Is the Application of SNOMED CT Concept Model sufficiently Quality Assured?

JM Rodrigues MD PhD1,2, S Schulz MD3, B Mizen MSc1, A Rector MD PhD4, S Serir MSc1

1INSERM LIMICS UPMC UP 13 Paris, France; 2University of Saint Etienne, CHU, Department of Public Health and Medical Informatics, Saint Etienne, France; 3Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria; 4University of Manchester, UK

Abstract

The terminological content of SNOMED CT, the world’s largest clinical terminology is linked to description logics expressions, which give support to consider SNOMED CT a formal ontology. The Terminology Quality Assurance (TQA) of such a terminology resource is hampered by errors in modeling, which act as a barrier for the successful use of electronic health records to ensure semantic interoperability. One application case is the new version of ICD, now in its pre-final form, the content of which is based on a subset of SNOMED CT. The ongoing alignment exercise has highlighted significant modeling issues in more than one third of cases that contrasted SNOMED CT concept model instances with the intuitive meaning given by their Fully Specified Names or synonyms lexically mapped to ICD-11 class names. We recommend prioritizing SNOMED CT TQA on the subset of the core SNOMED CT content to constitute the always true common ontology between SNOMED CT and ICD-11.

Introduction

SNOMED CT1 is the world’s largest clinical terminology with about 300,000 representational units, called concepts. They are linked to terms in several languages, with English providing the highest coverage. All lexical entities in SNOMED CT are named Terms. They include Fully Specified Names, Preferred Terms, and Synonyms. SNOMED CT concept are described and defined by expressions following a formalism called Compositional Grammar (CG)2, which can be interpreted according to description logics (DLs) which allows SNOMED CT to be considered a formal ontology. SNOMED CT is introducing itself as “a terminological resource which consists of codes representing meanings expressed as terms, with interrelationships between the codes to provide enhanced representation of the meanings.”

Nearly half of SNOMED CT’s concepts constitute the large Procedure and Clinical Finding hierarchies. They are not covered by any Terminology Quality Assurance (TQA) mechanisms apart from an abstraction network derivation methodology which focuses on internal consistency only3,4,5.

We here propose to assess the quality of SNOMED CT concept model by assessing its ability to support a formalization based not only on the closed world of SNOMED or ICD but on the open world assumption6 by taking the external example of SNOMED CT ICD-11 semantic alignment work.

As explained in 7 this semantic alignment needs several steps: step 1 is a lexical map between the ICD-11 class name with textual definition and SNOMED CT Fully specified name or synonyms: step 2 is a tentative match between ICD-11 textual definition and SNOMED CT Compositional Grammar (CG) expressions2. Further steps as queries on pre-coordinated or post-coordinated SNOMED CT concepts to take care of ICD-11 inclusions and exclusions rules are out of the scope of this paper.

The purpose is to assess the quality of the inferred CG expressions candidates to the common ICD-11 SNOMED CT ontology by measuring the need of modification of CG expressions in order to represent ICD 11 class definitions if and only if there is a possible lexical map between ICD 11 class and pre-coordinated or post-coordinated SNOMED CT concepts. This is justified by the seemingly high rate of the same strings of characters in both systems: it allows to avoid the more complex issues when the same meaning is expressed by different strings of characters.
We do not consider the current pre-final version of ICD-11 as a gold standard for healthcare terminological resources and the total or partial omission of a SNOMED CT concept that seems necessary to ICD 11 external authority. We neither assess the clinical consistency of ICD 11 textual definitions. We assess only the ability of the existing CG expression(s) to represent the ICD-11 textual definition when lexically mapped to SNOMED CT concept description terms or if necessary to modify this concept model instance according to assumptions, rules, and standards that are specific to the SNOMED CT concept model.

Material

The ICD-11 Mortality Morbidity and Standard (ICD 11-MMS) is extracted from the WHO Browser (date: 31 Jan 2017). The semantic alignment principles explained above are performed with SNOMED CT content as displayed by the IHTSDO browser, corresponding to the SNOMED CT release of the same date, which includes:

1. Number of direct parents
2. Number of role groups
3. Number of children
4. Whether the concept is primitive of fully defined
5. Depth (distance from root)
6. Compositional grammar (CG) expression

The SNOMED CT specific compositional grammar (CG) is a set of rules that govern the way in which SNOMED CT expressions are represented as a plain text string. It is a syntax that supports a wide range of clinical meanings to be captured in a clinical record, without requiring the terminology to include separate concept for each case. Clinical expressions using SNOMED CT concepts can be of two types: pre-coordinated expressions, which use a single SNOMED CT concept identifier and post-coordinated expressions, which contain more than one SNOMED CT identifier.

The IHTSDO browser provides the representation in plain text and in a diagram. For example, the SNOMED CT concept 31978002 Fracture of tibia, a fully defined pre-coordinated concept, is defined using a CG expression as being equivalent to Injury of tibia and Fracture of lower leg, with Associated morphology Fracture and Finding site Bone structure of tibia. This definition is shown below in Table 1 using the SNOMED CT CG formalism, together with the transcription into DL Manchester syntax and in Figure 1 for the CG diagram displayed by the browser. To enhance readability, concepts are noted in Italic and relations are noted in Bold.

Table 1. Fracture of Tibia in Compositional Grammar (CG) and Manchester syntax

<table>
<thead>
<tr>
<th>SNOMED CT ID</th>
<th>CG Expression</th>
<th>Manchester Syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>31978002</td>
<td>Fracture of tibia</td>
<td>Injury of tibia (disorder) +</td>
</tr>
<tr>
<td></td>
<td>414292006 Fracture of lower leg (disorder) :</td>
<td></td>
</tr>
<tr>
<td></td>
<td>{ 363698007 Finding site (attribute) = 12611008 Bone structure of tibia (body structure), 116676008 Associated morphology (attribute) = 72704001 Fracture (morphologic abnormality) }</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fracture of tibia equivalentTo</td>
<td>Injury of tibia (disorder) and Fracture of lower leg (disorder) and RoleGroup some</td>
</tr>
<tr>
<td></td>
<td>Finding site (attribute) some Bone structure of tibia (body structure) and Associated morphology (attribute) some Fracture (morphologic abnormality)</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. The SNOMED CT concept Fracture of tibia (ID = 31978002) as a CG diagram

The purpose of this study is to assess the quality of the SNOMED CT CG inferred expressions of concepts candidates to constitute the common ontology between ICD-11 and SNOMED CT. We measure the rate of issues in
modeling when the SNOMED CT concept model instance is not aligned with the Fully Specified Name of the SNOMED CT concept.

The study is limited to 428 classes from ICD-11-MMS, covering the circulatory system, and 522 ones covering the digestive system. We exclude ICD-11 codes ending with a "Y" or "Z" code (other specified or unspecified term at the subchapter level): 206 of such residual codes are in the circulatory chapter and 250 in the digestive chapter (see Table 3). Residuals assure that in ICD hierarchies all sub-hierarchies are fully exhaustive, although they have no clear meaning outside the context of the ICD hierarchy.

For this study, we focused on SNOMED CT concepts that can be lexically mapped to ICD-11 MMS classes either by one pre-coordinated SNOMED CT concept or by more than one post-coordinated SNOMED CT concept as stated in the introduction and explained in the following subsection. We analyze only the CG expressions of these candidate SNOMED CT concepts lexically mapped to ICD-11 MMS classes.

The work was done by two different knowledge engineering master students one in charge of the circulatory chapter and one in charge of the digestive chapter. Both chapters were supervised by the same senior ICD-11 and SNOMED CT expert.

Methods

Semantic interoperability is a crucial point for using SNOMED CT and ICD in parallel, e.g. to address clinical documentation, epidemiology, research and decision support use cases.

The first step is to perform a lexical map and in a second step a match of meaning between ICD-11-MMS classes, supported by their subclass hierarchies and text definitions and SNOMED CT concepts. Both terminologies are accessed by the web browsers made available by the respective maintenance organizations (cf. Figure 2 and Table 2).

We introduce the following symbols for the mapping types: M (refined by M1 and M2), A (refined by A1 and A2), P and Z. We consider the mapping of a SNOMED CT Concept $SC_i$, described by terms $ST_{i1} \ldots st_{in}$, to an ICD class $IC_i$, described by a name $IT_i$.

Lexical map

The following rules apply for lexical map

- If there is a full lexical map between the ICD-11 class name $IT_i$ and one SNOMED CT description $ST_{i1} \ldots st_{in}$, considered as pre-coordinated in SNOMED CT it is classified as M (for Lexical map) type.
- If there is no lexical map between any $IT_i$ and $ST_{ik}$, but if mapping can be achieved to the post-coordination of two or more descriptions $ST_{i1} \ldots st_{in}$ of $SC_k$, it is classified as A (for Addition map) type.
- If only a part of $IT_i$ of $IC_i$ can be lexically mapped to any $ST_{ik}$ it is classified as P (for Partial) type.
- Finally, if not even a partial lexical mapping between any $IT_{i0}$ of $IC_i$ and $ST_{ik}$ is possible, it is classified as Z (for Zero) type.

Match of meaning

Subsequently, the CG inferred expressions $^2$ of $SC_i$ are analyzed to check whether they correspond to the totality of the textual definition and hierarchy inheritance of $IC_i$ and of defining and constraining axioms of one or more than one $SC_i$.

The following cases are distinguished:

- M (lexical match) type:
  1. This expression fully represents the meaning of $IC_i$, a complete match meaning is assumed: the classification is refined to M1.
  2. This expression does not fully represent the meaning of $IC_i$, a new expression is produced according to CG: the classification is refined to M2.

- A (addition map) type:
  1. These expressions fully represent the meaning of $IC_i$, a complete match meaning is assumed: the classification is refined to A1.
2. These expressions do not fully represent the meaning of \( IC_i \), a new expression is produced according to CG: the classification is refined to \( A_2 \).

- **P type:**
  For \( IC_i \), it is then necessary to create a logical representation based on one existing CG expression plus an extended de novo CG expression.

- **Z type:**
  For this \( IC_i \), it is necessary to create a logical expression in accordance with SNOMED CT CG.

In the following, only \( M \) and \( A \) types will be analyzed.

---

**Figure 2.** ICD-11 SNOMED CT semantic alignment principles

---

**Table 2.** The lexical maps types and meaning matches between the ICD-11 MMS classes and SNOMED CT compositional grammar.

<table>
<thead>
<tr>
<th>Lexical map and meaning match</th>
<th>Action</th>
<th>Compositional grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical map and full meaning match (M 1).</td>
<td>Take the representation expression of the SNOMED CT concept</td>
<td>The existing pre-coordinated inferred expression of</td>
</tr>
</tbody>
</table>
SNOMED CT concept

Lexical map and no full meaning match (M 2)  
Take the representation expression of the SNOMED CT concept
Modify the existing pre-coordinated inferred expression of SNOMED CT concept

Post-coordinated lexical map possible and full meaning match (A 1).  
Take the representation of two or more pre-coordinated existing representations of SNOMED CT concepts
Post-coordination of two or more pre-coordinated existing inferred expression of SNOMED CT concepts

Post-coordinated lexical map possible but no full meaning match (A 2).  
Take the representation of two or more pre-coordinated existing representations of SNOMED CT concepts
Post-coordination and modification of two or more pre-coordinated existing inferred expression of SNOMED CT concepts

Partial lexical map (P)  
Take the representation of one pre-coordinated existing representation of SNOMED CT concept
One pre-coordinated existing inferred expression of a SNOMED CT concept plus an extended de novo CG expression

No lexical map (Z).  
Create a logical CG expression
A new logical CG expression

Results
Table 3 provides an overview of the results.

Table 3. Numbers of codes in the circulatory chapter and Digestive chapter, from ICD 11 MMS 2017 to SNOMED CT 31 January 2017 release by map and meaning match types

<table>
<thead>
<tr>
<th>Map and meaning match types</th>
<th>Circulatory system count</th>
<th>Circulatory system rate (%)</th>
<th>Digestive system count</th>
<th>Digestive system rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>243</td>
<td>56.77</td>
<td>275</td>
<td>52.68</td>
</tr>
<tr>
<td>M2</td>
<td>94</td>
<td>21.96</td>
<td>85</td>
<td>16.28</td>
</tr>
<tr>
<td>A1</td>
<td>51</td>
<td>11.92</td>
<td>29</td>
<td>5.56</td>
</tr>
<tr>
<td>A2</td>
<td>13</td>
<td>3.04</td>
<td>104</td>
<td>19.92</td>
</tr>
<tr>
<td>P</td>
<td>27</td>
<td>6.31</td>
<td>10</td>
<td>1.92</td>
</tr>
<tr>
<td>Z</td>
<td>0</td>
<td>0.00</td>
<td>19</td>
<td>3.64</td>
</tr>
<tr>
<td>Total (M+A+P+Z)</td>
<td>428</td>
<td>100</td>
<td>522</td>
<td>100</td>
</tr>
<tr>
<td>Total (M+A+P+Z)</td>
<td>428</td>
<td>67.51</td>
<td>522</td>
<td>67.62</td>
</tr>
<tr>
<td>&quot;Other” and “Unspecified”</td>
<td>206</td>
<td>32.49</td>
<td>250</td>
<td>32.38</td>
</tr>
<tr>
<td>Total</td>
<td>634</td>
<td>100</td>
<td>772</td>
<td>100</td>
</tr>
</tbody>
</table>

The two most frequent lexical map types are M (M1 plus M2) for full lexical map with a pre-coordinated SNOMED CT concept and A (A1 plus A2) full lexical map with a post-coordinated SNOMED CT concept: 93.69 % for the circulatory chapter and 94.44% for the digestive chapter.

The most frequent type is M1 for for both
The less frequent types are Z for no possible lexical map for the circulatory chapter (0%) and P for partial lexical map for the digestive chapter (1.92%).

The other main differences between the two chapters are the higher rate for the match types A2 for the digestive chapter (19.92% against 3.04%): These differences can be explained by interrater differences (the work was done by two different knowledge engineering master students supervised by the same senior terminology expert) or quality differences between these two chapters either in WHO ICD 11 or in SNOMED CT or in both.

We found for these two chapters nearly the same percentage found on a previous work on the circulatory chapter with browsers based on WHO ICD-11 and SNOMED CT one year before the current study (31 January 2016). The goal of our study is to measure the quality of SNOMED CT concept model for the SNOMED CT concepts candidates for a lexical map with ICD 11 class names and definitions i.e lexical map types M and A which correspond to more than 90% of ICD-11 classes for the circulatory and the digestive chapters.

As an example for M1 type ICD-11 DA 40.4 *Perforation of esophagus* is defined by: “perforation of esophagus is a penetration or hole of the wall of the esophagus, resulting in luminal contents in esophagus flowing into the mediastinum and/or thoracic cavity”.

The full lexical map is with SNOMED CT 23387001 | *Perforation of esophagus (disorder)* | which is a fully defined concept with the following pre-coordinated SNOMED CT inferred expression in Figure 4 and the diagram in Fig. 5.

**Figure 4.** Pre-coordinated CG expression for *Perforation of Esophagus*

```
==== 37657006 [Disorder of esophagus (disorder)] + 51875005 [Gastrointestinal perforation (disorder)] + 300286002 [Lesion of esophagus (finding)] :
  { 363698007 [Finding site (attribute)] = 32849002 [Esophageal structure (body structure)], 116676008 [Associated morphology (attribute)] = 36191001 [Perforation (morphologic abnormality)] } 
```

**Figure 5.** Pre-coordinated CG diagram for *Perforation of Esophagus*

As an example for M2 type ICD-11 BA00 *Essential hypertension* is defined as “Essential (primary) hypertension, accounting for 95% of all cases of hypertension, is defined by ICD as high blood pressure for which a secondary cause cannot be found”.

The full lexical map is with SNOMED CT 59621000 | *Essential hypertension (disorder)* | which is not a fully defined concept but a primitive meaning that the SNOMED CT concept FSN has no full representation with the following pre-coordinated SNOMED CT inferred expression in Figure 6 and the diagram in Figure 7.

**Figure 6.** Pre-coordinated CG expression for *Essential hypertension*
This is only the representation of hypertensive disorder, systemic arterial (disorder) which has the same representation and there is no representation of the ICD-11 class definition statement: a secondary cause cannot be found. We will present later how CG can represent this statement.

Table 3 shows how SNOMED CT concept model represents logically ICD-11 classes

<table>
<thead>
<tr>
<th>Map and meaning match types</th>
<th>Circulatory system count</th>
<th>Circulatory system rate (%)</th>
<th>Digestive system count</th>
<th>Digestive system rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M 1</td>
<td>243</td>
<td>56.77</td>
<td>275</td>
<td>52.68</td>
</tr>
<tr>
<td>M 2</td>
<td>94</td>
<td>21.96</td>
<td>85</td>
<td>16.28</td>
</tr>
<tr>
<td>Rate M2/(M1+M2)</td>
<td>94/337</td>
<td>27.89</td>
<td>85/360</td>
<td>23.61</td>
</tr>
<tr>
<td>A 1</td>
<td>51</td>
<td>11.92</td>
<td>29</td>
<td>5.56</td>
</tr>
<tr>
<td>A 2</td>
<td>13</td>
<td>3.04</td>
<td>104</td>
<td>19.92</td>
</tr>
<tr>
<td>Rate A2/(A1+A2)</td>
<td>13/64</td>
<td>18.75</td>
<td>104/133</td>
<td>78.19</td>
</tr>
<tr>
<td>Poor quality Rate= (M2+A2)/(M1+M2+A1+A2)</td>
<td>107/401</td>
<td>26.68</td>
<td>199/493</td>
<td>40.36</td>
</tr>
</tbody>
</table>

To summarize the quality issue there is a rate of poor quality of 306/894 = 34.22 %

It is interesting to compare the poor quality rate to represent logically ICD-11 classes with the poor quality rate to represent SNOMED concepts which can be the rate of primitive SNOMED CT concept out of the candidates SNOMED CT concept to represent ICD 11 classes as shown in Table 4.

Table 4. Primitive SNOMED CT concepts by map and meaning match types

Figure 7. Pre-coordinated CG diagram for Essential hypertension
The types with full map and meaning match (M1 and A1) have a low rate of SNOMED CT primitive concepts (from 0% to 33%) and the types with no full match (M2 and A2) have a high rate of SNOMED CT primitive concepts (from 81% to 100%). It can be considered that the quality issue with SNOMED CT concept model concerns mainly the rate of primitives with few exceptions.

**Discussion**

The semantic alignment between ICD-11-MMS classes and SNOMED CT concepts provides a good opportunity to trace the quality issues with the SNOMED CT concept model. According to our study, 107 out of 401 SNOMED CT concepts (26.68%) in the circulatory chapter and 199 out of 493 SNOMED CT concepts (40.36%) in the digestive chapter from the Clinical finding hierarchy that lexically mapped to ICD-11 classes show modeling issues resulting in misalignments not only between ICD-11 MMS class meaning and SNOMED CT concept model instances, but as well between SNOMED CT Fully Specified Names and concept model instances. We here list some typical examples: As shown in Table 4 in most of the cases this is related to the high number of primitives, i.e. not fully defined SNOMED CT concepts for more than 80% of SNOMED CT concept candidates to lexically map with ICD-11 class names.

**Misalignment between SNOMED CT concept FSN and primitive representation**

In the example of Essential hypertension (ICD-11 class BA 00), which is the main category of Arterial hypertension which is one of the most frequent disease in the world, Figure 7 shows that the existing SNOMED CT expression does not represent the lack of secondary cause, which is the meaning of “essential” or “idiopathic”. The CG provides the possibility to represent the lack of secondary cause by adding the following attribute and domain value: 370135005 [Pathological process (attribute)] = 54690008 [Unknown (origin) (qualifier value)] and the diagram of Figure 8.

![Figure 8. Actual SNOMED CT CG candidate representation of ICD 11 class BA 00 and proposed fully defined representation](image-url)
There are several other cases with the wording “of unknown etiology” or, e.g. as 85598007 | Constrictive pericarditis (disorder) | with no representation of “constrictive”, 373945007 | Pericardial effusion (disorder) | with no representation of “effusion”, 706882009 | Hypertensive crisis (disorder) | with no representation of “crisis”.

**Misalignment between SNOMED CT concept FSN and fully defined representation**

ICD-11 DA52.51 Allergic gastritis due to IgE-mediated hypersensitivity can be fully represented by the SNOMED CT concepts 1824008 | Allergic gastritis (disorder) | and 422076005 | Immunoglobulin E-mediated allergic disorder (disorder) |, both of which are fully defined. The role of Immunoglobin E is not represented in the present version of the SNOMED CT concept model.

**Conclusion**

To answer the question of the title, *viz.* whether the application of the SNOMED CT concept model is sufficiently quality assured, we can state the following points as a route to an answer.

Semantic interoperability between a clinical terminology as SNOMED CT and a statistically aggregated classification as ICD-11 will be more and more necessary in the future.

The SNOMED CT concept model is a good opportunity to provide a logical basis to automate the coordination.

Problems in matching SNOMED and ICD are a way of highlighting quality issues when applying the SNOMED CT concept model.

We recommend prioritizing SNOMED CT TQA on the subset of the core SNOMED CT content by using ICD11 definition coming from WHO closed world.
This can contribute to develop a common SNOMED CT and ICD-11 ontology always true in an open world assumption.

References