Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

STRUCTURAL INDICATORS FOR EFFECTIVE QUALITY ASSURANCE OF SNOMED CT

by
Ankur Agrawal

The Standardized Nomenclature of Medicine - Clinical Terms (SNOMED CT - further abbreviated as SCT) has been endorsed as a premier clinical terminology by many national and international organizations. The US Government has chosen SCT to play a significant role in its initiative to promote Electronic Health Record (EHR) country-wide. However, there is evidence suggesting that, at the moment, SCT is not optimally modeled for its intended use by healthcare practitioners. There is a need to perform quality assurance (QA) of SCT to help expedite its use as a reference terminology for clinical purposes as planned for EHR use.

The central theme of this dissertation is to define a group-based auditing methodology to effectively identify concepts of SCT that require QA. As such, similarity sets are introduced which are groups of concepts that are lexically identical except for one word. Concepts in a similarity set are expected to be modeled in a consistent way. If not, the set is considered to be inconsistent and submitted for review by an auditor. Initial studies found 38% of such sets to be inconsistent. The effectiveness of these sets is further improved through the use of three structural indicators. Using such indicators as the number of parents, relationships and role groups, up to 70% of the similarity sets and 32.6% of the concepts are found to exhibit inconsistencies.

Furthermore, positional similarity sets, which are similarity sets with the same position of the differing word in the concept’s terms, are introduced to improve the

likelihood of finding errors at the concept level. This strictness in the position of the differing word increases the lexical similarity between the concepts of a set thereby increasing the contrast between lexical similarities and modeling differences. This increase in contrast increases the likelihood of finding inconsistencies. The effectiveness of positional similarity sets in finding inconsistencies is further improved by using the same three structural indicators as discussed above in the generation of these sets. An analysis of 50 sample sets with differences in the number of relationships reveal 41.6% of the concepts to be inconsistent.

Moreover, a study is performed to fully automate the process of suggesting attributes to enhance the modeling of SCT concepts using positional similarity sets. A technique is also used to automatically suggest the corresponding target values. An analysis of 50 sample concepts show that, of the 103 suggested attributes, 67 are manually confirmed to be correct.

Finally, a study is conducted to examine the readiness of SCT problem list (PL) to support meaningful use of EHR. The results show that the concepts in PL suffer from the same issues as general SCT concepts, although to a slightly lesser extent, and do require further QA efforts. To support such efforts, structural indicators in the form of the number of parents and the number of words are shown to be effective in ferreting out potentially problematic concepts in which QA efforts should be focused. A structural indicator to find concepts with synonymy problems is also presented by finding pairs of SCT concepts that map to the same UMLS concept.

STRUCTURAL INDICATORS FOR EFFECTIVE QUALITY ASSURANCE OF SNOMED CT

by
Ankur Agrawal

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

Department of Computer Science

August 2013

APPROVAL PAGE

STRUCTURAL INDICATORS FOR EFFECTIVE QUALITY ASSURANCE OF SNOMED CT

Ankur Agrawal

Dr. Yehoshua Perl, Dissertation Co-Advisor — Date
Professor, Computer Science Department, NJIT

Dr. Mei Liu, Dissertation Co-Advisor — Date
Assistant Professor, Computer Science Department, NJIT

Dr. Gai Elhanan, Committee Member — Date
Chief Medical Information Officer, Halfpenny Technologies

Dr. Michael Halper, Committee Member — Date
Program Director & Professor, Information Technology Program, NJIT

Dr. James Geller, Committee Member — Date
Chair & Professor, Computer Science Department, NJIT

Dr. Chunhua Weng, Committee Member — Date
Assistant Professor, Columbia University

BIOGRAPHICAL SKETCH

Author: Ankur Agrawal

Degree: Doctor of Philosophy

Date: July 2013

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2013

- Bachelor of Engineering in Computer Engineering,
  Purbanchal University, Biratnagar, Nepal, 2006

Major: Computer Science

Presentations and Publications:

Agrawal A, Perl Y, Chen Y, Elhanan G, Liu M, Identifying inconsistencies in SNOMED CT problem lists using structural indicators, To appear in AMIA annual symposium proceedings, Washington D.C., November 2013.

He Z, Ochs C, Agrawal A, Perl Y, Zeginis D, Tarabanis K, Elhanan G, A family-based framework for supporting quality assurance of biomedical ontologies in BioPortal, To appear in AMIA annual symposium proceedings, Washington D.C., November 2013.

Agrawal A, Perl Y, Elhanan G, Identifying problematic concepts in SNOMED CT using a lexical approach, To appear in MedInfo proceedings, Copenhagen, August 2013.

Agrawal A, He Z, Perl Y, Wei D, Halper M, Elhanan G, The readiness of SNOMED problem list concepts for meaningful use of electronic health records, Journal of Artificial Intelligence in Medicine, pp. 73-80, Jun 2013.

Ochs C, Agrawal A, Perl Y, Halper M, Tu SW, Carini S, Sim I, Noy N, Musen M, Geller J, Deriving an abstraction network to support quality assurance in OCRe, AMIA annual symposium proceedings, Chicago, IL, pp. 681-689, November 2012.

Agrawal A, Elhanan G, Halper M, Dissimilarities in the logical modeling of apparently similar concepts in SNOMED CT, AMIA annual symposium proceedings, Washington D.C., pp. 212-216, November 2010.

To My Beloved Parents,
Mr. Suresh Agrawal & Mrs. Kanta Agrawal

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my advisor Dr. Yehoshua Perl for his constant guidance, encouragement and feedback during the course of my PhD degree. I am most grateful to my co-advisor Dr. Mei Liu for patiently correcting my writing and to Dr. Gai Elhanan for guiding and mentoring my research and putting in a huge amount of effort in doing most of the auditing work. My sincere thanks also goes to my other dissertation committee members including Dr. Michael Halper, Dr. James Geller and Dr. Chunhua Weng for their continuous support and guidance. I would also like to thank Dr. Yan Chen for doing some of the auditing work during the course of my research.

I would like to acknowledge Dr. Marino Xanthos and Ms. Clarisa Gonzalez-Lenahan for being great advisors and friends of the Graduate Student Association. I am especially grateful to Mr. Jeff Grundy for all his help as the director of the Office of International Students and to Dr. George Olsen for writing me tens of recommendation letters during the process of my job search. I greatly appreciate Dr. David Nassimi for all his support as the doctoral program director of the Department of Computer Science and Ms. Angel Bell for all her effort as the department secretary.

Special thanks to my mom, Mrs. Kanta Agrawal, and my dad, Mr. Suresh Agrawal, who always encouraged me to go for higher education and pursue my goals as a researcher in computer science. I also greatly appreciate the support of my sister, Ekta Agrawal, and my brother, Pratik Agrawal. Last, but by no means least, I would like to thank all my friends at NJIT, especially Ishita Biswas, Kashif Qazi and Suresh Solaimuthu, who made my life here pleasant and joyful.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF TABLES
(Continued)

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Significance of SNOMED CT

SNOMED CT (SCT) [1] is slated to become an integral part of Health Information Technology (HIT) systems. Encoding patients' problems in Electronic Health Records (EHRs) by using concepts from SCT has been proposed as part of the "meaningful use" of such systems. The Health Information Technology for Economic and Clinical Health Act (HITECH) component of the American Recovery and Reinvestment Act [2] was designed to jumpstart the transition of medical providers to use electronic Health Information Systems (HISs) [3]. In the proposal for the initial HIT standards [4], SCT is to be used to "enable a user to electronically record, modify, and retrieve a patient's problem list for longitudinal care (i.e., over multiple office visits)." To accelerate the adoption and meaningful use of EHR by providers, incentives and penalties were defined [4, 5]. The use of SCT to encode problem lists of current and active diagnoses for at least 80% of all unique patients was proposed as one indication of meaningful use. Moreover, SCT is slated to become the exclusive coding system for problem lists by 2015.

1.2 Importance of SNOMED CT in Electronic Health Record

The past decade has witnessed uproar in the way information technology can be used in various sectors for storing and retrieving data. One sector that still lags behind in successfully embracing the information technology is the field of health informatics which serves hospitals, clinics and other health care facilities. EHRs have several significant advantages over the traditional method of using pen and paper to record

patient data [6]. One of them is the efficiency in the retrieval of records. Digging through the paper files can be bothersome and time consuming. Quick access to records can be lifesaving during emergencies. EHRs can be used to quickly retrieve patient information without any delay thus facilitating more efficient decision making process. Another important advantage of EHRs over paper records is the ease of sharing patient data. When a patient goes to a new doctor, he no longer has to worry about filling up forms with his medical history. The doctor can easily access the medical history by using the patient's personal identifying information. Furthermore, clinical notes are more legible as they are computer formatted instead of handwritten. In short, EHRs can help improve healthcare quality in many ways.

The benefits of EHRs depend on several factors. An important aspect is the reference terminology that it relies on to record patient data in a standard and consistent manner. The EHR can only be as good as the quality of the reference terminology being used. The reference terminology, as such, plays an important role in determining the large scale adoption of EHRs by healthcare providers. Since SCT has been touted as the primary reference terminology to be used in EHRs to record patient data, the success in the adoption of EHRs rely heavily on the quality of SCT.

1.3 Current State of SNOMED CT

SCT is regarded as a comprehensive, high quality terminology which can be used in EHRs to record a patient's clinical data. SCT contains more than 290,000 active concepts (July 2012 release) spread over 19 broad hierarchies like the *Clinical Finding*, *Body Structure*, *Specimen*, etc. As such, SCT provides a good coverage of clinical concepts. In fact, in [7], SCT was found to cover 88.4% of the diagnosis/problem list terms used by

clinicians within a computerized physician order entry (CPOE) system. Because of its comprehensiveness and good coverage, SCT has been proposed to be used in a variety of settings [8].

Despite the good coverage of the diagnosis/problem lists, SCT still exhibits deficiencies in its structure and modeling that can hamper its use in EHRs. As an example of a problem with relationship modeling, the fully-defined concept *Acute myocardial infarction* is not hierarchically linked to the concept *Ischemic heart disease*. There is also no physiological attribute linking it to the associated myocardial ischemic process. A physician attempting to encode *Acute myocardial infarction* in an EHR is likely to search for it in an ischemic heart disease subsection. Not being able to locate the desired concept is liable to result in frustration according to Rector et al. [74]. In the same paper, Rector et al. give several other examples showing various modeling problems in SCT such as diabetes being classified as a disease of the abdomen and arteries of the foot being placed in pelvis. The authors conclude that without further quality assurance, clinicians may not realize the implications of what they are saying; researchers may not realize what their queries should retrieve, and post-coordination cannot be expected to be reliable thus compromising the interoperability and meaningful use of EHR.

As for concept modeling, SCT lacks in sufficient synonyms that accompany the concepts. Only 36% of SCT concepts have synonyms. Besides, around 77% of SCT's concepts are primitive, i.e., they lack the necessary relationships for full definition. These deficiencies of SCT can adversely affect the applications dependent on it.

HITECH's meaningful use regulations include provisions for decision support [5] to improve performance on high priority health conditions and covers both clinical and

patient related aspects. The meaningful use regulations actually call for a context-sensitive form of decision support. For example, a documentation form is presented for patients with diabetes that includes a required section for the diabetic foot exam, where the same form would be presented for patients without diabetes and with the diabetic foot exam section removed. Similarly, Certified Electronic Health Record Technology (CEHRT) can suggest that a patient with diabetes should be referred to a diabetic foot screener. While such decision support can be achieved by simply hardcoding the linkage between a diagnosis to a specific form or activity, it is easy to see how hierarchical and especially lateral relationships in a controlled terminology might support a dynamic form of context-sensitive decision support. The use of controlled terminologies to enable such forms of decision support has been described in the past [9-16].

However, accurate and consistent modeling of hierarchical and attribute relationships is critical for dynamic, context sensitive secondary use of clinical controlled terminologies. Much of the decision support proposed in meaningful use regulations is diagnosis related. Thus, SCT's major role in encoding problem lists stands to directly affect the ability of CEHRT to provide dynamic, context-sensitive decision support. The fact that a concept is marked primitive indicates a potential deficiency in its relationship structure, which in turn may lead to its incorrect positioning in a hierarchy by a DL classifier. Missing or incorrect relationships influence the inheritance of properties. CEHRT, which takes advantage of SCT's inherent structure, can thus be negatively impacted. As such, quality assurance of SCT becomes very important.

An intensive auditing effort is urgently needed to ensure quality assurance in SCT. However, an extensive audit of all concepts of SCT requires extensive quality

assurance resources and will require a long time. A desired approach in coping with this urgent quality assurance need is to develop computational techniques for identifying subsets of SCT with higher concentration of errors. This will result in more errors being corrected for a given amount of QA effort and resources.

1.4 Impact of the Study

The dissertation presents semi-automatic and automatic techniques to identify subsets of SCT and problem list concepts that would be more prone to modeling errors. A group based auditing technique is presented that creates groups of concepts from SCT that are lexically similar but differ in their parameters. In addition, techniques using the number of parents and words as indicators of the concept complexity are presented. These techniques enable auditors to focus on those concepts that have been identified by the algorithm as being more prone to error.

The work in this document will, thus, allow for rapid discovery of those concepts needing improvement in SCT and will help improve the hierarchical and relational modeling of those concepts. This will lead to an improvement of the content of SCT and problem list and will result in better encoding of problem lists in EHRs. This should ultimately result in a quality and affordable healthcare, which is one ultimate goal of the current HITECH initiative.

1.5 Test Bed

The test bed for the studies performed in this document would be the *Procedure* hierarchy from SCT and the SCT problem lists (Clinical Observations Recording and Encoding (CORE) [17] and Veterans Health Administration and Kaiser Permanente

(VA/KP) [18]). However, the techniques developed are general enough to be applied to other SCT hierarchies. This will allow for the usability of the technique across wide range of applications, benefiting different classes of users.

1.6 Dissertation Overview

In Chapter 3, it is hypothesized that the lexically similar concepts should have similar modeling and if they are not modeled in a similar way, the modeling may contain inconsistencies. A study is conducted by partitioning the concepts into groups of lexically similar concepts. These concepts are then analyzed for inconsistencies in their modeling and the results are presented. The study further introduces the usage of structural indicators in grouping similar concepts. The underlying idea is that the introduction of such indicators into the generation of similarity sets would help in further exposing the inconsistent modeling among similar concepts thus increasing the likelihood of finding inconsistent concepts. A preliminary study has already been published in [19].

Chapter 4 builds on the work done in Chapter 3 with an aim to increase the likelihood of finding inconsistencies among the concepts of the similarity sets. Positional similarity sets are introduced and the methodology is further refined by using structural indicators such as number of parents, relationships and role groups. Again, a preliminary study was published in [20].

Chapter 5 introduces a methodology to algorithmically suggest attributes to enhance the modeling of SCT concepts. The methodology builds on the positional similarity sets introduced in Chapter 4. A technique is also identified to automatically suggest the corresponding target values.

Chapter 6 presents a study to examine the readiness of SCT problem lists (PL) to support meaningful use of EHRs. The study is conducted on two random sample sets of SCT concepts. The first consist of concepts strictly from the PL. The second contain general SCT concepts distributed proportionally to the PL's in terms of their hierarchies. Each of the two sample sets is analyzed for modeling errors. The result of the analyses is presented and two structural indicators are suggested to locate inconsistencies in hierarchical relationships with statistical significance. A third structural indicator is suggested to identify missing synonyms. A part of the study has been published in [21] and another part of the study has been accepted for publication in [22].

CHAPTER 2

BACKGROUND

2.1 SNOMED CT Overview

SCT [1] is a controlled clinical reference terminology with comprehensive coverage of clinical findings, diseases, procedures, therapies and outcomes intended for recording clinical data [7, 23]. This data can be made available to computer systems for clinical decision support [24] and improved patient safety [25-27].

SCT started as a pathology-specific nomenclature (SNOP) [28, 29] in 1965 and since then has extended into other medical fields. SCT got its current form after the merger of College of American Pathologists' (CAP) SNOMED RT (Reference Terminology) [30] and the UK National Health Service's (NHS) Clinical Terms Version 3 (also known as the Read codes) [31]. In 2007, the SCT intellectual property rights were transferred from the CAP to the SNOMED SDO in the formal creation of the International Health Terminology Standards Development Organization (IHTSDO) [32].

SCT can cross map to other international standards such as ICD-9-CM, ICD-10 and OPCS-4. It supports ANSI, DICOM, HL7, and ISO standards. SCT is currently available in American English, British English, Spanish, Danish, and Swedish with other translations under way.

A new version of SCT is released by IHTSDO every six months in January and in July. The content of SCT evolves with each release with changes in concepts, descriptions, and relationships. A history mechanism keeps track of these changes over time.

The January 2013 release of SCT consists of more than 310,000 active concepts with unique meanings and formal logic-based definitions organized into 19 hierarchies. The hierarchies are comprised of parent-child relationships, meaning that broader concepts are at the top of the hierarchy (parent) followed by more specific concepts (child). An example of a parent-child relationship would be *blood test* and *laboratory test* in which *laboratory test* is the parent and *blood test* is the child because blood test is a type of laboratory test.

There are different structural parameters associated with these concepts such as relationships (hierarchical and attribute) and groups. These parameters help in extending the meaning of these concepts. With more than 1.4 million relationships, part of SCT's power lies in the relationships built into its core clinical concepts. For example, the attribute *finding site* connects *acute subglottic laryngitis* to *subglottis structure*, conveying the knowledge that an *acute subglottic laryngitis* involves only that particular structure and not any other structure. SCT's technical documentation [33] outlines well defined rules for the domains and ranges of its defining attributes and attribute values.

Each SCT concept has a collection of descriptions (terms), including one fully specified name (FSN), along with a preferred term and possibly a set of synonyms. SCT's more than 1.1 million English language descriptions offer flexibility in expressing the concepts, thus enabling clinicians to say things in multiple ways and still be understood.

Each concept is further classified by its status of logical definition: fully-defined vs. primitive. A primitive concept is underspecified in the sense that not enough attributes are available to distinguish it from its parents (and siblings) and the automated detection of its sub-concepts is not allowed.

SCT and its precursors, with 50%–70% coverage of concepts of interest [34, 35], have consistently outperformed other sources. In [7], SCT was found to cover 88.4% of the diagnosis/problem list terms used by clinicians within a computerized physician order entry (CPOE) system. In 2004, the VA concluded that SCT has promise as a coding system for clinical problems [36]. A survey in 2010 indicated that 68% of users perceived SCT's coverage as satisfactory or better [37, 38]. In [39], SCT was deemed suitable to provide standardized representations of information created by two interface terminologies, noting that enriching SCT semantics would improve representation of the external terms.

The most commonly perceived use of terminologies such as SCT is the encoding of clinical data within electronic medical systems including EHRs and Clinical Information systems (CIS). SCT has been utilized or proposed for use in a variety of settings. The American Academy of Ophthalmology, for example, has chosen SCT as its official clinical terminology [40]. An extensive literature review regarding the use of SCT in clinical practice is presented in [8]. In [41], another literature search sought to identify SCT applications in critical care. The findings revealed investigations of SCT or its actual use in the representation of disorders of newborn infants [42], nursing flow-sheets [43], allergic diseases and associated problems [44], the representation of common patient problems [23], anesthesia patient safety [25], and intensive care [45]. SCT has also been used in the automatic grouping of adverse drug reactions terms [46] and in the annotation of tissue microarray data [47].

2.2 The Underlying Description Logic Model

SCT is designed to use Description Logic (DL) as the underlying knowledge representation model [48]. As such, operations like concept union, negation, intersection and subsumption are supported in SCT. SCT's DL underpinnings can be used by DL classifiers to ensure internal consistency. SCT has two main views: a stated, explicit view and the commonly available inferred view which is derived by a DL classifier. The same SCT infrastructure also supports semantic reasoners. Inferences derived by reasoners can form the basis for sophisticated decision-support tools and applications. However, the performance of classifiers and reasoners is directly related to the completeness and correctness of the logical formalism on which they rely.

2.3 The HITECH Initiative

Reaffirming convictions that electronic information systems are essential to improving healthcare [49], the HITECH component of the American Recovery and Reinvestment Act [2] was designed to jumpstart the transition of medical providers to use electronic health information systems (HISs) [3]. In the proposal for the initial HIT standards [4], SCT is to be used to "enable a user to electronically record, modify, and retrieve a patient's problem list for longitudinal care (i.e., over multiple office visits)." To accelerate the adoption and *meaningful use* of EHRs by providers, incentives and penalties were defined [4, 5]. The encoding of problem lists of current and active diagnoses for at least 80% of all unique patients was proposed as one indication of meaningful use. Moreover, SCT is slated to become the exclusive coding system for problem lists by 2015 [4].

Defining meaningful use of EHRs will involve problem lists encoded with SCT [4, 5]. To facilitate adoption, the NLM has posted the UMLS CORE [50]. Among the sources in the UMLS, SCT covers the highest percentage (81%) of the concepts. In the e-prescribing domain, the U S Food and Drug Administration (FDA) has adopted the VA/KP Problem List Subset of SCT as the terminology to represent indications in electronic labels [18]. In an evaluation of medication indication phrases, SCT, as a whole, covered 90.3%, while its Clinical Finding hierarchy covered 79.5%.

2.4 Problem Lists

A problem list is a "best practices" subset of clinical terms which is most commonly used by clinicians to record patient diagnosis. Problems lists are the essence of problem-oriented approach to medical records which was first introduced by Dr. Lawrence Weed more than 40 years ago [51, 52]. The usefulness and future of such approach was further discussed in [53, 54]. Problem lists are considered to be an important element of the EHR by several standards organization such as the Institute of Medicine, Joint Commission, American Society of Testing and Materials and Health Level Seven [50]. An encoded problem list is also considered one of the core objectives of the "meaningful use" regulation of EHR [5].

SCT is slated to become the standard terminology for EHR encoding of diagnoses and problem lists [4] by 2015. To facilitate the meaningful use of EHRs, the National Library of Medicine (NLM) has published two SCT problem list subsets, CORE and VA/KP to be used in coding patient data for EHR. The CORE subset comprises of datasets submitted by seven institutions - Beth Israel Deaconess Medical Center, Intermountain Healthcare, Kaiser Permanente, Mayo Clinic, Nebraska University

Medical Center, Regenstrief Institute and Hong Kong Hospital Authority [17]. The VA/KP subset was created for indexing Structured Product Labeling (SPL) [18] which is approved by Health Level Seven (HL7) and adopted by the FDA as a mechanism for exchanging medication information.

There are a total of 5,862 current concepts in the January 2012 version of the CORE Problem List Subset of SCT and a total of 16,622 current concepts in the September 2009 release of VA/KP problem list of SCT. The two lists have 4,004 concepts in common. For the purpose of this study, the two problem lists are combined to create a combined SCT problem list (PL). The PL consists of 18,472 unique and current concepts. The January 2012 release of SCT consists of 295,753 current concepts. As such, the PL covers 6.2% of the SCT concepts but over 81% of the most commonly used terms by the clinicians during patient diagnosis.

However, evidence suggests that the two problem lists have substantial quality problems and are not ready to serve the anticipated EHR meaningful use needs. In [21], a preliminary study of the concepts from both CORE and VA/KP problem lists has shown that they suffer from the same problems as SCT concepts in their corresponding hierarchies such as high percentage of erroneous and inconsistent relationships and substantial percentage of primitive concepts. There have also been some comparative studies on the CORE problem list, the VA/KP problem list and problem lists derived from other terminologies like ICD-9. A comparative analysis of CORE and VA/KP problem lists is presented in [55]. An evaluation of the VA/KP problem list is presented in [56]. A comparison between CORE subset and ICD-10-CM/PCS HIPAA code sets is

presented in [57]. The coverage and coding efficiency between CORE subset, the subset used by Mayo Clinic and a random SCT subset was evaluated in [58].

The problem list will play an important role in supporting the effective clinical recording of patient data in EHR to improve patient safety, health care quality, and health information exchange. Therefore, to gain support among its users, the problem list must be of the highest possible quality.

2.5 Auditing Techniques

Auditing is an essential part of SCT's maintenance. SCT is a large and complex clinical terminology containing hundreds of thousands of concepts that are linked by millions of relationships. As such modeling errors are unavoidable in its design which makes the QA of SCT extremely important. The importance of auditing in the design life cycle of a terminology along with its application in SCT was presented in [59]. Auditing SCT can be challenging due to limited resources. Computational techniques to help identify groups of concepts that are more likely to contain errors can lead to an efficient utilization of the limited QA resources.

A review of auditing methods applied to the content of controlled biomedical terminologies such as SCT was presented in [60]. The study presents techniques to measure quality factors related to different aspects of the terminology such as the synonyms and relationship modeling. A guest editorial in the form of a special issue on auditing terminologies appeared in the Journal of Biomedical Informatics in 2009 [61]. These studies explored different approaches in auditing terminologies such as abstraction network based methods, methods based on description logic and ontological principles, and natural language processing techniques.

The uses of description logic and ontological principles to audit SCT have been studied by various researchers such as Bodenreider, Cornet and Ceusters. In [62], seven ontological principles were defined and the properties of SCT was examined with respect to these principles. A method to use abstraction networks in complement to description logic to identify errors in SCT is presented in [63]. Methods for auditing DL-based terminologies like SCT through the detection of concepts with equivalent or inconsistent definitions are presented in [64-66]. Other ontology based techniques favoring the use of strict logical and ontological theories in SCT to help detect different types of errors have been presented in [67, 68]. A formal concept analysis (FCA) based model for auditing the semantic completeness of SCT was presented in [69] and a lattice based structural auditing method was presented in [70]. The use of semantic distance metrics to support auditing of SCT was presented in [71]. The use of evolutionary terminology auditing technique was applied to SCT in [72]. An assessment of the systematic use of linguistic phenomena to represent the lexical and semantic features in SCT was presented in [73]. In [74], Campbell et al. described a lexically suggested logical closure to track the quality of a terminology like SCT. Problems related to the use of grammatical conjunctions "and" and "or" in SCT was presented by Mendonca et al. [75]. A method to detect under-specification in SCT using a lexical technique was presented in [76].

The research group at NJIT's Structural Analysis of Biomedical Ontologies Center (SABOC), has been formulating automated structural methodologies to detect concepts that are likely to contain errors, as part of an effort to make terminology auditing more efficient [77]. Such methodologies have been successfully applied to SCT [77-79]. During the application of these methodologies, situations have been observed

where concepts that were similar in every aspect but were not modeled in the same manner. For example, while *Insertion of Kantrowitz heart pump* has the attribute *direct device* with a value of *Cardiac assist implant* (January 2010 release), its sibling *Removal of Kantrowitz heart pump* has *direct device* with a value of *Device*, a very broad concept. Other hierarchical differences were noted as well.

Most of the existing methods mentioned above have some limitations. Some rely on the assumption that concept definitions are non-primitive (i.e. they are regarded as providing necessary and sufficient conditions). Some require attribute relationships in order to be applicable on a terminology. The analysis in this dissertation focuses on both parts of SCT's definitional structures: hierarchical (i.e., IS-As) and attributes. The latter aspect can be further segmented as assignment of attribute and attribute value. The comparison between the modeling of lexically similar concepts concentrates on both of these aspects.

CHAPTER 3

SIMILARITY SETS

3.1 Introduction

SCT is built upon description logic (DL) principles [80], with each concept being defined by its hierarchical (IS-A) and lateral (attribute) relationships to other concepts in the terminology. From a clinical perspective, particularly from the point of view of human clinicians, the presentation format of concepts in the form of terms (e.g., fully-specified names and preferred names) is often of primary concern. On the other hand, computer programs—particularly those performing some kind of reasoning—are built around the concepts' DL formulations. One would expect that these two perspectives be highly consistent. In particular, terms exhibiting a similar word structure should have underlying DL modeling that is analogous in structure.

For algorithms to work reliably, the validity and consistency of the conceptual representations within CBTs is crucial. Rector *et al* [81] clearly demonstrated the issue utilizing the *Myocardial infarction* example. In SCT (January 2010 release), myocardial infarction is not classified as a type of ischemic heart disease due to incomplete formal logic definitions. As a result, a hypothetical research query that looks to gather all *Ischemic heart disease* patients, relying on SCT coded data, will exclude myocardial infarction patients unless the researchers had prior knowledge of the issue or run their query using an aggregate of all instances of ischemic heart disease. The example crystallizes the implications of incomplete, incorrect, and inconsistent modeling on healthcare applications down the road. Such inconsistencies may be perceived to have minimal implications regarding clinical coding. However, inconsistencies may

significantly affect the performance of reasoners and inference generation (e.g., in the context of error detection and decision support) as these explicitly rely on the completeness and consistency of formal definitions. Therefore, this study analyzes the conceptual representation of sets of concepts similar at the term-level in an attempt to characterize the consistency of the modeling across these concepts. Sets of concepts with similar terms are gathered through standard lexical techniques. Such an analysis is performed on SCT's *Procedure* hierarchy.

The *Procedure* hierarchy of SCT is the most semantically complex of the 19 hierarchies of SCT, with 28 potential defining attributes [82] (Table 3.1). For most attribute domains, SCT defines one or more ranges from which target values can be assigned. For example, the attribute *component* can be assigned target values from four ranges (Table 3.2). Concepts in the *Procedure* hierarchy have an average of 2.4 unique attributes and 1.9 parents per concept (compared with 1.8 and 1.7, respectively, for *Clinical finding*). This makes the *Procedure* hierarchy a prime target to examine methods to explore and detect issues with SCT's formal definitions.

Table 3.1 Defining Attributes for *Procedure* Hierarchy

access	has intent	procedure device	revision status
approach	has specimen	procedure morphology	scale type
component	indirect device	procedure site	time aspect
direct device	indirect morphology	procedure site - direct	using device
direct morphology	measurement method	procedure site - indirect	using access device
direct substance	method	property	using energy
has focus	priority	recipient category	using substances

Table 3.2 Ranges for the *Component* Attribute

Cell structure (cell structure)
Observable entity (observable entity)
Organism (organism)
Substance (substance)

A lexical methodology is used to identify sets of similar concepts and is applied to one of the most attribute-rich hierarchies, Procedure, to create similarity sets which acts as control sets. The methodology to generate the similarity sets is then slightly tweaked to generate four additional set types with at least one concept in the set having different number of parents, different number of relationships, different number of groups and different number of all three of the above. A sample of 50 sets from each of these five set types are examined in regard to hierarchical, definitional, attribute, attribute/value, and role-group aspects. The evaluation revealed that 38 (Control) to 70 percent (Different relationships) of similarity sets within the samples exhibited significant inconsistencies.

3.2 Method

The core assumption is that concepts whose descriptions are of a similar word structure are expected to have similar logical representations. Figure 3.1 displays the snapshot of the modeling of the concept *death due to radiotherapy toxicity* along with that of the concept *death due to chemotherapy toxicity* taken from CliniClue browser. The modeling of the concept *death due to radiotherapy toxicity*, on its own, provides very little help in determining anything wrong or missing for this concept. However, comparing it to a lexically similar concept *death due to chemotherapy toxicity*, as shown in Figure 3.1, makes the inconsistency pretty obvious. The first concept *death due to radiotherapy*

toxicity is missing the relationship *due to* with target value *toxicity due to radiotherapy*. Furthermore, it is highly likely that this concept is defined as primitive because of this missing relationship since the lexically similar concept with such a relationship is modeled as fully defined. It can be seen from this example that comparing lexically similar concepts with each other makes it easy to find incorrect or missing information for the concepts since similarly worded concepts should be modeled in a similar way.

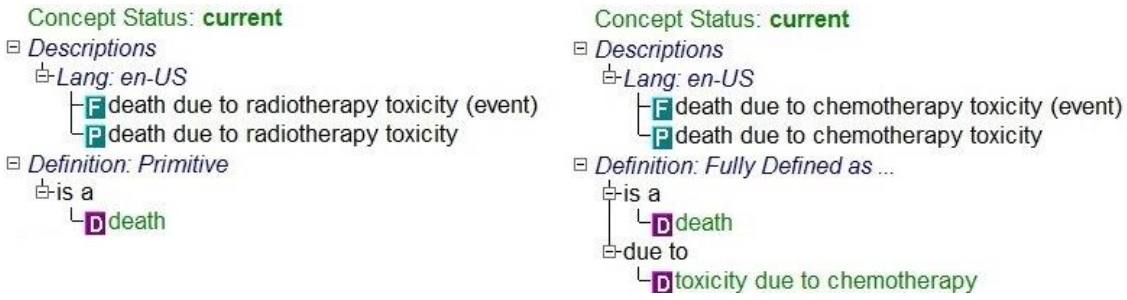

Figure 3.1 CliniClue snapshot of the modeling of two lexically similar concepts about death due to toxicity.

Source: The Clinical Information Consultancy Ltd, "CliniClue Xplore", 2009, Available from: http://www.cliniclue.com.

The methodology is thus based on the formation of groups of concepts where fully specified names (FSNs) are similar in their word structure. In particular, the focus is on FSNs that differ from each other by one word. Such groups are referred to as similarity sets (simply "*sets*," for short). For example, let t_1 and t_2 be five-word FSNs with t_1 = "w_1 w_2 w_3 w_4 w_5" and t_2 = "w_1 w_3 w_4 w_5 w_6", where each w_i is an individual word. Then the concepts of t_1 and t_2 are in a set together because these FSNs differ only by one word: w_2 versus w_6. Standard lexical variations as well as stop-words (like "a," "an," "the") are ignored. Based on preliminary results and for practical reasons, the analysis was limited to FSNs of five words or more, with the semantic tags included in the word

count. An example set of three concepts (with terms of length five) is shown in Table 3.3. The hyphenated “Sperm-cervical” is considered one word.

Table 3.3 Example of a Set containing Three Concepts

CID	FSN
252940006	Sperm-cervical mucus interaction test (procedure)
252942003	Sperm-cervical mucus slide test (procedure)
252943008	Sperm-cervical mucus contact test (procedure)

An inconsistency in a set is defined as any instance where at least one of its concepts could unequivocally incorporate conceptual modeling elements from any other concept in the set. Figure 3.2 depicts a similarity set of two concepts: *Conversion from uncemented total knee replacement* and *Conversion to uncemented total knee replacement*. Both concepts are somewhat ambiguous (arguably the "from" more than the "to" one) since they do not indicate to, or from (respectively), what the conversions occur. Both concepts involve a total knee replacement (TKR) procedure and both are revisions since their FSNs indicate a transition between different types of TKRs. As both concepts are primitives, it cannot be assumed that all the defining information is present. Nevertheless, significant modeling discrepancies are evident. Although both concepts have a single parent, its type is different. The "from" concept is only linked hierarchically to *Revision of knee arthroplasty* even though logically, it must be some form of TKR. The "to" concept, although a revision, is not linked hierarchically to any revision-type parent, not even through an attribute. The "to" concept lacks the *Revision status* attribute but has the *Procedure site - Indirect* and the *Direct device* attributes with their assigned values. As for the assigned attribute values, although both concepts have the attribute

Method, their respective assigned values differ: *Surgical action* for the "from" concept and *Surgical insertion - action* and *Repair - action* for the "to" concept. *Surgical action* is an ancestor of both *Surgical insertion - action* and *Repair - action*. The two possible *Method* values for the "to" concept also highlight that it has two attribute groups whereas the "from" concept has only one group. Thus, utilizing a similarity set of minimum size (two concepts), four different types of possible inconsistencies are demonstrated: hierarchical, attribute assignment, attribute values, and groups. These findings are only minimally affected by the vagueness of the concepts or the auditor's subjectivity.

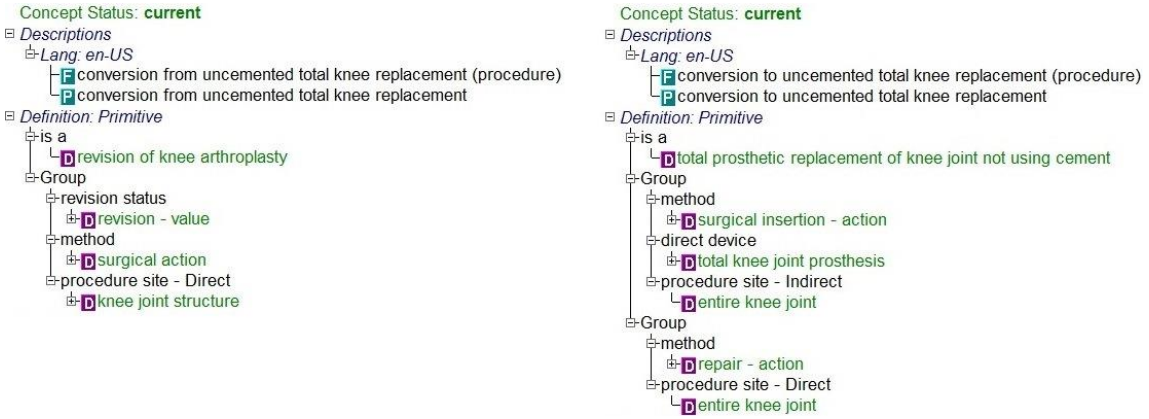

Figure 3.2 A two-concept similarity set concerning total knee replacement.

Source: The Clinical Information Consultancy Ltd, "CliniClue Xplore", 2009, Available from: http://www.cliniclue.com.

Based on the observations, the following four hypotheses are formulated:

Hypothesis 1: Similarity sets whose concepts exhibit different numbers of parents are more likely to harbor inconsistencies than randomly selected similarity sets.

Hypothesis 1.1: The inconsistency type is more likely to be hierarchical.

Hypothesis 2: Similarity sets whose concepts exhibit different numbers of attributes are more likely to harbor inconsistencies than randomly selected similarity sets.

Hypothesis 2.1: The inconsistency type is more likely to be attribute-related.

Hypothesis 3: Similarity sets whose concepts exhibit different numbers of role groups are more likely to harbor inconsistencies than randomly selected similarity sets.

Hypothesis 3.1: The inconsistency type is more likely to be role-group related.

Hypothesis 4: Similarity sets whose concepts exhibit different number of parents, relationships, and groups are more likely to harbor inconsistencies than randomly selected similarity sets.

Accordingly, and based on the inferred view of the January 2011 release of SCT, five samples from SCT's *Procedure* hierarchy are formulated. The first sample (Control) serves as a control sample, composed of concepts that differ from the base concept by one word without respect to the number of parents, relationships, or groups. The rest of the four samples correspond with hypotheses one through four: Diff-Par sample, Diff-Rel sample, Diff-Grp sample, and Diff-All sample. Each sample consists of randomly selected, 50 mutually exclusive similarity sets, controlled only for their respective parameter. In each similarity set (except Control), at least two concepts differ in the number of occurrences of the sample's main criteria.

The samples are presented (non-blinded, single spreadsheet) to, and evaluated by, a single auditor (Dr. Gai Elhanan), a physician with extensive background in controlled biomedical terminologies. The auditor was not looking for errors but rather for clear inconsistencies between the inferred views of the concepts in a similarity set: hierarchical, definitional, attribute assignment, attribute target values, and groups. Within each set, the auditor looked for all types of inconsistencies. The default interface setting of CliniClue Xplore was used for the presentation of concepts for evaluation.

3.3 Results

The *procedure* hierarchy in SCT's January 2011 release contains 52,011 concepts. This makes it the second largest SCT hierarchy after the *Clinical finding* hierarchy. Table 3.4 provides additional information regarding the top five SCT hierarchies by size. For instance, the average number of parents per concept for the *procedure* hierarchy is 1.9, the average number of unique relationships per concept is 2.4 and the average number of groups per concept is 0.8. Also, 40.5% of the concepts in *procedure* hierarchy are primitive and 67.9% are leaf concepts.

Table 3.4 Top Five SNOMED Hierarchies by Size

Hierarchy	#Concepts	Avg #Parents/ Concept	Avg #Unique Relationships/ Concepts	Avg #Groups/ Concept	%Non-Primitives	%Leaf Concepts
Clinical Finding	97538	1.7	1.8	0.5	40.6	68.2
Procedure	52011	1.9	2.4	0.8	40.5	67.9
Organism	32225	1.0	0	0	0	78.4
Body Structure	31142	1.5	0.1	0	2.5	56.2
Substance	23752	1.2	0	0	0	79.0

After removing stop words and selecting FSNs of five remaining words or more, the algorithm utilized 26,980 unique concepts from the hierarchy (51.9%) for similarity sets. Overall, 4886 unique concepts were included in the 2111 similarity sets generated for the *Procedure* hierarchy, representing 9.4 percent of all concepts in the hierarchy, and 18.1 percent of all eligible concepts in the hierarchy. The five samples included 250 sets containing 797 unique concepts. Table 3.5 provides general set information for the *Procedure* hierarchy while Table 3.6 summarizes the characteristics of each sample.

None of the samples' similarity sets was excluded due to irrelevant association between the concepts.

Table 3.5 Overall Similarity Sets for the *Procedure* Hierarchy

Set type	#Sets	Avg #concepts	%Concepts covered	Largest set	Median
All similarity sets in hierarchy	2111	2.9	9.4	61	2
Different # of parents	573	3.4	3.2	50	2
Different # of relationships	352	3.5	2.0	61	2
Different # of groups	224	3.1	1.3	27	2
Different # of parents, relationships and groups	99	3.3	0.6	20	3

Table 3.6 Sample Characteristics for the Similarity Sets

Set type	#Sets	#Cpts	%Non-prim	%Leaf	Avg #par/cpt	Avg #rel/cpt	Avg #grp/cpt
Control sample	50	128	22.6	79.7	1.3	2.3	0.4
Diff-Par sample	50	149	29.5	71.8	1.9	2.8	0.5
Diff-Rel sample	50	222	40.0	64.8	1.6	2.7	0.4
Diff-Grp sample	50	148	39.2	71.6	1.6	3.1	1.5
Diff-All sample	50	150	38.0	67.3	1.8	3.1	1.2

Table 3.7 summarizes the findings across the five samples. The Control sample exhibited inconsistencies in 38% of the similarity sets. The non-Control samples exhibited inconsistency rates of 52 to 70 percent. For Diff-Rel, with 70% of inconsistent sets, this was a statistically significant difference compared to Control (Fisher's exact test, two-tailed). Thus, these findings strongly confirm the second hypothesis: Concepts in a similarity set with different number of relationships have a higher likelihood of

inconsistency. The use of a strict statistical test was chosen for this study. In fact, under the Chi-square test, the findings in all samples are statistically significant compared to the Control sample.

Table 3.7 Summary of Findings in the Five Similarity Set Samples

Sample	Sets	Inconst sets		Concepts	Inconst cpts		P-value (two-tailed) Fisher's exact test
	#	#	%	#	#	%	
Control	50	19	38	128	27	21.1	
Diff-Par	50	29	58	149	48	32.2	0.07
Diff-Rel	50	35	70	222	54	24.3	0.002
Diff-Grp	50	26	52	148	38	25.7	0.2
Diff-All	50	28	56	150	49	32.6	0.1

The auditing process strictly looked for the five inconsistency types within each similarity set, namely, hierarchical, attribute assignment, attribute value, role groups and definitional. Table 3.8 breaks down the inconsistency types found within concepts for all the five different samples. Set concepts from Diff-Par predominantly exhibited hierarchical inconsistencies (95.8% $p<0.001$), whereas set concepts from Diff-Rel predominantly exhibited inconsistencies involving attribute assignments (98.1%, $p<0.001$) thus confirming Hypotheses 1.1 and 2.1, respectively. The results also demonstrate a meaningful correlation in the Diff-Par, Diff-Rel, and Diff-All sample concepts between hierarchical and attribute assignment issues.

Table 3.8 Breakdown of Inconsistency Types within Concepts of Inconsistent Sets

Sample	Inconst cpts	Hierarchical		Attrb assgn		Attrb value		Groups		Definitional	
		#	%	#	%	#	%	#	%	#	%
Control	27	10	37.0	14	51.8	3	11.1	2	7.4	5	18.5
Diff-Par	48	46	95.8	17	35.4	14	29.2	4	8.3	5	10.4
Diff-Rel	54	24	44.4	53	98.1	7	13	9	16.7	0	0
Diff-Grp	38	24	63.2	6	15.8	10	26.3	21	55.3	4	10.5
Diff-All	49	20	40.8	24	49.0	8	16.3	46	93.9	0	0

The following example illustrates some of the issues summarized above. It involves a set containing five concepts, three of which are *Primary cemented total ankle replacement*, *Primary cemented total hip replacement*, and *Primary cemented total knee replacement* (Figure 3.3). A hierarchical discrepancy can be seen in the fact that although the three procedures differ only in the joint involved, each is anchored to a conceptually different sub-hierarchy parent(s). The first has the parent *Prosthetic cemented total ankle replacement*, the second has *Insertion of hip prosthesis, total*, and the third has the two parents *Arthroplasty of knee* and *Implantation of joint prosthesis into knee joint*. Of the four role-groups utilized in the modeling of the three concepts, none comprises the exact same set of attributes. While the rationale for the two different role-groups for the knee replacement procedure is most likely explained by inheritance from its two parents, none of the parents has the identical role- groups, and they are clearly inconsistent with the modeling of the hip and ankle procedures.

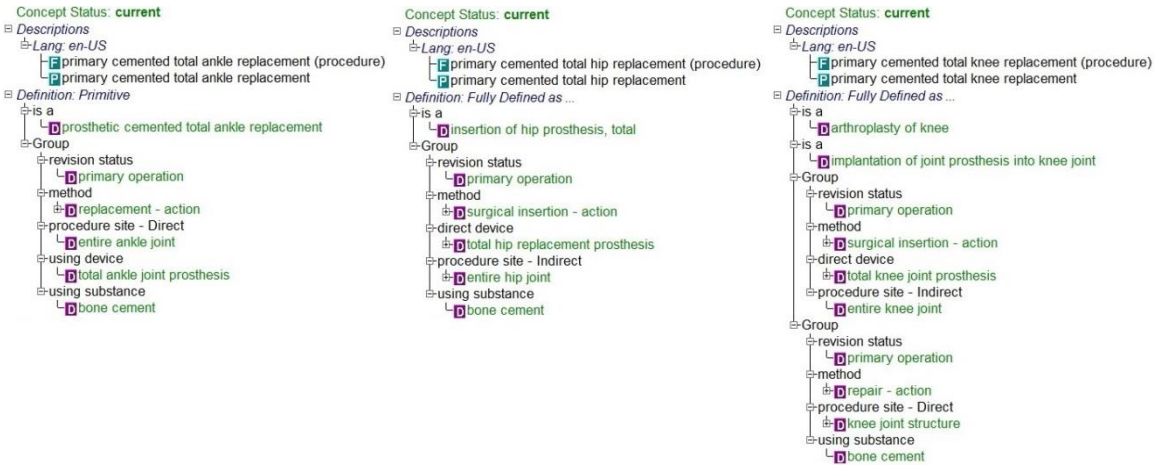

Figure 3.3 Three similar concepts depicting modeling inconsistencies.

Source: The Clinical Information Consultancy Ltd, "CliniClue Xplore", 2009, Available from: http://www.cliniclue.com.

Attribute/ value discrepancies can be seen for the attribute *method* which has the three different target values: *Replacement – action*, *Surgical insertion – action*, and *Repair – action*. The *using device* attribute is present only for the ankle procedure although it should apply to all procedures due to their nature. The definitional status of the ankle procedure is primitive (underspecified), while the knee and hip procedures are fully-defined. It is not clear that the presence of the *direct device* and *procedure site - indirect* attributes for the knee and hip procedures truly fully defines them. (Note that not all inconsistencies present in this example set have been discussed.)

3.4 Discussion

Terminologies, such as SCT, emphasize conceptual definitions over textual definitions. SCT itself does not contain any textual definitions for its concepts, and fully relies on its DL definitions. Thus, it might be surprising to note that in the *Procedure* hierarchy, one of the more semantically complex in SCT, more than 60% of concepts are primitives (i.e.,

under-defined). This phenomenon, obviously, can be used to justify many of the findings of the present study. If concepts are under-defined, there is no guarantee that the modeling of similar concepts should even be the same; attributes and role-groups may differ. For most basic uses of a terminology, such as concept searches, coding, and subset extraction, this phenomenon may seem of small significance. However, when one considers that hierarchical aspects also play a part in conceptual definitions and that attribute and attribute/value assignments may affect even the most basic terminological queries, under-defined concepts may inflict significant practical damage. The ability of DL classifiers to operate is directly related to the robustness of the underlying logical formulations. Inconsistencies, as described in this work, combined with the fairly inexpressive logic underlying SCT, are bound to escape detection [63]. Moreover, the ability of reasoners and classifiers to detect other errors, properly classify, or draw other inferences is severely limited under such circumstances.

The study starts with the premise that lexically similar concepts are expected to exhibit similar modeling. The control sample validates that many of the similarly worded concepts in SCT's *Procedure* hierarchy, are not modeled in a consistent manner (38% overall). Furthermore, the study indicates that concepts in similarity sets with difference in attributes are much more likely to be inconsistently modeled (70%). Moreover, the findings suggest that algorithmic detection and resolution of inconsistencies is feasible, as will be discussed later.

The results indicate that in the authoring process of SCT, very little attention is given to identify similar concepts, with little consideration to the importance of modeling them in a consistent manner. While the vast majority of the inconsistencies cannot be

considered errors, as each individual concept conforms to SCT's guidelines, they may pose significant obstacles to reasoning engines based on SCT's modeling structure. This is not a trivial manner as Rector *et al* [81] so amply demonstrated. Revisiting their *Myocardial infarction* example clearly illustrates how such deficiencies can interfere with meaningful utilization of data collected in clinical repositories: research queries may not return all relevant cases, decision support opportunities may be missed, analytics may be skewed, and clinical care can be affected. Campbell *et al* discuss similar issues in [74].

Although in the current context of HITECH and MU, SCT serves mostly as a source for subsets and lists, it is hard to imagine that it was chosen only due to its lexical comprehensiveness. Naturally, the next step beyond using SCT's concept descriptions in lists is taking advantage of SCT's hierarchical structure and formal definitions. The true potential of any controlled biomedical terminology is embodied in the knowledge captured within its semantic network [9-11, 15, 83]. SCT faces expectations to serve as an interface terminology and not only as a reference terminology [37-39, 60, 63]. In its current state, SCT cannot serve "as-is" in clinical applications even as a reference for limited sets [84, 85]. It is expected that for use within clinical applications vendors will use well-curated subsets and that dedicated extensions will be developed. However, not all CEHRT vendors can purchase or invest resources to develop such subsets and extensions may diverge from each other in a manner that will be counter-productive for data interoperability. The IHTSDO invests significant effort in formulating SCT with DL for computational purposes. However, incomplete and inconsistent application results in a structure that is questionable for use except for the generation of SCT's inferred view from the stated one.

Rector *et al* [81] suggest that a comprehensive auditing effort is urgently needed, estimated at up to two years for the CORE subset. However, the CORE subset is just a small portion of SCT's *Clinical finding* hierarchy. A broader auditing effort will require a much larger coordinated effort that may be beyond the reach of the IHTSDO. As SCT continues to grow, delays will complicate matters further. Therefore, it is essential to develop and implement a variety of auditing methodologies that can be incorporated into the authoring process or routinely executed after the fact with high yield. As Wei and Bodenreider [63] concluded, DL classifiers cannot detect that which is not defined. Other methods are needed to complement the classifiers. The analysis in this study is independent of SCT's DL-based infrastructure as it inspects, holistically, modeling elements of one concept and compares them to those of similar concepts. Inconsistencies of the types described in this study must be evaluated outside the realm of DL since, ultimately, SCT's usefulness from an algorithmic and individual perspective will be judged by the consistency and sufficiency of its conceptual definitions [37-39, 60, 63].

This study demonstrates that a simple lexical algorithm can very effectively detect similar concepts that are inconsistent in their logical modeling utilizing differences in attributes as an indicator. Moreover, this methodology can be applied to other semantically rich SCT hierarchies such as the *Clinical finding* (16 attributes) hierarchy with similar effectiveness. It is reasonable to expect that other such hierarchies harbor similar inconsistencies but that the effectiveness and yield of this method will decline with declining semantic complexity. Other algorithms, utilizing different and more sophisticated lexical methods and word length selection may improve on the results. However, additional methodologies could introduce noise and reduce specificity as

discussed by Campbell *et al* [74] and, with the current yield described throughout this study, an immediate need is not seen to employ such methodologies.

The present study opens the possibility for algorithmic enhancement of SCT's formal definitions utilizing an indicator that was used to identify sets, i.e. different attribute assignments in similarly worded concepts. Although more than half of SCT's concepts are not fully defined, it can be reliably assumed that the vast majority of them are not erroneous. Thus, it is posited that most of the additional attributes and attribute target values (when the attribute target value is not directly associated with the specific word that differentiates between the similar concepts) can be reasonably assigned to the other similarity set member concepts that lack them.

This work also emphasizes the importance of external auditing of such large bodies of knowledge. As SCT's use expands, the significance and implications of such "imperfections" will only increase. Eventually, these are bound to manifest themselves in patient care. While the emphasis during the last decade was mostly on expanding content coverage, more effort should now be applied to quality assurance. This is too big an effort to be solely tasked to the IHTSDO or third-party entities. Only a collaborative, open process can ensure, under the umbrella of the IHTSDO, effective results.

Consider the example in Figure 3.4. For the purpose of this discussion, the differences in hierarchical modeling are ignored. The concept on the right lacks the *has specimen* attribute. Adding this attribute with its target value to create the hypothetical concept as depicted in Figure 3.5 will be correct, improve the consistency of the modeling, and potentially contribute toward qualifying the concept as a fully specified concept. Other algorithmic approaches to identify possible missing attributes can be

employed. For example, a method can detect that certain FSN words are not represented as an attribute target in the formal definition. In this case, "serum" is not present as an attribute target value. However, such a method may be less effective in proposing a possible resolution.

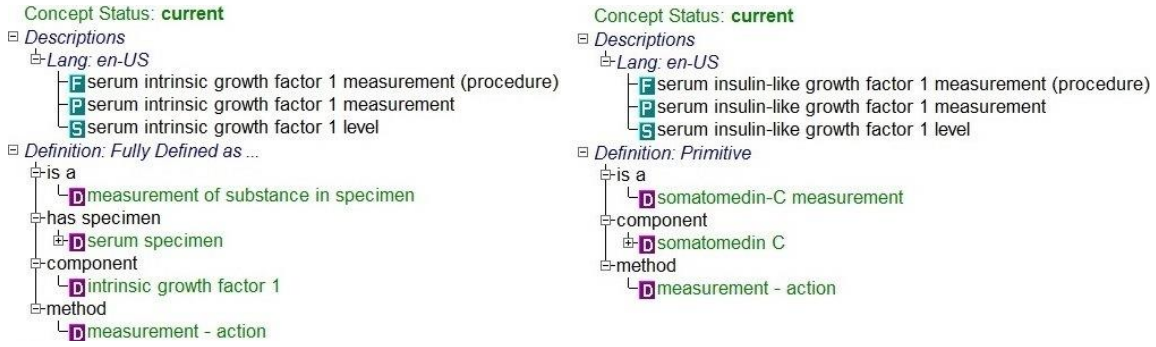

Figure 3.4 A two-concept similarity set from *procedure* hierarchy of SNOMED with differences in the assignment of attributes.

Source: The Clinical Information Consultancy Ltd, "CliniClue Xplore", 2009, Available from: http://www.cliniclue.com.

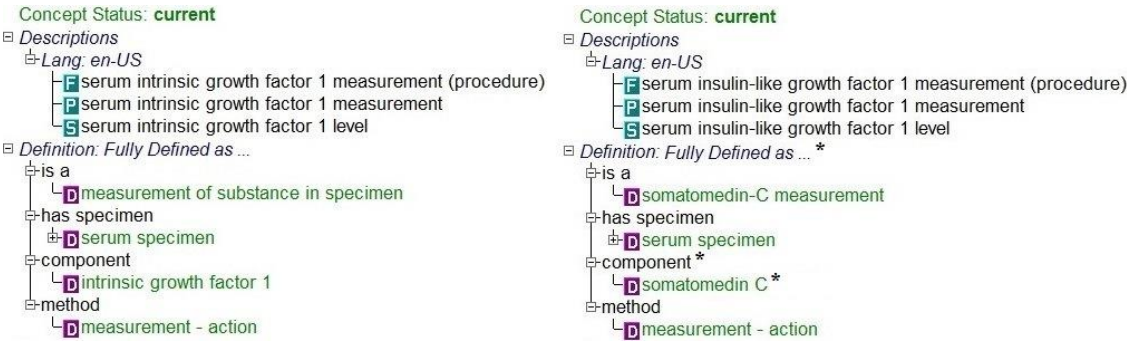

Figure 3.5 A hypothetical enhancement using added attributes marked with an asterisk.

Source: The Clinical Information Consultancy Ltd, "CliniClue Xplore", 2009, Available from: http://www.cliniclue.com.

This study was limited due to the use of a non-blinded, single auditor. However, it is considered that the nature of the evaluation for inconsistencies is only minimally subjective, if at all, due to the definition of an inconsistency. For example, the

consideration of a missing attribute, a yes/no type of decision, is algorithmically detectable. Furthermore, it is not likely that this study identified missing attributes as false positives. It is more likely that the review process included a certain degree of false negatives as missed findings. Any bias towards a specific inconsistency type in its respective sample would have affected each sample in a similar manner while the auditor was instructed to exhaustively document all types of inconsistencies in each and every similarity set.

In light of scarce auditing resources, it is believed that this methodology is suitable for use by a single reviewer and can be easily utilized during the authoring process. It is proposed that this and other complementary lexical and non-classifier methodologies be adopted by the IHTSDO as part of the editing process in conjunction with current methodologies as well as for the routine maintenance of the inferred view of SCT.

3.5 Summary

The *Procedure* hierarchy of SCT exhibits various significant modeling inconsistencies. Based on additional preliminary studies, there is reason to believe that other attribute-rich hierarchies may exhibit similar issues. Such inconsistencies cannot be detected strictly by DL classifiers and may propagate to affect clinical care. Lexical methods can help detect such inconsistencies during the editing process, thus preventing their inclusion in new releases. As SCT becomes more prevalent in the clinical care domain, it is time to step up the auditing effort.

CHAPTER 4

POSITIONAL SIMILARITY SETS

4.1 Introduction

The study in Chapter 3 provided a methodology to improve the correctness and consistency of SCT concepts by introducing similarity sets. An analysis of a sample of such sets identified inconsistencies in up to 70% of the sets and 32.6% of the concepts. This study builds on the notion of having a consistent modeling between lexically similar concepts as discussed in Chapter 3 and introduces positional similarity sets which are groups of lexically similar concepts that differ from each other by one word of their fully specified names and the differing words occupy the same position in their names. Applying strictness in the position of differing words results in an increased lexical similarity between the concepts in a set thus increasing the contrast between the lexical similarities and modeling differences. This increase in contrast increases the likelihood of finding inconsistencies.

The efficiency of the positional similarity sets is further improved by introducing the use of three structural indicators in the form of the number of parents, relationships and role groups in the formation of such sets. The results show that the use of positional similarity sets further improves the likelihood of finding inconsistencies with up to 41.6% of the concepts found to have one or more of the following kinds of inconsistencies – hierarchical, role group, attribute assignment and attribute value. SCT concepts suffer from inconsistent modeling and the positional similarity sets can be an effective way of finding such inconsistencies thus improving the correctness and completeness of SCT concepts.

4.2 Method

A methodology was defined in Chapter 3 to create similarity sets based on the lexical similarity between concepts. Similarity sets are groups of concepts that differ from the seed concept by only one word of their fully specified name. The seed concept is the first concept that is selected to lexically match with the other concepts to form a set. The methodology for generating similarity sets takes each concept of an SCT hierarchy as the seed concept and lexically matches it with every other concept in the hierarchy to form such sets. Algorithmic precautions are taken to prevent redundancy in the formation of such sets, i.e., no two sets consist of exactly the same concepts.

The situation of similarity sets may get more complex when more than two concepts participate in a set. In Chapter 3, a similarity set included all the concepts for which there was exactly one word difference from the seed concept that was used to generate the sets. So, one concept C2 may differ from the seed concept C1 in the first word of the FSN, while another concept C3 may differ from the seed concept C1 in the fifth word. However, C2 and C3 which are both in the same similarity set, created with the seed concept C1, may not be as mutually similar as they differ from one another in two words of the FSN, the first word and the fifth word. Such a situation is illustrated in Table 4.1.

Table 4.1 A Similarity Set containing Three Concepts regarding Ligament Procedure

179875006	Primary arthroscopic xenograft ligament replacement (procedure)
179885007	Revision arthroscopic xenograft ligament replacement (procedure)
179879000	Primary arthroscopic xenograft ligament augmentation (procedure)

As can be seen in Table 4.1, the second concept differs from the seed concept in the first position of their FSN in the replacement type involved, primary vs. revision. The

third concept differs from the seed concept in the fifth position of their FSN in the type of procedure involved, replacement vs. augmentation. However, the second and the third concepts are less similar as they differ in two aspects, revision vs. primary procedure and replacement vs. augmentation procedure.

As mentioned earlier, the expectation is that lexically similar concepts will also be similar in their modeling. However, there are cases of differences in modeling of the concepts in a similarity set. When the modeling of concepts in a similarity set display differences, there is a contrast between the similarity of the concepts as expressed in the lexical way and the differences expressed in their modeling. This contrast points out to a high likelihood of inconsistencies. This tendency was illustrated by a high percentage of inconsistencies in a sample of similarity sets in Chapter 3

With the purpose to enhance the similarity among all the concepts in a set, the notion of positional similarity sets is introduced in this study. In a positional similarity set, all the concepts in a set mutually differ by one word and at the same position in their FSN. To generate such sets, the algorithm picks a seed concept *s* and a position *p* of a word in it, and the set includes all the concepts with the same number of words which differ from the seed concept in one word at the position *p*. Using this strictness in position for the concepts in the set of Table 4.1, two positional similarity sets are created as shown in Table 4.2. The seed concept is the same in both these sets. However, in the first set, the differing word is in the first position whereas in the second set, the differing word is in the fifth position of the concept FSNs.

Table 4.2 Two Positional Similarity Sets Generated using Different Positions for the Same Seed Concept

179875006	Primary arthroscopic xenograft ligament replacement (procedure)
179885007	Revision arthroscopic xenograft ligament replacement (procedure)
179875006	Primary arthroscopic xenograft ligament replacement (procedure)
179879000	Primary arthroscopic xenograft ligament augmentation (procedure)

Stop words such as “a”, “an” and “the” are ignored in the creation of the positional similarity sets. The methodology considers concepts where FSNs are of length five words or more including the semantic tag to form the sets. The number “five” is chosen because preliminary studies showed that using concepts, of length less than five words, often generated sets where similarity between the concepts was meaningless as a result of their short word length. Besides, use of concepts with five words or more also helps to keep the average set size to around 2.5. This means that most sets have two or three concepts. Such set size makes it easy for the auditor to compare the concepts side by side. The FSNs are chosen instead of synonyms or preferred terms as the FSNs best describe the meaning of a concept in SCT.

Strictness in the position of the differing words among all concepts in a set enhances the lexical similarity between the concepts in a set as compared to a general similarity set which may have differences in more than one word between some pairs of concepts in a set. This greater similarity between all the concepts of a positional similarity set makes the contrast between the lexical similarity and the modeling differences sharper in cases of modeling differences. This contrast points out to a high likelihood of inconsistencies. Besides, the strictness in position also results in sets having smaller number of concepts on average. For instance, instead of a similarity set of three concepts as shown in Table 4.1, there will be two positional similarity sets of two

concepts each as shown in Table 4.2. Small sized sets are important from auditing efficiency perspective as an auditor is not overwhelmed by the sheer number of concepts that needs to be audited at the same time. Instead, s/he can focus each time on a small set of concepts for which it is easier and faster to detect an inconsistency.

The study aims at finding four different kinds of inconsistencies. The first kind is hierarchical, i.e., inconsistencies in the number and types of parents. The second type of inconsistency is the one in the assignment of attribute, i.e., inconsistencies in the number and types of attributes. The third inconsistency type is the one in attribute values, i.e., inconsistencies in the targets of the attributes. The fourth inconsistency type is the one associated with the role groups, i.e., the number of role groups associated with a concept and the number of attributes within a role group.

Consider the positional similarity set with two concepts differing just in their first words of the FSN as shown in the Table 4.3. Both the concepts define the morphology of the blood cells and only differ in the type of blood cell involved, red vs. white blood cell. Since the concepts are lexically similar, they are expected to be modeled in a similar way. If the two similar concepts are not modeled in a similar way, it is assumed, in general, that if the modeling of one concept is more comprehensive than the modeling of the other concept with regards to the attribute types, the more comprehensive modeling is likely the correct one and therefore the additional modeling features can be applied to the other concept. Figure 4.1 displays the modeling of these two concepts.

Table 4.3 A Positional Similarity Set with Two Concepts regarding Cell Morphology

82461003 Red blood cell morphology (procedure)
44190001 White blood cell morphology (procedure)

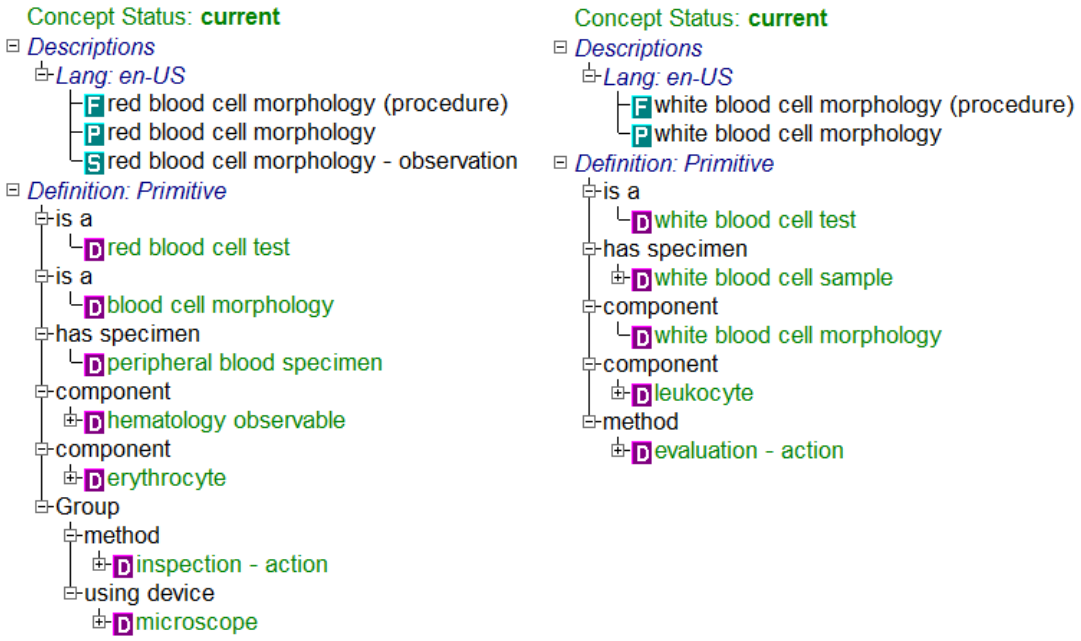

Figure 4.1 CliniClue snapshot of the modeling of two lexically similar concepts representing blood cell morphology.

Source: The Clinical Information Consultancy Ltd, "CliniClue Xplore", 2009, Available from: http://www.cliniclue.com.

A comparison between the modeling of these two concepts show inconsistencies of all four types as mentioned above. First, consider the hierarchical structure of the two concepts as shown in Figure 4.1. It can be seen that the *red blood cell morphology* has two parents as compared to one for the *white blood cell morphology*. The additional parent *blood cell morphology* for the red concept should also be a parent of the white concept.

Now, consider the attribute assignments for the two concepts. It is reasonable to assume that these two procedures are conducted using similar methodologies and equipment. While the red cells concept has an attribute *using device*, it is missing in the modeling of the white cells concept and could reasonably be added to the modeling of the latter.

Next, looking at the attribute values of the two concepts, several inconsistencies can be seen. The *has specimen* attribute has different target values for the two concepts with different level of specificity. While the value *white blood cell sample* from the white cells concept is more specific, it could be argued that the *peripheral blood specimen* target value may actually be more appropriate. Additionally, one of the *component* attributes of the red cells concept has the target value *hematology observable* whereas the white cells concept's *component* attribute has the target value *white blood cell morphology*. Both these target values are *observable entities* but the latter is a grandchild of the former and hence more specific. Accordingly, the red cells concept should also have the *component* attribute with a more specific value which would be *red blood cell morphology* from the *observable entities* hierarchy. This example also shows that the two concepts enrich one another with regards to more accurate modeling of the attribute value and not only that the concept with more detailed modeling enriches the other. The *method* attributes of the two concepts also have different target values, *inspection-action* vs. *evaluation-action* with the former being a grandchild of the latter and hence more specific. Hence, the white cells concept should also have the same more refined target.

In terms of the modeling with respect to role groups, it can be seen that the red cells concept has one role group whereas the white cells concept lacks any role group which again shows the inconsistent modeling of the two concepts. The presence of the additional attribute and group in the red cells concept is due to inheritance from the parent *blood cell morphology* that is missing from the white cells concept. Thus, in this case, a hierarchical inconsistency contributes to the missing attribute and group.

Apart from the modeling inconsistencies, this contextual comparison of the two concepts also highlighted the missing synonym for the white cells concept which was added accordingly. Figure 4.2 shows the revised corresponding modeling of these two concepts after rectifying the above described inconsistencies. The changes in the modeling have been marked with an asterisk. The auditing process also demonstrated the ease with which such inconsistencies could be detected using the positional similarity sets.

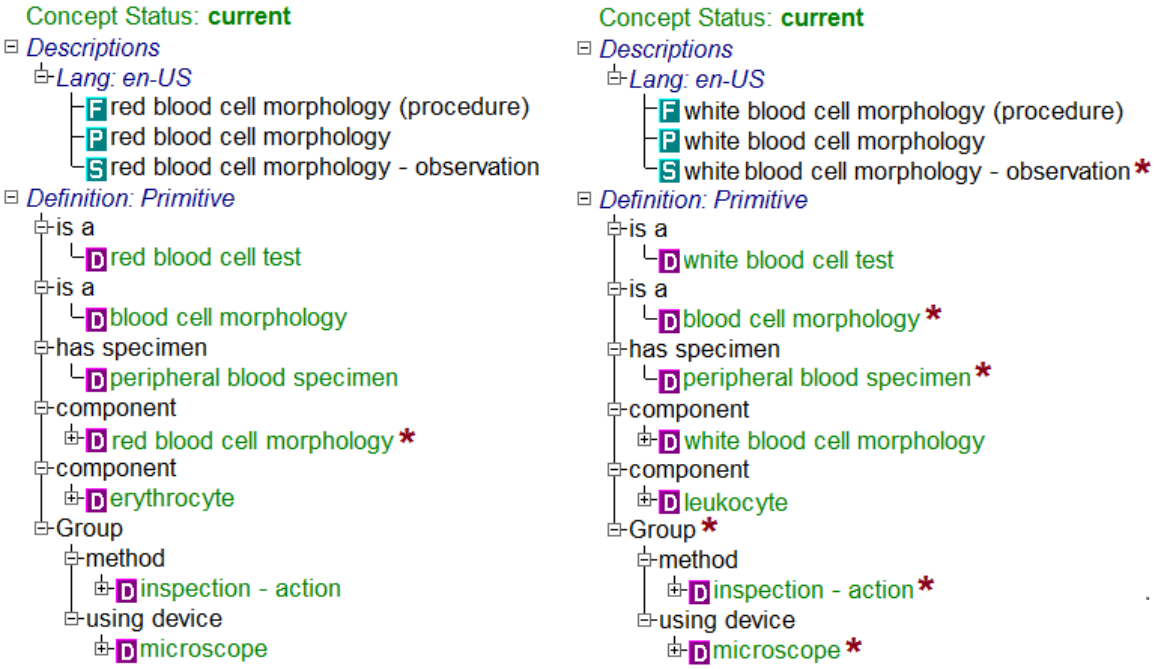

Figure 4.2 Proposed corresponding modeling of the two concepts from Figure 4.1.

Source: The Clinical Information Consultancy Ltd, "CliniClue Xplore", 2009, Available from: http://www.cliniclue.com.

The study further focuses on utilizing a technique based on identifying specific positional similarity sets for which a higher percentage of inconsistently modeled concepts is expected to be found. Once such sets are identified, the auditor reviews them

with an expectation for a more effective auditing than for the control positional similarity sets, raising the yield of the auditing measure by the percentage of the reviewed concepts found modeled inconsistently. The study stresses on the variation between observed structural differences, say in the number of parents, versus a decision obtained from an auditor that this difference is an unjustified modeling error or inconsistency which should be corrected. It is noted that some lexically similar concepts in a positional similarity set may have justified structural modeling differences due to the semantic difference implied by the differing word.

To implement this methodology, a positional similarity set is considered for an auditor's review only if some concept within this set has a different number of one of the structural parameters in the form of parents, relationships or groups. The reason being that with differences in these structural parameters, the contrast between the lexical similarity and differences in structural modeling is increased versus the control positional similarity sets, thus, tending to increase the likelihood of errors among such concepts. To test such observations, the following three hypotheses are formulated. For each of these hypotheses, a sub hypothesis is also formulated to characterize the nature of the inconsistencies.

Hypothesis 1: If some concepts within a positional similarity set have different number of parents than the other concepts; the concepts of such a set are more likely to be inconsistent compared to a control positional similarity set.

Hypothesis 1.1: The inconsistencies found in positional similarity sets with some concepts having different number of parents are most likely to be hierarchical.

Hypothesis 2: If some concepts within a positional similarity set have different number of attributes than the other concepts; the concepts of such a set are more likely to be inconsistent compared to a control positional similarity set.

Hypotheses 2.1: The inconsistencies found in positional similarity sets with some concepts having different number of attributes are most likely to be attribute-related.

Hypothesis 3: If some concepts within a positional similarity set have different number of role groups than the other concepts; the concepts of such a set are more likely to be inconsistent compared to a control positional similarity set.

Hypotheses 3.1: The inconsistencies found in positional similarity sets with some concepts having different number of role groups are most likely to be role group-related.

To test these three hypotheses, appropriate sample sets are created. To create positional similarity sets as per the requirements of Hypothesis 1, only those sets that have at least one concept with different number of parents from some other concept in the set are considered. Such sets were called the Diff-Par sets. Similarly to test Hypothesis 2, only those sets that have at least one concept with different number of relationships from some other concept in the set are considered. Such sets are called Diff-Rel sets. To test Hypothesis 3, only those sets that have at least one concept with different number of role groups from some other concept in the set are considered. Such sets are called the Diff-Grp sets. The sets formed without considering the number of parents, relationships and groups are called Control sets.

For the purpose of testing these hypotheses, there are four different set types created for the *procedure* hierarchy of SCT based on these hypotheses. A group of randomly selected 50 sets are taken from each of these four positional similarity set types

and provided to an auditor to check if the indicators actually helped in finding more inconsistencies. The auditing is blindfolded as the auditor is not informed of the methodology being used to generate the sets in order to avoid any kind of bias in the results. The sets are audited by Dr. Gai Elhanan. The *procedure* hierarchy of January 2011 release of SCT is utilized for the purpose of this study.

4.3 Results

Table 4.4 summarizes the results of the auditing of the four sample set types. Each of the four sample sets, namely, Control, Diff-Par, Diff-Rel and Diff-Grp have randomly selected 50 sets from the *procedure* hierarchy of SCT. The control sample consists of 102 unique concepts of which 18.6% are found to be inconsistent. In comparison to the control sample, 39.6% of the concepts for Diff-Par sample, 41.6% of the concepts for Diff-Rel sample and 33.0% of the concepts for Diff-Grp sample are found to be inconsistent supporting Hypotheses 1, 2 and 3. The increase in the number of inconsistent concepts for all the three Diff-Par, Diff-Rel and Diff-Grp samples as compared to the Control sample is found statistically significant according to the two-tailed Fisher's test. The first two hypotheses are even found highly statistically significant (see Table 4.4).

Table 4.4 Summary of the Auditing of Four Positional Similarity Set Types

Sample Type	Unique concepts	Inconsistent concepts		p-value two-tailed Fisher's test	Comments
	#	#	%		
Control	102	19	18.6		
Diff-Par	111	44	39.6	0.0009	Highly statistically significant
Diff-Rel	125	52	41.6	0.0003	Highly statistically significant
Diff-Grp	115	38	33.0	0.0202	Statistically significant

Table 4.5 displays the breakdown of the inconsistency types for each of the four samples. For the Diff-Par sample, 84.1% (37 out of 44) of the inconsistent concepts were found to have hierarchical problems. This value was regarded as statistically significant by Fisher's exact test as compared to the combined data for the other three samples for the hierarchical inconsistencies (62 out of 109) thus confirming Hypothesis 1.1. For the Diff-Rel sample, 86.5% (45 out of 52) of the inconsistent concepts had problems with the assignment of attributes. Fisher's exact test gave a statistically significant result when compared with the combined data of the other three samples for the same inconsistency type (41 out of 101) thus confirming Hypothesis 2.1. For the Diff-Grp sample, 60.5% (23 out of 38) of the inconsistent concepts exhibited problems with their role groups. Again, a statistically significant result was obtained when compared with the combined data of the other three sample types for inconsistencies in role groups (20 out of 115) thus confirming Hypothesis 3.1.

Table 4.5 Breakdown of Inconsistency Types among the Four Positional Set Types

Sample Type	Unique cpts	Inconst cpts		Hierarchical		Attrb assgn		Attrb value		Groups	
	#	#	%	#	%	#	%	#	%	#	%
Control	102	19	18.6	12	63.2	11	57.9	2	10.5	4	21.1
Diff-Par	111	44	39.6	37	84.1	20	45.5	9	20.5	4	9.1
Diff-Rel	125	52	41.6	25	48.1	45	86.5	7	13.5	12	23.1
Diff-Grp	115	38	33.0	25	65.8	10	26.3	14	36.8	23	60.5

4.4 Discussion

SCT was formed as a result of the merger between SNOMED Reference Terminology and United Kingdom's Clinical Terms Version 3 (CTV3). This merge is likely one of the factors resulting in an incomplete and inconsistent modeling of SCT. Ideally, such

inconsistencies could and should have been detected in the merging process, where the modeling could have been made consistent for similar concepts. SCT is also constantly evolving with a new version released every six months by IHTSDO which makes it further difficult to keep SCT concepts free from inconsistencies. The methodology described offers an opportunity to improve the consistency of the modeling of SCT concepts using the ease and efficiency of comparing lexically similar concepts to identify the inconsistencies in modeling.

This study provided a contextual auditing of SCT concepts based on their lexical similarity with other concepts of SCT. A positional similarity set provides a context so that the modeling of similarly worded concepts can be compared. Similarly worded concepts are typically expected to be modeled in a similar way. The assumption of the study was that the positional similarity sets can help identify inconsistencies in SCT concepts with ease as they offer the display of the contrast between the lexical similarities and modeling differences. The results of the study supported this assumption as the auditor was able to detect inconsistencies in SCT concepts using such sets which would otherwise likely go unnoticed. The auditor found 18.6% of the concepts as being inconsistent using the positional similarity sets as shown in Table 4.4.

The study further presented techniques to increase the likelihood of finding inconsistent concepts using three structural indicators in the form of the number of parents, relationships and groups in the formation of positional similarity sets. Grouping together lexically similar concepts, some of which differ in the number of their parents, relationships or groups, increases the contrast between the lexical similarity and the structural differences which makes the inconsistencies much more obvious. The Diff-Par

sample and the Diff-Rel sample, consisting of positional similarity sets in which some concepts have different number of parents or relationships, contained 39.6% and 41.6% inconsistent concepts respectively. The Diff-Grp sample, consisting of positional similarity sets in which some concepts have different number of role groups, contained 33.0% inconsistent concepts which is a lower yield as compared to the Diff-Par and the Diff-Rel sample. The increase in the number of inconsistent concepts was found to be statistically significant as compared to the Control sample for all three indicators.

The results of this study have also shown support for the three refined hypotheses regarding the inconsistency types that can be found using the positional similarity set types with different structural indicators. As a result of this finding, a methodology to automatically detect inconsistencies of a particular type has emerged. If, for example, one is interested in inconsistencies in the relationships of the concepts, positional similarity sets with some concepts differing in the number of relationships should be selected.

Future work will involve enhancing the algorithm that generates positional similarity sets. It would be interesting to study the effect of the definition of the concepts (primitive vs. fully defined), the downward hierarchical level of the concepts (leaf concepts vs. non-leaf concepts) and the sibling relationships between the concepts in a set. Another goal would be to incorporate methods to correct inconsistencies algorithmically. For example, studies will be performed to effectively identify a missing relationship such as the one shown in Figures 4.1 and 4.2 algorithmically without the need of a manual review. Further studies will also involve incorporating other concept descriptors besides the FSN such as the preferred terms and synonyms in the generation

of similarity sets especially when the concept FSNs are not similar but the lexically similarity can be captured with these alternate descriptors.

Future work will also involve generating a web interface which can display the modeling of similar concepts side-by-side so that the auditor can easily compare the concepts to one another. Such a system will also be able to suggest changes to the modeling of these concepts to make the concept more consistent with other similar concepts. The reviewer will just need to check if such suggestions are valid.

4.5 Summary

The study presented the notion of a positional similarity set which is a group of concepts that are similar with respect to the word structure of their FSNs except for one word in a specific position to support quality assurance technique to effectively identify inconsistencies in SCT. The contextual auditing of lexically similar concepts was shown to be effective in identifying inconsistencies which would otherwise go unnoticed. The use of the three structural indicators in the form of the number of parents, relationships and groups along with the positional similarity sets was shown to be effective in increasing the likelihood of identifying inconsistencies. Quality assurance techniques such as this can be used to complement the efforts of IHTSDO to improve the quality of SCT thus making it a more viable product to be used in EHRs and other medical applications.

CHAPTER 5

ALGORITHMIC SUGGESTION OF ATTRIBUTES TO ENHANCE THE MODELING OF SNOMED CONCEPTS

5.1 Introduction

SNOMED CT (SCT) has gained widespread acceptance as a clinical terminology, supported by the HITECH Act of 2009. Increased usage brought with it increased scrutiny of SCT's content. Although regarded as the most comprehensive clinical terminology available, SCT has also been found to suffer from spotty coverage, errors and inconsistencies in the modeling of the concepts, possibly due to lack of sufficient quality assurance. In Chapters 3 and 4, a lexical technique was presented to check for inconsistencies in the modeling of SCT concepts by grouping similar concepts together. The studies were based on the premise that lexically similar concepts should be modeled in a similar way. This study builds on those previous studies and presents an algorithmic technique that can automatically suggest attributes and their target values for SCT concepts. A sample of 50 concepts, each with one or more algorithmically suggested attributes and target values, is audited for correctness by an experienced auditor. The results are analyzed and presented in this study.

5.2 Method

A methodology is devised based on the notion that lexically similar concepts should also be modeled in a similar way. The methodology starts with first creating groups of lexically similar concepts known as positional similarity sets as described in Chapter 4. Two concepts C1 and C2 appear together in a positional similarity set if:

- Both concepts C1 and C2 have the same number of words in their fully specified names (FSNs)
- The number of words in their FSN is five or more without considering the stop words
- There is only one differing word in both the concepts
- The differing word is at the same position in their FSN

Table 5.1 displays an example of a positional similarity set. The two concepts in the set are lexically similar and only differ in the substance involved, urobilinogen vs. blood. The differing word is also in the same position, i.e., the first position of their FSN. All such possible positional similarity sets are generated for the *procedure* hierarchy of SCT. For each such set, a list of unique attributes for every concept in the set is recorded. Each concept in the set is then checked to see if it has all those attributes. If it doesn't, the attribute is suggested to be added to the concept. This study does not take into consideration the notion of groups when trying to find new attributes to enhance the modeling of the concepts. Instead all the unique attributes from all groups are collected into a single list to suggest attributes to be added to similar concepts.

Table 5.1 A Positional Similarity Set with Two Concepts regarding Test Strip Measurement

250416001	Urobilinogen concentration, test strip measurement (procedure)
250414003	Blood concentration, test strip measurement (procedure)

The algorithm automatically analyzes the attributes of the two concepts of Table 5.1 and found that while the urobilinogen concept had three attributes, the blood concept had none. A CliniClue snapshot of the modeling of these two concepts is shown in Figure 5.1.

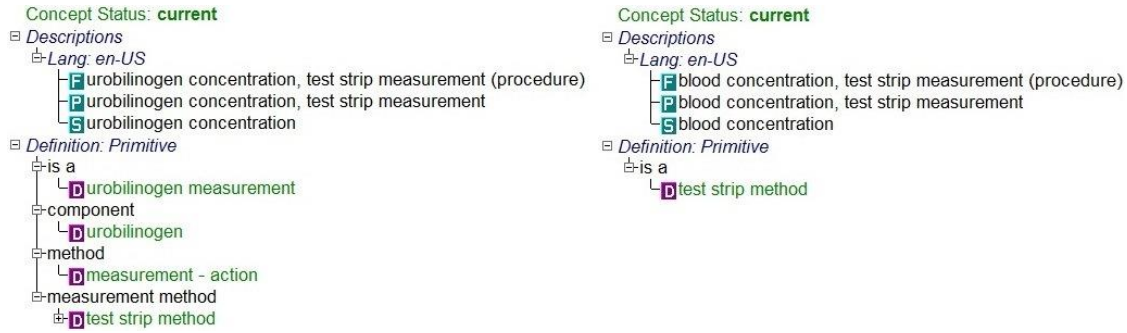

Figure 5.1 A CliniClue snapshot of the modeling of two similar concepts regarding test strip measurement.

Source: The Clinical Information Consultancy Ltd, "CliniClue Xplore", 2009, Available from: http://www.cliniclue.com.

The algorithm then automatically suggests the three attributes as possible addition to the blood concept, namely, *component*, *method* and *measurement method*. Besides suggesting attributes, the algorithm can also automatically suggest the corresponding target value for the attributes. The target value comes from the concept from which the attribute was picked and recorded. A refined modeling of the pre-operative concept along with that of the post-operative concept is shown in Figure 5.2.

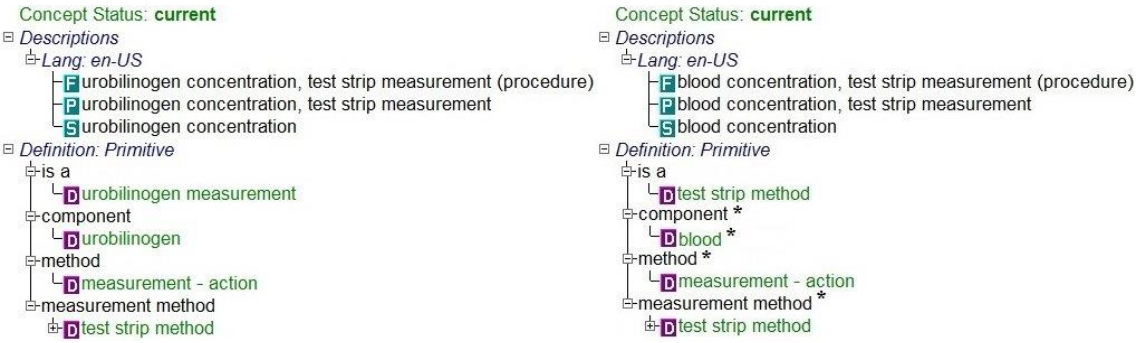

Figure 5.2 Modeling of two procedures regarding test strip measurement after the addition of suggested attributes (marked with an asterisk).

Source: The Clinical Information Consultancy Ltd, "CliniClue Xplore", 2009, Available from: http://www.cliniclue.com.

In some cases, the algorithm changes the wording of the target value to better suit the current concept for which the attribute is being suggested. For instance, while suggesting *component* as an attribute for the blood concept, the suggested target value was changed to *blood* from *urobilinogen* to better suit the modeling of the current concept.

If the modeling of a concept has a more refined version of an attribute that comes from a similar concept, the attribute is not suggested for addition. For instance, if the concept has an attribute *procedure site – direct*, then *procedure site* which may be present in a similar concept is not suggested for addition. Consider the modeling of two similar concepts *trapping of intracranial aneurysm (procedure)* and *ligation of intracranial aneurysm (procedure)* as shown in Figure 5.3.

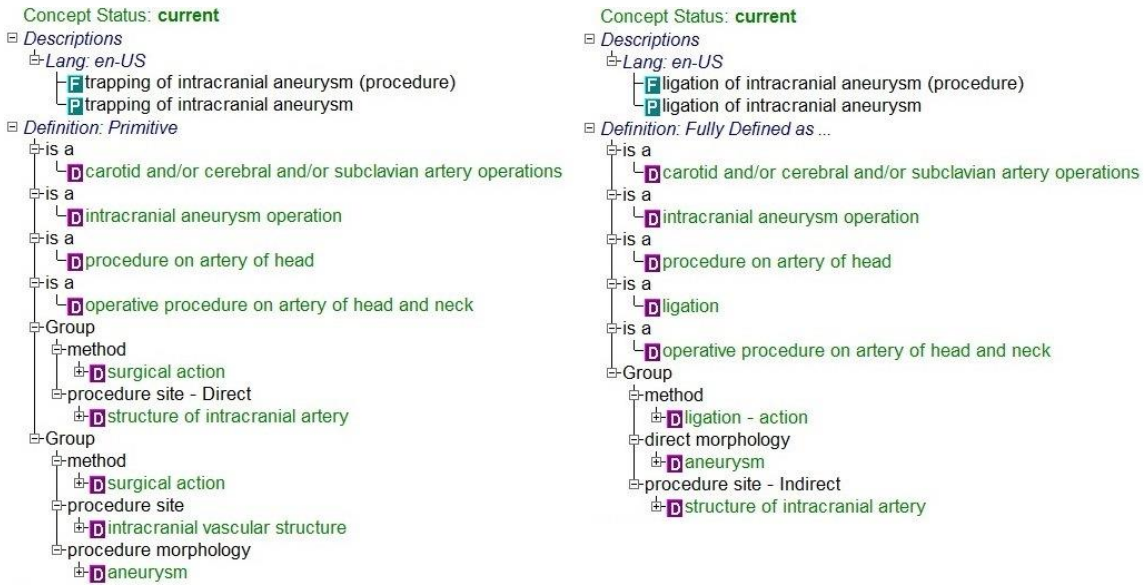

Figure 5.3 Modeling of two concepts related to intracranial aneurysm.

Source: The Clinical Information Consultancy Ltd, "CliniClue Xplore", 2009, Available from: http://www.cliniclue.com.

The algorithm suggested *direct morphology* to be added to the trapping concept as it is more refined version of *procedure morphology*. The algorithm also suggested *procedure site – indirect* to be added to the trapping concept. On the other hand, for the ligation concept, *procedure morphology* was not suggested for addition as it already has a more refined attribute in the form of *direct morphology*. Similarly, *procedure site* was also not suggested since a more refined attribute (*procedure site - indirect*) is already present in its modeling. The only attribute suggested for the ligation concept was *procedure site – direct*. The new modeling of these two concepts with the new suggested attributes (marked with an asterisk) is shown in Figure 5.4. This example will be further discussed in the Discussion section.

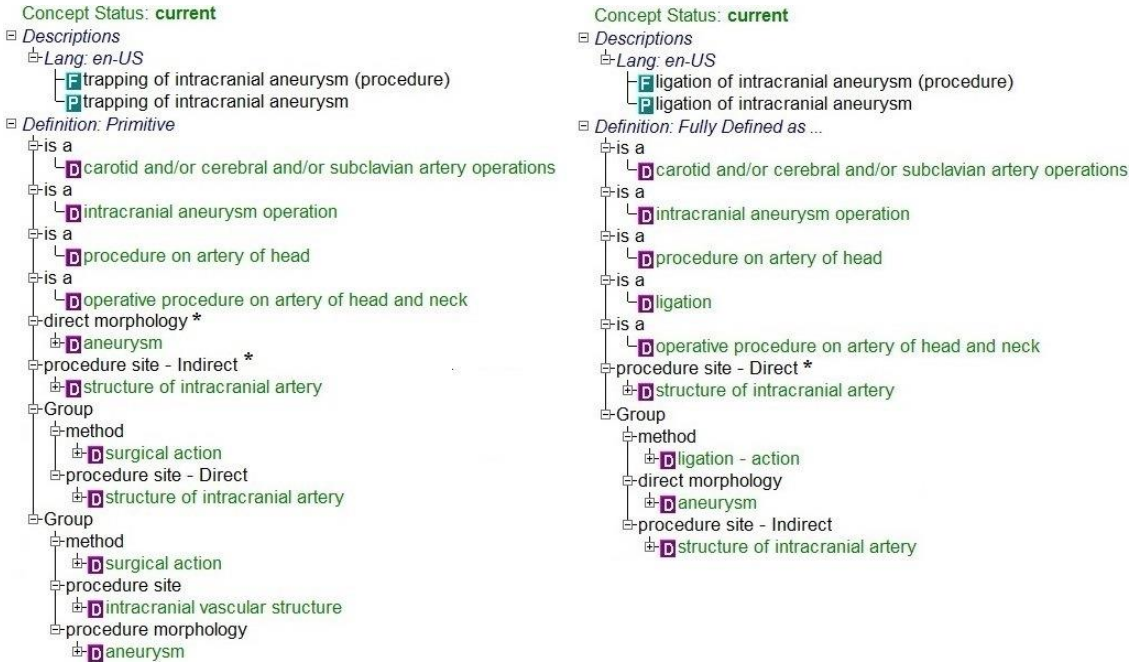

Figure 5.4 Modeling of two concepts related to intracranial aneurysm after the addition of suggested attributes (marked with an asterisk).

Source: The Clinical Information Consultancy Ltd, "CliniClue Xplore", 2009, Available from: http://www.cliniclue.com.

All concepts along with their suggested attributes and target values are recorded. Also recorded is the set from which this concept originated. A randomly selected sample of 50 such concepts along with their originating sets and suggested attributes are then analyzed by an experienced auditor (Dr. Gai Elhanan) to check for correctness and accuracy of the suggested attributes and their targets. The January 2013 release of SCT is used for the purpose of this study. The CliniClue browser is used by the auditor for reference purpose to check the concept modeling.

5.3 Results

Table 5.2 displays some general data for the *procedure* hierarchy from the January 2013 release of SCT. The hierarchy consists of 53147 current concepts with an average number of 2.4 unique attributes per concept.

Table 5.2 Summary of Data for *Procedure* Hierarchy from January 2013 Release of SNOMED

Current concepts in *procedure* hierarchy	53147
%Primitives	59
%Leaf concepts	68
Average #parents	1.8
Average #unique attributes	2.4
Average #groups	0.8

Table 5.3 displays data related to the concepts in the *procedure* hierarchy for which possible additional attributes were identified by the methodology. A total of 10451 positional similarity sets were generated for the hierarchy. Of these, 2624 sets were identified as having one or more concepts with suggested attributes. A total of 5518 concepts were suggested one or more attributes of which 5056 were unique concepts. A total of 28 unique attributes were suggested for these concepts.

Table 5.3 Summary of Concepts with Suggested Attributes for *Procedure* Hierarchy

#Sets	10451
#Sets with concepts for which attributes are suggested	2624
#Unique concepts for which attributes are suggested	5056
#Total concepts for which attributes are suggested	5518
#Total attributes suggested	8686
#Average attributes per concept suggested	1.57
#Unique attributes suggested	28

Table 5.4 displays some statistics related to the target values of the suggested attributes. For the suggested 8686 attributes for the *procedure* hierarchy, the wording of a total of 4458 target values were changed of which 71.4% were found to be present in SCT as concepts whereas 28.6% were suggested as new concepts. In terms of unique target values, 48.7% of them were found to be present in SCT as concept whereas 51.3% of them were suggested as new concepts.

Table 5.4 Summary of Target Value Data for *Procedure* Hierarchy

	#	%
#Total target values suggested	8686	
#Target values changed before suggestion	4458	51.3
#Target values in SCT as a concept	3182	71.4
#Target values not in SCT as a concept	1276	28.6
#Unique target values changed before suggestion	1364	
#Unique target values in SCT as a concept	665	48.7
#Unique target values not in SCT as a concept	699	51.3

Table 5.5 displays the characteristics of the 50 concepts that were randomly selected for evaluation. The data is, in general, comparable to that of the entire *procedure* hierarchy as shown in Table 5.2. Of the 50 concepts, 54% are primitive and 44% are leaf nodes. The average number of parents is 1.88, the average number of attributes is 2.16 and the average number of groups is 0.84.

Table 5.5 Characteristics of the 50 Sample Concepts

#Concepts	50
%Primitives	54
%Leaf	44
Average #parents	1.88
Average #attributes	2.16
Average #groups	0.84

Table 5.6 lists the statistics for the suggested attributes for the sample concepts. There were 31 concepts (out of 50) for which all the suggested attributes were found to be correct. A total of 103 attributes were suggested for the 50 concepts with an average of 2.06 attributes per concept. Of these 103 attributes, 67 were found by the auditor to be correctly suggested which gives a yield of 65%.

Table 5.6 Summary of Suggested Attribute Data for the Sample of 50 Concepts

	#	%
Concepts for which all the suggested attributes were correct	31	62
Suggested unique attributes	18	
Suggested total attributes	103	
Suggested average attributes per concept	2.06	
Suggested attributes found to be correct by an auditor	67	65

The algorithm was designed in a way such that if a more refined attribute was already present in the modeling of a concept, a more general attribute was not suggested. For instance, the concept *revision to open reduction of fracture and locked reamed intramedullary nail fixation* has the attribute *procedure site - direct*. A similar concept *revision to closed reduction of fracture and locked reamed intramedullary nail fixation* has the more general attribute *procedure site*. In this case, the algorithm does not suggest *procedure site* as an attribute to be added to the modeling of *open reduction of fracture and locked reamed intramedullary nail fixation* since a more refined attribute *procedure*

site – direct is already present in its modeling. Table 5.7 displays the four general attributes along with their more refined versions that are used in *procedure* hierarchy.

Table 5.7 List of Attributes and their Refined Versions used in *Procedure* Hierarchy

General attribute	Refined attribute
Procedure site (attribute)	Procedure site - Indirect (attribute)
	Procedure site - Direct (attribute)
Procedure device (attribute)	Using device (attribute)
	Direct device (attribute)
	Indirect device (attribute)
Using device (attribute)	Using access device (attribute)
Procedure morphology (attribute)	Direct morphology (attribute)
	Indirect morphology (attribute)

There were seven instances where a sibling attribute of the one already present in the modeling of the concept was suggested. Four of such instances involved the suggestion of *procedure site – indirect* when *procedure site – direct* was already present in the modeling and three instances involved the suggestion of *procedure site – direct* when *procedure site – indirect* was already present in the modeling.

Besides, there were seven instances where two sibling attributes were suggested for a concept based on other similar concepts. The sample had one case of direct morphology/ indirect morphology, one case of direct device/ using device and five cases of procedure site – direct/ procedure site – indirect. In five of the seven instances, one or both of the attributes in the suggested sibling pair was incorrect.

Additionally there were two instances where attribute pair with parent-child relationship was suggested as possible addition to the modeling of a concept. One of the instance involved suggesting both procedure site and procedure site – indirect and the

other case involved suggesting both using device and using access device. One of these two instances was found to be incorrectly suggested.

In terms of target values, of the suggested 103 target values, five were found to be irrelevant to the corresponding concept. For instance, the concept *cementoplasty using fluoroscopic guidance* was suggested the attribute *using device* with target value *angioscope*. However, *angioscope* is a cardiovascular endoscope whereas cementoplasty is related to bones thus making the attribute value irrelevant for this concept. The suggestion was because of the fact that similar concept *angioscopy using fluoroscopic guidance* has that attribute and target value.

Table 5.8 lists the statistics with regards to the target values of the suggested attributes of the 50 sample concepts. Of the 103 suggested target values, the wording in the names of 21 target values were changed to make it relevant to the corresponding concept. Of these 21 target values, two were present in SCT as concepts whereas 19 were suggested as new concepts. The two suggested target values that were present as valid SCT concepts are *structure of male bladder neck* and *blood*.

Table 5.8 Summary of Target Value Data for the Sample of 50 Concepts

	#	%
Suggested target values	103	
Suggested target values found relevant	98	95.1
Target values with changed wording	21	20.4
Target values not in SCT as a concept	19	90.5
Unique target values with changed wording	18	
Unique target values not in SCT as a concept	16	88.9

In terms of unique values, 16 of the 18 suggested unique target values were not a concept in SCT. Of these 16 unique target values that were created as new concepts, two

were found to be inappropriate. The first one was the concept *thorax* which was suggested as a target value of the attributes indirect morphology and direct morphology for the concept *incision and drainage of thorax*. The second one was the concept *excision – value* which was suggested as a target value of the attribute *revision status* for the concept *excision of mastectomy scar*.

Of the 50 sample concepts, there were three instances where attributes were suggested based on lexically similar children concepts. For instance, the algorithm suggested adding the attribute *using device* with the target value of *angiography catheter* for the concept *fluoroscopic angiography of splenic artery*. This was suggested as a result of a lexically similar child concept *fluoroscopic angioplasty of splenic artery* having the attribute *using device* with the target value *angioplasty catheter*. In each of the three instances, one attribute was suggested based on the lexically similar child concept. Two of them were found to be correctly suggested.

Besides, there were 46 attributes borrowed from sibling concepts of which 33 were found to be correct (72%). On the other hand, 56 attributes were borrowed from non-sibling concepts of which 34 were found to be correct (61%). For instance, the concept *excision of mastectomy scar* was suggested the attribute *revision status* with the target value *excision - value* which was found to be incorrect. This attribute was suggested because it was present in the modeling of a lexically similar non-sibling concept *revision of mastectomy scar*.

5.4 Discussion

The study used a lexical technique to automatically suggest attributes and target values for concepts in SCT. The methodology was based on the premise that lexically similar

concepts are modeled in a similar way as discussed in Chapter 4 (Positional Similarity Sets). A technique was used to change the wording of the target value to suit the requirements of the corresponding concept for which the attribute was being suggested. For 62% (31 of the 50) of the sample concepts, all the attributes that were suggested for them were correct. Overall, 65% (67 out of 103) of the suggested attributes were found to be correct.

This was a preliminary study aimed at introducing a methodology to algorithmically identify attributes with minimal human intervention, analyze and assess the results of the study and identify further rules to improve the algorithm. One of the observations made was with regards to sibling attributes. The current algorithm suggested attributes for a concept which already had a sibling attribute in its modeling. Consider the modeling of the two similar concepts *trapping of intracranial aneurysm (procedure)* and *ligation of intracranial aneurysm (procedure)* as shown in Figure 5.3. The algorithm suggested the attribute *procedure site – indirect* for the trapping concept. But, the trapping concept already has the attribute *procedure site – direct* which is a sibling of *procedure site – indirect*. The current algorithm cannot distinguish between which one of such sibling attributes is the correct one. In fact, on many occasions, it is even tough for a human auditor to make a decision between the sibling attributes such *as procedure site – direct* and *procedure site - indirect*. There were seven such cases in the sample as discussed in the Results section.

Besides, in sets of three or more concepts, there are cases where two sibling attributes may be suggested for a concept. Consider the positional similarity set in Table 5.9. The set consists of 8 concepts of which five have the attribute *procedure site – direct*

whereas two of them have the attribute *procedure site – indirect*. The eighth concept *evaluation of gastrointestinal tract (procedure)* has the attribute *procedure site* which is the parent of the attributes *procedure site – direct* and *procedure site – indirect*.

Table 5.9 A Positional Similarity Set with Eight Concepts regarding Gastrointestinal Tract Procedure

	Procedure site - indirect	Procedure site - direct
intubation of gastrointestinal tract (procedure)	yes	
imaging of gastrointestinal tract (procedure)		yes
extubatoin of gastrointestinal tract (procedure)	yes	
biopsy of gastrointestinal tract (procedure)		yes
Radiography of gastrointestinal tract (procedure)		yes
Ultrasound of gastrointestinal tract (procedure)		yes
Fluoroscopy of gastrointestinal tract (procedure)		yes
evaluation of gastrointestinal tract (procedure)		

The current algorithm suggested both the attributes *procedure site – direct* and *procedure site – indirect* for this eighth concept as possible additions since these are more refined form of the attribute *procedure site*. There were seven such instances in the sample where two sibling attributes were suggested as possible additions as mentioned in the Results section. In such cases, a way can be identified to only suggest one of the sibling attributes. For instance, only the more frequently used attribute among similar concepts can be suggested. In the example of Table 5.9, that would be the attribute *procedure site – direct* as it appears five times in the set. In case of a tie, both the sibling attributes can be suggested. The situation can, however get tricky when the modeling of a concept has two or more groups. In such cases, the sibling attributes can be present in the modeling of the same concept but in different groups.

Apart from suggesting attributes, suggesting appropriate target value also forms an integral part of this study. The algorithm changes the wording of the target value as described in the Methods section to make it relevant to the corresponding concept. Of the 21 suggested target values where the wording was changed, two were found to be present in SCT as concepts. One of them was shown in Figure 5.2 where the target value for *method* attribute was changed to *blood* to make it relevant to the corresponding concept *blood concentration, test strip measurement (procedure)*. However, in 19 of the cases, the suggested target values were not found to be a concept in SCT. For instance, consider the modeling of two similar concepts as shown in Figure 5.5.

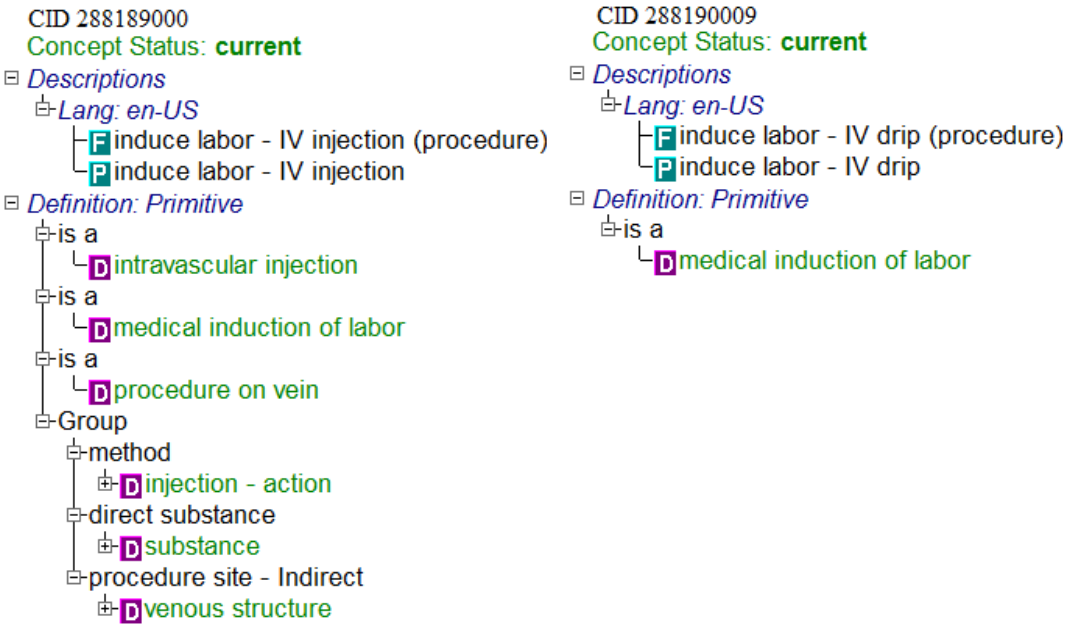

Figure 5.5 A CliniClue snapshot of the modeling of two similar concepts regarding induction of labor procedure.

Source: The Clinical Information Consultancy Ltd, "CliniClue Xplore", 2009, Available from: http://www.cliniclue.com.

The algorithm automatically suggested three attributes for the drip concept of Figure 5.5, namely, *method*, *direct substance* and *procedure site – direct* based on the similar injection concept. However, while suggesting the attribute *method* for the drip concept, the suggested target value was changed to *drip – action* from *injection - action* to better suit the modeling of the current concept. But the target concept *drip – action* or a similar concept is not present in SCT although the concept *injection – action* is present in SCT. Of the 16 unique target values that were created as new concepts, 14 (87.5%) were found to be relevant. Considering the relevance of the new target concepts, the method can also act as an effective technique to identify missing concepts in SCT.

In this study, the algorithm first created the positional similarity sets and then went through all the concepts of each set to suggest attributes by comparing it to the similar concepts of the sets. However, a concept can be in multiple sets as shown in Table 5.10 which can lead to a concept being presented with suggested attributes multiple times. Future studies will take each concept, find all concepts similar to that concept, aggregate all the attributes that are identified as suggestions and present them all together for the concept thus avoiding repetition.

Table 5.10 An Example of a Case where the Same Concept Appears in Two Different Positional Similarity Sets

Percutaneous insertion of pulmonary valve using fluoroscopic guidance (procedure) *Percutaneous replacement of pulmonary valve using fluoroscopic guidance (procedure)*
Percutaneous insertion of pulmonary valve using fluoroscopic guidance (procedure) *Percutaneous insertion of aortic valve using fluoroscopic guidance (procedure)*

Further studies will involve auditing larger samples and by multiple auditors in a blinded manner. This study will be used to further assess the cases that were discussed

above. This will include studying the effect of suggesting attributes based on lexically similar sibling concepts and lexically similar descendent concepts. The effect of suggesting sibling attributes and attributes with parent-child relationship will also be further studied to identify a way such that only the more appropriate attribute may be suggested. The current study does not consider role groups while suggesting attributes as discussed in the Methods section. Future studies will involve identifying ways to take into account these role groups while suggesting attributes. Besides, further studies will also involve improving the effectiveness of the positional similarity sets which provides the basis to algorithmically detect attributes to enhance the modeling of concepts. Towards this end, the effect of the hierarchical attributes and sibling attributes between the concepts in a set, and the effect of the concepts being primitive or non-primitive and leaf or non-leaf will be studied as improved sets will result in improved suggestion for attributes.

5.5 Summary

The study provided a technique to algorithmically suggest attributes to enhance the modeling of SCT concepts. An experiment was performed to validate the effectiveness of the method and the results showed 65% of the suggested attributes being identified correctly. With limited availability of resources, automatic techniques such as the one presented in this study will help in achieving consistency and correctness in the modeling of SCT concepts, thus leading towards the goal of an improved terminological content and consequently better health care delivery.

CHAPTER 6

PROBLEM LISTS

6.1 Introduction

By 2015, SCT will become the standard terminology for EHR encoding of diagnoses and problem lists in the USA [4]. To facilitate encoding of problem lists, the National Library of Medicine has extracted a collection of UMLS concepts dealing specifically with health problems [50]. As it happens, SCT was found to be the UMLS source offering the best coverage of this so-called UMLS clinical observations recording and encoding (CORE) problem list. The SCT portion (amounting to 81% coverage) was posted on-line in July 2011 as the "SCT CORE" problem list [17], comprising 5,862 active concepts. It was accompanied by the alternative "Veterans Health Administration and Kaiser Permanente (VA/KP)" problem list [18], consisting of 16,622 active SCT concepts. The two lists have 4,004 concepts in common.

While PL encoding is SCT's primary and immediate contribution toward meaningful use of EHRs, the use of SCT, due to its inherent structure, stands to support patient education and advanced clinical data repository queries which are amongst the meaningful use objectives. Such intended use of clinical terminologies had been suggested previously in numerous studies [9-16]. However, there are indications that at the moment, SCT as a clinical terminology is not optimally structured for such use [81]. Also, on the IHTSDO Special Interest Group discussion board [86] in 2010, there was a general agreement that "as is," SCT is not suitable for use in patient-facing applications. Such issues are barriers to its successful deployment.

This study aims to look at the proposed problem lists with an eye toward their readiness for use in EHRs. The focus is on the combination (union) of SCT CORE Problem List and Veteran Administration and Kaiser Permanente (VA/KP) Problem List (collectively as the "PL") containing a total of 18,480 SCT concepts. A study is performed to examine the quality of the PL. In particular, the study aims to determine if the modeling of the PL's concepts has reached a stable and correct state due to frequent use and increased scrutiny on them—or whether the PL requires further quality-assurance (QA) efforts. Equally sized random samples of concepts are extracted from two different concept populations for analysis: the first consists of concepts strictly from the PL and the second contains general SCT concepts distributed proportionally to the PL's in terms of their hierarchies.

The results of the analysis show that PL concepts suffer from the same issues as general SCT concepts, although to a slightly lesser extent, and thus indeed require further QA. This additional QA is especially warranted in view of the intended role of PL concepts for the meaningful use of EHRs. Towards this end, two structural indicators in the form of the number of parents and words are analyzed as an effective way to ferret out concepts with a high probability of error. A third structural indicator to identify errors in synonyms is also investigated.

6.2 Method

6.2.1 Comparative Analysis of PL and SNOMED Concepts

A study is conducted to assess various qualities and properties of the PL and determine how well its concepts are currently modeled in comparison with the rest of SCT. Two

samples, the PL sample and the proportional sample, are used to evaluate the correctness of the modeling of the concepts. The second sample serves as a control sample. Each concept is evaluated based on the following properties: (a) IS-A structure, (b) descriptions, i.e., FSN, PT, and synonyms (if they exist), (c) conceptual modeling, i.e., relationships, relationship targets, and relationship groups. The CliniClue browser is used to visualize the concepts and navigate SCT for the review. Findings are recorded according to a four-point scale of presumed significance: none, mild, moderate, and severe. Specifically, "severe" is assigned in cases of obvious errors in the hierarchy or relationship targets. "Moderate" is used for correct but overly broad or redundant parents that could be removed or replaced by more specific values. An assignment of "mild" denotes relationship targets of too general a nature, where more specific ones could readily be used. It is to be noted that these levels are not based on clinical significance but rather the degree of deviation from appropriate modeling compared to other SCT concepts.

6.2.2 Analysis of Concept Synonyms

An important characteristic of PL concepts that is examined in this study is the concept synonyms. Three random samples of 50 concepts each are examined. The first sample consists of concepts strictly from the PL. The second is composed of non-PL SCT concepts, but these are chosen directly in proportion to the distribution of concepts in the PL sample across hierarchies. That is, if 5% of the PL sample is from the Procedure hierarchy, then 5% of this second, so-called "proportional," sample is randomly chosen from that hierarchy. The third sample comprises concepts chosen from the population of SCT concepts at large, without any consideration of the PL. The need for the proportional

sample arises in the case where the results of the PL sample differ from those for the third sample (of the entire SCT). It helps to determine whether the difference stems from them being PL concepts or is due to the properties of their hierarchies (mainly, Clinical Finding), which may differ from those of the whole SCT.

For each one of the three samples, a count of the number of concepts with synonyms, and the average numbers of synonyms is presented. For comparison purposes, the count of the number of concepts with UMLS synonyms as well as the average number of UMLS synonyms per concept is also presented. All synonyms of the PL sample's concepts are reviewed manually. Additionally, a potential structural indicator of concepts having erroneous synonyms is examined by looking for two or more SCT concepts mapped to the same UMLS concept. For example, the two PL concepts *Dermatitis* and *Eczema* are both mapped to the UMLS concept *Dermatitis*. Indeed, the SCT concept *Dermatitis* has erroneous synonym "Eczema," since *Eczema* is a child of *Dermatitis*. This is a case where two related PL concepts of different granularity are erroneously modeled as synonyms in SCT. In another example, the siblings *Simple goiter* and *Endemic goiter* (each a PL concept) both map to the UMLS concept *Endemic goiter*. On closer examination, it can be seen that *Endemic goiter* has the term "simple goiter" as a (erroneous) synonym. The integration of SCT into the UMLS helps to unearth such erroneous SCT synonyms. It was found that 569 concept pairs from the PL exhibited a duplicate mapping to a UMLS concept. Furthermore, there are 2,056 pairs of such SCT concepts where only one is in the PL. A manual review is done to check for erroneous synonyms in a random sample of 50 such pairs from the 569 pairs.

6.2.3 Analysis of Number of Parents of a Concept

A theme that was discovered in previous research on the auditing of SCT is that "complex" concepts typically have higher likelihood of errors than "regular" concepts. One example of a type of complex concept for which studies confirmed such a tendency is a concept residing in a *region of strict inheritance* (of relationships) [87], defined in the context of an abstraction network for an SCT hierarchy [77]. Another example is *overlapping concepts*, defined with respect to elements of another kind of SCT abstraction network [88, 89].

The number of parents of a concept can also be taken as a parameter of complexity in the sense that a concept with multiple parents is a specialization of each and inherits their properties, making it a build-up of knowledge coming from multiple paths. Such a concept reflects multiple identities, being "a kind of this and a kind of that." Such an observation leads to a thought if multiple parents are an indicator of problems. That is, do concepts with multiple parents in the PL tend to have more problems than those with one parent? A study is conducted to analyze any errors found in the above mentioned review for the PL sample and the proportional sample with respect to the number of parents to see if multiple parents are indeed an indicator. The following two hypotheses are formulated:

Hypothesis 1: PL concepts with multiple parents are more likely to be in error than concepts with one parent.

Hypothesis 2: The likelihood of errors in PL concepts increases with the number of the parents.

Figure 6.1 lists the distribution of the PL concepts according to their numbers of parents. For example, 7,823 concepts (42.3%) in the PL have exactly one parent. There are 53 concepts with 8-15 parents. There is one PL concept, *Granuloma inguinale (disorder)*, with 15 parents, and another, *Menkes kinky-hair syndrome (disorder)*, with 12. The average number of parents for a PL concept is 1.90.

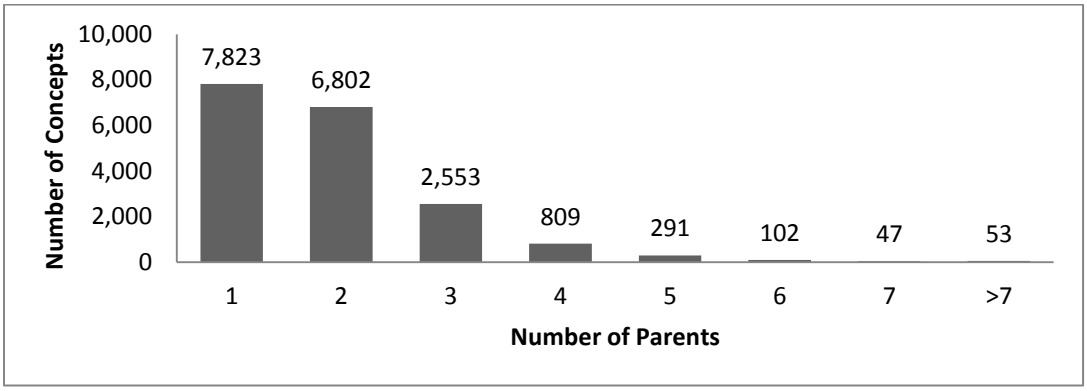

Figure 6.1 Distribution of concepts in problem list according to the number of parents.

To test the two hypotheses, eight samples of PL concepts are randomly generated according to their numbers of parents as shown in Figure 6.2. For the first six samples, for a given n ($1 \leq n \leq 6$), sample n has 50 concepts, each having n parents. There are only 47 concepts with seven parents in the PL, so all those are audited in a seventh sample. The eighth sample consists of all concepts with eight or more parents, the number of which is 53. In order to assess Hypothesis 1, an additional random sample of 250 PL concepts with only one parent each is also studied.

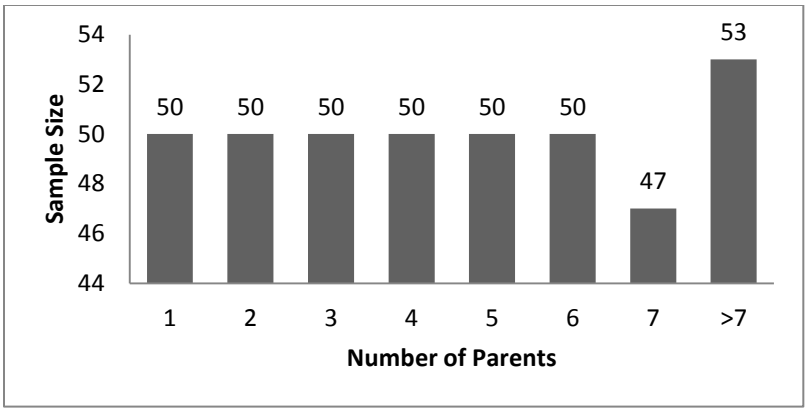

Figure 6.2 Distribution of concepts in sample sets according to the number of parents.

6.2.4 Analysis of Concept Net Word Length

The net word length of a concept is defined as the length of the fully specified name of a concept excluding the insignificant words also known as the stop words. A concept with more number of net words will possibly be highlighting multiple dimensions of the concept relating to multiple attributes which may result in the concept being more complex than concepts with less number of net words. From here on, "net words" will simply be referred to as "words." Based on the notion that more complex concepts are more likely to contain errors, the following two hypotheses are formed.

Hypothesis 3: Error concentration among concepts increases with the increase in the number of words in the concept fully specified name.

Hypothesis 4: Concepts with large word length and large number of parents tend to be more complex resulting in greater error concentration.

In order to investigate these hypotheses, a study is conducted on a random sample of the PL. The January 2012 release of SCT is used which brings down the total number of current PL concepts to 18,472 as compared to 18,480 concepts in the above studies which used the Jul 2011 release of SCT. The concepts of the PL are classified based on

the word length of the concept. The list of stop words used in this research work is as recommended in the SCT Developer Toolkit Guide 2012 [90] and has been clubbed together with some other frequently occurring words used by PubMed [91] and by members of UMLS MetaMap [92] to form a list of 157 stop words. Figure 6.3 displays the distribution of the PL concepts arranged by the word count. The concept word length in PL is in the range of 2 to 20 with approximately 90% of the concepts being in the range of 2-6 words.

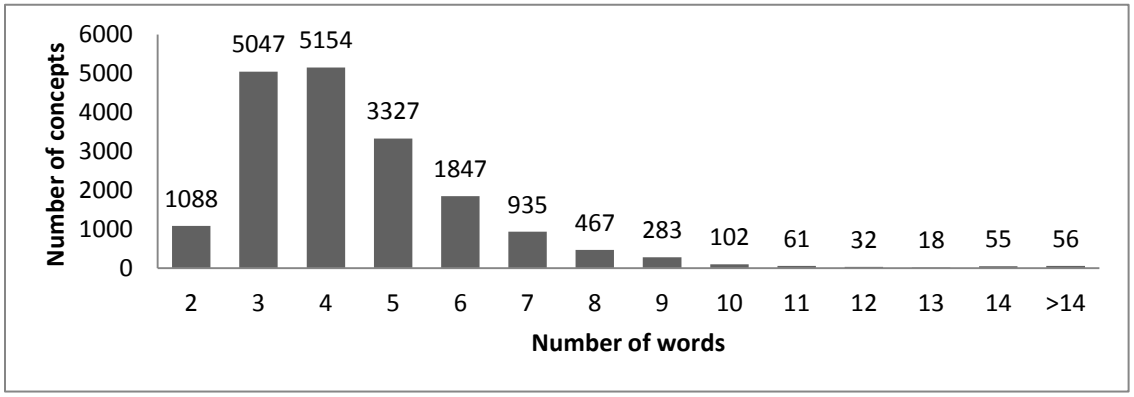

Figure 6.3 Distribution of concepts in problem list according to the number of words.

Moreover, a two-dimensional distribution of the PL concepts is displayed in Table 6.1. The rows denote the number of words which ranges from 2 to 20. The columns represent the number of parents ranging from 1 to 14. Each cell represents the number of concepts in PL with certain word length indicated by the row header and the number of parents which is indicated by the column header. For e.g., there are 569 concepts with two words and one parent whereas there is only one concept with 12 words and four parents. The last row and column in the table displays the total number of concepts with a

certain number of parents and certain word length respectively. For example, there are 7821 concepts with one parent and 1088 concepts with two words and in the entire PL.

Table 6.1 Word Length vs. Number of Parents for the Entire Problem List

Words/Parents	1	2	3	4	5	6	7	>7	Total
2	569	332	122	45	15	1	3	1	1088
3	2182	1781	673	268	86	33	11	13	5047
4	2181	1895	712	222	83	38	8	15	5154
5	1346	1337	460	127	32	14	7	4	3327
6	702	713	303	87	27	7	6	2	1847
7	374	344	137	51	18	5	3	3	935
8	197	192	53	17	6	1	1		467
9	92	112	70	5	4				283
10	48	32	17	2	1	2			102
11	39	18	4						61
12	15	16		1					32
13	10	4	2			2			18
14	46	7	2						55
>14	20	32	2	1	1				56
Total	7821	6815	2557	826	273	103	39	25	18472

For experimental analysis, a sample of concepts is randomly selected from the PL with a sizeable number of concepts from different word length category. For every word length that has more than 50 concepts, 50 randomly selected concepts are taken and for every word length that has less than or equal to 50 concepts, all of those concepts are taken to form the sample concept set. Figure 6.4 displays the distribution of the 656 concepts in this sample set.

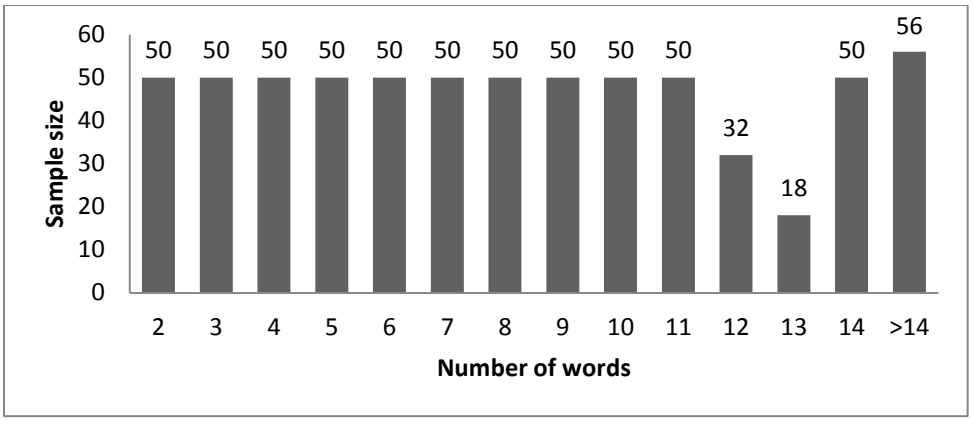

Figure 6.4 Distribution of concepts in sample sets according to the number of words.

The review of the samples for Sections 6.2.1 and 6.2.2 was done by Dr. Gai Elhanan whereas that of the samples for Sections 6.2.3 and 6.2.4 was done by Dr. Yan Chen. Both the auditors are trained in medicine and terminologies and have vast experience in auditing terminologies. The auditors were given the sample to audit without any information on the methodology used and the hypotheses being tested.

6.3 Results

6.3.1 Modeling Errors

Table 6.2 displays the results of the auditing of the modeling correctness of the PL and proportional samples. In total, 17 problems were found for each of the two samples. There were no severe problems found for the PL sample. Eleven concepts displayed moderate issues, and six exhibited mild ones for the PL sample. From the proportional sample, four concepts displayed severe problems, six exhibited moderate issues, and seven displayed mild issues.

An example of a "moderate" finding is exhibited by the concept *Benign neoplasm of skin of umbilicus (disorder),* from the PL sample, having three parents. Two of the

parents are *Benign neoplasm of skin of trunk, excluding scrotum* and *Benign neoplasm of skin of abdomen*. Considering that the focus of the concept is the umbilicus, the trunk is not an appropriate parent, especially since *Umbilical structure* is considered a *Structure of central region of abdomen* in SCT.

Table 6.2 Results of Auditing PL and Proportional Samples' Concepts

	Problem			
	Mild	Moderate	Severe	Total
PL	6	11	–	17
Proportional Sample	7	6	4	17

Another PL sample concept with moderate finding is *Lumbosacral spondylosis without myelopathy*. The concept has four parents: *Lumbosacral spondylosis, Disorder of trunk, Degenerative disorder*, and *Spondylosis without myelopathy*. Some of these are clearly related, while other seems to be defined at the wrong level. For example, why is *Disorder of trunk* a parent at this refined level? Should it not be defined at a higher level as parent of *Lumbosacral spondylosis*? The same can be said of *Degenerative disorder* as a parent of the grandparent *Spondylosis*. See Figure 6.5 for the modeling of the parents of *Lumbosacral spondylosis without myelopathy*, before and after QA.

A "mild" finding can be seen for the PL sample concept *Complication of reimplant (disorder)*. The concept has the relationship *associated with* that targets the *Surgical procedure*. However, *Surgical procedure* is an overly broad concept, really a container class for all types of procedures, reimplants, and whatnot. On the other hand, *Reimplantation (procedure)* does exist in SCT as a child of *Surgical procedure* and is a much more appropriately refined target.

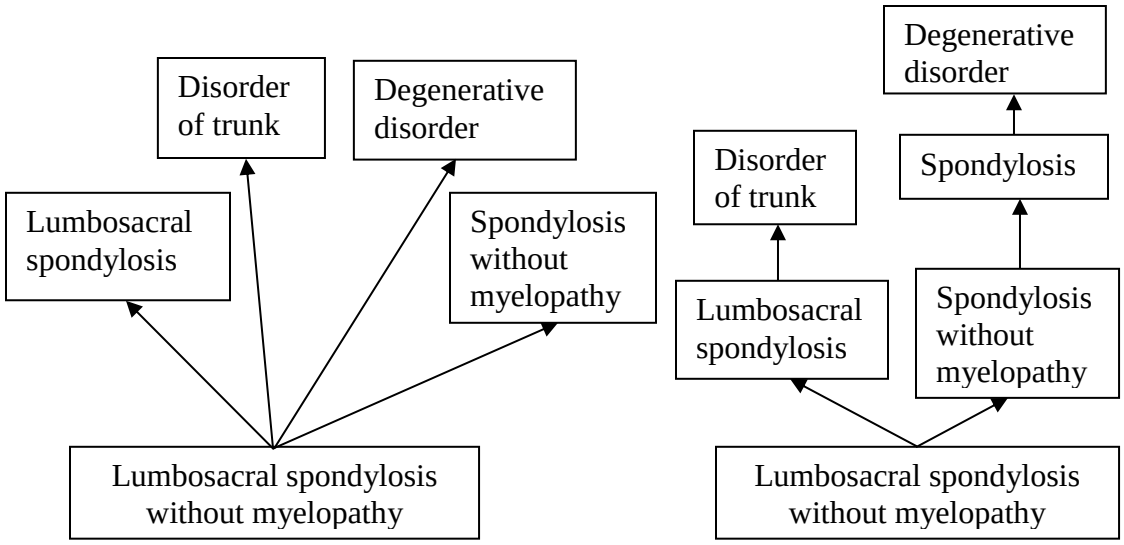

Figure 6.5 Parents of *Lumbosacral spondylosis without myelopathy* before and after QA.

An example of severe problems from the proportional sample can be found with *Cystic adventitial disease of popliteal artery (disorder)*, which has the following parents: *Vascular disease of abdomen and Systemic arterial finding*. However, the popliteal artery is not an abdominal artery, and this is not a systemic finding. Thus, these two parents were considered inappropriate. Additionally, the concept has the relationship *associated morphology* with the target *Cystic medial necrosis*. But cystic adventitial disease, a rare disorder, is not characterized by cystic medial necrosis but rather by mucinous cysts in the outer media or adventitia that progressively compromises the arterial lumen. Thus, this target value was also considered inappropriate. Altogether, these findings resulted in a "severe" rating. It is noted that in SCT such concepts are not generally associated with a laterality attribute. However, from a clinical perspective (and aside from the clinical knowledge used in this review or that of a presumed clinical user), the SCT content does not provide any clues to the fact that this concept should be associated with "left," "right," or "bilateral" modifiers. From this perspective, it should be considered as a

"moderate" issue as it applies to the clinical usefulness and the possibility of disseminating accurately coded information.

6.3.2 Synonym Errors

Table 6.3 shows various measures for the three samples. The SCT's July 2011 release was used for this study. The first two samples are similar with regard to the number of primitive concepts, about 60%, which is high, but lower than for the third sample. With regards to synonyms, the PL sample is clearly better than the other two, both in the number of concepts with synonyms and in the average number of synonyms per concept. But the numbers are still low when compared to the UMLS synonyms for these concepts. The extremely high number of UMLS synonyms for the PL sample (12.80 per concept, on average) is due to the popular concepts' occurrences in many UMLS sources, each with its own set of synonyms.

Table 6.3 Properties for Three Random Samples of Concepts

Property type	PL Sample	Proportional Sample	SCT Sample
# Primitive concepts	29 (58%)	31 (62%)	42 (84%)
# Concepts with SCT synonyms	27 (54%)	17 (34%)	17 (34%)
# Concepts with UMLS synonyms	45 (90%)	44 (88%)	45 (90%)
Average # SCT synonyms	1.16	0.46	0.40
Average # synonyms for concepts with SCT synonyms	2.15	1.36	1.17
Average # UMLS synonyms	12.80	2.60	2.84
Average # parents	2.02	1.84	1.52
Average # words in preferred term	4.58	5.00	5.32

The next two properties, the number of parents and the number of words in the preferred term are related to the complexity of the concepts rather than the quality of their modeling. For those two properties, there is no significant difference between the PL

sample and the proportional sample. Thus, this preliminary study indicates that probably due to their frequent use, the PL concepts are to some extent better in their properties than other concepts in their respective hierarchies. They are still far from satisfying the expected needs of coding diagnoses and problem lists for longitudinal care in EHRs due to their low synonym coverage and the high percentage of primitive concepts.

From the review of all synonyms of the PL sample's 50 concepts, only two erroneous synonyms were found and both were for the same concept. Specifically, *Premenopausal menorrhagia* has a total of four synonyms: two correct (*preclimacteric menorrhagia*, *excessive bleeding at onset of menopause*) and two erroneous ones (*climacteric menorrhagia*, *menopausal menorrhagia*). Note that this low error rate is at least partially due to the low average number of synonyms for the PL sample and SCT in general. Looking only at the 12 PL sample concepts with multiple synonyms, the two erroneous synonyms constitute 5% of the 43 synonyms. Looking only at the five PL concepts with at least four synonyms, the two erroneous synonyms constitute 7% of their 27 synonyms. In a similar trend, reviewing the UMLS synonyms for the PL sample concepts (where the average is high with 12.8 synonyms per concept), eight of them were found to have erroneous synonyms.

While it is seen that the number of erroneous synonyms for PL concepts is, in general, low, the situation is much different for the 569 pairs of PL concepts where both pair members were mapped as duplicates to a UMLS concept. From the random sample of 50 such pairs of PL concepts, 26 pairs (52%) were found to have a synonym error. Hence, by auditing such pairs, which can be identified automatically, it is possible to further lower the percentage of erroneous synonyms for PL concepts with a relatively

small effort. For example, in the PL, *Endemic cretinism* and its parent *Congenital iodine deficiency syndrome* are both mapped to the UMLS concept *Endemic cretinism* (CUI C0342200). However, endemic cretinism is just one type of cretinism and is not necessarily synonymous with it. Nevertheless, both SCT concepts have the term "cretinism" as a synonym, which may have contributed to the confusion. Hence, "cretinism" should be removed as a synonym from SCT's *Endemic cretinism*.

6.3.3 Number of Parents as an Indicator of Errors

Tables 6.4 and 6.5 display the distribution of errors in the PL sample and the proportional sample with regard to number of parents. The moderate and severe errors are combined into one category; the mild errors are in a separate category to facilitate a comparison with the proportional sample. For example, in the PL sample, there are 24 concepts with exactly two parents. Five of them had mild errors, while another five had moderate or severe errors. In total, 41.7% of such concepts exhibited errors. For the 35 concepts with multiple parents, there are a total of 15 (42.9%) in error versus just two (13.3%) among the 15 concepts with one parent. There are five erroneous concepts (45.4%) for those with at least three parents. Interestingly, all these errors were moderate; none were mild. These results are in line with Hypothesis 1. Another interesting observation is that three out of four concepts with at least four parents have moderate errors (75% error rate). However, this sample is too small for the evaluation of Hypothesis 2.

For the proportional sample, the findings are, in general, similar to those for the PL sample (with the exception of some errors being severe) with a 21.7% error-rate for concepts with one parent, 44.4% for concepts with two or more parents, and 70% for concepts with at least three parents as shown in Table 6.5. The error-rate for concepts

with four or more parents is 50%, with one of the two concepts exhibiting a severe error level and the other exhibiting a moderate error level. This distribution also supports Hypothesis 1. But, again, there is not enough data for the evaluation of Hypothesis 2.

Table 6.4 Errors in the PL Sample of 50 Concepts

#Parents	#Concepts	Mild errors	Moderate + severe errors	#Errors	%Errors
1	15	1	1	2	13.3
2	24	5	5	10	41.7
3	7	–	2	2	28.6
4	3	–	2	2	66.7
5	1	–	1	1	100.0

Table 6.5 Errors in the Proportional Sample of 50 Concepts

#Parents	#Concepts	Mild errors	Moderate + severe errors	#Errors	%Errors
1	23	3	2	5	21.7
2	17	1	4	5	29.4
3	6	3	2	5	83.3
4	3	–	1	1	33.3
5	1	–	1	1	100.0

Table 6.6 presents the results of auditing the six samples of 50 random concepts of the PL derived based on the number of parents n ($1 \leq n \leq 6$), the collection of 47 concepts having seven parents, and the collection of 53 concepts with eight or more parents. For example, among the 50 concepts with one parent examined, four were found to be in error. Moreover, a total of six errors were discovered in these four concepts, yielding a rate of 1.50 errors per erroneous concept. Similarly, for the 53 concepts with eight or more parents, 27 were found to be erroneous. Besides, these 27 concepts were found to have a total of 62 errors, yielding a rate of 2.29 errors per erroneous concept.

Table 6.6 Error Concentration in Concepts with Different Number of Parents

#Parents (n)	#Concepts	#Erroneous concepts	%Erroneous concepts	#Errors	Avg #errors per erroneous concept
1	50	4	8.0	6	1.50
2	50	3	6.0	3	1.00
3	50	4	8.0	9	2.25
4	50	9	18.0	15	1.66
5	50	24	48.0	46	1.91
6	50	28	56.0	75	2.67
7	47	15	31.9	23	1.53
≥ 8	53	27	50.9	62	2.29

In addition, a random sample of 250 PL concepts with only one parent was also audited. Table 6.7 compares the results of this auditing to those for the sample of 250 PL concepts obtained by aggregating the results of Samples 2-6 from Table 6. The purpose of this comparison is to further test Hypothesis 1. The results in Table 7 support this hypothesis in two ways. First, there is a much higher percentage of erroneous concepts in the case of multiple parents versus one parent: 27.2% and 8%, respectively. Second, the multi-parent concepts display about 50% more errors per erroneous concept on average than do single-parent concepts. The difference is statistically significant according to the Chi Square test (with $p < 0.0001$).

Table 6.7 Error Concentration in Concepts with One vs. 2-6 Parents

#Parents	#Concepts	#Errns cpts	%Errns cpts	#Errors	Avg #errors per errns cpt
1	250	20	8.0	29	1.45
2–6	250	68	27.2	148	2.17

Even though, as seen in Table 6.6, there is a trend of increasing errors with the number of parents, it is not strictly monotonic. For example, the sample of two-parent concepts shows three erroneous concepts as compared to four for the sample of single-

parent concepts. Similarly, the sample of 47 concepts with seven parents exhibits 15 erroneous concepts as compared to the 28 for the sample of 50 concepts with six parents. To capture this general trend, the results of Table 6.6 are aggregated as shown in Table 6.8.

Table 6.8 Error Concentration for Aggregate Concepts

#Parents	#Concepts	#Errns cpts	%Errns cpts	#Errors	Avg #errors per errns cpt
1–3	150	11	7	18	1.64
4–6	150	61	40	136	2.23
7–15	100	42	42	85	2.02

The samples of concepts with 1-3 parents are combined in the first row, those with 4-6 parents are in the second row, and 7-15 are at the bottom in Table 6.8. It can be seen that the percentages of erroneous concepts for the second group (4-6 parents) and the third group (7-15 parents) are much higher than for first group (1-3 parents). As a matter of fact, the number of erroneous concepts for the 150 concepts of the "4-6 parent" sample is statistically significantly higher than for the 150 concepts of the "1-3 parent" sample according to the Chi Square test (with $p < 0.0001$). Furthermore, the average number of errors per erroneous concept in the former sample is larger, with a ratio between the two of 1.36 (= 2.23/1.64). Similarly, the number of erroneous concepts for the "7-15 parents" sample is statistically significantly higher than for the "1-3 parents" sample (Chi Square test with $p < 0.001$). Again, the average number of errors per erroneous concept in the former sample is larger, with a ratio of 1.23 (= 2.02/1.64). Hence, Hypothesis 2 is confirmed as a low granularity measure (but not as a high granularity measure). Also, the high percentage of errors does not continue to grow for the extremely high number of parents.

6.3.4 Number of Words as an Indicator of Errors

Figure 6.6 shows the distribution of error among the concepts with different word length from the 656 sample concepts that were randomly picked for auditing. As shown, 8% and 4% of the concepts with word lengths 2 and 3, respectively are erroneous. For concepts with word length 4 or more, the error percentage increases to double digits with the exception of the concepts with word length 14 where the error rate stands at 8%. For concepts with more than 14 words, the error percentage increases to 41.1%. This distribution of error is in line with Hypothesis 3.

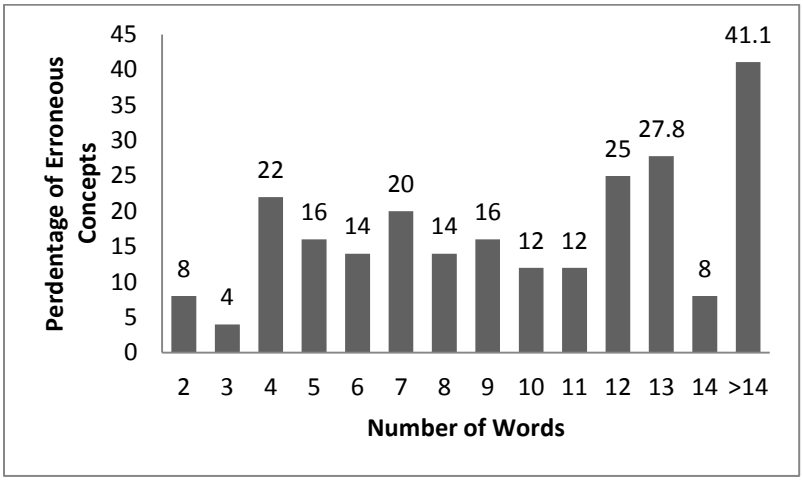

Figure 6.6 Distribution of error percentage among sample concepts with different word length.

In Figure 6.7, the 656 sample concepts have been aggregated into three groups based on their word length. Concepts with 2-3 words are the small length concepts, concepts with 4-14 words are the mid length concepts and concepts with 15 or more words are the large length concepts. As can be seen in Figure 6.7, small length concepts have the least percentage of error which is 6%. As compared to this, mid length concepts were found to have a 16% error rate. On the other hand, large length concepts exhibited a

41% error rate. This again demonstrates that the error rate increases with the increase in word length of the concepts which supports Hypothesis 3.

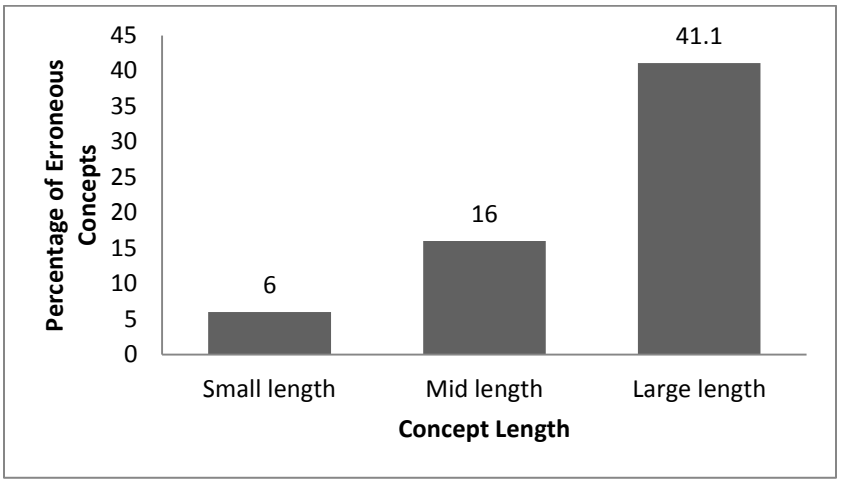

Figure 6.7 Distribution of error percentage among aggregated sample concepts with different word length.

Fisher's exact test is used to calculate the two-tailed P value to determine if the association between the three groups is statistically significant. The association between the small length and mid length groups was found to be statistically significant ($p = 0.0075$). Furthermore, the association between the mid length and large length groups and between the small length and large length groups were found to be extremely significant ($p < 0.0001$).

Table 6.9 presents the two-dimensional view of error percentage found in the sample of 656 concepts distributed by their number of words and number of parents. For instance, 8% of the concepts, that are of length two and have one parent, exhibit error whereas 100% of concepts of length five and having seven parents are found to be erroneous. A general trend that is observed here is that the percentage of concept error

tends to increase with the increase in the number of parents as well as with an increase in the number of net words. This error distribution is in line with Hypothesis 4.

Table 6.9 Two Dimensional Distribution of Error Percentage among Concepts

Words/Parents	1	2	3	4	5	6	7
2	8	7.2	0	25			
3	18.2	0	0	0	0	0	
4	23.6	20	33.4	20	50	0	
5	6.7	13.1	0	33.4	50		100
6	0	14.3	11.2	0	100	100	
7	30	11.8	0	16.7	50		
8	20	0	40	100			
9	5.3	23.6	20	0	50		
10	8	7.7	11.2	50	100		
11	6.7	18.8	25				
12	26.7	18.8		100			
13	10	0	100			100	
14	0	66.7	0				
>14	90	9.4	50	0	100		

In order to give a compact view of the general trend of error being exhibited by the concepts with different number of words and parents, the concepts are aggregated into three groups with respect to word length and into two groups with respect to the number of parents. Based on the word length, the concepts have been divided into the three groups as discussed above: groups with words length of 2-3, 4-14 and >14. Based on the number of parents, the concepts have been divided into two groups: group with small number of parents (1-3) and group with large number of parents (>3).

Table 6.10 presents the percentage error among the aggregate concept groups and shows a general trend of increasing error percentage from top to bottom and left to right. Only 6.4% of the concepts with 2-3 words and 1-3 parents are erroneous as compared to 50% of the concepts with more than 14 words and 4-7 parents. These findings support Hypothesis 4.

Table 6.10 Distribution of Error Percentage for Aggregate Concepts

Words/Parents	1-3	4-7
2-3	6.4	4.8
4-14	13.3	42.6
>14	40.8	50

6.4 Discussion

This study investigated whether and to what extent the concepts of the PL, the problem lists derived from SCT, suffer from the deficiencies that SCT in general is known to suffer from, regarding its support for primary and secondary meaningful use of EHRs. Of particular interest was whether due to frequent use and increased scrutiny, the concepts of PL are of better quality. According to the results of this study, the PL concepts show better synonym coverage and less severe modeling errors than those concepts in the proportional sample, derived with an eye toward the major hierarchies covered in the PL. This may be due to more attention paid in their initial modeling or improvement resulting from users' feedback.

However, even with higher synonym coverage and better modeling, the PL concepts still suffer from the same problems as those in the general SCT population, just to a slightly lesser extent. As was shown, the synonym coverage of PL concepts is still

poor and definitely falls short of the level required for proper support in the primary meaningful use of EHRs, namely, problem-list encoding of 80% of patients—as an incentive for practitioners. A concerted effort to increase synonym coverage at least for the PL concepts is needed to fulfill the requirement of the HITECH initiative. Also, extensive efforts to improve the relationship modeling need to be made. Such efforts are complex and are expected to demand more editorial resources than the efforts to increase synonym coverage. Note that progress in accurate relationship modeling is also expected to manifest itself in a decrease in the number of primitive PL concepts, currently amounting to approximately 60% of its overall content.

As partial remedies for the findings of this study, three structural indicators are presented that can help to optimize the effectiveness of QA work on the PL concepts. The first, dealing with synonym problems, targets pairs of concepts when there is a duplicate mapping to a UMLS concept. Erroneous synonyms may result in the reporting of incorrect concepts from the problem lists in EHRs. The results of auditing the 50 sample pairs, out of the 569 total pairs, show a 52% error rate. Extrapolating to the 569 pairs, one can expect to find quite a few (i.e., 296) erroneous synonyms.

With regards to synonyms, a deeper study into the UMLS synonyms for the sample concepts will be needed. It is, however, observed that many of the UMLS atoms are simple lexical permutations. For example, *Diabetic Nephropathy* has 36 synonyms, including "*Diabetic Nephropathies*," "*Diabetic nephropathies*," and "*Nephropathy - diabetic*." These duplicates should be filtered out when comparing with the SCT synonyms. There should also be a focus on synonyms that are truly different phrases and not just lexical variations. Such a comparison will give a more realistic target for the

desired amount of synonyms for the PL concepts. With regard to the 2,056 PL-concept/non-PL-concept pairs with duplicate UMLS mappings, further studies are needed to determine their synonym error percentages (specifically for the PL-concept member).

The second indicator, namely, the simple measure of the number of parents, deals with general modeling problems. Note that according to a recent study of SCT users [38], about 85% found "severe" errors (i.e., obvious errors in the hierarchy or relationship targets) to be somewhat bothersome. Out of these, 60% were very much bothered about incorrect parents. This second indicator can guide the ordering of the QA efforts starting with concepts having extremely high numbers of parents and working downward from there.

In the auditing of a sample of 400 concepts with various numbers of parents, summarized in Table 6.8, 103 erroneous concepts out of the 250 concepts (41.2%) were found with at least four parents. There were just 11 erroneous concepts out of the 150 concepts (7.3%) with less than four parents. The ratio of errors per erroneous concept for those concepts with at least four parents was 2.15 (= 221/103) versus a ratio of 1.64 (= 18/11) for those with less than four parents.

Based on the percentages of erroneous concepts and error ratios reported in Table 6.6, one can calculate a weighted estimate of the expected findings for the entire set of 1,302 PL concepts with at least four parents (see Figure 6.1). Using these data, one can calculate the estimate as follows:

$$809 \cdot 0.180 + 291 \cdot 0.480 + 102 \cdot 0.560 + 47 \cdot 0.319 + 53 \cdot 0.509 = 146 + 140 + 57 + 15 + 27 = 385$$

Therefore, one can expect 29.6% (= 385/1302) of the concepts with four or more parents to be in error. The weighted estimate of the number of errors for these expected 385 erroneous concepts can be calculated from the above formula and the data in the last column of Table 6.6 as:

$$146 \cdot 1.66 + 140 \cdot 1.91 + 57 \cdot 2.67 + 15 \cdot 1.53 + 27 \cdot 2.29 = 746$$

Hence, their expected error ratio is 1.94 (= 746/385). This estimate emphasizes the power of this indicator to deliver a high yield of severe errors for a moderate amount of auditing effort.

In contrast, if one were to audit all 17,178 PL concepts with less than four parents, the weighted estimate of the number of erroneous concepts is just 1,237 (7.2%), with only 1,805 total errors and an error ratio of 1.46. Considering the extensive efforts required for such a large audit, this expected yield would likely not warrant it.

On the other hand, one can limit the audit to the 493 concepts with at least five parents. The weighted expected number of erroneous concepts in this case is 239 (48.5%). The corresponding number of errors is 504, with an error ratio of 2.11.

A third structural indicator in the form of word length is also shown to help isolate groups of concepts with high likelihood of error. The combination of word length and number of parents is also shown to be an effective structural indicator to ferret out the problematic concepts. By focusing on such concepts, a more limited effort can provide a relatively high yield in terms of the number of errors though the total number of errors is smaller.

The method described in this study has a trade-off between the extent of the QA efforts and the results obtained. It is aimed towards making use of the limited resources in

an efficient way. In this study, a single auditor reviewed individual concepts one at a time using the CliniClue Browser and using the knowledge of related concepts such as parents, children and siblings. Future work will include more than one auditor to improve the authenticity of the auditor's report. It has been shown in past studies [19, 93] that group auditing can be a more efficient way to expose errors which otherwise may be difficult to find. Future work will involve identifying and applying such appropriate group-based auditing along with the methods used in this study to come up with more evolved QA methods. The study showed a way to combine two structural indicators to ferret out the more vulnerable concepts from the rest of the SCT. Future work will involve combining more of such indicators to get a more sophisticated method that can help in the QA efforts.

6.5 Summary

A study is performed to examine the readiness of the concepts in the problem lists for their intended meaningful use in EHRs. It is found that these concepts tend to suffer from the same problems as the concepts found throughout the general SNOMED CT content, just to a slightly lesser degree. Such problems include a high percentage of primitive concepts (likely to be missing relationships), and deficient and inconsistent modeling of relationships. The conclusion is that further QA efforts are needed for the problem lists' concepts. Leaving the problems unaddressed will have deleterious effects on secondary meaningful use of EHRs, a hallmark of the HITECH initiative. To help guide such QA efforts, three straightforward structural indicators that can be used to ferret out concepts with potential errors are examined. One was shown to be good in dealing with synonym problems, and the other two with hierarchical and attribute relationship problems.

CHAPTER 7

CONCLUSION

The usage of large standard terminologies like SCT is highly influenced by quality assurance issues. Past studies have identified instances of inconsistent modeling in SCT which could act as a barrier for the successful use of SCT in EHRs. An intensive auditing effort is needed to improve the quality of SCT concepts. However, an audit of all concepts of SCT requires extensive quality assurance resources and will require a long time. A desired approach in coping with this urgent quality assurance need is to develop techniques for identifying subsets of SCT with expected higher concentration of errors.

This dissertation presents one such approach which analyzed the conceptual representation of sets of concepts that are lexically similar at the term-level in an attempt to characterize the consistency of the modeling across these concepts. Similarity sets were introduced and a sample of 60 such sets was audited by an experienced auditor. As many as 30% of the sample sets were found with inconsistent modeling of concepts. The dissertation then presented a way to utilize three structural indicators to improve the efficiency of the similarity sets. These structural indicators included the number of parents, relationships and groups between the concepts of a similarity set. The method was proven to be effective with up to 70% of the audited sample sets found to be inconsistent.

Since the idea of group auditing is to present the auditor with a small set of concepts with high likelihood of inconsistencies, it is important to improve the likelihood of finding inconsistent concepts in similarity sets. A study was conducted along this line and positional similarity sets, which are similarity sets with strictness imposed on the

location of the differing word in concept FSN, were introduced. The use of such sets improved the likelihood of finding inconsistent concepts to approximately 22% as compared to 13% with general similarity sets. Moreover, the efficiency of positional similarity sets was enhanced by introducing the same three structural indicators as above. The use of such indicators increased the likelihood of finding inconsistent concepts to 42%.

Furthermore, a study was conducted to algorithmically suggest attributes to enhance the modeling of SCT concepts without an auditor having to manually identify them. The technique was based on the framework of the positional similarity sets. The results showed the method to be effective with one or more suggestions attributes to be valid for 45 out of a sample of 50 concepts. The methodology suggested 103 attributes for these 50 concepts of which 67 were found to be correctly suggested.

The dissertation also presented a study conducted on SCT problem list concepts to examine their readiness for their intended meaningful use in EHRs. It was found that these concepts tend to suffer from the same problems as the concepts found throughout the general SNOMED CT content, just to a slightly lesser degree. Such problems include a high percentage of primitive concepts (likely to be missing relationships), and deficient and inconsistent modeling of relationships. The conclusion is that further QA efforts are needed for the problem lists' concepts. To support such efforts, two straightforward structural indicators in the form of the number of parents and words were shown to effectively ferret out concepts with high likelihood of inconsistencies. A structural indicator was also presented to deal with the synonymy problems by identifying pairs of concepts in SCT that map to the same UMLS concept.

REFERENCES

[1] SNOMED CT. Available at: http://www.ihtsdo.org/snomed-ct [accessed 23 October 2012]

[2] American Recovery and Reinvestment Act. 111 United States Congress, 2009. Available at: http://www.recovery.gov/About/Pages/The_Act.aspx [accessed 15 September 2012]

[3] Blumenthal D. Launching HITECH. New England Journal of Medicine. 2010;362:382-5.

[4] Office of the National Coordinator for Health Information Technology, Department of Health and Human Services. Health information technology: Initial set of standards, implementation specifications, and certification criteria for electronic health record technology. Final rule. Federal Register. 2010;75:44589-654.

[5] Medicare and Medicaid Programs; Electronic health record incentive program; final rule. In: Department of Health and Human Services, Centers for Medicare and Medicaid Services; 2010. p. 276.

[6] Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, Scoville R, et al. Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. Health Affairs. 2005;24:1103-17.

[7] Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. AMIA Annual Symposium Proceedings. 2003:699-703.

[8] Cornet R, de Keizer N. Forty years of SNOMED: a literature review. BMC Medical Informatics and Decision Making. 2008;8 Supplement 1:S2.

[9] Cimino JJ, Elhanan G, Zeng Q. Supporting infobuttons with terminological knowledge. AMIA Annual Symposium Proceedings. 1997:528-32.

[10] Elhanan G, Cimino JJ. Controlled vocabulary and design of laboratory results displays. AMIA Annual Symposium Proceedings. 1997:719-23.

[11] Elhanan G, Socratous SA, Cimino JJ. Integrating DXplain into a clinical information system using the World Wide Web. AMIA Annual Symposium Proceedings. 1996:348-52.

[12] Zeng Q, Cimino JJ. Automated knowledge extraction from the UMLS. AMIA Annual Symposium Proceedings. 1998:568-72.

[13] Zeng Q, Cimino JJ. A knowledge-based, concept-oriented view generation system for clinical data. Journal of Biomedical Informatics. 2001;34:112-28.

[14] Zeng Q, Cimino JJ, Zou KH. Providing concept-oriented views for clinical data using a knowledge-based system: an evaluation. Journal of American Medical Informatics Association. 2002;9:294-305.

[15] Cimino JJ. From data to knowledge through concept-oriented terminologies: experience with the Medical Entities Dictionary. Journal of American Medical Informatics Association. 2000;7:288-97.

[16] Mendonca EA, Cimino JJ, Johnson SB, Seol YH. Accessing heterogeneous sources of evidence to answer clinical questions. Journal of Biomedical Informatics. 2001;34:85-98.

[17] The CORE Problem List Subset of SNOMED CT. Available at: http://www.nlm.nih.gov/research/umls/Snomed/core_subset.html [accessed 15 September 2012]

[18] UMLS enhanced VA/KP Problem List Subset of SNOMED. Available at: http://www.nlm.nih.gov/research/umls/licensedcontent/vakpproblemlist.html [accessed 15 September 2012]

[19] Agrawal A, Elhanan G, Halper M. Dissimilarities in the Logical Modeling of Apparently Similar Concepts in SNOMED CT. AMIA Annual Symposium Proceedings. 2010;2010:212-6.

[20] Agrawal A, Perl Y, Elhanan G. Identifying problematic concepts in SNOMED CT using a lexical approach. To appear in Studies in Health Technology and Informatics. 2013.

[21] Agrawal A, He Z, Perl Y, Wei D, Halper M, Elhanan G, et al. The readiness of SNOMED problem list concepts for meaningful use of electronic health records. Artificial Intelligence in Medicine. 2013;58:73-80.

[22] Agrawal A, Perl Y, Chen Y, Elhanan G, Liu M. Identifying inconsistencies in SNOMED CT problem lists using structural indicators. To appear in AMIA Annual Symposium Proceedings. 2013.

[23] Elkin PL, Brown SH, Husser CS, Bauer BA, Wahner-Roedler D, Rosenbloom ST, et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clinic Proceedings. 2006;81:741-8.

[24] Ciolko E, Lu F, Joshi A. Intelligent clinical decision support systems based on SNOMED CT. Conference Proceedings of the IEEE Engineering in Medicine and Biology Society. 2010;2010:6781-4.

[25] Elevitch FR. SNOMED CT: electronic health record enhances anesthesia patient safety. Journal of American Association of Nurse Anesthetists. 2005;73:361-6.

[26] Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in Health Technology and Informatics. 2006;121:279-90.

[27] Farfán Sedano FJ, Terrón Cuadrado M, García Rebolledo EM, Castellanos Clemente Y, Serrano Balazote P, Gómez Delgado A. Implementation of SNOMED CT to the medicines database of a general hospital. Studies in Health Technology and Informatics. 2009;148:123-30.

[28] Sommers SC. Systematized nomenclature of pathology. Pathologia et microbiologia. 1967;30:826-7.

[29] Wells AH. Systematized nomenclature of pathology. Conversion to the computer language of medicine. Minnesota Medicine. 1972;55:585-90.

[30] Spackman KA, Campbell KE, Côté RA. SNOMED RT: a reference terminology for health care. AMIA Annual Symposium Proceedings. 1997:640-4.

[31] O'Neil M, Payne C, Read J. Read Codes Version 3: a user led terminology. Methods of Information in Medicine. 1995;34:187-92.

[32] History of SNOMED CT. Available at: http://www.ihtsdo.org/snomed-ct/history0 [accessed 15 October 2012]

[33] IHTSDO SNOMED CT technical reference guide. Available at: http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/SNOMED_CT_Publications/SNOMED_CT_Technical_Reference_Guide_20090131.pdf [accessed 23 October 2012]

[34] Campbell JR, Carpenter P, Sneiderman C, Cohn S, Chute CG, Warren J. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity. CPRI Work Group on Codes and Structures. Journal of American Medical Informatics Association. 1997;4:238-51.

[35] Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. Journal of American Medical Informatics Association. 1997;4:484-500.

[36] Penz JF, Brown SH, Carter JS, Elkin PL, Nguyen VN, Sims SA, et al. Evaluation of SNOMED coverage of Veterans Health Administration terms. Studies in Health Technology and Informatics. 2004;107:540-4.

[37] Elhanan G, Perl Y, Geller J. A survey of direct users and uses of SNOMED CT: 2010 status. AMIA Annual Symposium Proceedings. 2010:207-11.

[38] Elhanan G, Perl Y, Geller J. A survey of SNOMED CT direct users, 2010: impressions and preferences regarding content and quality. Journal of American Medical Informatics Association. 2011;18 Supplement 1:i36-44.

[39] Rosenbloom ST, Brown SH, Froehling D, Bauer BA, Wahner-Roedler DL, Gregg WM, et al. Using SNOMED CT to represent two interface terminologies. Journal of American Medical Informatics Association. 2009;16:81-8.

[40] Hoskins HD, Hildebrand PL, Lum F. The American Academy of Ophthalmology adopts SNOMED CT as its official clinical terminology. Ophthalmology. 2008;115:225-6.

[41] Shahpori R, Doig C. Systematized Nomenclature of Medicine-Clinical Terms direction and its implications on critical care. Journal of Critical Care. 2010;25:364 e1-9.

[42] James AG, Spackman KA. Representation of disorders of the newborn infant by SNOMED CT. Studies in Health Technology and Informatics. 2008;136:833-8.

[43] Kim H, Harris MR, Savova G, Chute CG. Content coverage of SNOMED-CT toward the ICU nursing flowsheets and the acuity indicators. Studies in Health Technology and Informatics. 2006;122:722-6.

[44] Simpson CR, Anandan C, Fischbacher C, Lefevre K, Sheikh A. Will Systematized Nomenclature of Medicine-Clinical Terms improve our understanding of the disease burden posed by allergic disorders? Clinical and Experimental Allergy. 2007;37:1586-93.

[45] Bakhshi-Raiez F, Cornet R, de Keizer NF. Cross-mapping APACHE IV "reasons for intensive care admission" classification to SNOMED CT. Studies in Health Technology and Informatics. 2008;136:779-84.

[46] Alecu I, Bousquet C, Jaulent MC. A case report: using SNOMED CT for grouping adverse drug reactions terms. BMC Medical Informatics and Decision Making. 2008;8 Supplement 1:S4.

[47] Shah NH, Rubin DL, Supekar KS, Musen MA. Ontology-based annotation and query of tissue microarray data. AMIA Annual Symposium Proceedings. 2006:709-13.

[48] Nadkarni PM, Marenco LA. Implementing description-logic rules for SNOMED-CT attributes through a table-driven approach. Journal of American Medical Informatics Association. 2010;17:182-4.

[49] Blumenthal D. Stimulating the adoption of health information technology. New England Journal of Medicine. 2009;360:1477-9.

[50] Fung KW, McDonald C, Srinivasan S. The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institutions. Journal of American Medical Informatics Association. 2010;17:675-80.

[51] Weed LL. The problem oriented record as a basic tool in medical education, patient care and clinical research. Annals of clinical research. 1971;3:131-4.

[52] Weed LL. Medical records that guide and teach. New England Journal of Medcine. 1968;278:652-7.

[53] Salmon P, Rappaport A, Bainbridge M, Hayes G, Williams J. Taking the problem oriented medical record forward. AMIA Annual Symposium Proceedings. 1996:463-7.

[54] Hayes GM. Computers in the consultation. The UK experience. Proceedings of the Annual Symposium on Computer Application in Medical Care. 1993:103-6.

[55] Wright A, Feblowitz J, McCoy AB, Sittig DF. Comparative analysis of the VA/Kaiser and NLM CORE problem subsets: an empirical study based on problem frequency. AMIA Annual Symposium Proceedings. 2011;2011:1532-40.

[56] Mantena S, Schadow G. Evaluation of the VA/KP problem list subset of SNOMED as a clinical terminology for electronic prescription clinical decision support. AMIA Annual Symposium Proceedings. 2007:498-502.

[57] Steindel SJ. A comparison between a SNOMED CT problem list and the ICD-10-CM/PCS HIPAA code sets. Perspectives in Health Information Management. 2012;9:1b.

[58] Fung KW, Xu J, Rosenbloom ST, Mohr D, Maram N, Suther T. Testing three problem list terminologies in a simulated data entry environment. AMIA Annual Symposium Proceedings. 2011;2011:445-54.

[59] Min H, Perl Y, Chen Y, Halper M, Geller J, Wang Y. Auditing as part of the terminology design life cycle. Journal of American Medical Informatics Association. 2006;13:676-90.

[60] Zhu X, Fan JW, Baorto DM, Weng C, Cimino JJ. A review of auditing methods applied to the content of controlled biomedical terminologies. Journal of Biomedical Informatics. 2009;42:413-25.

[61] Geller J, Perl Y, Halper M, Cornet R. Special issue on auditing of terminologies. Journal of Biomedical Informatics. 2009;42:407-11.

[62] Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in SNOMED CT: an exploration into large description logic-based biomedical terminologies. Artificial Intelligence in Medicine. 2007;39:183-95.

[63] Wei D, Bodenreider O. Using the abstraction network in complement to description logics for quality assurance in biomedical terminologies - a case study in SNOMED CT. Studies in Health Technology and Informatics. 2010;160:1070-4.

[64] Cornet R, Abu-Hanna A. Auditing description-logic-based medical terminological systems by detecting equivalent concept definitions. International Journal of Medical Informatics. 2008;77:336-45.

[65] Cornet R, Abu-Hanna A. Description logic-based methods for auditing frame-based medical terminological systems. Artificial Intelligence in Medicine. 2005;34:201-17.

[66] Cornet R, Abu-Hanna A. Two DL-based methods for auditing medical terminological systems. AMIA Annual Symposium Proceedings. 2005:166-70.

[67] Ceusters W, Smith B, Kumar A, Dhaen C. Mistakes in medical ontologies: where do they come from and how can they be detected? Studies in Health Technology and Informatics. 2004;102:145-63.

[68] Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED-CT. Studies in Health Technology and Informatics. 2004;107:482-6.

[69] Jiang G, Chute CG. Auditing the semantic completeness of SNOMED CT using formal concept analysis. Journal of American Medical Informatics Association. 2009;16:89-102.

[70] Zhang GQ, Bodenreider O. Large-scale, exhaustive lattice-based structural auditing of SNOMED CT. AMIA Annual Symposium Proceedings. 2010;2010:922-6.

[71] Wang J, Day R, Visweswaran S, Hogan W. The use of semantic distance metrics to support ontology audit. AMIA Annual Symposium Proceedings. 2010;2010:842-6.

[72] Ceusters W. Applying evolutionary terminology auditing to SNOMED CT. AMIA Annual Symposium Proceedings. 2010;2010:96-100.

[73] Bodenreider O, Burgun A, Rindflesch TC. Assessing the consistency of a biomedical terminology through lexical knowledge. International Journal of Medical Informatics. 2002;67:85-95.

[74] Campbell KE, Tuttle MS, Spackman KA. A "lexically-suggested logical closure" metric for medical terminology maturity. AMIA Annual Symposium Proceedings. 1998:785-9.

[75] Mendonça EA, Cimino JJ, Campbell KE, Spackman KA. Reproducibility of interpreting "and" and "or" in terminology systems. AMIA Annual Symposium Proceedings. 1998:790-4.

[76] Pacheco E, Stenzhorn H, Nohama P, Paetzold J, Schulz S. Detecting Underspecification in SNOMED CT concept definitions through natural language processing. AMIA Annual Symposium Proceedings. 2009;2009:492-6.

[77] Wang Y, Halper M, Min H, Perl Y, Chen Y, Spackman KA. Structural methodologies for auditing SNOMED. Journal of Biomedical Informatics. 2007;40:561-81.

[78] Wei D, Halper M, Elhanan G, Chen Y, Perl Y, Geller J, et al. Auditing SNOMED relationships using a converse abstraction network. AMIA Annual Symposium Proceedings. 2009:685-9.

[79] Wang Y, Halper M, Wei D, Perl Y, Geller J. Abstraction of complex concepts with a refined partial-area taxonomy of SNOMED. Journal of Biomedical Informatics. 2012;45:15-29.

[80] Spackman KA, Dionne R, Mays E, Weis J. Role grouping as an extension to the description logic of Ontylog, motivated by concept modeling in SNOMED. AMIA Annual Symposium Proceedings. 2002:712-6.

[81] Rector AL, Brandt S, Schneider T. Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications. Journal of American Medical Informatics Association. 2011;18:432-40.

[82] SNOMED CT User Guide July 2012 International Release (US English). Available at: http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_UserGuide_Current-en-US_INT_20120731.pdf [accessed 26 December 2012]

[83] Cimino JJ. Linking patient information systems to bibliographic resources. Methods of Information in Medicine. 1996;35:122-6.

[84] Curbside Consult with Dr. Jayne 12/17/12. Available at: http://histalk2.com/2012/12/17/curbside-consult-with-dr-jayne-121712 [accessed 9 January 2013]

[85] Jamoulle M Some views about SNOMED-CT by a general practitioner. Available at: http://docpatient.net/onto/doc/SNOMED_CT_study_MJ_2010.pdf [accessed 9 January 2013]

[86] IHTSDO concept model special interest group. Available at: https://thecap.basecamphq.com/projects/388747/posts/31226264/comments#6375 6401 [accessed 29 April 2012]

[87] Halper M, Wang Y, Min H, Chen Y, Hripcsak G, Perl Y, et al. Analysis of error concentrations in SNOMED. AMIA Annual Symposium Proceedings. 2007:314-8.

[88] Wang Y, Wei D, Xu J, Elhanan G, Perl Y, Halper M, et al. Auditing complex concepts in overlapping subsets of SNOMED. AMIA Annual Symposium Proceedings. 2008:273-7.

[89] Wang Y, Halper M, Wei D, Gu H, Perl Y, Xu J, et al. Auditing complex concepts of SNOMED using a refined hierarchical abstraction network. Journal of Biomedical Informatics. 2012;45:1-14.

[90] SNOMED CT developer toolkit guide. Available at: http://www.ihtsdo.org/fileadmin/user_upload/doc/download/doc_DeveloperToolkitGuide_Current-en-US_INT_20120731.pdf [accessed 13 Aug 2012]

[91] PubMed Stopwords. Available at: http://www.ncbi.nlm.nih.gov/books/NBK3827/?rendertype=table&id=pubmedhelp.T43 [accessed 13 Aug 2012]

[92] UMLS MetaMap. Available at: http://metamap.nlm.nih.gov [accessed 13 Aug 2012]

[93] Chen Y, Gu HH, Perl Y, Geller J. Structural group-based auditing of missing hierarchical relationships in UMLS. Journal of Biomedical Informatics. 2009;42:452-67.