Selective Retrieval of Pre- and Post-coordinated SNOMED Concepts

Robert H. Dolin, MD[1] (Robert.H.Dolin@kp.org); Kent A. Spackman, MD, PhD[2,3]; David Markwell, MB BS[4]

[1]Kaiser Permanente; [2]College of American Pathologists; [3]Oregon Health Sciences University; [4]The Clinical Information Consultancy, UK

ABSTRACT

In general, it is very straightforward to store concept identifiers in electronic medical records and represent them in messages. Information models typically specify the fields that can contain coded entries. For each of these fields there may be additional constraints governing exactly which concept identifiers are applicable. However, because modern terminologies such as SNOMED CT are compositional, allowing concept expressions to be pre-coordinated within the terminology or post-coordinated within the medical record, there remains the potential to express a concept in more than one way. Often times, the various representations are similar, but not equivalent.

This paper describes an approach for retrieving these pre- and post-coordinated concept expressions: (1) Create concept expressions using a logically-well-structured terminology (e.g., SNOMED CT) according to the rules of a well-specified information model (in this paper we use the HL7 RIM); (2) Transform pre- and post-coordinated concept expressions into a normalized form; (3) Transform queries into the same normalized form. The normalized instances can then be directly compared to the query.

Several implementation considerations have been identified. Transformations into a normal form and execution of queries that require traversal of hierarchies need to be optimized. A detailed understanding of the information model and the terminology model are prerequisites. Queries based on the semantic properties of concepts are only as complete as the semantic information contained in the terminology model. Despite these considerations, the approach appears powerful and will continue to be refined.

INTRODUCTION

It is commonplace to send SNOMED concept identifiers in messages and/or store them in medical records[1]. Typically, each field that can hold a SNOMED concept identifier will have some constraints on the allowable set of codes. Despite these constraints, there remains the potential to express the same concept in more than one way. For instance, the concept "**excision of pituitary gland**" could be represented by a single pre-coordinated code for "**hypophysectomy**", or by the post-coordination of codes for "**brain excision**" and "**pituitary gland**". This paper describes an evolving approach to querying for pre- and post-coordinated SNOMED concepts, which would enable retrieval of either "**hypophysectomy**" or "**brain excision**" + "**pituitary gland**".

Perhaps more realistic is the case where pre- and post-coordinated concepts are not equivalent, but still need to retrieved by a particular query. The approach described here enables the retrieval of "**partial excision of pituitary gland by transfrontal approach**" or "**hypophysectomy**" + "**transfrontal approach**" or "**brain incision**" + "**pituitary posterior lobe**" when querying for "**pituitary operations**". In other words, selective retrieval is not the same as a determination of pre- vs. post-coordinated concept equivalency.

Basically, the approach converts pre- and post-coordinated concepts into a common form[2-3]. The query itself is also converted into this common form. The common form of the instances can then be directly compared to the common form of the query. The conversion process is facilitated by the use of guidelines that recommend one of several possible representations.

SELECTIVE RETRIEVAL APPROACH

1. Work within an information model.

This approach assumes that SNOMED concept identifiers are stored or transmitted in a structure provided by an information model. In this paper we use the HL7 Reference Information Model (RIM)[4] because it is a standardized healthcare information model with well-defined rules for the inclusion of coded concepts[5-6]. While the

examples in this paper use the RIM to illustrate the retrieval process, the general principles hold regardless of the specific information model.

The RIM is an object-oriented model comprised of classes and their attributes. Figure 1 shows a portion of RIM Version 1.15's Procedure class. Each RIM attribute has a specified data type. Data types that can carry SNOMED concept identifiers include: CODED with EQUIVALENTS (CE), which carries a code, the name of the coding scheme the code is drawn from, the concept's display name, and alternative equivalent identifiers; and CONCEPT DESCRIPTOR (CD), which builds on the CE by adding an internal grammar that supports the post-coordination of codes.

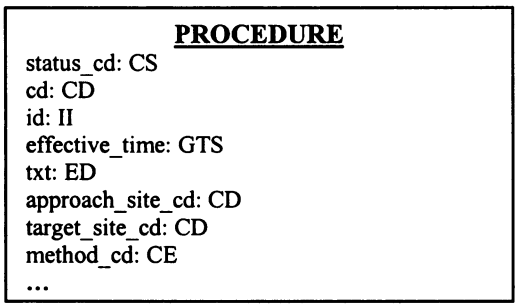

Figure 1. A portion of RIM Version 1.15's Procedure class.

RIM attributes of type CE or CD can also have a specified vocabulary domain of allowable values. These domains can include HL7-defined concepts or can be drawn from HL7-recognized coding systems such as LOINC or SNOMED CT. The vocabulary domain specifications stated in the RIM can be further constrained based on realm of use and/or a particular coding scheme. A constrained domain is known as "value set". Figure 2 shows the Procedure class along with suggested SNOMED CT value sets. These SNOMED CT value sets constrain the allowable set of SNOMED concept identifiers that can occur in these fields. SNOMED CT can express complex value sets as a "Context Concept Subset" that, for instance, excludes values from one subtype descent and includes values from another descent.

2. Transform instances into a common canonical form

Note that even with the constraints shown in Figure 2, it is possible to say the same thing in several ways or to say slightly different things. Some examples are shown in Table 1.

#	Procedure.cd	Procedure. approach_site_cd	Procedure. target_site_cd
1	Partial excision of pituitary gland by transfrontal approach		
2	Hypophysectomy	Transfrontal approach	
3	Excision	Transfrontal approach	Pituitary gland structure
4	Brain excision	Transfrontal approach	Pituitary posterior lobe
5	Brain incision		Pituitary gland structure

Table 1. Various possible instantiations of the model shown in Figure 2.

The next step in retrieval is to convert these representations into a common form. The common form described here is known as the SNOMED CT "long canonical form" (further described in Spackman[2]), which maximally decomposes concepts into their most proximate primitive supertypes, and their most specific values for each defining relationship. This conversion requires there to be some combination of mapping rules and guidelines.

Table 2 shows a mapping from RIM attributes to SNOMED defining attributes (meaning those attributes used by SNOMED CT in the pre-coordinated definition of concepts).

RIM attribute	SNOMED CT defining attribute
Procedure.cd	IsA
Procedure.approach_site_cd	APPROACH
Procedure.target_site_cd	PROCEDURE-SITE
Procedure.method_cd	METHOD

Table 2. Mapping from RIM attributes to SNOMED CT defining attributes.

Given this mapping, the values in Table 1 can be mapped into a SNOMED CT representation, and from there, converted into the SNOMED CT long canonical form, as seen in Table 3. The transformation into long canonical form, where each concept becomes fully decomposed into its atomic parts, is computed by looking at each concept's complete definition in the SNOMED CT database[7].

3. Transform queries into the same common canonical form

It's worth noting that the approach to selective retrieval being discussed is not the same as determination of pre- vs. post-coordination

PROCEDURE

status_cd: CS
cd: CD <= Descendant of Procedure (71388002)
id: II
effective_time: GTS
txt: ED
approach_site_cd: CD <= Descendant of Procedural approach (103379005)
target_site_cd: CD <= Descendant of Body structure (123037004)
method_cd: CE <= Descendant of Action (129264002)
...

Figure 2. A portion of RIM Version 1.15's Procedure class, showing SNOMED CT value sets.

concept equivalency, which isn't necessary. It is clear that the concepts in Table 3 are not equivalent, yet some of them should be retrieved when searching for patients who have had an "**excision of the pituitary gland**" or an "**operation on the pituitary gland**".

Selective retrieval is similar to classification in that instances satisfying a query are those that would classify as a child concept of the query concept. The query itself is converted into a canonical form, and the query results are determined by comparing the canonical form of the instances to the canonical form of the query. Think of the query as itself being a SNOMED concept (as opposed to, say, a SQL statement), and think of the results as those concepts that would classify as descendants of the query concept.

An example will help illustrate. Assume the goal is to search for all patients who have had an "**excision of the pituitary gland**" or to find all patients who have had an "**operation on the pituitary gland**". These queries, expressed in long canonical form, are shown in Table 4.

The retrieval process now checks to see which of the concepts in SNOMED long canonical form would classify as a descendent of the query concept.

Concepts that aggregate under "**excision of the pituitary gland**" would be those procedures matching ALL the following criteria:

- At least one supertype is a kind of Procedure.
- At least one METHOD is a kind of Excision action.
- At least one PROCEDURE-SITE is a kind of Pituitary structure.

Concepts that aggregate under "**operation on the pituitary gland**" would be those procedures matching ALL the following criteria:

- At least one supertype is a kind of Procedure.
- At least one METHOD is a kind of Surgical action.
- At least one PROCEDURE-SITE is a kind of Pituitary structure.

Given this, instances 1-4 would aggregate under "**excision of the pituitary gland**", and instances 1-5 would aggregate under "**operation on the pituitary**". In addition, patient records containing the pre-coordinated SNOMED concept "**total excision of pituitary gland by transfrontal approach**" would aggregate under both queries, whereas "**incision of pituitary gland**" would only aggregate under the second query.

GUIDELINES TO MINIMIZE VARIABILITY

As noted above, transforming instances into a common form requires a mapping between the structures and semantics of the information model and the semantics of the SNOMED CT terminology model. While exhaustive mapping rules can achieve a transformation, it is likely that a combination of mapping rules plus guidelines to minimize variability will result in easier conversions. In some cases, such as the case with HL7, the information model itself has a declared set of rules for where and how to use coded concepts. In addition, SNOMED can issue additional rules to further constrain the variability.

1. Where to post-coordinate?

In the HL7 RIM, it is possible to post-coordinate by using additional RIM attributes and/or by using the expressive power of the CD data type, which contains a grammar for the post-coordination of SNOMED concept identifiers. The post-coordination grammar allows you to assign modifiers (defining attributes and their values, where values can themselves be further

Instance	Mapped into SNOMED CT	Converted into Long Canonical Form
1	IsA Partial excision of pituitary gland by transfrontal approach	IsA Partial hypophysectomy APPROACH Transfrontal approach METHOD Excision action PROCEDURE-SITE Pituitary gland structure
2	IsA Hypophysectomy APPROACH Transfrontal approach	IsA Procedure APPROACH Transfrontal approach METHOD Excision action PROCEDURE-SITE Pituitary gland structure
3	IsA Excision APPROACH Transfrontal approach PROCEDURE-SITE Pituitary gland structure	IsA Procedure APPROACH Transfrontal approach METHOD Excision action PROCEDURE-SITE Pituitary gland structure
4	IsA Brain excision APPROACH Transfrontal approach PROCEDURE-SITE Pituitary posterior lobe	IsA Procedure APPROACH Transfrontal approach METHOD Excision action PROCEDURE-SITE Pituitary posterior lobe
5	IsA Brain incision PROCEDURE-SITE Pituitary gland structure	IsA Procedure METHOD Incision action PROCEDURE-SITE Pituitary gland structure

Table 3. Instances in Table 1 can be mapped into the SNOMED CT representations in column 2, and from there into the long canonical forms shown in column 3.

following concept can be constructed in a CD data type and contained as the value for the Procedure.cd attribute: (Excision (PROCEDURE-SITE :: Pituitary structure)). The concept could also be post-coordinated using a combination of the Procedure.cd and Procedure.target_site_cd fields.

Evolving guidelines for the use of SNOMED concept identifiers in the RIM stipulate that:

- A user should not break apart a pre-coordinated concept in order to maximally populate applicable RIM attributes. For instance, if the procedure site is pre-coordinated in the procedure concept, the user does not need to redundantly populate the Procedure.target_site_cd attribute.

- Applicable RIM attributes must be used where possible. Users can choose to redundantly post-coordinate in the CD data type. For instance, if the procedure site is not pre-coordinated in the procedure, the user must convey body site information in the Procedure.target_site_cd attribute, and can optionally post-coordinate the same body site in the CD data type.

- SNOMED guidelines can recommend against post-coordinating in certain fields. For instance, post-coordination in Procedure.approach_site_cd is not recommended because SNOMED CT does not have any defining attributes for approach concepts.

2. Where to use SNOMED CT Qualifiers?
SNOMED CT defines a set of "qualifiers" – which are a subset of sensible modifiers that can be post-coordinated to a concept to refine (as opposed to alter) its meaning. For instance, the SNOMED concept "**appendectomy**" has PRIORITY qualifiers, and therefore when used in a medical record, can be modified with (PRIORITY :: elective) or (PRIORITY :: emergent). Where SNOMED CT Qualifiers are defined, their use is highly recommended. The preferred way of instantiating SNOMED CT Qualifiers in the HL7 RIM is within the CD data type.

3. What are the allowable post-coordination combinations?
Unconstrained post-coordination can lead to an unwieldy number of possible expressions. The allowable expressions can be constrained by limiting the roles available. Those defining attributes used by SNOMED CT in the pre-coordinated definition of concepts (e.g., "APPROACH", "METHOD", "PROCEDURE-SITE") should be used preferentially when post-coordinating. Some additional roles are used solely as qualifiers, and these are also allowable. Each of these roles has a declared set of concepts where they are applicable, and a declared set of modified) to a concept. For instance, the concepts that are legal values (e.g., the

Query	Converted into Long Canonical Form
Excision of the pituitary gland	IsA Procedure METHOD Excision action PROCEDURE-SITE Pituitary structure
Operation on the pituitary gland	IsA Procedure METHOD Surgical action PROCEDURE-SITE Pituitary structure

Table 4. Queries expressed in long canonical form

PROCEDURE-SITE role can only be used to modify procedure concepts, and its value must be a body site concept).

CONCLUSION

The approach to selective retrieval described in this report appears promising and will continue to be investigated and refined. Several implementation considerations and aggregation limitations have been identified.

The logistics of the approach require a transformation into a canonical form, and require the use of queries that traverse the SNOMED subsumption hierarchies. When transformations to alternative representations are used to enhance the efficiency of retrieval, they should be stored as secondary supporting tables, and not replace the original patient record. Optimization of subtype testing (e.g. is "excision action" a kind of "surgical action") also needs to be considered.

Application of this approach requires a thorough understanding of both the information model and the semantic structure of SNOMED CT. Queries should retrieve concepts and concept specializations, and should not retrieve negated concepts or concepts whose context is different than intended (where for instance a finding is present in a family member vs. the patient).

This approach is based on the shared use of definitional roles among concepts that are all expressed in a common form. Where SNOMED CT has yet to define sufficiently fine-grained definitional roles (e.g. "HAS-COLOR") and apply them to all relevant pre-coordinated concepts (e.g. "red rash"), it will not be possible to directly query based on this particular characteristic (e.g. it wouldn't be possible right now to directly query for all patients with "red findings").

The process for selectively retrieving pre- and post-coordinated SNOMED CT concepts relies on a syntactic transformation into a standard form. The sections above used the SNOMED long canonical form as the standard form, but there are other standard forms – such as a description logic formalism[8]. One advantage to using a description logic formalism as the standard form is that classification algorithms (including the same ones used to classify SNOMED CT) can perform additional retrieval, above and beyond that described above, fully utilizing the logic-based definitions of SNOMED CT concepts. Selective retrieval using a classifier is likely the most powerful technique of all, and should possibly serve as the gold-standard for completeness, by which other techniques are to be compared.

REFERENCES

1. SNOMED Clinical Terms. College of American Pathologists. Jan 2002.
2. Spackman KA. Normal forms for description logic expressions of clinical concepts in SNOMED RT. Fall AMIA 2001.
3. Brown PJB, Sonksen P. Evaluation of the quality of information retrieval of clinical findings from a computerized patient database using a semantic terminological model. JAMIA 2000;7:392-403.
4. HL7 Reference Information Model (www.hl7.org/library/data-model/RIM/modelpage_non.htm)
5. Beeler GW Jr. HL7 version 3 - an object-oriented methodology for collaborative standards development. Int J Med Inf. 1998 Feb;48(1-3):151-61.
6. Bakken S, Campbell KE, Cimino JJ, Huff SM, Hammond WE. Toward vocabulary domain specifications for HL7-coded data elements. JAMIA 2000;7(4):333-42.
7. SNOMED Clinical Terms, Technical Implementation Guide. College of American Pathologists. Jan 2002.
8. Campbell KE, Das AK, Musen MA. A logical foundation for representation of clinical data. JAMIA 1994;1:218-232.