Research Paper ■

Auditing the Semantic Completeness of SNOMED CT Using Formal Concept Analysis

Guoqian Jiang, PhD, Christopher G. Chute, MD, DrPH

Abstract **Objective:** This study sought to develop and evaluate an approach for auditing the semantic completeness of the SNOMED CT contents using a formal concept analysis (FCA)–based model.

Design: We developed a model for formalizing the normal forms of SNOMED CT expressions using FCA. Anonymous nodes, identified through the analyses, were retrieved from the model for evaluation. Two quasi-Poisson regression models were developed to test whether anonymous nodes can evaluate the semantic completeness of SNOMED CT contents (Model 1), and for testing whether such completeness differs between 2 clinical domains (Model 2). The data were randomly sampled from all the contexts that could be formed in the 2 largest domains: *Procedure* and *Clinical Finding*. Case studies (n = 4) were performed on randomly selected anonymous node samples for validation.

Measurements: In Model 1, the outcome variable is the number of fully defined concepts within a context, while the explanatory variables are the number of lattice nodes and the number of anonymous nodes. In Model 2, the outcome variable is the number of anonymous nodes and the explanatory variables are the number of lattice nodes and a binary category for domain (*Procedure/Clinical Finding*).

Results: A total of 5,450 contexts from the 2 domains were collected for analyses. Our findings revealed that the number of anonymous nodes had a significant negative correlation with the number of fully defined concepts within a context ($p < 0.001$). Further, the *Clinical Finding* domain had fewer anonymous nodes than the *Procedure* domain ($p < 0.001$). Case studies demonstrated that the anonymous nodes are an effective index for auditing SNOMED CT.

Conclusion: The anonymous nodes retrieved from FCA-based analyses are a candidate proxy for the semantic completeness of the SNOMED CT contents. Our novel FCA-based approach can be useful for auditing the semantic completeness of SNOMED CT contents, or any large ontology, within or across domains.

■ **J Am Med Inform Assoc.** 2009;16:89–102. DOI 10.1197/jamia.M2541.

Introduction

The structure of modern terminologies has advanced well beyond simple 1-dimensional subsumption relationships through the introduction of composite expressions (see The International Organization for Standardization formal definition[1]); these relationships have long been sought in controlled medical terminologies.[2–5] SNOMED-CT, the most comprehensive clinically oriented medical terminology system, now provides a platform where composite expressions have the potential to be used in clinical situations.[6] SNOMED CT adopted a description logic (DL) foundation[7,8] that has allowed its curators to formally represent concept meanings and relationships.[9]

Because compositional variants and precomposed concepts may introduce redundancy and hamper information retrieval, SNOMED CT defines a concept "normal form" or canonical representation as the maximal decomposition of concepts into a set of primitive defining supertypes. These normal forms are proposed as the formal representation of clinical meaning for SNOMED CT concepts to support authoring tasks, distribution, and other purposes, such as the comprehensive retrieval of precoordinated and postcoordinated SNOMED CT expressions from clinical records.[10,11]

Given the size, complexity, and sophistication of SNOMED, the need arises for automated and reliable means to algorithmically assess the completeness, correctness, consistency, and competency[12] of the vocabulary and its adherence to good terminology practices. We introduce one such method in this article, based on formal concept analysis (FCA).

Formal concept analysis is a generic structure of lattice-building algorithms, based on mathematical lattice theory, which permits visualizing partial or incomplete order in an information lattice and its consequences.[13,14] It provides a

Affiliations of the authors: Division of Biomedical Statistics and Informatics, Mayo Clinic College of Medicine, Rochester, MN.

Supported in part by R01 LM07319.

The authors thank Harold R. Solbrig, Thomas M. Johnson, and James D. Buntrock for their critical input and support. The authors also thank the reviewers for their insightful and constructive criticism, which materially improved this work.

Correspondence: Dr. Guoqian Jiang, Division of Biomedical Statistics and Informatics, Mayo Clinic College of Medicine, Rochester, MN 55905. e-mail: <jiang.guoqian@mayo.edu>.

Received for review 06/23/07; accepted for publication: 09/23/08.

candidate mechanism for developing completeness-evaluation algorithms. Our hypothesis is that reformulating the rules of composition and compositional transformations associated with SNOMED normal forms, in the language of lattice theory, may provide a novel approach for auditing large terminologies such as SNOMED CT. The objective of this study is to develop and evaluate an approach for auditing the semantic completeness of the SNOMED CT contents using a FCA-based model.

Background

Auditing the Medical Terminological Systems

A number of generalized approaches have been designed for auditing medical terminologies.[12,15] Here we review those approaches that have mainly been applied to SNOMED CT auditing.

Spackman[16] summarized the overall size of the SNOMED CT and its rates of change over a period of 3 years, and found that awareness of the rate of change in the terminology can help both terminology and application developers. However, simply invoking DL formalisms for terminology development may not prevent incorrect or incomplete representations in a medical terminology or ensure compliance with all principles of a sound ontology.[17] Indeed, Bodenreider et al.[18] pointed out that the descriptions of many concepts in SNOMED CT are minimal or incomplete, with possible "detrimental consequences on inheritance." Investigators have suggested that using more complete logical and ontological practices would prevent certain families of modeling mistakes and improve the quality of a large terminology such as SNOMED CT.[18,19]

Halper et al.[20] developed an ad-hoc approach for analysis of error concentrations in SNOMED CT through investigating the area taxonomy, derived from a partition of SNOMED's concepts based on their respective sets of relationships. Spackman et al.[21] examined SNOMED CT from the perspective of formal ontological principles. While they showed the usefulness of formal ontological principles for improving consistency of design decisions of SNOMED CT, they argued that "the applicability of some of the formal ontology principles in providing consistent guidance in the very large areas like clinical finding and procedure is still not as clear, and appears to require further elaboration and study."

In addition, several studies developed auditing approaches by applying SNOMED CT to actual problems and use cases. For example, Green et al.[22] developed and evaluated a method for the structured representation of heart murmur findings using SNOMED CT postcoordination. Similarly, Richesson et al.[23] used SNOMED CT to represent consistently clinical research data based on the semantic characterization of data items on Case Report Forms.

Semantic Completeness

Semantic completeness is a property of any logical system, including well-formed terminologies, that invokes 3 components: (1) a procedure for constructing formulas that will yield propositions when interpreted using specific rules (well-formed formulas), (2) a definition of truth that relates to interpretations and models of logical systems (expressive completeness), and (3) a proof procedure that allows new well-formed formulas to be derived from old ones (deductive completeness).[24] A logical system is logically complete if every true, well-formed formula can be derived.

Applying the notion of completeness to ontologies, we find that it can be interpreted in a number of ways. For example, Fox et al.[25] defined the "functional completeness" of ontology as its ability to represent the information necessary for a function to perform its task, i.e., the completeness of an ontology is determined by its competency. Over a decade ago, Devanbu and Jones[26] asserted that a terminology system should satisfy 4 requirements: completeness, correctness, consistency, and competency. Completeness as indicated by Devanbu indicates that a terminology system should have the knowledge necessary to represent a domain. They used competency to indicate that the system should have efficient algorithms to perform the inferences needed for the application. Given that medical concept representation in modern medical terminologies may be regarded from the perspectives of both breadth of coverage and depth of representation,[27] we consider that the completeness of a medical terminology system should include 2 parts: complete content coverage (*coverage completeness*) and complete semantics (*semantic completeness*), which supports its competency. The former has historically been well addressed,[28–32] whereas few studies address the latter. In this study, we focus on the semantic completeness of the SNOMED CT contents.

Relevant Constructs of SNOMED CT

Precoordination and Postcoordination

Precoordination is the use of composite expressions of coded concepts within a terminology to define a new coded concept. For example, "hand joint pain" is a composite expression defined by [[is a = joint pain] and [finding site = hand joint structure]], manifest as a precoordinated expression in SNOMED CT. The ability to define composite expressions within a terminology opens a whole new realm of expressive possibility. The same techniques used to create precoordinated expressions can also be used to describe new external classes that have no equivalent concept code within the terminology. The creation of new classes externally is called *postcoordination*. Given that it is neither practical nor desirable to define precoordinated terms for every conceivable clinical situation, postcoordination becomes one promising solution to the problem of clinical content completeness.[33,34] For example, for all possible types and severity of fractures of all possible bones and their subdivisions,[34] or for all potential points on the chest that can be considered for point of maximum intensity and area of radiation of a heart murmur[22]; the enumerated possibilities may be finite but are awkwardly large.

Primitive and Fully Defined Concepts

Due to the fact that composite expressions are built using concept codes and the basic fact that not all classes are amenable to formal definition, some of the concept codes in any terminology will remain implicitly defined. Concept codes that fall into this category are referred to as *primitive*. SNOMED CT concepts are either primitive or fully defined. In the language of description logics, the asserted conditions of a primitive concept are necessary but not sufficient, and the asserted conditions of a fully defined concept are both necessary and sufficient. All members of the set of sufficient

conditions are also necessary conditions. For example, "gastric ulcer" is a fully defined concept given that the definition [[associated morphology = ulcer] and [finding site = stomach structure]] is asserted as both necessary and sufficient, whereas the aforementioned "hand joint pain" is a primitive concept given that its definition is asserted as necessary but not sufficient. We assert that the higher the proportion of fully defined concepts within a domain, the more complete are the semantics of that domain. We use fully defined concepts as an independent index to represent the semantic completeness of SNOMED CT in this study.

Normal (or Canonical) Forms

As introduced in the technical documentation of SNOMED CT,[11] a normal form is a view that can be generated by maximally decomposing any valid expression by applying a set of logical transformation rules. The purpose of generating normal forms is to facilitate complete and accurate retrieval of precoordinated and postcoordinated SNOMED CT expressions from clinical records or other resources. Two alternative normal forms are proposed: the long normal form and the short normal form. Both normal form transformation algorithms are described in the technical document. For example, the long normal form for "hypophysectomy" is shown in Figure 1.

Basic Notions of FCA

Many published articles and books describe the features of FCA in detail.[13,35,36] Here we briefly introduce some basic notions and features to help explain the modeling process in the next section. A (1-valued) *formal context* is defined as a triple comprising a set of *formal objects*, a set of *formal attributes*, and binary relations expressing which attributes describe each object. Usually, a formal context can be represented by a cross table. In many use cases, we may find that the relations between the objects and the attributes are a set of values rather than binary relations. Thus a *many-valued formal context* could also be expressed in a cross table. However, for the FCA application, many-valued formal contexts are transformed into a 1-valued context by *transformational scaling*.[13,35]

Graphically, a formal context could be visualized by a line diagram of a *concept lattice*. A concept lattice consists of the set of *formal concepts* of a formal context and the subconcept–superconcept relations between the formal concepts. Each node in a concept lattice represents a formal concept for which the meaning is interpreted by a set of formal objects (*extension*) and a set of formal attributes (*intension*). In other words, the extension covers all objects belonging to this concept and its child nodes, while the intension comprises all defining attributes for this concept and its parent nodes. The labels for each node are usually displayed on the lattice; the FCA literature refers to these labels as *own objects* and *own attributes*, respectively. Retrieving the extension and the intension of a node (i.e., a formal concept) from a concept lattice is achieved by a trace down or trace up using well-specific rules. A node without a label for its own object in a concept lattice is called an *anonymous node.*

[52699005 | hypophysectomy]
[11668003 | isa] = [71388002 | procedure]
{[260686004 | method] = [129304002 | excision-action],
[363704007 | procedure site] = [56329008 | pituitary structure]}

Figure 1. A long normal form for "hypophysectomy." The semantics of the long normal form may be interpreted as that hypophysectomy is a subtype of procedure which is defined by the conditions "method = excision-action" and "procedure site = pituitary structure." The square brackets indicate pair of a concept identifier with its preferred name (separated by bar "|"). The curly brackets indicate the conditions used for defining "procedure."

The notion of anonymous nodes is not specific to our FCA approach. For example, the Generalized Architecture for Languages, Encyclopedias, and Nomenclatures in medicine (GALEN) common reference model, developed by the University of Manchester, does not enumerate all sanctioned variants, e.g., it does not pre-enumerate all possible left-handed and right-handed variants of anatomical structures. Instead, GALEN defines anonymous concepts by expressions such as (SolidStructure which <isPairedOrUnpaired leftRightPaired>) (representing a bilateral solid structure).[37–39] Similarly, the Web Ontology Language (OWL) supports the representation of an anonymous class (i.e., anonymous concept), which may be or may not be an equivalent class of a named class.[40] However, we argue that the FCA-based approach is DL-language independent, and it encodes the problem of multiple relations in the definition of (many-valued) multicontexts and allows the transformation of a multicontext into a meaningful structure of concept lattices.[41] Thus FCA provides a flexible and scalable way to capture the semantic completeness (represented by anonymous nodes) of modern terminologies.

Formal concept analysis has been advocated as a mechanism to represent and process context knowledge in domains such as the description of patient cases, interpretation of therapeutic decisions, and the representation of rules.[42] It provides a means to represent the semantics underlying the meaning of a concept definition[35] and has been applied to many knowledge representation areas, such as ontology building,[43,44] ontology mapping and merging,[45,46] lexical databases, and taxonomy modeling.[47,48]

Methods

An FCA-based Model for the Normal Forms of the SNOMED CT Expressions

Expanding the Semantic Space for a Specific Domain

To do retrieval or analysis of stored clinical data, the stored expressions and the query expression would both be compared in their normal form when evaluating equivalence or subsumption.[10] However, in many real use cases, we found that the attributes (or slots) of a normal form for a specific expression were not sufficient to meet the requirements of postcoordination in a clinical statement. For instance, consider the composition expression "[*hypophysectomy*] + [*approach = transfrontal approach*]" retrieved from a clinical record. When converting this expression into normal form, we find that the normal form of "hypophysectomy" only contains 3 attributes—"isa", "procedure site", and "method;" there is no slot for "approach" (Fig. 1).

The absence of an expected normal form to support information retrieval and other use cases suggests that all subconcepts within a domain should be "semantically completed"; specifically, the domain should be expanded

or populated with the additional concepts necessary to create the normal forms needed for retrieval. For our purposes in this report, we regard any SNOMED CT concept that contains subconcepts as a candidate domain. Still using "hypophysectomy" as an example, when we retrieve the normal forms of its 20 subclasses we find that there are 7 attributes used in these normal forms (Table 1). For convenience, according the children of "hypophysectomy" the status of a domain, these attributes could be shared to describe any postcoordinated expressions retrieved from clinical statements within that domain. Using the expanded semantic space, we find that this domain does contain a slot "approach" that could be used for converting the composition expression "[hypophysectomy] + [approach = transfrontal approach]" into a normal form.

Modeling the Normal Forms Using the FCA

Invoking the description-logic structures of SNOMED CT, the attribute name-value model was applied to describe the composite expression in SNOMED CT. The name part of the attribute name-value pair is a conceptId that refers to a concept and the value part of the attribute name-value pair is an expression.[11] A normal form, in fact, is a decomposed structure for a nested expression using a set of specific transformation rules. Table 1 shows the long normal forms of all concepts in the domain "hypophysectomy" defined in the SNOMED CT.

In the language of the FCA, the data in Table 1 can be interpreted as a formal context, which actually is referred to as a many-valued formal context. Table 1 may be understood as a structure that contains a set of objects (whose names are heading of rows, i.e., the concept names of a domain), a set of attributes (whose names are heading of columns, i.e., the name part of the attribute name-value pairs), and a set containing all attribute values described by the entries in the table cells (i.e., the value part of the attribute name-value pairs).

For the FCA application, the many-valued formal context can be transformed to a 1-valued context, or Boolean form, by transformation scaling. Here, we take a 2-step approach for the transformation. The first step, called plain scaling, substitutes each attribute in the original many-valued context with a set of columns representing each one of the allowed values for the attribute. This corresponds to the notion that is called reification in the DL community, i.e., transforming one relation and its object value into a relationship.[49] Still using the data in Table 1 as our example, we obtain 22 Boolean columns containing the same information that substitutes for the 7 original attributes or column headings in Table 1 (Table 2). The Xs in Table 2 indicate when an object has a defined attribute value. We concatenated the attribute value and the original attribute name to synthesize a readable column name, e.g., "Procedure(isA)".

The second step in transformation scaling is to complete the context using hierarchical knowledge within SNOMED CT. For example, in the column "Procedure site" of Table 1, there are 3 different values, including "Pituitary structure," "Pituitary part," and "Entire pituitary gland." When we retrieve the relationships (including both direct and indirect relationships using the transitive closure of "isa" relationship) among them, we may find that "Pituitary part" and "Entire pituitary gland" are subconcepts of "Pituitary structure." Thus, for the transformation to a 1-valued context, those objects having the values "Pituitary part" and "Entire pituitary gland," besides being X'ed for "Pituitary part(Procedure site)" and "Entire pituitary gland(Procedure site)" as taken in the first step, are also X'ed for "Pituitary structure(Procedure site)." For each column of Table 1, the same transformation using the transitive closure is applied to partially complete the context. In addition, the relationships among all the column headings are also retrieved. For Table 1, we may find that "Procedure site - Direct" and "Procedure site - Indirect" are the subproperties of "Procedure site." Thus, for the transformation to 1-valued context, those objects having the attributes "Procedure site - Direct" and "Procedure site - Indirect" are included for completing the context for the attribute "Procedure site." In Table 1, only 1 concept, "Excision of lesion of pituitary gland," has 2 attributes defined. Besides being X'ed for "Pituitary structure(Procedure site - Direct)" and "Pituitary structure(Procedure site - Indirect)," the concept is also X'ed for "Pituitary structure(Procedure site)" because the former 2 properties are both subproperties of "Pituitary structure(Procedure site)." These inferred Xs that derive from completing the context using hierarchical associations in the base terminologies are shaded in Table 2 for exposition, although they are not further distinguished in FCA analyses.

Visualizing the Modeled Domain Using Concept Lattice

Besides the cross table representation, there is a graphical representation of formal contexts using the line-diagram form for concept lattice. Figure 2 shows a line diagram of the concept lattice for the context given in Table 2. The lattice contains exactly the same information as the cross table. Each node in the diagram represents a formal concept of the context and the ascending paths of line edge between the 2 nodes represent the subconcept and superconcept relations. For the readability of the lattice, we only display the labels for objects (i.e., the concept names of the domain "hypophysectomy" defined in the SNOMED CT) in Figure 2.

Retrieving the Information about Anonymous Nodes

Inspection of Figure 2 reveals 5 nodes without a label attached, indicated by arrows. These nodes are the anonymous nodes. We propose that the anonymous nodes provide interesting and important information about the semantic (in)completeness of SNOMED CT contents. Table 3 shows the information about the anonymous nodes retrieved from the concept lattice given in Figure 2; note that the column Own Object is populated by None for every row in Table 3, which indicates that these nodes are anonymous by definition. For example, consider Node 2, which includes 3 extension objects. When we analyze the 3 objects, we find that they share the common attribute "Transfrontal approach(Approach)," indicated by the attribute label in this anonymous node. This suggests that an object label "Transfrontal hypophysectomy" is missing, which would cause a kind of semantic incompleteness of the domain contents. By analyzing the other 4 anonymous nodes, it is not difficult to conclude that they share common attributes, such as "Total excision," "Partial excision," "Excision biopsy(Method)," etc., some of which are indicated by the attribute labels and some of which are missing from both their object labels and attribute labels.

Table 1 ■ The Normal Forms of All Concepts in the Domain "Hypophysectomy"

Expression	isA	Procedure Site	Procedure Site, Indirect	Procedure Site, Direct	Direct Morphology	Method	Approach
Hypophysectomy	Procedure	Pituitary structure	–	–	–	Excision - action	–
Excision of lesion of pituitary gland	Procedure	–	Pituitary structure	Pituitary structure	Morphologically abnormal structure	Excision - action | Surgical action	–
Excision of pituitary gland NOS	Excision of pituitary gland NOS	Pituitary structure	–	–	–	Excision - action	–
Excisional biopsy of hypophysis	Excisional biopsy of hypophysis	Pituitary structure	–	–	–	Excision biopsy	–
Excisional biopsy of pituitary gland by transfrontal approach	Procedure	Pituitary structure	–	–	–	Excision biopsy	Transfrontal approach
Excisional biopsy of pituitary gland by transsphenoidal approach	Procedure	Pituitary structure	–	–	–	Excision biopsy	Transsphenoidal approach
Hypophysectomy NEC	Hypophysectomy NEC	Pituitary structure	–	–	–	Excision - action	–
Other specified excision of pituitary gland	Other specified excision of pituitary gland	Pituitary structure	–	–	–	Excision - action	–
Partial excision of pituitary gland by transfrontal approach	Procedure	Pituitary part	–	–	–	Excision - action	Transfrontal approach
Partial excision of pituitary gland by transsphenoidal approach	Procedure	Pituitary part	–	–	–	Excision - action	Transsphenoidal approach
Partial hypophysectomy	Procedure	Pituitary part	–	–	–	Excision - action	–
Removal of normal pituitary gland	Removal of normal pituitary gland	Pituitary structure	–	–	–	Excision - action	–
Selective transsphenoidal pituitary adenomectomy	Selective transsphenoidal pituitary adenomectomy	Pituitary structure	–	–	–	Excision - action	Transsphenoidal approach
Sublabial hypophysectomy	Procedure	Pituitary structure	–	–	–	Excision - action	Transsphenoidal approach | Sublabial approach
Total excision of pituitary gland by transfrontal approach	Procedure	Entire pituitary gland	–	–	–	Excision - action	Transfrontal approach
Total excision of pituitary gland by transsphenoidal approach	Procedure	Entire pituitary gland	–	–	–	Excision - action	Transsphenoidal approach
Total hypophysectomy	Procedure	Entire pituitary gland	–	–	–	Excision - action	–
Transcranial hypophysectomy	Procedure	Pituitary structure	–	–	–	Excision - action	Transcranial approach
Transethmoidal hypophysectomy	Procedure	Pituitary structure	–	–	–	Excision - action	Transsphenoidal approach | Transethmoidal approach
Transseptal hypophysectomy	Transseptal hypophysectomy	Pituitary structure	–	–	–	Excision - action	Transsphenoidal approach
Transsphenoidal hypophysectomy	Procedure	Pituitary structure	–	–	–	Excision - action	Transsphenoidal approach

NEC = not elsewhere classified; NOS = not otherwise specified.

Table 2 ■ A Completed Formal Context of the Domain "Hypophysectomy"

	Procedure(isA)	Hypophysectomy NEC(isA)	Removal of normal pituitary gland(isA)	Transseptal hypophysectomy(isA)	Excision of pituitary gland NOS(isA)	Excisional biopsy of hypophysis(isA)	Selective transsphenoidal pituitary adenomectomy(isA)	Other specified excision of pituitary gland(isA)	Pituitary structure(Procedure site)	Pituitary part(Procedure site)	Entire pituitary gland(Procedure site)	Pituitary structure(Procedure site–Direct)	Pituitary structure(Procedure site–Indirect)	Morphologically abnormal structure (Direct morphology)	Excision - action(Method)	Surgical action(Method)	Excision biopsy(Method)	Transfrontal approach(Approach)	Transcranial approach(Approach)	Transsphenoidal approach(Approach)	Sublabial approach(Approach)	Transethmoidal approach(Approach)
Hypophysectomy	X								X						X	X						
Excision of lesion of pituitary gland	X								X			X	X	X	X	X						
Excision of pituitary gland NOS	X				X				X						X	X						
Excisional biopsy of hypophysis	X					X			X						X	X	X					
Excisional biopsy of pituitary gland by transfrontal approach	X								X						X	X	X	X	X			
Excisional biopsy of pituitary gland by transsphenoidal approach	X								X						X	X	X		X	X		
Hypophysectomy NEC	X	X							X						X	X						
Other specified excision of pituitary gland	X							X	X						X	X						
Partial excision of pituitary gland by transfrontal approach	X								X	X					X	X		X	X			
Partial excision of pituitary gland by transsphenoidal approach	X								X	X					X	X			X	X		
Partial hypophysectomy	X								X	X					X	X						
Removal of normal pituitary gland	X		X						X						X	X						
Selective transsphenoidal pituitary adenomectomy	X						X		X						X	X			X	X		
Sublabial hypophysectomy	X								X						X	X			X	X	X	
Total excision of pituitary gland by transfrontal approach	X								X		X				X	X		X	X			
Total excision of pituitary gland by transsphenoidal approach	X								X		X				X	X			X	X		
Total hypophysectomy	X								X		X				X	X						
Transcranial hypophysectomy	X								X						X	X			X			
Transethmoidal hypophysectomy	X								X						X	X			X	X		X
Transseptal hypophysectomy	X			X					X						X	X			X	X		
Transsphenoidal hypophysectomy	X								X						X	X			X	X		

Xs in shade indicate those relations completed by the hierarchical knowledge of the SNOMED CT.
Abbreviations as in Table 1.

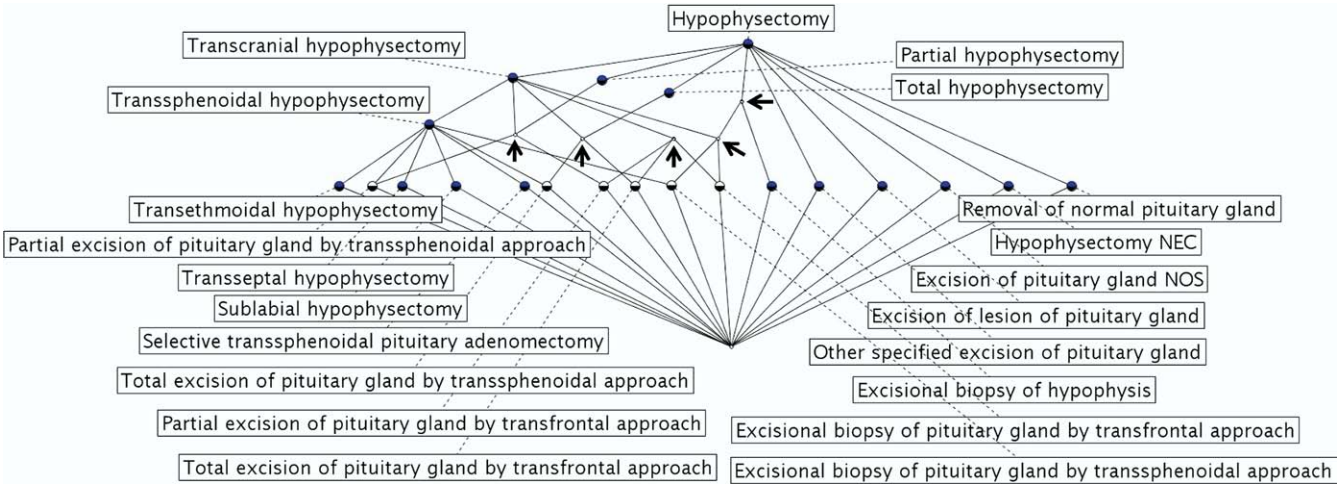

Figure 2. A line diagram of concept lattice for the domain "hypophysectomy." The arrows indicate those nodes that were called "anonymous node."

A Protégé plug-in on FCA-based ontology visualization is available at: http://informatics.mayo.edu/LexGrid/index.php?page= fca.

An Evaluation of the FCA-based Approach

We performed 2 modalities of evaluation: (1) quantitative and (2) inspection and interpretation of results.

Evaluation by Statistical Model

Our hypothesis in this study is that the anonymous nodes retrieved from the FCA model would be an index of the semantic (in)completeness of the SNOMED CT contents. Therefore, the research questions here are: (1) Can the number of anonymous nodes characterize the semantic completeness of SNOMED CT domains? (2) Can our approach using the number of anonymous nodes audit the difference of semantic completeness across domains?

We used the January 2006 version of SNOMED CT files as provided directly by CAP (College of American Pathologists, Northfield, IL. http://www.cap.org), the original producer of the nomenclature. We implemented a method for transforming SNOMED CT expressions into normal forms, consistent with the algorithms described in SNOMED CT technical document.[11] In this study, we focused on the 2 largest domains of SNOMED CT: *Clinical Finding* (SCTID_404684003) and *Procedure* (SCTID_71388002). Through a stratified sampling, 2 sets of sample contexts were collected randomly from the direct subbranches of the 2 domains. Table 4 shows the total number of subconcepts in that version of SNOMED CT, the total number of contexts that could be formed, and the number of contexts (the sample size) that we randomly selected for each subbranch. Because computing FCA context is computationally intensive, we randomly selected about 10% of the total number of contexts for each subbranch.

Because a concept in SNOMED CT may have multiple superconcepts, and a concept could be selected more than 1 time from a different subbranch of a domain in the sampling process, we removed these repeated sample contexts. In addition, many primitive SNOMED CT concepts have very minimal definitions (e.g., limited to 1 isa relation), so we also

Table 3 ■ The Information about the Anonymous Nodes Retrieved from the Concept Lattice of the Domain "Hypophysectomy"

Anonymous Node	Own Object	Own Attribute	Extension
Node 1	None	None	1. Total excision of pituitary gland by transsphenoidal approach 2. Total excision of pituitary gland by transfrontal approach
Node 2	None	Transfrontal approach(Approach)	1. Total excision of pituitary gland by transfrontal approach 2. Partial excision of pituitary gland by transfrontal approach 3. Excisional biopsy of pituitary gland by transfrontal approach
Node 3	None	None	1. Partial excision of pituitary gland by transsphenoidal approach 2. Partial excision of pituitary gland by transfrontal approach
Node 4	None	None	1. Excisional biopsy of pituitary gland by transsphenoidal approach 2. Excisional biopsy of pituitary gland by transfrontal approach
Node 5	None	Excision biopsy(Method)	1. Excisional biopsy of pituitary gland by transsphenoidal approach 2. Excisional biopsy of pituitary gland by transfrontal approach 3. Excisional biopsy of hypophysis

"None" indicates that no label was found for own object(s) or own attribute(s) of an anonymous node.

Table 4 ■ The Stratified Sampling for 2 Largest Domains: *Clinical Finding* and *Procedure*

SCTID	Concept Name	SubCls Num	Context Num	Sample Size
404684003	Clinical finding			
64572001	Disease (disorder)	74769	15573	1557
118234003	Finding by site	67355	13635	1527
250171008	Clinical history and observation findings	20314	3138	314
118240005	Finding by method	8397	1013	101
225552003	Wound finding	4791	1061	106
102957003	Neurological finding	4170	703	70
307824009	Administrative statuses	2396	278	28
417893002	Deformity	888	163	20
419026008	Effect of exposure to physical force	628	87	20
365860008	General clinical state finding	613	70	20
267038008	Edema	370	57	20
418799008	Finding reported by subject or history provider	330	41	20
127357005	Finding related to physiologic substance	88	3	3
207577006	[X]Additional symptom, signs and abnormal clinical and laboratory findings classification terms	86	13	13
285153007	Sequelae of external causes and disorders	62	13	13
384740007	Finding of grade	53	11	11
80631005	Clinical stage finding	28	5	5
217020008	Medical and surgical procedures as the cause of abnormal reaction of patient or later complication, without mention of misadventure at the time of procedure	16	5	5
69449002	Drug action	11	0	0
365858006	Prognosis/outlook finding	6	0	0
405533003	Adverse incident outcome categories	5	0	0
Subtotal		185376	35869	3853
71388002	Procedure			
128927009	Procedure by method	38404	7781	778
362958002	Procedure by site	33848	7298	730
363691001	Procedure by device	6647	1057	106
108252007	Laboratory procedure	9531	946	95
362961001	Procedure by intent	4157	680	68
243120004	Regimes and therapies (regime/therapy)	3034	436	44
127777001	Provider-specific procedure	1981	321	32
14734007	Administrative procedure	2446	308	31
408767007	Procedure with a clinical finding focus	1223	154	20
373311009	Procedure by approach	1099	135	20
399248000	Procedure related to anesthesia and sedation	666	114	20
408766003	Procedure with a procedure focus	1296	114	20
410533009	Procedure by priority	216	22	22
389067005	Community health procedure	24	4	4
225288009	Environmental care procedure	20	5	5
266705004	Preoperative/postoperative procedures	12	4	4
389084004	Staff related procedure	11	3	3
371883000	Outpatient procedure	0	0	0
373111004	Procedure in coronary care unit	0	0	0
410606002	Social service procedure	6	0	0
7922000	General treatment	0	0	0
Subtotal		104621	19382	2002
Total		289997	55251	5855

Context Num = number of the contexts that could be formed in each subbranch; Sample Size = number of contexts randomly selected; SubCls Num = number of subconcepts in each subbranch.

removed those contexts only having the isa relations as the formal attributes.

For each sample context, we transformed the data into 1-valued contexts and completed the context—i.e., we remodeled the data using the FCA methods above. For each context in the resultant matrix and concept lattice, we counted the number of fully defined concepts within the objects (*definedObjNum*), the number of lattice nodes (*latticeNodeNum*), and the number of anonymous nodes (*anonymousNodeNum*), and recorded its domain type (*domain*: Procedure/Clinical Finding). For example, consider the context of the domain "hypophysectomy" given in Figure 2 in the sampling process. The set of data related with this context is *definedObjNum* = 14, *latticeNodeNum* = 27, *anonymousNodeNum* = 5, *domain* = "Procedure."

Two quasi-Poisson regression models were developed to answer our research questions.[43] One is to test whether the anonymous nodes can explain the semantic complete-

ness of the SNOMED CT contents (Model 1), and the other is to test whether semantic completeness differs between domains (Model 2). Poisson regression assumes that a process or outcome occurs infrequently following the Poisson distribution, determined by a dependent variable (x) and a response variable (Y) which has an expected value of 1: log(E(Y)) = a+bx; it is well suited to low-frequency count data, which is the nature of the data in this study. Quasi-Poisson regression differs from Poisson in that its expected value need not be 1; or frequency counts on average are a bit larger than 1. We used a Poisson model for anonymous nodes because they are rare (non-Gaussian); we used the quasi-Poisson variant because our observed occurrences were not singular (i.e., did not have an expected value of 1). Furthermore, we used Poisson regression techniques rather than simple proportion comparisons so that we could adjust for confounding factors, such as the number of nodes in a sublattice, or compute different point estimates of effect across domains, i.e., so that we could build a multivariate model.

We created 2 regression models for our published analyses. In Model 1, the outcome variable was *definedObjNum*, which is used here as a proxy to represent the semantic completeness of the SNOMED CT contents, with the explanatory variables being log-transformation *latticeNodeNum* (to dampen skewing) and *anonymousNodeNum*. In Model 2, the outcome variable is *anonymousNodeNum* and the explanatory variables are log-transformed l*atticeNodeNum* and *domain* (binary: *Procedure*/*Clinical Finding*). The log transformation of variable *latticeNodeNum* empirically optimized the goodness of fit statistics (Akaike Information Criterion)[50] for model selection relative to other possible transformations. We performed the regression analyses using the open-source statistical software R, version 2.3.1.[51]

Validation by Case Studies

For providing inspection and interpretation evidence, a small set of sample contexts was randomly selected from those contexts that have 1 anonymous node. The authors of this article reviewed and analyzed the anonymous nodes and described their findings. The latest version (20070131) of SNOMED CT was used for validation.

Results

Statistical Results

A total of 3,853 contexts in the domain of Clinical Finding and 2002 contexts in the domain of Procedure were computed and collected (Table 4). By removing repeated sample contexts and those contexts only having the isa relations as the formal attributes, we obtained 3,586 contexts from the domain *Clinical Finding* and 1,864 contexts from the domain *Procedure*, yielding 5,450 unique contexts for analyses. Table 5 shows the basic characteristics of the dataset.

Table 5 ■ Basic Characteristics of the Dataset

Variable	Mean	Min	Max	SD
definedObjNum	20.7	0	4674	148.2
latticeNodeNum	159.3	1	66985	1848.2
anonymousNodeNum	101.1	0	56673	1475.6

anonymousNodeNum = number of anonymous nodes; *definedObjNum* = number of fully defined concepts within the objects; *latticeNodeNum* = number of lattice nodes; SD = standard deviation.

The regression results of Model 1 are detailed in Table 6. The results showed that, after normalizing for the number of lattice nodes, the number of anonymous nodes had significant negative effects on the number of fully defined concepts within the objects of a context ($p < 0.001$). The dispersion parameter (expected value) of the model was 7.96. The finding reveals that the number of anonymous nodes may explain the semantic completeness of the SNOMED CT contents. In other words, the larger the number of anonymous nodes within a specific domain, the smaller the number of fully defined concepts within that domain; we suggest this indicates that SNOMED CT contents are quantifiably semantically incomplete.

The regression results of Model 2 are detailed in Table 7. The results showed that, adjusting for the number of lattice nodes, the contexts from the domain *Clinical Finding* have fewer anonymous nodes than those from the domain *Procedure* ($p < 0.001$). The dispersion parameter of the model was 4.26. The finding reveals that the semantic completeness is significantly different between the domains *Clinical Finding* and *Procedure* when the number of anonymous nodes is used as the representation of the semantic completeness of the SNOMED CT contents.

Case Study Results

Table 8 provides a list of top 20 contexts (i.e., domains) ranked by the proportion of anonymous nodes (i.e., the ratio of the number of anonymous nodes over the number of lattice nodes); the anonymous nodes are identified.

Four domains (i.e., contexts) that have 1 anonymous node were randomly selected for human-based review. Two domains are from *Clinical Finding* domain, and 2 are from *Procedure* domain. Table 9 shows the information of 4 anonymous nodes from 4 specific domains. Of note, while 3 of the potential missing relationships shown remain missing in the current version of SNOMED CT, one of the discovered anonymous nodes from the 2006 version has been corrected in the 2007 version.

Table 6 ■ Results of the Quasi-Poisson Regression Model: Model 1

Dependent Variable	Independent Variable	Coefficients	Standard Error	t Value	p
definedObjNum	Intercept	−0.77	0.028	−27.13	<0.001
	Log(*latticeNodeNum*)	0.86	0.0044	192.96	<0.001
	anonymousNodeNum	−9.16e-6	8.06e-7	−11.37	<0.001
	Overall model: residual deviance: 44587 on 5447 degrees of freedom (dispersion parameter for quasi-Poisson family taken to be 7.96)				

Table 7 ■ Results of the Quasi-Poisson Regression Model: Model 2

Dependent Variable	Independent Variable	Coefficients	Standard Error	t Value	p
anonymousNodeNum	Intercept	−1.92	0.014	−136.33	<0.001
	Log(*latticeNodeNum*)	1.18	0.0015	790.85	<0.001
	domain(ClinicalFinding)	−0.20	0.0059	−34.13	<0.001
	Overall model: residual deviance: 29276 on 5447 degrees of freedom (dispersion parameter for quasi-Poisson family taken to be 4.26)				

The variable *domain* was coded as binary (*Procedure/ClinicalFinding*).

- Sample domain 1: *Open wound of shoulder region and upper limb with tendon involvement (SCTID_269176007)*

  This is a subdomain of *Clinical Finding* and it contains 8 concepts. One of them is a fully defined concept. The anonymous node identified has an own attribute "Upper arm structure (body structure)" and 2 extensions: the concept "Open wound of upper arm with tendon involvement (disorder)" and the concept "Multiple open wounds of upper arm with tendon involvement (disorder)." This may imply that a super concept of the 2 concepts is missing and worth adding as a first-class concept. In addition, we found that the sibling concepts of these 2 concepts are not consistently distinguished by the "single" and "multiple" properties. This representation persists in the latest version of SNOMED CT (Fig. 3).
- Sample domain 2: *Phlebitis of intracranial venous sinus (SCTID_18058007)*

  This is a subdomain of *Clinical Finding* and it contains 13 concepts; 4 of them are fully defined concepts. The anonymous node identified has an own attribute "Superior sagittal sinus structure (body structure)," and 2 extensions: the concept "Phlebitis of superior sagittal sinus (disorder)" and the concept "Endophlebitis of superior sagittal sinus (disorder)." We found that the assignment of the latter concept as the subconcept of the former one is missing. This is particularly striking since such a relationship is asserted for all of its siblings. This singular exception persists in the latest version of SNOMED CT. (Fig. 4).
- Sample domain 3: *Operation on vas deferens (SCTID_23304006)*

  This is a subdomain of "Procedure" and contains 12 concepts, 7 of which are fully defined concepts. The anonymous node identified has an own attribute "Excision - action (qualifier value)" and 2 extensions: the concept "Removal of valve of vas deferens (procedure)" and the concept "Bilateral vasectomy (procedure)." This may imply that a superconcept of the 2 concepts is missing. While we consider that the concept "vas deferens excision" is worth adding, we found that there is an existing superconcept "vas deferens excision (SCTID_120013000)" for the 2 concepts, which, however, is not a subconcept of "Operation on vas deferens (SCTID_23304006)." This is obviously an error, which indeed has been fixed in the latest version (Fig. 5).
- Sample domain 4: *Serologic test for herpes virus (SCTID_14421005)*

  This is a subdomain of Procedure, and it contains 7 concepts. None of them are fully defined concepts. The anonymous node identified has an own attribute "Human herpes simplex virus type 2 antibody (substance)" and 2 extensions: the concept "Herpes simplex virus 2 antibody pattern determination (procedure)" and the concept "Herpes simplex virus 2

Table 8 ■ Top 20 Contexts with the Anonymous Nodes Identified (Ranked by the Proportion of Anonymous Nodes)

SCTID	Concept Name	Lattice NodeNum (LNN)	Anonymous NodeNum (ANN)	Proportion (ANN/LNN)	Domain
129233004	Procedure on bone (organ)	30678	27613	90.0%	Procedure
118699001	Procedure on pelvis	42631	36885	86.5%	Procedure
118710009	Procedure on lower extremity	18977	16262	85.7%	Procedure
118745001	Procedure on joint	15793	13483	85.4%	Procedure
928000	Disorder of musculoskeletal system	66985	56673	84.6%	Clinical finding
129152004	Procedure on back	1828	1524	83.4%	Procedure
417163006	Traumatic and/or nontraumatic injury	59739	48249	80.8%	Clinical finding
71861002	Implantation	8162	6440	78.9%	Procedure
118943001	Disorder of pelvis	34179	26666	78.0%	Clinical finding
414252009	Finding of back	12198	9508	77.9%	Clinical finding
118712001	Procedure on thigh	1040	810	77.9%	Procedure
373196008	Operative procedure on bone of lower extremity	1519	1179	77.6%	Procedure
76069003	Disorder of bone	23408	18156	77.6%	Clinical finding
230896003	Intracranial vascular operation	519	399	76.9%	Procedure
118953000	Bone finding	24216	18516	76.5%	Clinical finding
38629001	Operative procedure on the arteries of the thorax and abdomen	2796	2130	76.2%	Procedure
112698002	Operation on joint	6062	4614	76.1%	Procedure
2119009	Repair of blood vessel	5139	3832	74.6%	Procedure
120166004	Mediastinum repair	2249	1661	73.9%	Procedure
239364005	Maxillofacial bone operation	1336	979	73.3%	Procedure

anonymousNodeNum (ANN) = number of anonymous nodes; *latticeNodeNum (LNN)* = number of lattice nodes; *proportion(ANN/LNN)* = the ratio of *anonymousNodeNum* over *latticeNodeNum*.

Table 9 ■ The Information of 4 Anonymous Nodes from 4 Specific Domains

Sample Domains		Anonymous Node Information	
SCTID	Concept Name	OwnAttributes	Extensions
Samples from clinical finding domain			
269176007	Open wound of shoulder region and upper limb with tendon involvement	Upper arm structure (body structure)(Finding site (attribute))	1. Open wound of upper arm with tendon involvement (disorder) 2. Multiple open wounds of upper arm with tendon involvement (disorder)
18058007	Phlebitis of intracranial venous sinus	Superior sagittal sinus structure (body structure)(Finding site (attribute))	1. Phlebitis of superior sagittal sinus (disorder) 2. Endophlebitis of superior sagittal sinus (disorder)
Samples from procedure domain			
23304006	Operation on vas deferens	Excision - action (qualifier value)(Method (attribute))	1. Removal of valve of vas deferens (procedure) 2. Bilateral vasectomy (procedure)
14421005	Serologic test for herpes virus	Human herpes simplex virus type 2 antibody (substance)(Component (attribute))	1. Herpes simplex virus 2 antibody pattern determination (procedure) 2. Herpes simplex virus 2 antibody assay (procedure)

antibody assay (procedure)." This implies that a superconcept of the 2 concepts is missing and the superconcept may be named as "Serologic test for herpes simplex virus 2." This is still not fixed in the latest version (Fig. 6).

Discussion

About the FCA Model

In this study, we used a high-level SNOMED CT concept and its subconcepts as a proxy to select a specific domain, within which all subconcepts were transformed into their normal forms. As a consequence, the semantic space of the domain in question is expanded (by completing the context) and the details of the semantic definitions of all concepts within the domain could be collected to support the normalization process for instances of postcoordinated expressions of the domain.

The normal form is the central structure for the formal representation of SNOMED CT concepts. We consider that the normal form is also the intermediate layer for the tasks of semantic integration, such as comparing, merging, and classifying precoordinated and postcoordinated expressions. By formalizing the normal forms of a specific domain using the language of the FCA, the model provides the potential to establish an automatic way to perform these semantic integration tasks.

A transformational scaling is needed to transform a many-valued formal context into a 1-valued or Boolean context. By this kind of transformation—in particular, completing the context—the semantic space of the domain is expanded further and made more complete. We consider that it is a feature of the FCA scaling model that every fine-grained element defined in SNOMED CT is exploited. In addition, we found that a step to "complete the context" from relationships that are not otherwise exhaustively asserted is also required by using the hierarchical knowledge of the SNOMED CT. By this kind of completion, the model acquires a robust representation of the semantics explicitly and implicitly contained in SNOMED CT.

As a consequence of transforming the 1-value table into the concept lattice, new entities are synthesized that lack any "own object" labels; these new entities are called anonymous nodes. When formal structures are represented graphically, they induce associative structures in a user's mind,

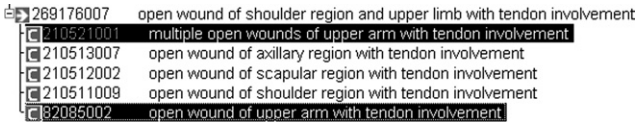

Figure 3. The sample domain "Open wound of shoulder region and upper limb with tendon involvement (SCTID_269176007)" in 20070131 version of SNOMED CT. This figure is a part of screenshot of CliniClue 2006—Terminology Browser (http://www.clinical-info.co.uk).

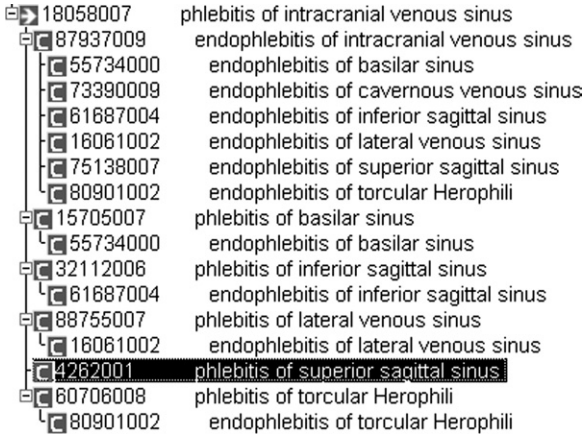

Figure 4. The sample domain "Phlebitis of intracranial venous sinus (SCTID_18058007)" in 20070131 version of SNOMED CT. This figure is a part of screenshot of CliniClue 2006—Terminology Browser (http://www.clinical-info.co.uk).

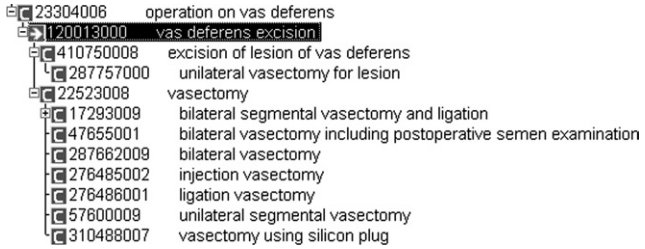

Figure 5. The sample domain "Operation on vas deferens (SCTID_23304006)" in 20070131 version of SNOMED CT. This figure is a part of screenshot of CliniClue 2006—Terminology Browser (http://www.clinical-info.co.uk).

which could provide an opportunity for the discovery of tacit knowledge, emergent structure, and the continuing evolution of fully defined meanings.[47] In this study, when the formal structure of the normal forms was visualized as a concept lattice, the anonymous nodes attracted our attention. As shown in Figure 2, 5 anonymous nodes emerge and their information (Table 3) could be retrieved and provided to the SNOMED CT curators and developers for consideration. We suggest that it would be difficult to acquire this kind of knowledge from other existing approaches. For instance, while a node labeled by "Transsphenoidal hypophysectomy" exists in the lattice, an object label "Transfrontal hypophysectomy" is probably missing from an anonymous node in the same level. We suggest that this kind of specific knowledge is a practical mechanism to show foci of incompleteness within SNOMED CT, and thus could be useful for auditing SNOMED CT or any other large terminology.

While the anonymous nodes may represent missing concepts, an interesting question is whether it is sensible or necessary to represent all anonymous nodes as concepts in a biomedical terminology. Clearly the appropriateness will vary with the purpose of the terminology and the navigability of hierarchies. We argue that explicit representation of the knowledge retrieved from anonymous nodes with the support of lattice-based visualization may provide a mechanism for terminology curators to define rules that may inform which anonymous nodes would deserve representation. For example, the fact that there is a concept "Transsphenoidal hypophysectomy" may be an argument for a rule to represent "Transfrontal hypophysectomy." These kinds of issues have also been discussed in a study of SNOMED CT about classifying diseases with respect to anatomy[52] and in a work on the compositionality of the Gene Ontology by Ogren et al.[53]

Regression Models and Results

We developed a quasi-Poisson regression model to test whether the number of anonymous nodes could explain the semantic completeness of SNOMED CT contents. The number of fully defined concepts within a context was used as the outcome variable to represent the semantic completeness of the SNOMED CT contents. We believe that this is reasonable because the modeling of a fully defined concept expresses the full meaning of the concept. In other words, the higher the percentage of the SNOMED CT concept codes that were fully defined, the more complete are the semantics of the SNOMED CT contents. Improving this kind of logic definition has become one of main goals of the SNOMED CT curators.[16] The regression results confirm our hypothesis that the number of anonymous nodes correlates negatively with the number of fully defined concepts within a context (i.e., the semantic completeness) after normalizing for the number of lattice nodes of the context. We consider that the adjusting variable here is necessary because the size of the contexts indicated by the number of lattice nodes was different and obviously confounding.

While the FCA model proposed in this study could provide specific information about the anonymous nodes for auditing the semantic completeness of a specific domain in SNOMED CT, we also developed an approach for measuring the differences in the semantic completeness among different domains. In our Model 2, the number of anonymous nodes was used as the outcome variable. The results show that the contexts from the domain *Clinical Finding* have fewer anonymous nodes than the domain *Procedure*, after adjusting for the size of those contexts. Furthermore, the findings indicate that the semantic completeness of the 2 largest domains in the SONMED CT is significantly different, i.e., the contexts from the domain *Clinical Finding* is more semantically complete than those from the domain *Procedure*. Thus we believe that the approach also could be used to audit or compare the semantic completeness of any arbitrarily defined domains or subdomains.

We used quasi-Poisson regression models in this study rather than the more conventional Poisson regression model because we found that an overdispersion (having an expected value >1) existed in the data for outcome variables *definedObjNum* and *anonymousNodeNum*, which are count data, nonnegative and highly skewed. A possible reason for the overdispersion in these data is that both the number of objects that are fully defined concepts and the number of anonymous nodes do not occur independently. In fact, these sampled data were extracted from the contexts that were nested. The dispersion parameters in the 2 models were 7.96 and 4.26, indicating that use of a quasi-Poisson regression model increased the standard error metric by the square root of each dispersion parameter,[50] penalizing significance measures. However, this means that our finding of significant difference in semantic completeness across domains is more likely to be real, since this significance metric is highly conservative by absorbing the full "penalty" of dispersion estimates higher than 1.

Practical Significance

We consider that the proportion of anonymous nodes (i.e., the ratio of the number of anonymous nodes over the number of lattice nodes) in a specific domain may provide a practical measure for the semantic completeness of the domain. In our sampled contexts (i.e., domains), the domain

Figure 6. The sample domain "Serologic test for herpes virus (SCTID_14421005)" in 20070131 version of SNOMED CT. This figure is a part of screenshot of CliniClue 2006—Terminology Browser (http://www.clinical-info.co.uk).

"Procedure on bone (organ)" has the highest proportion of anonymous nodes, up to 90% (Table 8). This indicates the curators may need to pay more attention to the domain for further investigation.

We performed case studies and reviewed 4 domains that have 1 anonymous node identified. This analysis demonstrated that the anonymous nodes are useful not only to identify missing concepts from a domain, but also to identify the semantic inconsistency and errors within a domain. Through referencing the latest version of SNOMED CT, we validated that some errors identified by our approach have been fixed in the latest version (e.g., sample domain 3), and some errors have not been fixed. We consider that this analysis provides anecdotal evidence for the effectiveness of our FCA approach to auditing the SNOMED CT.

Limitations

There are several limitations in this study. First, our approach using the anonymous nodes retrieved from the FCA-based model should complement other auditing approaches for SNOMED CT. In addition, the approach was based on a model specific to SNOMED CT. Validation of the approach in other DL-based clinical terminologies will be necessary in the future. Second, the FCA approach should not be applied for evaluating semantic completeness to terminological systems that have very minimal definitions (e.g., limited to one isa relation). In this study, we have removed those contexts only having the isa relations as the formal attributes. Third, while the number of anonymous nodes correlates well with semantic completeness of SNOMED CT contents, we reviewed 4 domains to demonstrate the practical significance of our FCA approach. A more systematic review beyond our anecdotal examination would be the next step for future study. Fourth, SNOMED CT uses "role groups" for grouping "attribute-value pairs" to simplify the terminology's concept model.[8] The model in this study dealt only with the attribute-value pairs and did not consider the issue of role groups. We consider that the semantics of role groups is a higher level abstraction for representing the knowledge of SNOMED CT expressions and further studies are needed to address how to formalize the role groups using FCA as a tool. In addition, the FCA model only formalized the long normal forms, and did not include a representation of SNOMED CT context model. Finally, the sample size in this study was only about 10% of all contexts that could be formed, and while efficient for testing our model, it is possible that the sample is not representative of all contexts.

Conclusion

In this study, we developed a novel approach for auditing the semantic completeness of SNOMED CT using an FCA-based model. We demonstrate that the anonymous nodes retrieved from the FCA model can explain the semantic completeness of SNOMED CT contents indicated by the fully defined concepts. Our novel FCA-based approach may be useful for auditing the semantic completeness of SNOMED CT not only for a specific domain, but also for different domains.

References ■

1. ISO 17115: Health Informatics—Vocabulary for Terminological Systems. 1st edition. 2007. The International Organization for Standardization (Geneva, Switzerland. http://www.iso.org).
2. Chute CG, Cohn SP, Campbell JR. A framework for comprehensive health terminology systems in the United States: development guidelines, criteria for selection, and public policy implications. ANSI Healthcare Informatics Standards Board Vocabulary Working Group and the Computer-Based Patient Records Institute Working Group on Codes and Structures. J Am Med Inform Assoc 1998;5:503–10.
3. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37:394–403.
4. ISO 1087-1: Terminology Work—Vocabulary, Part 1: Theory and Application: Technical Committee TC 37/SC 1; ISO Standards—Terminology (Principles and Coordination). 1996. The International Organization for Standardization (Geneva, Switzerland. http://www.iso.org).
5. ISO 1087-2: Terminology Work—Vocabulary, Part 2: Computer Applications: Technical Committee TC 37/SC 1; ISO Standards—Computer Applications for Terminology. 1996. The International Organization for Standardization (Geneva, Switzerland. http://www.iso.org).
6. SNOMED. Available at: http://www.snomed.org/. Accessed February 20, 2007.
7. Spackman KA, Campbell KE. Compositional concept representation using SNOMED: towards further convergence of clinical terminologies. Proc AMIA Symp 1998:740–4.
8. 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. Proc AMIA Symp 2002:712–6.
9. Yu AC. Methods in biomedical ontology. J Biomed Inform 2006;39:252–66.
10. Spackman KA. Normal forms for description logic expressions of clinical concepts in SNOMED RT. Proc AMIA Symp 2001: 627–31.
11. SNOMED Clinical Terms Guide: Transforming Expressions to Normal Forms. July 2005 release. The International Health Terminology Standards Development Organization. Copenhagen, Denmark. Available at http://www.ihtsdo.org. Accessed November 2008.
12. Cornet R, Abu-Hanna A. Two DL-based methods for auditing medical terminological systems. Proc AMIA Symp 2005:166–70.
13. Granter B, Wille R. Formal Concept Analysis: Mathematical Foundations. New York: Springer, 1999.
14. Kalfoglou Y, Dasmahapatra S, Chen-Burger Y. FCA in knowledge technologies: experiences and opportunities. In: Concept Lattices: Second International Conference on Formal Concept Analysis, ICFCA 2004, Sydney, Australia, February 23–26, 2004, Proceedings. Eklund P (Ed.), Vol. 2961, 2004, ISBN 978-3-540-21043-6.
15. Cimino JJ. Auditing the Unified Medical Language System with semantic methods. J Am Med Inform Assoc 1998;5:41–51.
16. Spackman KA. Rates of change in a large clinical terminology: three years experience with SNOMED CT clinical terms. Proc AMIA Symp 2005:714–8.
17. Ceusters W, Smith B, Flanagan J. Ontology and medical terminology: why description logics are not enough. In: Proceedings of TEPR 2003—Towards an Electronic Patient Record. San Antonio, Texas, May 10–14, 2003. CD-ROM publication.
18. Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in DL-based terminologies: a case study in SNOMED CT. In: Hahn J, editor. KR-MED 2004 Proceedings. Wilstler, Canada: AMIA, 2004:12–20.
19. Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED-CT. Medinfo 2004;11:482–6.

20. Halper M, Wang Y, Min H, Chen Y, Hripcsak G, Perl Y, Spackman KA. Analysis of error concentrations in SNOMED. AMIA Symp Proc 2007:314–8.
21. Spackman K, Reynoso G. Examining SNOMED from the perspective of formal ontological principles. In: Hahn J, editor. KR-MED 2004 Proceedings. Wilstler, Canada: AMIA, 2004:81–7.
22. Green JM, Wilcke JR, Abbott J, Rees LP. Development and evaluation of methods for structured recording of heart murmur findings using SNOMED-CT post-coordination. J Am Med Inform Assoc 2006;13:321–33.
23. Richesson RL, Andrews J, Krischer J. Use of SNOMED CT to represent clinical research data: a semantic characterization of data items on Case Report Forms in vasculitis research. J Am Med Inform Assoc 2006;13:536–46.
24. Cordi V, Mascardi V. Checking the completeness of ontologies: a case study from the semantic web. In: Proceedings of the Italian Conference on Computational Logic (CILC-2004). 2004. Panegai and G. Rossi eds., Quaderno del Dipartimento di Matematica, vol. 390, University of Parma, 2004.
25. Fox MS, Gruninger M. On ontologies and enterprise modeling. In: International Conference on Enterprise Integration Modelling Technology 97. New York: Springer-Verlag, 1997.
26. Devanbu PT, Jones MA. The use of description logics in KBSE systems: experience report. In: Fadini B, Osterweil L, Lamsweerde AV, Proceedings of the 16th International Conference on Software Engineering. Los Alamitos, CA: IEEE Computer Society Press, 1994:23–5.
27. Rassinoux AM, Miller RA, Baud RH, Scherrer JR. Compositional and enumerative designs for medical language representation. Proc AMIA Annu Fall Symp 1997:620–4.
28. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR, for the Computer-Based Patient Record Institute's Work Group on Codes & Structures. The content coverage of clinical classifications. J Am Med Inform Assoc 1996;3:224–33.
29. 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. J Am Med Inform Assoc 1997;4:238–51.
30. 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. J Am Med Inform Assoc 1997;4:484–500.
31. Penz JF, Brown SH, Carter JS, et al. Evaluation of SNOMED coverage of Veterans Health Administration terms. Medinfo 2004;11:540–4.
32. Elkin PL, Brown SH, Husser CS, et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clin Proc 2006;81:741–8.
33. Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc 2006;13:277–88.
34. Elkin PL, Brown SH, Lincoln MJ, Hogarth M, Rector A. A formal representation for messages containing compositional expressions. Int J Med Inform 2003;71:89–102.
35. Diaz-Agudo B, Gonzalez-Calero PA. Formal concept analysis as a support technique for CBR. Knowledge-Based Syst 2001;14: 163–71.
36. Priss U. Formal concept analysis in information science. Annual Review of Information Science and Technology 2006;40:521–43.
37. Rector AL, Bechhofer S, Goble CA, Horrocks I, Nowlan WA, Solomon WD. The GRAIL concept modelling language for medical terminology. Artif Intell Med 1997;9:139–71.
38. Rector AL, Rogers JE. Ontological and Practical Issues in Using a Description Logic to Represent Medical Concepts: Experience from GALEN. Technical Reports, School of Computer Science PrePrint, University of Manchester: CSPP-35:1-35, 2006. Available at: http://www.opengalen.org. Accessed November 2008.
39. Zhang S, Bodenreider O. Aligning representations of anatomy using lexical and structural methods. Proc AMIA Annu Fall Symp 2003:753–7.
40. W3C Recommendation. OWL Web Ontology Language Reference. Available at: http://www.w3.org/TR/owl-ref/. Accessed November 2008.
41. Cimiano P, Hotho A, Stumme G, Tane J. Conceptual Knowledge Processing with Formal Concept Analysis and Ontologies. In: Proceedings of the Second International Conference on Formal Concept Analysis (ICFCA 04). New York: Springer, 2004:189–207.
42. Schnabel M. Representing and processing medical knowledge using formal concept analysis. Methods Inf Med 2002;41:160–7.
43. The IEEE P1600.1 Standard Upper Ontology Working Group (SUO WG). Home page. Available at: http://suo.ieee.org/. Accessed February 20, 2007.
44. Jiang G, Ogasawara K, Endoh A, Sakurai T. Context-based ontology building support in clinical domains using formal concept analysis. Int J Med Inform 2003;71:71–81.
45. Kalfoglou Y, Schorlemmer M. IF-Map: an ontology-mapping method based on information-flow theory. In: LNCS 2800 Journal of Data Semantics 1. 98–127. New York: Springer, 2003.
46. Stumme G, Maedche A. Merging ontologies by means of formal concept analysis. First International Workshop on Databases, Documents, and Information Fusion. Magdeburg, Germany: April 2001.
47. Priss U, Old LJ. Modeling lexical databases with formal concept analysis. J Universal Comput Sci 2004;10:967–84.
48. Priss U. Formalizing botanical taxonomies. In: De Moor A, Lex W, Ganter B (eds). Conceptual Structures for Knowledge Creation and Communication. Proceedings of the 11th International Conference on Conceptual Structures. New York: Springer Verlag, 2003:309–22, LNAI 2746.
49. W3C Technical Report. Design Issues: Architectural and philosophical points. Available at: http://www.w3.org/DesignIssues/Reify.html. Accessed March 13, 2008.
50. Maindonald J, Braun J. Data analysis and graphics using R: an example-based approach. Cambridge: Cambridge University Press, 2003.
51. The R Project for Statistical Computing. Home Page. Available at: http://www.r-project.org/. Accessed February 20, 2007.
52. Burgun A, Bodenreider O, Mougin F. Classifying diseases with respect to anatomy: a study in SNOMED CT. AMIA Annu Symp Proc 2005:91–5.
53. Ogren PV, Cohen KB, Hunter L. Implications of compositionality in the gene ontology for its curation and usage. Pac Symp Biocomput 2005:174–85.