Unambiguous data modeling to ensure higher accuracy term binding to clinical terminologies

Rahil Qamar, (PhD), MSc, Jay (Subbarao) Kola, MBBS, (PhD), Alan L. Rector, MD, PhD
BHIG, University of Manchester, Manchester, U.K.

Abstract
Work in the field of recording standard, coded data is important to reduce medical errors caused by misinterpretation and misrepresentation of data. The paper discusses the need to ensure that the source of the data i.e. the clinical data model is unambiguous to increase the quality and accuracy of the data mapping to terminology codes. The study chooses one especially ambiguous data model and remodels it to make clearer both the structure of the data, as well as its intended use and semantics. By ensuring an unambiguous model, results of the data mapping increased in accuracy from 64.7% to 80.55%. The clinical experts evaluating the models found it easier working with the revised model and agreed on the mappings 93.1% times as against 48.57% times previously. The aim of the study is to encourage good modeling practice to enable clinicians to record and code patient data unambiguously and accurately.

Introduction
One of the key objectives of most national strategies for developing IT solutions for health care is to ensure that all Electronic Health Records (EHRs) in all hospitals share a common information model and standards for data, classifications, and coding systems [1]. For instance, the NHS has agreed to adopt SNOMED CT terminology as the standard used to code all data being recorded [14]. The aim is to control the vocabulary used to record patient data to reduce the possibility for differing interpretation of information and the possibility of errors resulting from traditional paper records [14].

Efforts are also underway to promote the use of data entry forms for structured data capture in EHRs. Clinical data models help in defining the structure of the information to be stored in the EHRs [2] through the data entry forms. If clinical data models are to be used increasingly to render structured data entry forms and guide the entry process, it is critical to extend the codification of data to these models as well. Examples of clinical data models are HL7 CDA documents and messages, and the European openEHR Archetypes. A semi-automated mapping mechanism called the Model Standardisation using Terminology (MoST) System was developed to match clinical model terms to codes in a terminology system. Discussion of the methodology and process are beyond the scope of this paper but have been covered in a previously [3].

Based on the results and feedback gathered during the evaluation of the MoST system results, the paper discusses the problems faced by the evaluators when mapping terms from an ambiguous data model to terminology codes. One especially ambiguous model was then redesigned and the MoST process was repeated to determine whether the improved data model resulted in better mappings. The central idea is to highlight the need for authoring unambiguous and usable data models to have any chance of success at codifying the clinical data represented in them unambiguously. Dealing with terminology issues is a problem that requires addressing once issues with the source of the data i.e. the data models are resolved.

Methods
Data Source – openEHR Archetypes: The data models chosen in the study to provide the source of the data were the openEHR Archetype Models. Archetypes are essentially nested hierarchies of data elements required for recording a particular clinical event. These models specify pre-defined constraints on the data recorded. The archetype approach is being standardized by the openEHR community for wider acceptance in the European healthcare domain. The openEHR Archetypes conform to the openEHR Reference Model [15]. The term ‘Archetypes’ is used in the paper to refer to openEHR Archetypes.

The archetype selected for critical review has been categorized as an observation by the author of the model. The other types of clinical data entry models are evaluation, instruction, and action.

The terminology separation feature of Archetypes was the main reason for selecting the model. This feature separates the data expressions from the terminology used to identify the data. There is an internal label for the data terms along with an annotation to describe their intended meaning. In
addition, there is an external terminology which can be mapped to the data in the ‘term binding’ section of the Archetype. The ease of performing term binding at any stage of the model authoring process made it especially attractive to adopt the model for the mapping exercise in the research study.

**Terminology Source - SNOMED CT:** SNOMED CT aims to be a comprehensive clinical terminology that provides clinical content and expressivity for clinical documentation [8][9].

The reasons for selecting SNOMED CT, referred to as SNOMED in the paper, to code the data were: (i) the size of the terminology. With over 370000 concepts there was a greater scope of finding a semantically correct match for an archetype term, (ii) ease of using the terminology. The data files can be imported into database tables with minimum effort, and (iii) ease to query concepts. Information extracted consisted of concept definitions and their relationship to other concepts in the terminology.

**Mapping Methodology:** The aim of the mapping methodology is to firstly provide an automated system to generate a list of semantic and lexical SNOMED matches to archetype data terms. This automated task is performed by the MoST system [3]. The second part of the methodology is manual and requires clinical experts to map the data terms to the most relevant SNOMED code(s).

The term ‘modeler’ will be used for the original author of the archetype model, ‘clinical expert’/‘evaluator’ for the person performing the mapping, and ‘mapping’ for the term binding process.

**Test Case Analysis:** Archetype data terms from the ‘histology pap’ model, as seen in Figure 1, were sent to the MoST system which returned a number of SNOMED codes per term. The model was taken from the official openEHR website’s published list of archetypes [11]. The model snapshot in Figure 1 was taken from the ‘Terminology’ section of the Archetype Editor developed at LiU, Sweden [12]. The search results were evaluated by 10 clinical experts to determine the relevancy of the codes to perform mapping. They were also asked to provide feedback on the quality of the data model, the terminology, as well as the search mechanism. However, all the experts agreed that the primary reason for the lower accuracy and confidence ratings was the ambiguities in the data model.

**Evaluation:** The ‘histology pap’ model had approximately 40 terms which were sent to the MoST system. The evaluators were required to score, on a scale of 0 to 10, the relevancy of all the SNOMED codes, as shown in Table 1, returned before and after the filtering process. The matches found were of the order 1:n such that 1 archetype term had n SNOMED codes returned as appropriate matches.

<table>
<thead>
<tr>
<th>Archetype data terms (1:n search)</th>
<th>SNOMED CT codes matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before filtering</td>
<td>40</td>
</tr>
<tr>
<td>After filtering</td>
<td>34</td>
</tr>
</tbody>
</table>

Table 1: Total number of archetype terms and the corresponding number of SNOMED codes returned

A code was deemed relevant if the total score given by the 10 evaluators was above 50%. The mapping accuracy which was 34.5% before filtering increased to 64.7% (22/34*100). In addition to accuracy, the rating was also based on the ‘confidence’ quotient. This indicated the confidence the evaluators placed on a particular SNOMED code they thought was relevant for mapping the archetype term to. The higher the score (max. 10) given to a particular code, the higher the level of confidence and vice versa. The ambiguities present in the model resulted in an average confidence rating with a low of 6/10 (6 out of 10) given for each code deemed relevant.

Although the accuracy rate of the results may appear to be sufficiently high at 64.7% indicating that of every 10 archetype terms atleast 6 terms found a relevant SNOMED code to map to, it still highlights problems with the adequacy of the data model, the terminology, as well as the search mechanism. However, all the experts agreed that the primary reason for the lower accuracy and confidence ratings was the ambiguities in the data model.

**Critique by Clinical Experts**

**Issues with the data source model:** The ‘histology pap’ model is the source of the data required for the study. The central issues with the model as stated by the experts revolve around two main concerns. First, what kind of information is the model trying to capture? Second, can the model in its present state be safely used to map to a terminology system without resulting in conflict or error of the intended purpose of the model? Based on the two main concerns following are the issues that were raised:

1) Ambiguous categorisation of top level data terms in model: Although the ‘histology pap’ model belongs to the openEHR observation type, it is intended to record ‘cytological findings of a pap test’. In addition, the children of the root concept ‘cervical smear’ are used to represent different aspects of a test, which would categorise it as a procedure. E.g. the children include ‘specimen collection’, ‘quality (and all its children)’, ‘identification (and all its
children’, and ‘process date’. It would have been less ambiguous for the clinical experts to have worked with a procedure type archetype instead although it is unclear whether the archetype would then belong to the openEHR action or instruction clinical entry type.

2) Conflicting categorisation with terminology system: The intended purpose of the archetype led to different opinions amongst evaluators regarding the basic category to which the archetype belonged. Some experts preferred to map the root term ‘cervical smear’ to a SNOMED observable (although no appropriate codes were available in this category in SNOMED) while most preferred SNOMED codes from the ‘finding’ and ‘procedure’ categories, as shown in Table 2. Other archetype terms in the model found appropriate matches with ‘finding’ codes. Some other models had more terms with conflicting categorisation, and are discussed in [13].

3) Ambiguously named terms: In more than one place the labels for the archetype terms are inadequate to convey the intended meaning of the modeler. E.g. the term ‘clinical’ is defined as ‘clinical notes sent with request’ in the model. Given that ‘clinical’ is nested under the term ‘specimen collection’, as shown in Figure 1, it is difficult to comprehend the intent of the modeler. It is quite easy for a clinician to interpret ‘clinical’, in such a case, to refer to the clinical specimen collection process, especially since there is a sibling term ‘specimen’ with ‘description’ as its child. The ambiguity could be easily avoided by using a more explicit label such as ‘clinical notes’. Some other poorly labeled terms in the model are ‘macroscopic’ and its child ‘appearance’, ‘microscopic’, ‘process’, ‘identification’, ‘comment’, ‘laboratory’, and ‘preservative’, shown in Figure 1.

4) Similar or duplicate labeling of terms with inadequate definitions to resolve ambiguity: On several occasions in the model, the labels assigned to the terms were either duplicated or very similar to other labels in the model. Although the context in which they had been used was different it was difficult to precisely ascertain the intended use as the text definitions lacked in clarity. E.g. ‘specimen’ was used twice in the same model. In the first instance, it meant to record ‘details of specimen’. Later it was used to record the ‘problem with the specimen’. Also there were other uses of the word specimen such as ‘specimen collection’, ‘specimen unlabeled’, ‘insufficient’ and ‘damaged’ specimen. Most of the terms were placed in different contexts in the hierarchy. However the context itself was not clear which did not justify its stated position in the hierarchy. The other term was ‘image’ used for macroscopic and microscopic findings.

5) Insufficient separation of meta data information from core clinical recording information: The model does not sufficiently separate data that is meta (or general) to the recording of the ‘cervical smear’ and the data that is specific to this particular kind of recording. Given that 79% of the evaluators agreed

<table>
<thead>
<tr>
<th>Archetype term</th>
<th>Appropriate matches in SNOMED CT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Papanicolaou smear test (procedure)</td>
<td>[119252009]</td>
</tr>
<tr>
<td>Cervical cytology test (procedure)</td>
<td>[416107004]</td>
</tr>
</tbody>
</table>

Table 2: Sample matches in SNOMED CT for archetype terms. Matches include SNOMED category and code.
that the model represents the recording of a procedure, our comments will be based on procedure types. Terms such as ‘specimen collection’, ‘satisfactory assessment’, ‘quality’, ‘labeling’, and ‘preservative’ are common to any procedure and are therefore meta to the model. These terms need to be placed separately from data that is specific to the recording of cervical smear such as ‘low grade’ and ‘high grade’ epithelial abnormality, ‘negative for dysplasia or malignancy’, and ‘specific cytological finding’. In fact, since archetypes can reuse whole or parts of other archetypes, the meta data could be modeled in a separate archetype and included for use in this model. Even if the meta data is included with the more specific clinical data, it should be clearly separated to reduce ambiguity and enhance simplicity of use.

6) Using post-coordination of terms adequately and in the right place: Some terms in the model were needlessly split into fine-grained representations of a concept to the extent that their semantics were lost in the process. E.g. the term ‘specimen’ and its child ‘description’ both had the same definition of ‘details of the specimen’. Individually they did not make much sense to the expert trying to infer the semantics of both the terms. However, when combined together to form a composite term ‘specimen description’ the intent was clearer. It would also be useful for systems like MoST to lookup well-defined terms for matches in SNOMED. Table 3 below shows the matches returned by MoST given the current model representation. Had the term been in a more explicit pre-coordinated form i.e having one code to represent a single concept, the expert might have selected only the code 115267000 (code shown in Table 3).

<table>
<thead>
<tr>
<th>Archetype term</th>
<th>Post-coordinated SNOMED match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low grade epithelial abnormality</td>
<td>Epithelial cell abnormality (morphologic abnormality) [373886001] + Low grade (qualifier value) [349915008]</td>
</tr>
<tr>
<td>Histological grade finding</td>
<td>(finding) [373372005] + Low grade (qualifier value) [349915008]</td>
</tr>
</tbody>
</table>

Table 4: Suggested post-coordination of the archetype term w.r.t. SNOMED CT

Issues with the terminology source model: Despite the wide coverage of clinical areas by SNOMED CT several problems persist with its content. The problems relate to their definitions and subsumption relationships in the hierarchy, as well as their categorisation and naming, discussed in [13]. Despite being based on formal logic, SNOMED has problems with its concepts being classified consistently especially the concept hierarchies.

The focus of this paper is primarily to address issues arising from ambiguous archetype models. Issues with SNOMED, though discussed briefly, are part of a much larger problem and will be discussed in a separate paper. In this paper we briefly discuss some of the concepts that were missing in SNOMED and were considered important for inclusion. Most of the missing concepts were common to clinical tests or procedures. Some of the codes found missing were:

- ‘Technically unsatisfactory’ to represent technical problems in any kind of test performed.
- ‘Appearance’ specific to histopathology examinations rather than general physical examinations. Also concepts for ‘macroscopic’ and ‘microscopic’ appearance of the tissue.
- ‘Adequate numbers of squamous cells present’ localizing it to PAP smear. The alternative is to post-coordinate Squamous epithelial cell (cell) [80554009] with a new general concept ‘adequate number of cells present’.
- ‘Slide’, ‘container’, and ‘specimen’ unlabelled instead of the generic ‘Sample unlabelled (finding) [125159002]’.
- Modifiers on ‘specimen’ such as ‘insufficient’, ‘damaged’, and ‘contaminated’ to qualify the quality of the specimen.

Table 3: SNOMED codes returned for the split terms.

<table>
<thead>
<tr>
<th>Archetype term</th>
<th>SNOMED CT Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specimen</td>
<td>Form and specimen details different (finding) [281322007]</td>
</tr>
<tr>
<td>Description</td>
<td>Description of specimen character (procedure)[115597007]</td>
</tr>
<tr>
<td></td>
<td>Specimen description (procedure) [115267000]</td>
</tr>
</tbody>
</table>

Table 3: SNOMED codes returned for the split terms. On the contrary, at times post-coordination i.e composing terms of two or more codes to convey the correct semantics was more suitable as compared to pre-coordinating them. E.g ‘low grade epithelial abnormality’ might have been better represented in the form of post-coordinated terms shown in Table 4.

Despite the ability to post-coordinate as above, grading of the cytological abnormality on PAP smear can be represented better than the present low grade and high grade. This has been achieved in the revised model discussed later.
• Concept for ‘preservation of specimen’ with modifiers ‘insufficient’, and ‘incorrect’ etc. to code problems with the preservation of the specimen.

It is interesting to note that most of the missing concepts relate to general (meta) information on pathology procedures although SNOMED started out with an initial focus on pathology (SNOP)[10]. This list will possibly grow in size as more archetypes are evaluated for finding term mappings in SNOMED.

**Revising the data source model**

Based on the issues with the archetype model presented by the evaluators, the existing model was modified to address a majority of the issues.

(i) The data model was specified as describing elements of a procedure and its related findings.

(ii) Meta data related to any procedure/test was separated from the rest of the model specific to recordings of the PAP smear.

(iii) The model hierarchy was changed to reflect a general pattern to any clinical procedure or test with some data specific to the PAP smear procedure.

(iv) The labels and descriptions of the archetype terms were modified to reflect more clearly their intent. Suggestions from the evaluators were taken into consideration when resolving ambiguous terms. E.g. the term ‘Clinical’ was modified to ‘Clinical Notes’ to describe the ‘Clinical history sent with the specimen’. A tabulated comparison of some of the changes made when revising the model is stated in Table 5 below.

**Revised model:** Figure 2 presents the revised version of the original ‘histology pap’ archetype model. The meta data common to any procedure was included in the same model to maintain some level of consistency with the original model.

**Summary of changes:**

- Any event
  - [  ] Lab test meta data
  - [  ] Quality
    - [  ] Specimen
      - [  ] Insufficient specimen
      - [  ] Damaged specimen
      - [  ] Contaminated specimen
    - [  ] Labelling
      - [  ] Side unlabelled (container labelled)
      - [  ] Container unlabelled (side labelled)
    - [  ] Specimen
      - [  ] Preparative
        - [  ] Incorrect preservative
        - [  ] Insufficient preservative
  - [  ] Procedure meta data
  - [  ] Sample identification
  - [  ] Data received by laboratory
  - [  ] Methodology

- [  ] Archetype term
  - [  ] Disease entity
  - [  ] Macroscopic
  - [  ] Microscopic
  - [  ] Clinical history and examination
  - [  ] Test results
  - [  ] Observational findings
  - [  ] Specific pathological report

**Table 5: Sample of changes made to the original model in the revised version of the histology pap model**

<table>
<thead>
<tr>
<th>Original Model</th>
<th>Revised Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low grade epithelial abnormality</td>
<td>Mild epithelial abnormality</td>
</tr>
<tr>
<td>High grade epithelial abnormality</td>
<td>Moderate epithelial abn.</td>
</tr>
<tr>
<td>Macroscopic</td>
<td>Macroscopic findings</td>
</tr>
<tr>
<td>Appearance</td>
<td>Macroscopic appearance</td>
</tr>
</tbody>
</table>

**Table 6: Improved matches in SNOMED CT for archetype terms. Matches include SNOMED category and code.**

<table>
<thead>
<tr>
<th>Archetype term</th>
<th>Appropriate matches in SNOMED CT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mild epithelial abnormality</td>
<td>Cervical intraepithelial neoplasia grade 1 (disorder) [285836003]</td>
</tr>
<tr>
<td></td>
<td>Mild epithelial dysplasia (morphologic abnormality) [33288004]</td>
</tr>
<tr>
<td>Clinical notes with specimen</td>
<td>Clinical history and observation findings (finding) [250171008]</td>
</tr>
<tr>
<td></td>
<td>Clinical history/examination observable (observable entity) [363788007]</td>
</tr>
<tr>
<td>Microscopic findings</td>
<td>Microscopic specimen observation (finding) [395538009]</td>
</tr>
</tbody>
</table>

**Re-evaluation of revised model:** The revised model was sent to the MoST system and the results were evaluated by 5 clinical experts. The accuracy of the results increased from 64.7% to 80.55% (29/36*100). In addition, the scores given by the experts were higher than in the original model, thereby increasing the confidence rating from an average of 6/10 to 9/10. There was also much better agreement on the mappings between the experts with all of them agreeing on the same term 93.1% of the times, as
against 48.57% times previously. The high ratings were primarily due to the unambiguous structure and use of terms in the model, as well as because of much clearer term definitions, which clearly stated the intended use of the term in the model. Some of the matches returned are shown in Table 6.

Discussion

The paper highlighted the problems encountered when attempting to semantically integrate clinical content of two different models i.e. the data models and the terminology models. More importantly the paper focused on the importance of good modeling techniques which should be adopted by data modelers to ensure that the ‘source’ of the integration process establishes sound and safe mapping results. Therefore, before being overly critical of terminology models it is imperative to take stock of the validity and semantic appropriateness of the data models that will be increasingly used to capture clinical data in health setups.

Although a specific data modeling formalism and terminology system were chosen for the study, the concerns are common to any other modeling formalism such as HL7 messages, ICD, and others. The formalism isn’t the key focus of the study. Rather it is the modeling strategy and intended use and semantics of the data terms, irrespective of the syntax and schema utilised for their representation, which is of prime concern.

Conclusion

The issue of ensuring good modeling designs and strategies which are unambiguous in representation and intent was discussed in the paper. The main reason for the discussion was to encourage good modeling practice to (i) enable clinicians’ record patient data unambiguously, and (ii) establish reliable and accurate coding of data to standardise the data recorded in EHRs.

Upon re-evaluation of both the archetype models and the mapping results, the experts found it simpler to understand the revised version of the ‘histology pap’ archetype model. Also the mapping accuracy increased from 64.7% to 80.5%. The higher accuracy score meant that not only did the term matching tool, called MoST, return more appropriate SNOMED codes but also that more clinical experts agreed on the relevancy of the codes. A formal guideline needs to exist which specifies the salient features to follow when modeling procedure and finding type data models to help guide the model authoring process. Therefore, it is critical for any work which brings together two or more structurally and compositionally different models to concentrate not only on the end result but also the source and the means to that end to ensure quality and precision.

Acknowledgement

This work is supported in part by the EU Funded Semantic Mining project and the UK MRC CLEF project.

References

13. Qamar R, Rector AL. Semantic Issues in Integrating Data from Different Models to Achieve Data Interoperability. Medinfo Conference 2007, Brisbane, Australia (Accepted).