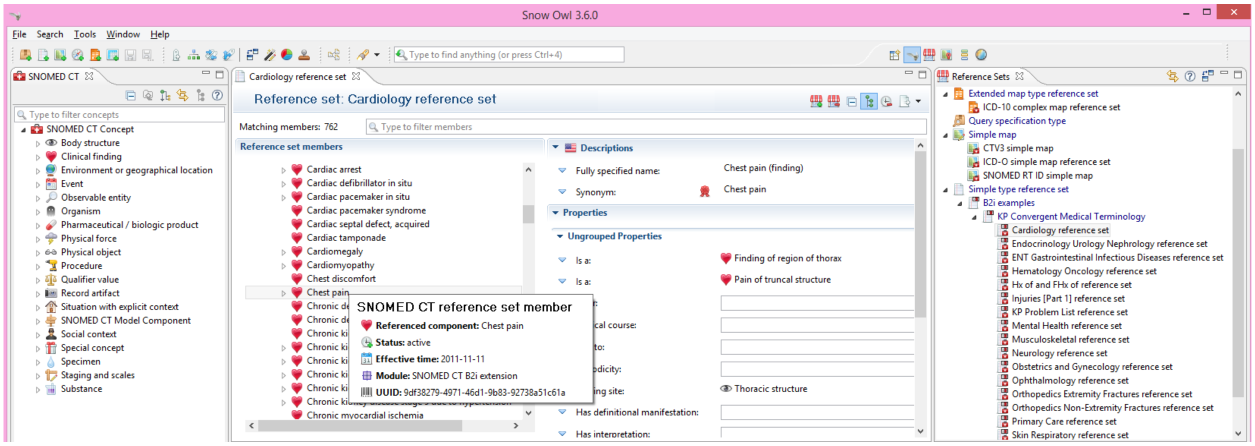
10.2.1 TERMINOLOGY QUERY INTERFACES

When querying clinical data, it may be necessary to first define a subset of SNOMED CT concepts (e.g. disorders or procedures) that may then be compared against values in a patient record. A number of different options exist for creating these SNOMED CT subsets, including:

1. Selecting individual SNOMED CT concepts (i.e. extensional definition)
2. Authoring queries directly using a query language (i.e. intensional definition)
3. Authoring queries using a structured form (i.e. a form which generates an intensional definition)

10.2.1.1 SELECTING INDIVIDUAL CONCEPTS

This approach uses a SNOMED CT browser to allow individual SNOMED CT concepts to be searched, selected and added to a subset. For large subsets this can be quite time consuming, however it is quite suitable for smaller subsets. A number of commercial tools are available which help to perform this task, including Apelon’s Distributed Terminology System (case study 2.4) and B2i’s Snow Owl terminology server (case study 2.5). Figure 10-1 below illustrates Snow Owl’s authoring interface for Simple reference sets.



SNOMED CT Concept
- Body structure
- Clinical finding
- Environment or geographical location
- Event
- Observable entity
- Organism
- Pharmaceutical / biologic product
- Physical force
- Physical object
- Procedure
- Qualifier value
- Record artifact
- Situation with explicit context
- SNOMED CT Model Component
- Social context
- Special concept
- Specimen
- Staging and scales
- Substance

Reference set: Cardiology reference set

Matching members: 762

Reference set members
- Cardiac arrest
- Cardiac defibrillator in situ
- Cardiac pacemaker in situ
- Cardiac pacemaker syndrome
- Cardiac septal defect, acquired
- Cardiac tamponade
- Cardiomegaly
- Cardiomyopathy
- Chest discomfort
- Chest pain
- Chronic de
- Chronic de
- Chronic ki
- Chronic ki
- Chronic ki
- Chronic ki
- Chronic ki
- Chronic kidney disease stage 5 due to hypertension
- Chronic myocardial ischemia

SNOMED CT reference set member
- Referenced component: Chest pain
- Status: active
- Effective time: 2011-11-11
- Module: SNOMED CT B2i extension
- UUID: 9df38279-4971-46d1-9b83-92738a51c61a

Descriptions
- Fully specified name: Chest pain (finding)
- Synonym: Chest pain

Properties

Ungrouped Properties
- Is a: Finding of region of thorax
- Is a: Pain of truncal structure
- r:
- ical course:
- to:
- odicity:
- ing site: Thoracic structure
- Has definitional manifestation:
- Has interpretation:

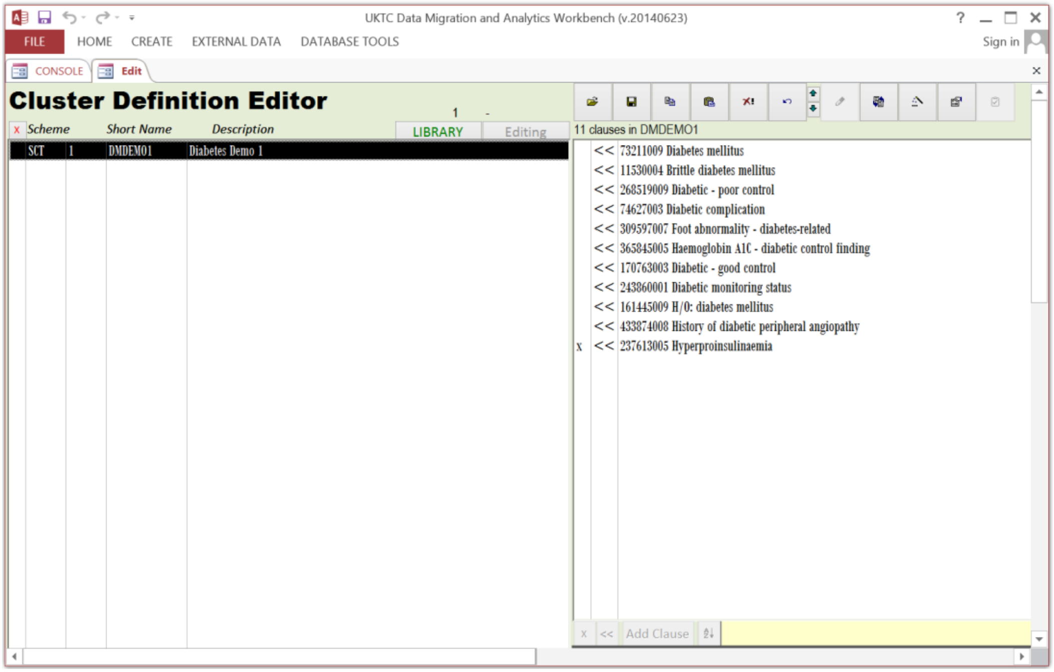
Reference Sets
- Extended map type reference set
  - ICD-10 complex map reference set
- Query specification type
- Simple map
  - CTV3 simple map
  - ICD-O simple map reference set
  - SNOMED RT ID simple map
- Simple type reference set
  - B2i examples
    - KP Convergent Medical Terminology
      - Cardiology reference set
      - Endocrinology Urology Nephrology reference set
      - ENT Gastrointestinal Infectious Diseases reference set
      - Hematology Oncology reference set
      - Hx of and FHx of reference set
      - Injuries [Part 1] reference set
      - KP Problem List reference set
      - Mental Health reference set
      - Musculoskeletal reference set
      - Neurology reference set
      - Obstetrics and Gynecology reference set
      - Ophthalmology reference set
      - Orthopedics Extremity Fractures reference set
      - Orthopedics Non-Extremity Fractures reference set
      - Primary Care reference set
      - Skin Respiratory reference set

Figure 10-1: B2i's Snow Owl interface for authoring Simple Reference Sets

10.2.1.2 AUTHORING QUERIES USING A QUERY LANGUAGE

Other user interfaces allow a subset to be defined using a text-based query written using a predefined query language (e.g. SNOMED CT Expression Constraint Language, or SNOMED CT Query Language). These interfaces tend to be for the more technical user. However, some clinical users may be taught to use these interfaces if required.

Two examples of this style of interface are illustrated in Figure 10-2 and Figure 10-3. Figure 10-2 shows the NHS Data Migration Workbench query interface, while Figure 10-3 shows B2i's Snow Owl query interface.



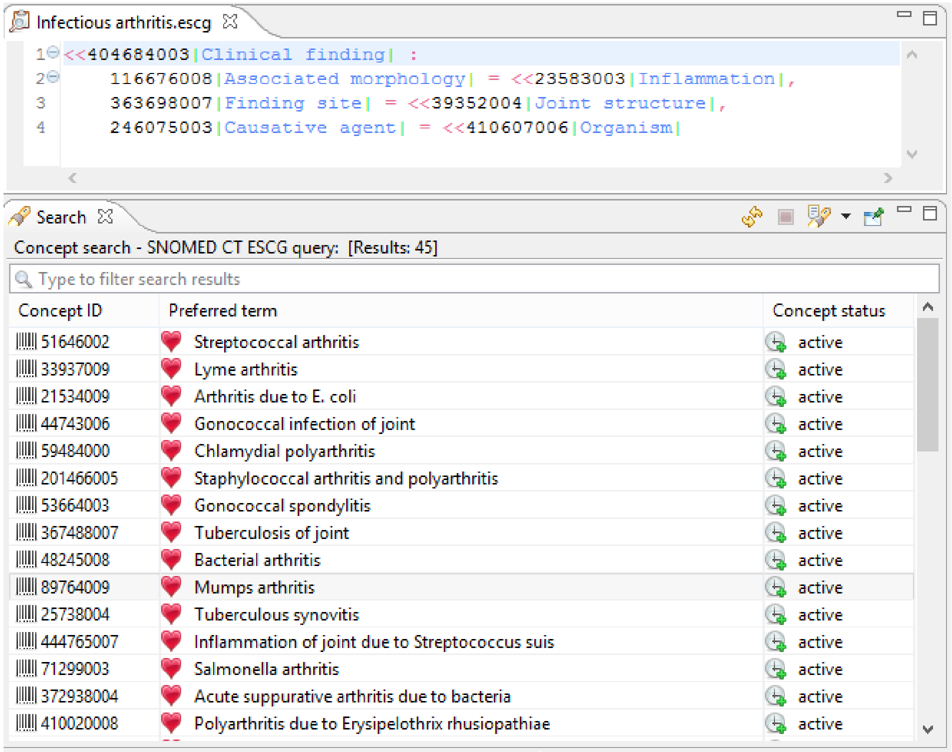
Cluster Definition Editor

Scheme	Short Name	Description
SCT	1	DMDEMO1	Diabetes Demo 1

11 clauses in DMDEMO1

```
  << 73211009 Diabetes mellitus
  << 11530004 Brittle diabetes mellitus
  << 268519009 Diabetic - poor control
  << 74627003 Diabetic complication
  << 309597007 Foot abnormality - diabetes-related
  << 365845005 Haemoglobin A1C - diabetic control finding
  << 170763003 Diabetic - good control
  << 243860001 Diabetic monitoring status
  << 161445009 H/O: diabetes mellitus
  << 433874008 History of diabetic peripheral angiopathy
x << 237613005 Hyperproinsulinaemia
```

Figure 10-2: NHS Data Migration Workbench interface for authoring queries



Infectious arthritis.escg

```
<<404684003|Clinical finding| :
    116676008|Associated morphology| = <<23583003|Inflammation|,
    363698007|Finding site| = <<39352004|Joint structure|,
    246075003|Causative agent| = <<410607006|Organism|
```

Search

Concept search - SNOMED CT ESCG query: [Results: 45]

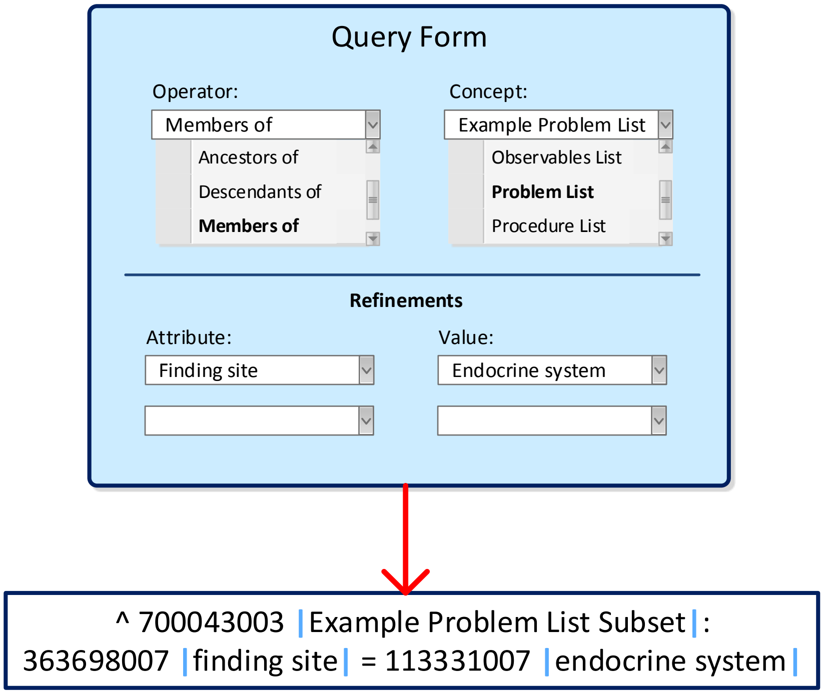
Type to filter search results

Concept ID	Preferred term	Concept status
51646002	Streptococcal arthritis	active
33937009	Lyme arthritis	active
21534009	Arthritis due to E. coli	active
44743006	Gonococcal infection of joint	active
59484000	Chlamydial polyarthritis	active
201466005	Staphylococcal arthritis and polyarthritis	active
53664003	Gonococcal spondylitis	active
367488007	Tuberculosis of joint	active
48245008	Bacterial arthritis	active
89764009	Mumps arthritis	active
25738004	Tuberculous synovitis	active
444765007	Inflammation of joint due to Streptococcus suis	active
71299003	Salmonella arthritis	active
372938004	Acute suppurative arthritis due to bacteria	active
410020008	Polyarthritis due to Erysipelothrix rhusiopathiae	active

Figure 10-3: B2i's Snow Owl interface for authoring text-based queries

10.2.1.3 AUTHORING QUERIES USING A STRUCTURED FORM

A third style of user interfaces for authoring SNOMED CT subsets uses a structured form. A form-driven query tool may allow the user to select an operator (e.g. 'memberOf', 'descendantOf'), the concept or subset to which this operator is applied (e.g. 'Example problem list', 'Disorder'), and then one or more attribute values to limit the set of concepts returned. (Note: The attribute name may either be selected from a list, or hard coded on the form). Once the form is completed, a text-based query is automatically constructed from the selected values, and executed against SNOMED CT. This style of interface can be designed to allow users to exploit the rich semantics of SNOMED CT, while shielding them from the underlying technical details. Figure 10-4 illustrates how a generic form-driven interface for authoring SNOMED CT queries works. Vendor products which implement form-driven interfaces for authoring SNOMED CT queries include B2i's Meaningful Query web interface (as shown in Figure 10-5).



Query Form

Operator: Members of

Concept: Example Problem List

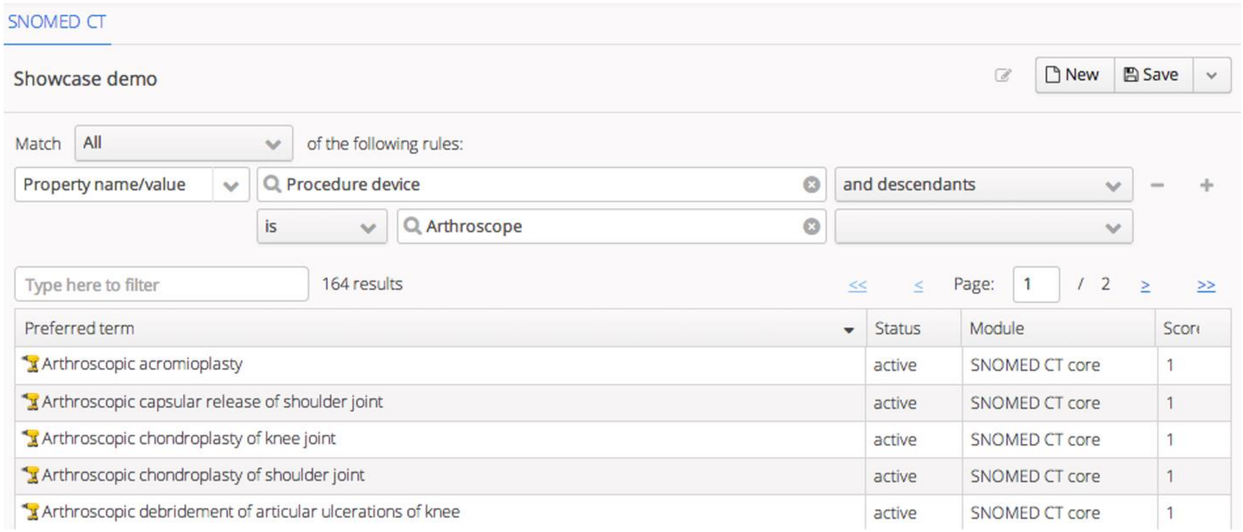
Refinements

Attribute: Finding site

Value: Endocrine system

^ 700043003 |Example Problem List Subset| :
363698007 |finding site| = 113331007 |endocrine system|

Figure 10-4: A generic form-driven interface for authoring SNOMED CT queries



Match All of the following rules:

Property name/value | Procedure device | and descendants

is | Arthroscope

Type here to filter 164 results Page: 1 / 2

Preferred term	Status	Module	Scor
Arthroscopic acromioplasty	active	SNOMED CT core	1
Arthroscopic capsular release of shoulder joint	active	SNOMED CT core	1
Arthroscopic chondroplasty of knee joint	active	SNOMED CT core	1
Arthroscopic chondroplasty of shoulder joint	active	SNOMED CT core	1
Arthroscopic debridement of articular ulcerations of knee	active	SNOMED CT core	1

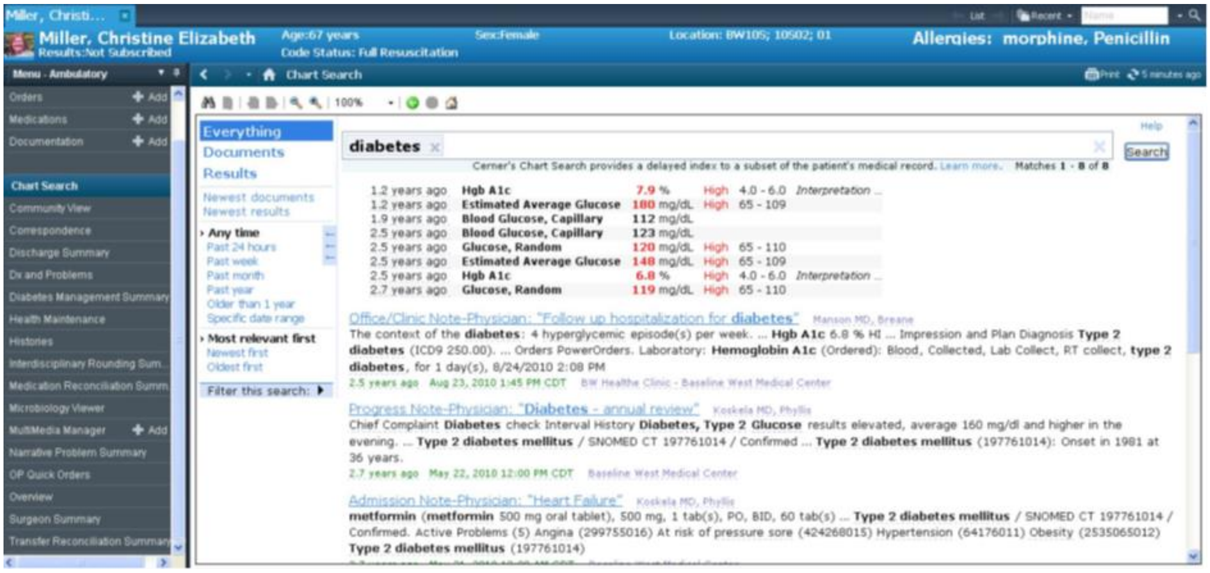
Figure 10-5: B2i Snow Owl's Meaningful Query web interface

10.2.2 PATIENT RECORD QUERY INTERFACES

When SNOMED CT queries are integrated (or embedded) into patient records queries, additional constraints are often added across demographic data (e.g. age, address) and episode of care data (e.g. healthcare provider, dates). These data items are often referred to as 'concrete values' and are typically not included in a terminology. A number of styles of interfaces are used to author patient record queries that include SNOMED CT content, including:

1. Free text semantic search
2. Queries using a predefined language (e.g. SQL, XQL, OQL or AQL)
3. Queries using a structured form (including both SNOMED CT and concrete value criteria)

Figure 10-6 shows an example of a search for ‘diabetes’ using Cerner’s Semantic Search tool (case study 2.8). This tool enables clinicians at the point of care to search in real time through a patient’s multiple charts, pathology reports and other documents for topics such as ‘heart disease’ and ‘diabetes’, using SNOMED CT’s hierarchical and non-hierarchical relationships.



Miller, Christine Elizabeth
Results:Not Subscribed
Age:67 years
Code Status: Full Resuscitation
Sex:Female
Location: BW105; 10502; 01
Allergies: morphine, Penicillin

diabetes

Cerner's Chart Search provides a delayed index to a subset of the patient's medical record. Learn more. Matches 1 - 8 of 8

1.2 years ago	**Hgb A1c**	**7.9** %	High	4.0 - 6.0	*Interpretation ...*
1.2 years ago	**Estimated Average Glucose**	**180** mg/dL	High	65 - 109	
1.9 years ago	**Blood Glucose, Capillary**	**112** mg/dL			
2.5 years ago	**Blood Glucose, Capillary**	**123** mg/dL			
2.5 years ago	**Glucose, Random**	**120** mg/dL	High	65 - 110	
2.5 years ago	**Estimated Average Glucose**	**148** mg/dL	High	65 - 109	
2.5 years ago	**Hgb A1c**	**6.8** %	High	4.0 - 6.0	*Interpretation ...*
2.7 years ago	**Glucose, Random**	**119** mg/dL	High	65 - 110	

Office/Clinic Note-Physician: "Follow up hospitalization for **diabetes**" Manson MD, Breane
The context of the **diabetes**: 4 hyperglycemic episode(s) per week. ... **Hgb A1c** 6.8 % HI ... Impression and Plan Diagnosis **Type 2 diabetes** (ICD9 250.00). ... Orders PowerOrders. Laboratory: **Hemoglobin A1c** (Ordered): Blood, Collected, Lab Collect, RT collect, **type 2 diabetes**, for 1 day(s), 8/24/2010 2:08 PM
2.5 years ago Aug 23, 2010 1:45 PM CDT BW Healthe Clinic - Baseline West Medical Center

Progress Note-Physician: "**Diabetes** - annual review" Koskela MD, Phyllis
Chief Complaint **Diabetes** check Interval History **Diabetes, Type 2 Glucose** results elevated, average 160 mg/dl and higher in the evening. ... **Type 2 diabetes mellitus** / SNOMED CT 197761014 / Confirmed ... **Type 2 diabetes mellitus** (197761014): Onset in 1981 at 36 years.
2.7 years ago May 22, 2010 12:00 PM CDT Baseline West Medical Center

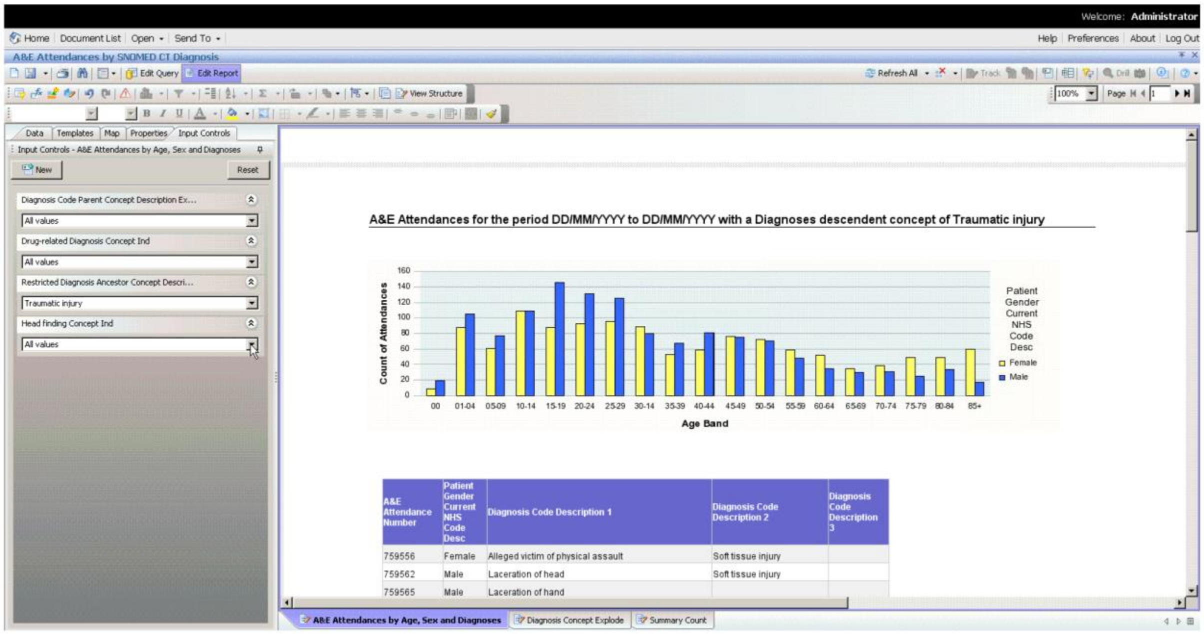
Admission Note-Physician: "Heart Failure" Koskela MD, Phyllis
metformin (**metformin** 500 mg oral tablet), 500 mg, 1 tab(s), PO, BID, 60 tab(s) ... **Type 2 diabetes mellitus** / SNOMED CT 197761014 / Confirmed. Active Problems (5) Angina (299755016) At risk of pressure sore (424268015) Hypertension (64176011) Obesity (2535065012) **Type 2 diabetes mellitus** (197761014)

Figure 10-6: User interface of Cerner’s Chart Search/Semantic tool

10.3 RESULTS VISUALIZATION

When a SNOMED CT enabled query over patient records is executed, the results of this query can be visualized in a number of ways, including tables, charts, scatter diagrams and colored epidemiology maps. While some of these results visualization techniques can be used with any coding system, others are able to utilize the unique features of SNOMED CT in powerful ways.

For example, Figure 10-7 shows a report produced by Cerner’s data warehouse query tool. This tool uses a simple graphical interface which directly creates powerful reports using the SNOMED CT hierarchy content. The screenshot below in Figure 10-7 shows a report of attendances with diagnoses which are a descendant of the SNOMED CT concept 417746004 |traumatic injury|.


Welcome: Administrator
Home | Document List | Open | Send To
Help | Preferences | About | Log Out
A&E Attendances by SNOMED CT Diagnosis
Edit Query
Edit Report
Refresh All
Track
Drill
View Structure
100%
Page 1
Data | Templates | Map | Properties | Input Controls
New
Reset
160
140
120
100
80
60
40
20
0
Count of Attendances
00
01-04
05-09
10-14
15-19
20-24
25-29
30-14
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-84
85+
Age Band
Patient Gender Current NHS Code Desc
Female
Male
A&E Attendances by Age, Sex and Diagnoses
Diagnosis Concept Explode
Summary Count

Input Controls - A&E Attendances by Age, Sex and Diagnoses

Diagnosis Code Parent Concept Description Ex...: All values

Drug-related Diagnosis Concept Ind: All values

Restricted Diagnosis Ancestor Concept Descri...: Traumatic injury

Head finding Concept Ind: All values

A&E Attendances for the period DD/MM/YYYY to DD/MM/YYYY with a Diagnoses descendent concept of Traumatic injury

A&E Attendance Number	Patient Gender Current NHS Code Desc	Diagnosis Code Description 1	Diagnosis Code Description 2	Diagnosis Code Description 3
759556	Female	Alleged victim of physical assault	Soft tissue injury	
759562	Male	Laceration of head	Soft tissue injury	
759565	Male	Laceration of hand		

Figure 10-7: Report produced by Cerner's data warehouse query tool

SNOMED CT's rich polyhierarchy provides a vast number of potential 'aggregators' for analytics, and possible views of SNOMED CT encoded data. This polyhierarchy can be exploited by visual exploratory data analysis tools to enable the visual inspection of complex datasets.

For example, the NHS have been using the Gephi open-source network analysis and visualization software, to explore SNOMED CT encoded renal datasets.

The first representation (in Figure 10-8) shows a projection of all concepts *directly* coded in the patient data, with the node size reflecting the frequency of each code. 36689008 |acute pyelonephritis| has a high frequency in the data and is therefore represented by a big node, while 254915003 |clear cell carcinoma of kidney| has a low frequency in the data and is therefore represented by a small node.

Using a simple concentration algorithm, which aggregates subsumed concepts up to a given threshold, the representation in Figure 10-9 is achieved. In this representation, the size of the purple nodes reflects the frequency of each code *plus* its subtypes, the size of the blue nodes reflects the frequency of each code's subtypes, and the size of the red nodes reflects the frequency of each code on its own. This enables trends to be visually detected – for example, 36171008 |glomerulonephritis| and 36171008 |acute pyelonephritis| - even when the frequency of these concepts themselves is relatively low.

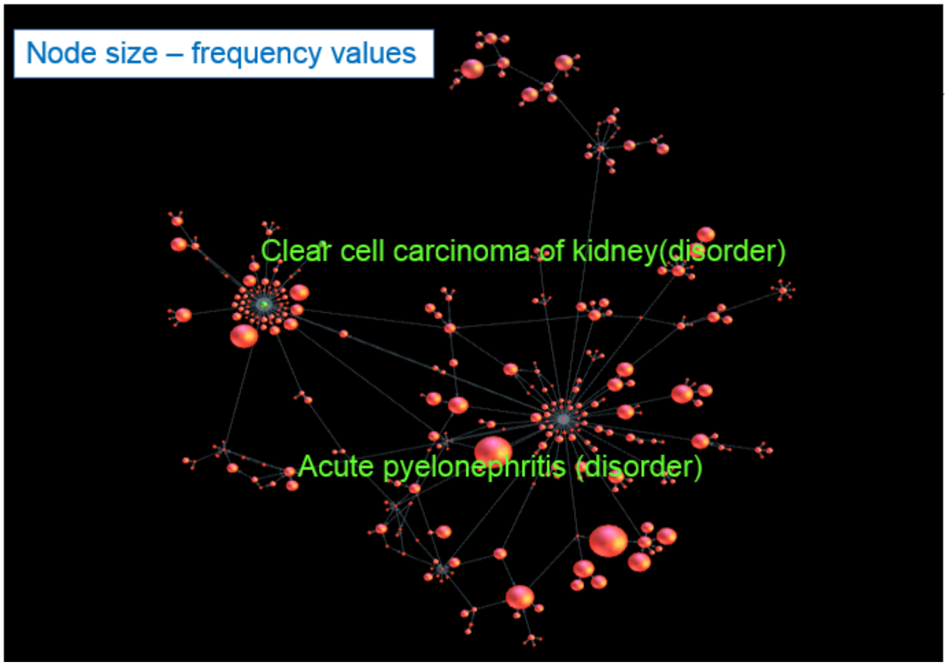

Node size – frequency values
Clear cell carcinoma of kidney(disorder)
Acute pyelonephritis (disorder)

Figure 10-8: Gephi representation of renal dataset showing direct code usage

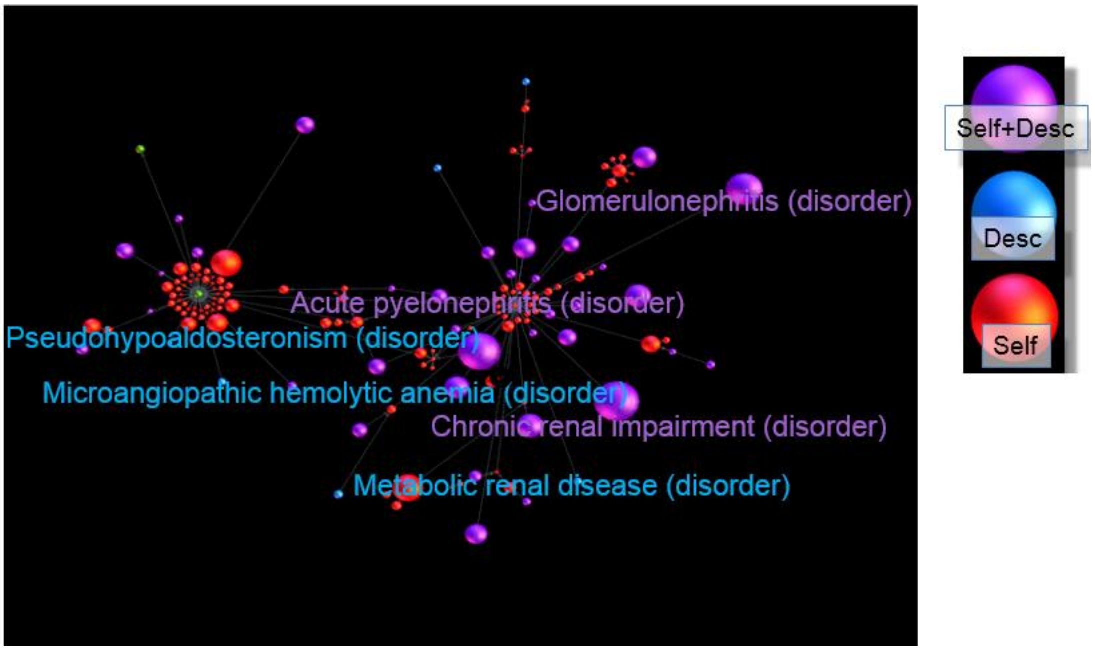

Glomerulonephritis (disorder)
Acute pyelonephritis (disorder)
Pseudohypoaldosteronism (disorder)
Microangiopathic hemolytic anemia (disorder)
Chronic renal impairment (disorder)
Metabolic renal disease (disorder)
Self+Desc
Desc
Self

Figure 10-9: Gephi representation of renal dataset with direct and inherited code usage

Innovative data visualization and analysis tooling is expected to become much more widespread as the powerful capabilities of SNOMED CT content are increasingly utilized.

11 CHALLENGES

11.1 OVERVIEW

This section discusses some of the challenges which should be considered when performing analytics over clinical data. Most of these challenges result from the fundamental nature of health record information, and therefore exist irrespective of the code system used. Many of these challenges are able to be mitigated using the unique features of SNOMED CT. The challenges fall into four broad categories:

- Reliability of patient data
- Terminology / information model boundary issues
- Concept definition issues
- Versioning

SNOMED CT offers significant advantages, compared to other code systems, in both performing powerful clinical analytics, and in mitigating many of these challenges.

11.2 RELIABILITY OF PATIENT DATA

High quality data collection is imperative to the quality and accuracy of analytics results, irrespective of the terminology used. Whether the focus is decision support, business intelligence, research or a mixture of all three - data quality is critical. High quality information is not the consequence of collecting as much data as possible. Instead, it is the product of intentionality and process design.

The factors that may impact the quality of patient data include:

- The design of user interfaces used to capture data

  Clinical user interfaces should be designed to make it as easy as possible to find the most appropriate code, and as difficult as possible to enter the wrong code. There are a variety of ways to improve the ease and effectiveness of data entry using SNOMED CT – such as searching over all synonyms, confirming the selected concept using the preferred term or fully specified name, ordering value lists effectively using an ordered reference set, searching using navigation hierarchies, and constraining data entry using subsets [10]. These techniques can also help to reduce data entry errors by prohibiting invalid input, helping the user to understand the correct meaning of the code selected, and ordering value lists in a clinically safe order (e.g. ordering medications by strength, rather than alphabetically).

- Use of diagnostic criteria to standardize data capture

  Diagnostic criteria and their application tends to vary widely according to care setting, patient status and healthcare professional. The consistent ascertainment and recording of even common diagnoses, such as asthma and myocardial infarction is often non-trivial. High quality prospective research studies require that diagnostic criteria for the condition being studied are understood, rigorously applied and accurately documented. In routine clinical practice doing this for potentially thousands of diagnoses in dozens of care settings is normally infeasible. Divergence and inconsistencies in criteria for diagnosis capture can undermine the validity of any conclusions which may be drawn from analytics. SNOMED CT mitigates this issue by allowing the query author to choose a reliable aggregating concept from SNOMED CT's extensive content.

- Consistency of data capture with analytics requirements
  Pick lists and constraints should be consistent with both clinical data collection needs and analytic requirements and these should never be in conflict. The presence or absence of particular concepts in value sets within different applications can cause data collection to be inconsistent. SNOMED CT mitigates this by allowing the query author to choose a reliable aggregating concept.
- Loss of meaning during data transformations
  Clinical data often undergoes a number of structural transformations and code mappings prior to data analytics being performed, during the process of preparing the data for messaging and/or loading into a data warehouse. In each of these transformations, care must be taken to ensure that the quality of the process is high, and that there is no incremental shift in the clinical meaning of the data. For example, mapping local codes to an alternate code system using non-equivalence maps (e.g. narrow to broad or broad to narrow) will change the clinical meaning of these codes to some degree. Any changes that effect the clinical meaning of the data may have an impact on the quality of data analytics. SNOMED CT helps to mitigate this by supporting the representation of equivalence maps, which can be used when the use case requires.

11.3 TERMINOLOGY / INFORMATION MODEL BOUNDARY ISSUES

When performing data analytics over clinical data, it is important to understand the interdependency between the terminology and the structural information model. For example, it is not sufficient to find a diagnosis of 56265001 |heart disease|, and make the assumption that the patient has heart disease. Instead, the surrounding information model must be considered to discover whether this is, for example, a confirmed diagnosis for the patient themselves, a suspected or preliminary diagnosis for the patient, or perhaps a family history of heart disease in the patient's paternal grandfather. Contextual or qualifying information about a code may appear in a variety of places, including:

- Within the information model – for example, a section heading titled "Family History"
- In the same coded data element – for example, precoordinated as "394886001|suspected heart disease|" or postcoordinated as "56265001 |heart disease|: 408729009 |finding context| = 415684004|suspected|"
- In a separate coded data element – for example, Diagnosis = 56265001 |heart disease|, Type = 148006 |preliminary diagnosis|

By understanding where and how this contextual or qualifying information is represented, more appropriate queries can be created.

When the same semantics may be represented in both the information model *and* the terminology, there is also a risk of ambiguity as to how these two representations should be combined. This is clearly demonstrated by models in which both the information model and the terminology can represent 'negation' or 'absence'. Does the combination of 'negation' in the information model and 'absence' in the terminology indicate:

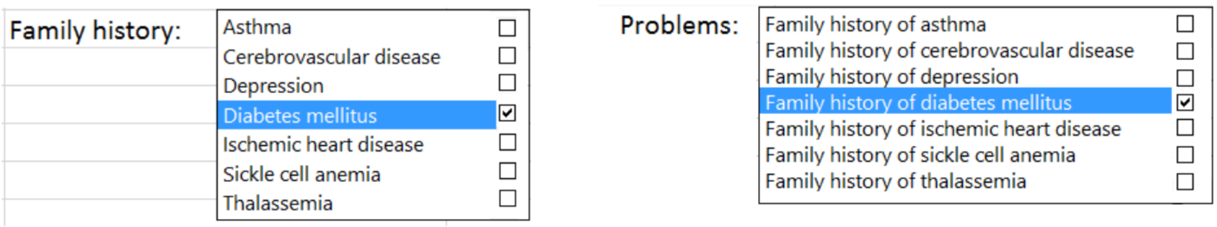
- Double negative,

- Redundant restatement of the negative, or
- Additional emphasis of the negative?

It is important in these situations to have clear rules about how the semantics in the information model and the terminology should be combined.

The challenge often becomes even greater when heterogeneous data sources are integrated. When different information models represent the same semantics using different combinations of structure versus terminology, retrieval and reuse may miss similar information. To avoid false negatives or false positives in the query results, the integration and/or analytics processes must resolve these differences.

For example, in Figure 11-1 below, the system on the left uses the 'Family history' structural heading to indicate that the selected disease is a family history, while the system on the right precoordinates this within the terminology. When integrating or querying across these data sources, these semantics need to be harmonized to ensure accurate queries can be performed.



Family history:

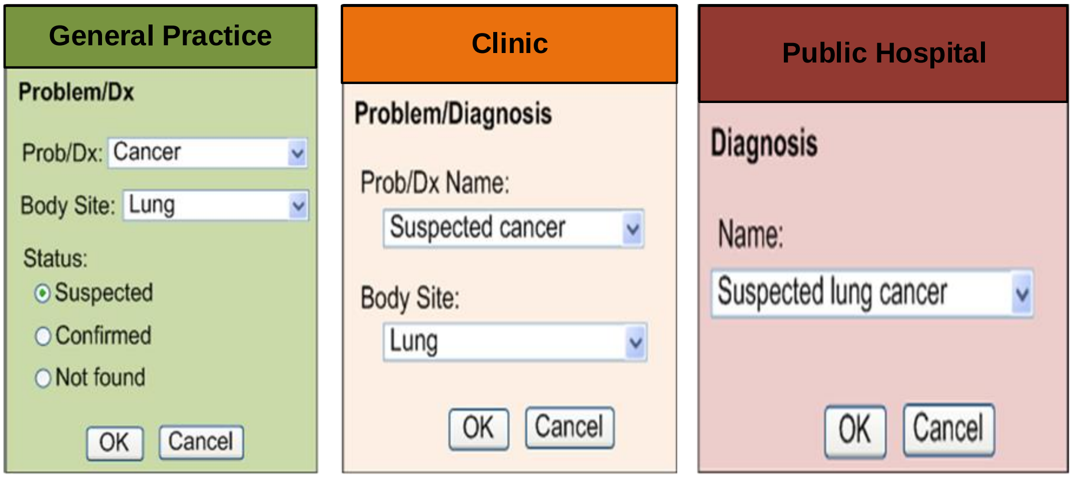
Asthma	□
Cerebrovascular disease	□
Depression	□
Diabetes mellitus	☑
Ischemic heart disease	□
Sickle cell anemia	□
Thalassemia	□

Problems:

Family history of asthma	□
Family history of cerebrovascular disease	□
Family history of depression	□
Family history of diabetes mellitus	☑
Family history of ischemic heart disease	□
Family history of sickle cell anemia	□
Family history of thalassemia	□

Figure 11-1: Two ways of recording family history of diabetes mellitus

Even when the same information model is used, different systems may populate this model with differing levels of precoordination. For example, the three clinical systems shown below in Figure 11-2 each collect data about a 'suspected lung cancer' diagnosis in a different way. For this reason, when given a common data model (as shown in Figure 11-3), different systems may populate this in different ways. When this occurs, queries must be careful to consider all possible representations of the data, to ensure that contextual and qualifying information about each code is correctly interpreted.



General Practice

Problem/Dx

Prob/Dx: Cancer

Body Site: Lung

Status:
● Suspected
○ Confirmed
○ Not found

Clinic

Problem/Diagnosis

Prob/Dx Name: Suspected cancer

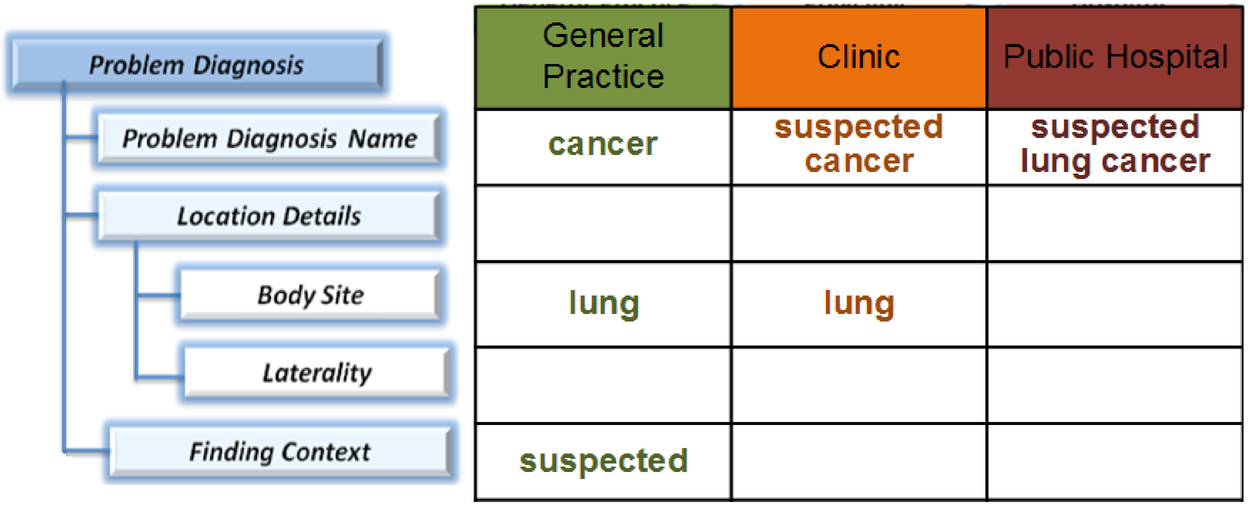
Body Site: Lung

Public Hospital

Diagnosis

Name: Suspected lung cancer

Figure 11-2: Three ways of recording suspected lung cancer



Problem Diagnosis	General Practice	Clinic	Public Hospital
Problem Diagnosis Name	cancer	suspected cancer	suspected lung cancer
Location Details			
Body Site	lung	lung	
Laterality			
Finding Context	suspected		

Figure 11-3: Three ways of populating a common Problem Diagnosis model

SNOMED CT is in the unique position to be able to resolve many of these challenges, using the techniques described in section 6.5 and 0. For example, SNOMED CT enables the computation of equivalence and subsumption between alternative representations of data. For example, the postcoordinated expression

22253000 |pain|: 363698007 |finding site| = 56459004 |foot|

(which can be represented either in a single data element or using two separate data elements for 22253000 |pain| and 56459004 |foot|) can be automatically determined to be equivalent to the precoordinated concept 47933007 |foot pain| (stored in a single data element).

Some cases exist, however, where SNOMED CT is not currently able to automatically establish equivalence. These cases primarily relate to concepts for which the SNOMED CT concept model does not yet fully model their meaning. For example, the two approaches for representing a ‘twin pregnancy’ shown below (Figure 11-4) are currently not able to be computed as equivalent using SNOMED CT.

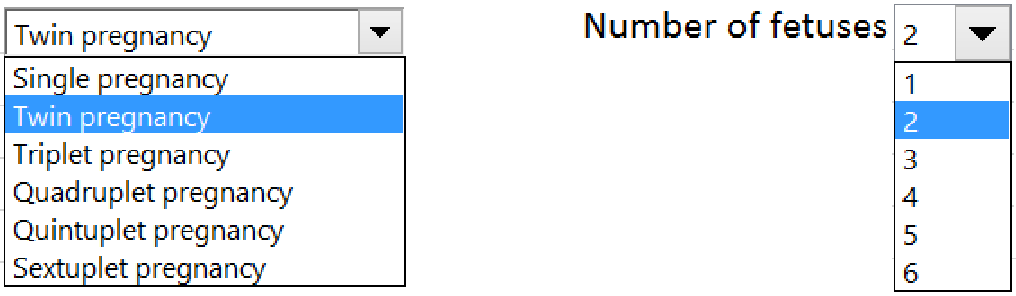


Figure 11-4: Two non-equivalent ways of recording a twin pregnancy using SNOMED CT

The SNOMED CT concept model continues to be extended to support equivalence and subsumption testing within an increasing number of hierarchies of SNOMED CT.

11.4 CONCEPT DEFINITION ISSUES

11.4.1 OVERVIEW

While SNOMED CT is the most comprehensive clinical terminology in the world, containing an extensive set of logic-based definitions which enable a broad range of powerful analytics, some challenges still exist, including:

- Logical versus vernacular
- Minimum sufficient sets
- Incomplete modelling

These challenges are described in more detail in this section.

11.4.2 LOGICAL VERSUS VERNACULAR

In some cases, the strict logical meaning of a term may differ somewhat from the local vernacular (or common) use of that term. For example, the assertions below in SNOMED CT are logically sound but may be counterintuitive to clinicians:

- |insect bite of nose| **is a subtype of** |head injury|
- |laceration of radial artery| **is a subtype of** |cardiovascular disease|.

Examples, such as these, exist in which the formal logical definitions of these concepts may lead to hierarchies that differ from what may be expected by some clinicians.

11.4.3 MINIMUM SUFFICIENT SETS

SNOMED CT definitions include the set of necessary and sufficient conditions that define the given concept. However, SNOMED CT does not currently distinguish the minimum sets which are sufficient to define these concepts. For example, the defining relationships of 154283005 |pulmonary tuberculosis| are:

116680003 |is a| = 64572001 |disease|
246075003 |causative agent| = 113858008 |mycobacterium tuberculosis complex|
116676008 |associated morphology| = 6266001 |granulomatous inflammation|
363698007 |finding site| = 39607008 |lung structure|

While the associated morphology of 'granulomatous inflammation' is necessarily present, the following set of defining relationships are sufficient to infer 154283005 |pulmonary tuberculosis|:

116680003 |is a| = 64572001 |disease|
246075003 |causative agent| = 113858008 |mycobacterium tuberculosis complex|
363698007 |finding site| = 39607008 |lung structure|

As a consequence if the following expression was recorded in a health record:

```
64572001 |disease|:
    246075003 |causative agent| = 113858008 |mycobacterium tuberculosis complex|
    363698007 |finding site| = 39607008 |lung structure|
```

This expression would *not* be returned by the following query:

<< 154283005 |pulmonary tuberculosis|

However, the query:

< 64572001 |disease|:
246075003 |causative agent| = << 113858008 |mycobacterium tuberculosis complex|
363698007 |finding site| = << 39607008 |lung structure|

would correctly return both the concept "154283005 |pulmonary tuberculosis|" and the above expression as required. In this way, the design of appropriate queries can help to mitigate this issue.

11.4.4 INCOMPLETE MODELLING

The SNOMED CT Concept Model continues to evolve to allow more concepts to be fully defined. For example, the 'Observable Entity' and 'Substance' hierarchies each have new concept models being developed, which will allow these concepts to be more fully defined in future releases of SNOMED CT. When the concept models for these hierarchies are incorporated, SNOMED CT's expressive power and analytics capabilities will be further expanded.

In those hierarchies for which the concept model has been established for some time (e.g. Clinical finding), ongoing expansion to SNOMED CT's formal logical definitions continues. However, there still remains some concepts which do not yet have all possible defining relationships included. This issue will be mitigated over time as more of SNOMED CT's concepts continue to be modelled.

11.5 VERSIONING

A new version of the International Edition of SNOMED CT is released twice a year (in January and July). National extensions mostly follow this cycle (albeit typically with a three month delay). However, some extensions (notably those including medication related concepts) are released more frequently.

When a longitudinal health record is populated with clinical data over a number of years, it is quite possible that the following may occur:

1. SNOMED CT concepts that were active at the time of recording have since been made inactive
2. SNOMED CT concepts that were primitive at the time of recording have since been defined
3. Reference sets that were used to populate pick lists may have changed
4. The SNOMED CT Concept Model that was used to construct expressions may have changed

To mitigate these versioning issues, SNOMED CT provides the following:

1. Each new version of the SNOMED CT International Edition that is released (in Release Format 2 - RF2) includes a set of Delta files (containing all changes to the content since the last release), a set of Snapshot files (containing the most recent version of every component that has ever been released in SNOMED CT), and a set of Full files (containing every version of every component that has ever been released in SNOMED CT). These files allow implementations to either incrementally adapt to new versions of SNOMED CT, or alternatively load a complete current snapshot of SNOMED CT content (with or without old versions). When longitudinal clinical records containing inactive concepts are queried, all prior descriptions and relationships of

these inactive concepts can still be queried using these snapshot files. SNOMED CT's RF2 distribution files also record the reason that each inactive component was inactivated, using 'historical association' reference sets (see section 5.6.2.11.4 of the Technical Implementation Guide [4] for more details).

2. SNOMED CT is maintained on the principle that every SNOMED CT concept identifier should retain its semantic integrity over time, even when its logical definition changes. The semantics of a SNOMED CT concept is established through its Fully Specified Name, and all changes to a concept's defining relationships are intended to improve the machine-readable processing of these semantics. That said, it is possible if required to determine what the logical definition of a concept was at any prior point in time using a Full release of SNOMED CT.
3. SNOMED CT's reference sets and their members are all fully versioned in SNOMED CT's RF2. A Snapshot release of a reference set includes the current version of every row that has ever been released (including both active and inactive rows). A Full release of a reference set includes every version of every row that has ever been released. Using this information, it is possible to adapt queries to consider both current and former members of any given reference set.
4. The SNOMED CT Concept Model changes very rarely. When it does, however, any attributes that are retired are retained as inactive concepts in the Snapshot and Full releases of SNOMED CT. It is expected that a complete Machine Readable Concept Model (MRCM) of SNOMED CT will be published in the future, and that this MRCM will be versioned in a manner that is consistent with other RF2 components.

12 REFERENCES

[1] Centers for Disease Control and Prevention (CDC), *Principles of Epidemiology in Public Health Practice, 3rd Edition*, 1978, http://www.cdc.gov/osels/scientific_edu/ss1978/lesson1/Section5.html.

[2] IHSTDO, *SNOMED CT Starter Guide*, 22 Feb 2014, http://snomed.org/sg.pdf.html.

[3] Wikipedia, *Analytics*, 2014, http://en.wikipedia.org/wiki/Analytics.

[4] IHTSDO, *SNOMED CT Technical Implementation Guide*, http://www.snomed.org/tig.

[5] IHTSDO, *Building the Business Case for SNOMED CT*, 2014, http://www.ihtsdo.org/resource/resource/98.

[6] IHTSDO, *Vendor Introduction to SNOMED CT*, 2015, http://snomed.org/vendorintro.pdf.html.

[7] IHTSDO, *SNOMED CT Lexical Resources*, SNOMED CT Document Library, http://snomed.org/lexical_resources.zip.

[8] NHS Health and Social Care Information Centre, *Summary Care Records*, 2014, http://systems.hscic.gov.uk/scr.

[9] S. Sachdea, *SCR reaches 40m patients*, E-Health Insider, 2 July 2014, http://www.ehi.co.uk/news/EHI/9500/scr-reaches-40m-patients.

[10] IHTSDO, *Search and Data Entry Guide*, 2014, http://ihtsdo.org/collab/doc10211.

[11] O. Bodenreider, *Using SNOMED CT in combination with MedDRA for reporting signal detection and adverse drug reactions reporting*, AMIA Annu Symp Proc., p. 45–49, 2009.

[12] W. W. W. C. (W3C), *OWL 2 Web Ontology Language Document Overview (Second Edition)*, 2012, http://www.w3.org/TR/owl2-overview/ .

[13] Wikipedia, *Natural language processing*, 2014, http://en.wikipedia.org/wiki/Natural_language_processing .

[14] I2B2, *NLP Research Data Sets*, 2014, https://www.i2b2.org/NLP/DataSets/Main.php .

[15] Nyström M, Vikström A, Nilsson G, Örman H, Åhlfeldt H., *Visualization of disease distribution with SNOMED CT and ICD-10*, World Congress on Medical and Health Informatics 2010, Cape Town (South Africa), p. 1100-3.

[16] Stanford University, *Protégé*, http://protege.stanford.edu/ .