Clustering clinical models from local electronic health records based on semantic similarity

Kirstine Rosenbeck Goeg, Ronald Cornet and Stig Kjaer Andersen

Linköping University Post Print

N.B.: When citing this work, cite the original article.

Original Publication:
http://dx.doi.org/10.1016/j.jbi.2014.12.015
Copyright: Elsevier
http://www.elsevier.com/

Postprint available at: Linköping University Electronic Press
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-118874
Clustering local electronic health record content based on semantic similarity

Authors and affiliations:

Kirstine Rosenbeck Gøeg\textsuperscript{a}, Ronald Cornet\textsuperscript{b,c}, Stig Kjær Andersen\textsuperscript{a}

\textsuperscript{a}Aalborg University, Department of Health Science and Technology, Fredrik Bajers Vej 7D2, 9220 Aalborg, Denmark

\textsuperscript{b} Academic Medical Center – University of Amsterdam, Department of Medical Informatics, P.O. Box 22700, 1100 DE Amsterdam, The Netherlands

\textsuperscript{c} Linköping University, Department of Biomedical Engineering, SE-581 83 Linköping, Sweden

Corresponding Author:

Kirstine Rosenbeck Gøeg, PhD Fellow
Aalborg University
Department of Health Science and Technology
Fredrik Bajers Vej 7
Room: C1-217
DK - 9220 Aalborg Ø
Phone:+45 9940 3710

e-mail:kirse@hst.aau.dk

Keywords: Computerized medical records, Semantics, SNOMED CT, Medical Record Linkage/standards, Medical Record Linkage/methods, Algorithms
Abstract

[Background] Clinical models in Electronic Health Records (EHR) are typically expressed as templates which support the multiple clinical workflows in which the system is used. The templates are often designed using local rather than standard information models and terminology, which hinders semantic interoperability. Semantic challenges can be solved by harmonizing and standardizing Clinical models. However, methods supporting harmonization based on existing clinical models are lacking. One approach is to explore semantic similarity estimation as a basis of an analytical framework. Therefore, the aim of this study is to develop and apply methods for intrinsic similarity-estimation based analysis that can compare and give an overview of multiple clinical models.

[Method] For a similarity estimate to be intrinsic it should be based on an established ontology, for which SNOMED CT was chosen. In this study, Lin similarity estimates and Sokal and Sneath similarity estimates were used together with two aggregation techniques (average and best-match-average respectively) resulting in a total of four methods. The similarity estimations are used to hierarchically cluster templates. The test material consists of templates from Danish and Swedish EHR systems. The test material was used to evaluate how the four different methods perform.

[Result&discussion] The best-match-average aggregation technique performed better in terms of clustering similar templates than the average aggregation technique. No difference could be seen in terms of the choice of similarity estimate in this study, but the finding may be different for other datasets. The dendrograms resulting from the hierarchical clustering gave an overview of the templates and a basis of further analysis.

[Conclusion] Hierarchical clustering of templates based on SNOMED CT and semantic similarity estimation with best-match-average aggregation technique can be used for comparison and summarization of multiple templates. Consequently, it can provide a valuable tool for harmonization and standardization of clinical models.
Introduction

Semantic interoperability is a highly desired characteristic of Electronic Health Record Systems. To this end, standardization of information models and terminologies is needed. However, going from local customizability to global standardization is a challenge, especially in terms of modeling and managing Clinical Models (CMs) because this is the place where local clinical requirements are expressed in computerized form. CM is a relatively new construct resulting from the fact that modern EHR architectures separate reference information models from clinical models, these are called two-level modeling approaches [1,2]. CMs define documentation structures used in clinical situations such as physical examination, nutrition screening or vital signs measurement, and for each clinical situation CMs can be bound to relevant terminology [3]. CMs are often referred to as either templates or archetypes or both. In this study, the word template is used in its common meaning as a structure intended for data entry for a specific clinical situation, i.e. defining the fields on the interface level not at the database level. Consequently, “template” does not refer to any standard such as openEHR or HL7, who have their own definitions of templates. A variety of CMs are needed to handle clinical documentation needs which make modeling and managing CMs complex. Getting an overview of the complexity requires insight, which can be gained by analyzing semantic similarities of existing templates.

For example, a vital sign template at one hospital could contain pulse, blood pressure, temperature, oxygen saturation and respiration frequency, each being a text field where quantities as well as comments could be written. Another hospital could have a template where quantities, comments and protocol-related fields are kept separately. An example of a pulse excerpt is shown in Figure 1. Manual comparison of the templates gives an idea about the semantic content of a vital signs template, and we can characterize the differences between the templates in natural language. Based on this analysis, we would be able to give guidance to hospitals that want to create new vital signs templates or suggest changes to existing templates that would support harmonization. However, imagine the case where there are ten different vital sign templates possibly expressed in different languages and we want to analyze semantic content, similarities and differences and make suggestions for a national or an international standard. The complexity of the material and the labor of a manual analysis make the task overwhelming, given the large number of needed pair-wise comparisons and the challenge of synthesizing these. Consequently, analyzing existing CMs requires an automated or at least semi-automated method. If such a method could be developed, it would be valuable at a local, national as well as an international level.
Figure 1 – The pulse-section of two vital sign templates as they could be defined in two different organizations

At the local level, requirement engineering is difficult and time consuming due to the complexity of the health care domain [4]. Reusing CMs, like templates for physical examinations or nutrition screening, could speed up the requirement engineering process. However, overcoming the lack of acceptance of templates developed elsewhere, known as the “Not invented here” syndrome, is a challenge. Reuse might also be a challenge because EHR-system failure has been associated with inability to support the micro detail of clinical work [5]. The result is that there is an unknown diversity of CMs used in clinical practice. In this context, analysis of differences and similarities between hospitals and departments could provide insight on whether harmonization is beneficial and/or possible. Moreover, given a better overview, design of new templates could take its point of departure in existing ones. E.g. if a group of templates all intended for physical examinations are known a canonical model can be developed on this basis. The next time a physical examination template is designed the canonical model can be used as point of departure, hence ideally creating harmonization and avoiding duplication of effort. A canonical model can also be used as a point of reference for similarity of different templates.

Nationally, health provider organizations and medical societies strive to manage health care by balancing resource management and treatment quality. One approach is development and implementation of clinical guidelines and national integrated care pathways to ensure a high and uniform quality of care. The feasibility of guidelines and pathways depend on uniform documentation procedures and quality indicators, hence, harmonized templates are beneficial. Medical societies also have an interest in harmonized documentation, because, in many cases, clinical research depends on uniform information. Harmonization could be supported by overviews of existing templates on a national level. However, no such overview exists, and getting it requires a way to compare templates that are currently expressed using local proprietary information models.

Internationally, different approaches to clinical modeling exist. They are aimed at developing, refining, implementing, and evaluating information models to ensure clinical involvement as well as semantically-interoperable systems [1,2,6-10]. Recently, an analysis criticized that many existing clinical modeling
approaches violate good modeling practice since they fail to model the requirements of the health care domain using a consistent healthcare-specific ontology [11]. It can be questioned, whether the analysis takes into account that requirement engineering processes are not the main scope of all the different clinical modeling approaches. However, the general conclusion that standardized models maybe are too distant from health care practice and actual clinical information systems might be supported by the fact that the adoption of standards, apart from DICOM, is slow[12] and there is a limited progress towards full semantic interoperability [13]. Developing bottom-up approaches for international clinical modeling might help adoption of these models. As for the national level, this requires overview and comparison of existing clinical documentation templates. However, language barriers increase the complexity of the challenge. Beside bottom-up approaches, semantic similarity analysis might also be relevant in getting an overview of existing clinical models in internationally available repositories such as the openEHR clinical knowledge manager [14], the clinical element model browser from Intermountain Healthcare, the Australian clinical knowledge manager [15] and HL7 FHIR resources. Stakeholders in the international modeling community are also concerned with information model harmonization and have joined forces in CIMI (Clinical Information Modeling Initiative) [16]. In such harmonization efforts, overview of existing CMs could also be useful.

Summing up, semantic similarity analysis of CMs could be valuable for a number of local, national and international applications. Therefore, the aim of our study was to develop a method for CM comparison. The method should be able to compare and give an overview of multiple CMs whether these are local templates or standardized information models. Comparison is challenged by lexical differences. Therefore, it is necessary to base the comparison on stable concept definitions. In this study, SNOMED CT is chosen based on its coverage and flexibility compared to other terminologies [17-20]. In addition, SNOMED CT has been tested in different clinical fields [21-23]. This means that a common semantic reference can be obtained. To be able to automate the method, semantic similarity estimation is used as a means to analyze similarities and differences. This is expanded on in the background section.

Background: Semantic similarity estimation in biomedical informatics
A semantic-similarity estimate can be understood as a numerical value reflecting the closeness in meaning between two terms or two sets of terms [24]. Both term similarity and set-of-term similarity are examined in the following.
Semantic similarity between two terms

Generally, semantic-similarity estimates are classified according to the underlying theoretical principles and the knowledge sources used. [25] Knowledge sources can be domain corpora, ontologies/taxonomies and thesauri. Theoretical principles denote whether the estimate is based on edges or on information content (IC). Edge-based estimates are based on the number of edges between two terms and variations thereof. An edge is the links between two terms e.g. if cow and pig are both mammals then the number of edges between cow and pig would be two (1:pig-mammal, 2:mammal-cow). IC-based measures are based on the IC of the two terms in question and variations thereof. The IC of a term is the logarithm of the probability of finding the term in a given corpus.

More than in other domains, semantic similarity estimation is often based on ontology in biomedical informatics. Explanations are that general-purpose resources like WordNet have limited coverage of biomedical terms [26], and that biomedical informatics has many available concept systems (e.g. Read codes, LOINC and SNOMED CT) [25]. Even though some of the available concept systems are not ontologies in the strict sense, they are used as such in some similarity estimation research e.g. Read codes in [27].

An estimate based solely on an ontology is called intrinsic. Intrinsic methods were the focus of a combined study and review done by Sánchez et al in 2011 [25]. Their study focused on systematically reviewing and re-formulating edge-based and IC-based semantic similarity estimates in an intrinsic information-theoretical context. The estimates reviewed were both edge-based [28,29] and IC based [30,31]. They also developed a method so that they could approximate set-theory estimates in terms of IC. The similarity estimates were evaluated using SNOMED CT and a reference set of 30 medical term pairs. In a previous study, the reference term pairs had been rated by physicians and coders in terms of their similarity [26]. An average based on these ratings serves as “gold standard” in Sánchez et al’s study, because the ratings can be interpreted as a quantification of experts’ perception of similarity. Sánchez et al’s study shows that classic edge-based and IC-based semantic similarity estimates improve their correlation with the expert ratings when re-formulating them from corpora-based to intrinsic. In addition, some of the similarity estimates taken from set-theory outperform classic similarity estimates in terms of correlation with the expert ratings. The basis of most of Sánchez et al’s estimates is the IC shown in equation (1).

\[
IC(c) = -\log p(c) \cong -\log \left( \frac{|\text{leaves}(c)| + 1}{\frac{|\text{subsumers}(c)|}{\text{max_leaves}} + 1} \right)
\]
In this equation leaves(c) is the set of concepts found at the end of the taxonomical tree under concept c. This can also be expressed as the descendants of c that do not have any children themselves [32]. Subsumers(c) is the complete set of taxonomical ancestors of c including itself. Max_leaves is the number of leaves of the least specific concept (the root concept). In a SNOMED CT context this means the number of leaves of 138875005 | SNOMED CT Concept |

In Sánchez et al’s study, the best agreement between expert similarity scores and similarity estimates is obtained when applying information content (IC) based similarity measure re-formulated from the set-theory estimate first published by Sokal and Sneath[25]. This is shown in equation (2).

\[
sim(c_1, c_2) = \frac{IC(LCS(c_1, c_2))}{2 \times (IC(c_1) + IC(c_2)) - 3 \times IC(LCS(c_1, c_2))}
\]

(2)

In this equation c1 and c2 are the two concepts of interest and LCS is the least common subsumer which means the most specific taxonomical ancestor common to c1 and c2. IC is estimated using equation (1).

When comparing the estimate in equation (2) with classic IC-estimates like Lin’s [30], which is shown in equation (3), it can be noted that it consists of the same components namely the IC of the two concepts and IC of LCS.

\[
sim(c_1, c_2) = \frac{2 \times IC(LCS(c_1, c_2))}{IC(c_1) + IC(c_2)}
\]

(3)

The presented similarity estimates always result in a number in the range [0; 1].

One possibility when comparing two sets of concepts is to compare each concept in the first set with each concept in the second set. For two sets with a magnitude of 10-50 concepts, this result in a similarity matrix containing 100-2500 similarity estimates. If detailed analysis of differences and similarities are required, similarity matrices might be applicable; however, for overview purposes, simpler estimates are required. Therefore, semantic similarity estimation between sets of concepts is examined in the next section.

**Semantic similarity between two sets of concepts**

Pesquita et al. have reviewed techniques in gene product comparison based on Gene Ontology (GO) annotation, which is a specialization of the problem of semantic comparison of sets of concepts. Their
classification of methods to find gene product similarity helps getting an overview of possible approaches [24]. In the following, the classification is presented in general terms instead of GO-specific.

- **Group-wise (set, graph or vector approaches).** Sets of concepts are compared directly without calculating individual similarities between concepts. In set approaches, overlap between sets is used as an estimate of similarity. In graph approaches the concepts of each set are represented as sub graphs of the original ontology and graph matching or similar techniques are used for comparison. In vector approaches a set of concepts is represented as a vector with each dimension representing a concept in the original ontology. E.g. each coordinate of vectors can be binary, denoting absence or presence of a term.

- **Pair-wise (all pairs or best pair approaches).** Given a pair-wise comparison of concepts i.e. the similarity matrix, the pair-wise approaches propose ways to aggregate the similarity estimates in the similarity matrix. The all-pairs methods use MIN, MAX or AVG functions. The best-pairs methods takes the AVG of the maximum values in each set’s directions, see equation (4) as proposed among others by [33]. In other words, given a similarity matrix the maximum value of each row and each column is found. All maximum values are added and normalized using the number of concepts in the sets.

\[
sim(s_1, s_2) = \frac{1}{m + n} \left( \sum_{k=1..m} \text{MAX} (\text{sim}(c_k, c_p)) + \sum_{p=1..n} \text{MAX} (\text{sim}(c_k, c_p)) \right)
\]  

(4)

The method section will present how similarity estimation was used in the CM comparison.

**Material and methods**

In the following section the CM comparison method is presented. The comparison method consists of SNOMED CT representation, template comparison and hierarchical clustering. Four different similarity estimation techniques were used. To evaluate these alternatives an evaluation method is presented as well. In the evaluation method local templates are compared using the four techniques and dendrograms and receiver operating characteristic (ROC) curves are used as outcome measures.

**Clinical Model comparison method**

**Template comparison**

Choosing intrinsic semantic similarity estimation as technique requires a simplified view of a template specification. Templates were considered as sets of SNOMED CT concepts which meant disregarding
structural information, data type, interface terminology etc. Post-coordinated expressions were split into their source concepts ignoring the attribute relationship concept e.g. the postcoordinated expression 118236001 | ear and auditory finding |:418775008 | finding method | = 76517002 | endoscopy of ear | would be split to 118236001 | ear and auditory finding | and 76517002 | endoscopy of ear |. Concepts that could not be mapped to SNOMED CT were not subject of further analysis.

Two information-content-based similarity estimates, Lin, see equation (3), and Sokal & Sneath (SoSn), see equation (2), were chosen for this study. A pair-wise combination technique was chosen to ensure that comparison was based on all aspects of the template concepts, not just the best match or the worst match (MIN or MAX approaches). Both all-pair comparison (AllAVG) and best-pair comparison (BestAVG), equation (4), were used.

The template comparison was done for each template pair for each of the four chosen techniques: Lin/AllAVG, Lin/BestAVG, SoSn/AllAVG and SoSn/BestAVG. The template comparison was implemented in JAVA using NetBeans. The input was templates expressed as Sets of SNOMED CT concepts. The June 2012 release of SNOMED CT was used. The text files distributed by the Danish national release center were implemented in a MySQL database. To improve performance, “number of leaves” was calculated for all concepts in SNOMED CT and stored in the database in advance. The output of the template comparison was a template-similarity matrix for each of the four chosen techniques. For the pairwise comparison of n templates, the template-similarity matrix consists of n² cells, with the diagonal being the comparisons of templates with themselves (hence similarity = 1) and cells under the diagonal being duplicates, as similarity is symmetric. These template-similarity matrices were the point of departure for the hierarchical clustering.

**Hierarchical clustering and dendrograms**

The goal of the analysis was to describe sub-clusters, because groups of templates are typically characterised as such. E.g., a hospital can formulate a general physical examination template and make specialisations for departments with special needs like the children’s department or the psychiatric ward. This was the reason why a hierarchical clustering method, as described in [36], was chosen. Hierarchical clustering can be visualized using dendrograms, which are easy to interpret and powerful in terms of clustering similar content without assuming a defined number of clusters or defining a classifier. Hierarchical clustering is based on grouping the most similar templates first and continuing until all templates are joined together. Joining the first two templates based on a similarity estimate is straightforward. However, there are different methods for determining the similarity between the now formed subgroup and the rest of the templates. Typical methods are nearest neighbour, which uses the minimal distance, farthest neighbour, which uses the maximum distance, and compromises that use
average or mean distance. In this study, the average distance methodology was chosen, where, since the study was done in a similarity context, 1-sim was used as a distance measure. The average similarity was chosen because it is a reasonable approach when there is no particular assumption regarding the shape of the clusters. The concept of “cluster shape” is meaningless (or at least very difficult to interpret) in a template similarity context. The hierarchical clustering method and dendrogram visualisation were implemented in Matlab using built-in pattern recognition functionality. The template-similarity matrices were taken as input, and the output was a dendrogram for each of the four techniques.

**Evaluation method**

The aim of the evaluation was to compare the four approaches Lin/AllAVG, Lin/BestAVG, SoSn/AllAVG and SoSn/BestAVG when applied in EHR-content analysis. The approaches were compared based on their ability to group physical examination templates and discriminate them from other types of templates.

**Material: Templates from Danish and Swedish EHR systems**

It is not possible to study the templates directly since they are proprietary models, and therefore different between the EHR-systems. Therefore, screen forms and locally produced requirement specification material was acquired from five different sites. The screen forms for this study were chosen, so that they could be separated into two groups that would make it possible to evaluate the content analysis method. These two groups were: “physical examination templates” and “other”. First, we chose a group of physical examination templates from different organisation and different specialities, i.e. a group that we would expect would cluster together. Afterwards, we chose a group of templates where the clinical focus was distinct from physical examination and where each should be different from the others, i.e. creating different reference points that would not cluster very closely with either physical examination or each other. The templates are presented in Table 1.

<table>
<thead>
<tr>
<th>Label</th>
<th>Purpose</th>
<th>Organisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NordCOPD</td>
<td>Out-patient follow-up regarding Chronic obstructive pulmonary disease (COPD) including e.g. measurement of forced expired volume using spirometry, inhalation therapy education and body mass index. Documented by physicians.</td>
<td>Lung departments in Region Northern Jutland, Denmark</td>
</tr>
<tr>
<td>NordExam</td>
<td>Physical examination including e.g. finding of head and neck, cardiac auscultation and neurological</td>
<td>All departments, Region Northern Jutland,</td>
</tr>
<tr>
<td>Code</td>
<td>Description</td>
<td>Location</td>
</tr>
<tr>
<td>------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>NordOrgan</td>
<td>Organ system walkthrough including central nervous system and gastrointestinal findings. Documented by doctors as a part of the patient history interview on admission.</td>
<td>All departments, Region Northern Jutland, Denmark</td>
</tr>
<tr>
<td>NordSocialNurse</td>
<td>Social status of patient including e.g. partnership status, occupational history and language findings. Documented by nurses on admission.</td>
<td>All departments, Region Northern Jutland, Denmark</td>
</tr>
<tr>
<td>NordStatusNurse</td>
<td>Nursing status of patient including e.g. skin, pain and nutrition findings. Documented by nurses multiple times during admission.</td>
<td>All departments, Region Northern Jutland, Denmark</td>
</tr>
<tr>
<td>OdenseAdmission</td>
<td>Admission to hospital information including e.g. Consent status for record sharing and patient history interview. Documented by physicians.</td>
<td>All departments, Odense University Hospital, Denmark</td>
</tr>
<tr>
<td>OdenseExam</td>
<td>Physical examination</td>
<td>All departments unless a special template is developed, Odense University Hospital, Denmark</td>
</tr>
<tr>
<td>OdenseExamEye</td>
<td>Physical examination for an eye department. In addition to a general physical examination (see above) specialized eye-related findings can be documented by physicians on admission.</td>
<td>Eye department, Odense University Hospital, Denmark</td>
</tr>
<tr>
<td>ÖstergötlandExam</td>
<td>Physical examination</td>
<td>All departments unless a special template is developed, hospitals in Östergötland county, Sweden</td>
</tr>
<tr>
<td>ÖstergötlandExamChild</td>
<td>Physical examination for a paediatric department. In addition to a general physical examination (see above) specialized findings e.g. puberty state and birth weight can be documented by physicians on admission.</td>
<td>Children department, hospitals in Östergötland county, Sweden</td>
</tr>
<tr>
<td>ÖstergötlandExamNeo</td>
<td>Physical examination for a neonatal department.</td>
<td>Neonatal department,</td>
</tr>
<tr>
<td>Template</td>
<td>Description</td>
<td>Location</td>
</tr>
<tr>
<td>---------------------------</td>
<td>--------------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>ÖstergötlandExamPsy</td>
<td>Physical examination for a psychiatric department. In addition to the general physical examination from Östergötland specialized findings e.g. puberty state and birth weight can be documented by physicians on admission.</td>
<td>Hospitals in Östergötland county, Sweden</td>
</tr>
<tr>
<td>RandersExam</td>
<td>Physical examination</td>
<td>(General template)</td>
</tr>
<tr>
<td></td>
<td>Used in lung department, Randers hospital, Denmark</td>
<td></td>
</tr>
<tr>
<td>UppsalaExamHaema</td>
<td>Physical examination</td>
<td>(General template)</td>
</tr>
<tr>
<td></td>
<td>Used in haematological department, Uppsala, Sweden.</td>
<td></td>
</tr>
<tr>
<td>UppsalaExamOrth</td>
<td>Surgical departments. Including e.g. blood pressure and respiration findings. Documented by physicians.</td>
<td>Orthopaedic department, Uppsala hospital, Sweden.</td>
</tr>
</tbody>
</table>

**SNOMED CT representation of templates**

To be able to compare templates, they were structured in accordance with a clinical content format [34]. In *Fel! Hittar inte referenskälla.*, the clinical content format is simplified to the most important classes, relationships and cardinalities. In the clinical content format, a template can have a number of fields, each of which is assigned a data type and a SNOMED CT concept. We did not have semantic data types such as ISO 2109 [ref] available because our models came from local organisations and our analysis of their models was based on user interfaces and local documentation (word documents). The data type only distinguished whether it was a text, number or a value set. Each field can have only one data type, but due to post-coordination each field can have several SNOMED CT concepts. The structured template information was stored in a database, and the interface terminology was mapped to SNOMED CT. The interface terminology consisted of the terms found on the user interfaces in the EHR-systems. The mapping was performed while formulating a set of guidelines to ensure consistent mapping [35]. This meant that even though there were two coders, no inter-rater agreement score could be calculated. However, since the purpose of the
guideline study was to ensure consistency, the templates can be considered very similar in terms of mapping-approach. This ensured that the similarity estimation in fact measured differences in content and not differences in mapping approach.

Figure 2 - The structuring process from local material to a clinical content format

Outcome measures
The outcome of the analysis of the templates was four dendrograms, and they were compared based on a description of topology to see what semantic characteristics of the templates were emphasised by the different approaches. In general, dendrogram comparison can be based on labelling, topology and heights [37,38]. However, direct height comparison is a questionable method when the heights are based on different metrics or different algorithms [38], and labelling was not examined since this is merely interesting if the identity of entities is unknown. In addition to this semi-quantitative evaluation, a simple classification was performed aimed at separating physical examination templates from other templates. Using the hierarchical clustering, a “physical examination cluster” was identified for all possible cluster-configurations. The ROC-curves (1-specificity, sensitivity) of the 4 methods were plotted for comparison.
**Results**

In Table 2 the result of the SNOMED CT mapping of the 15 templates is illustrated.

Table 2 - Result of SNOMED CT mapping

<table>
<thead>
<tr>
<th>Label</th>
<th>Fields</th>
<th>mapped</th>
<th>Post coordinated expressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>NordCOPD</td>
<td>77</td>
<td>67</td>
<td>20</td>
</tr>
<tr>
<td>NordExam</td>
<td>16</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>NordOrgan</td>
<td>8</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>NordSocialNurse</td>
<td>12</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>NordStatusNurse</td>
<td>15</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>OdenseAdmission</td>
<td>53</td>
<td>41</td>
<td>2</td>
</tr>
<tr>
<td>OdenseExam</td>
<td>27</td>
<td>26</td>
<td>5</td>
</tr>
<tr>
<td>OdenseExamEye</td>
<td>74</td>
<td>55</td>
<td>21</td>
</tr>
<tr>
<td>ÖstergötlandExam</td>
<td>49</td>
<td>47</td>
<td>3</td>
</tr>
<tr>
<td>ÖstergötlandExamChild</td>
<td>72</td>
<td>66</td>
<td>9</td>
</tr>
<tr>
<td>ÖstergötlandExamNeo</td>
<td>56</td>
<td>50</td>
<td>8</td>
</tr>
<tr>
<td>ÖstergötlandExamPsy</td>
<td>50</td>
<td>43</td>
<td>5</td>
</tr>
<tr>
<td>RandersExam</td>
<td>18</td>
<td>17</td>
<td>2</td>
</tr>
<tr>
<td>UppsalaExamHaema</td>
<td>35</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td>UppsalaExamOrth</td>
<td>7</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>569</strong></td>
<td><strong>497</strong></td>
<td><strong>76</strong></td>
</tr>
</tbody>
</table>
Figure 3 - Lin/AllAVG

Figure 4 - SoSn/AllAVG
Figure 5 - Lin/BestAVG,

Figure 6 – SoSn/BestAVG
When comparing the dendrograms, it can be observed that the aggregation technique affects the result more than the similarity estimate chosen. At a glance, the AllAVG technique (Figure 3 and Figure 4) is outperformed by the bestAVG technique (Figure 5 and Figure 6). This is further highlighted by the area under the ROC-curve (AUC) which is illustrated in Figure 7. The area under the curve is much larger for the BestAVG than AVG.

In the best match average dendrograms the topology is almost the same. Both BestAVG dendrograms cluster physical examinations, only the UppsalaExamOrth connects with other templates before the physical examination template cluster. Looking at the template description in Table 1 and the mappings in Table 2, it can be seen that the UppsalaExamOrth only consists of a few fields with coarse-grained information content. In addition, actually looking at the dendrograms in *Fel! Hittar inte referenskälla* and *Fel! Hittar inte referenskälla* reveals that UppsalaExamOrth is grouped with other coarse-grained templates with few fields. Consequently, the grouping probably indicates that UpssalaExamOrth is not a very typical physical examination rather than UpssalaExamOrth being subject to an incorrect clustering. The only thing that separates SoSN/BestAVG from SoSn/BestAVG is that OdenseAdmission is grouped with the physical examination cluster before the above mentioned “coarse-grained” cluster for Lin/BestAVG and after the “coarse-grained” cluster for SoSn/BestAVG. Consequently, the SoSn/BestAVG performs slightly better from an AUC perspective because UpssalaExamOrth is in the “coarse-grained” cluster.
Figure 7 - ROC curve. From the bottom: Lin/AllAVG (turquoise, AUC=0.71), SoSn/AllAVG (red, AUC=0.78), Lin/BestAVG (green, AUC=0.96) and SoSn/BestAVG (blue, AUC=0.98).

Discussion

Our results showed that semantic similarity estimation with BestAVG aggregation technique was able to cluster similar templates using hierarchical clustering and dendrograms. The BestAVG technique outperformed AllAVG. Similarity estimation was based on SNOMED CT and intrinsic Lin and SoSn estimates respectively.

Strengths and weaknesses

We chose to simplify templates to make it possible to apply semantic similarity techniques. The simplification included ignoring information about the structure and data types of the templates, ignoring concepts that could not be mapped to SNOMED CT and splitting post coordinated expressions while ignoring the attribute relationships. In a similarity estimation perspective, information about data type does not make much sense to introduce in an analysis. Some structural issues may arise because CMs can be complex and have a highly nested structure which means that terminology bindings attached to inner fields may have their meaning changed by the data group definition. E.g. the data group "family history" would change the meaning of the inner field “diagnosis”. The evaluated templates were not highly nested, but for other CMs handling this axis modification problem might improve the precision of the comparisons.
One way of approaching this would be to take into account the SemanticHealthNet work on ontology patterns[39]. The terminology related simplifications may have introduced a bias in the study since 13% of the interface terms could not be mapped to SNOMED CT and 13% were post coordinated expressions. Instead of not mapping terms to SNOMED CT, we could have tried to map to more general concepts. This could give a more accurate result because super concepts carry many of the same semantic features as sub concepts, and also in terms of number of terms analyzed. However, choosing super concepts could result in overestimation e.g. if a granular concept e.g. “ECG findings” was mapped to a coarse grained concept like “heart findings”, and “heart findings” was found in other templates, a similarity of 1 would be wrongfully identified. An alternative would be to represent the unmapped concepts with the root concept, but this would result in similarity 1 when unmapped concepts are compared to each other. To make a conservative estimate, all unmapped concepts would have to be represented with a non-SNOMED CT identifier and every time this identifier was compared to any other concept the similarity should be manually set to zero.

A more accurate representation of post-coordinated expression would require the similarity estimation to analyze semantic features other than the SNOMED CT IS-A hierarchy. As explained in e.g. [40] both pre-coordinated and post coordinated terms can be translated to a normal form using the SNOMED CT content model and a number of rules and guidelines. Each SNOMED CT expression would then consist of a focus concept and a number of attribute relationships. Finding a meaningful semantic similarity estimate based on normal form would be challenging because similarity of each attribute depend on the focus concept e.g. endoscopy of ear and endoscopy of gastric track is not similar in any normal sense just because they are both are endoscopies. Consequently, adding semantic features to the similarity analysis would increase the complexity of the analysis considerably.

The similarity estimate was chosen in accordance with the findings of Sánchez [25], showing that the SoSn estimate performed better than other estimates in terms of accordance with human perception of similarity. However, the use of the SoSn estimation in a biomedical informatics context was new and we questioned whether the SoSn correlation with human perception of similarity would make a difference in our study. Therefore Lin’s estimate, equation (3), was chosen as well. Even though the topology was almost the same for the two BestAVG dendrograms it cannot be concluded from this study that it does not matter whether Lin or SoSn similarity estimates are chosen. The heights of the dendrograms vary, the AUC is slightly better for SoSn, and for other applications or aggregation techniques there may be larger differences in topology, as it can be seen from the AllAVG dendrograms. Similar performance of Lin and SoSn estimates could be explained by the strong correlation given that they are both IC based.
In this study, we chose two aggregation techniques all-pair AVG and Best-pair AVG. In a GO-specific context, best-pair average methods tend to outperform other pair-wise combination strategies [24]. However, in a Read Code based study [27], the MAX and AVG functions using Lin and Resnik similarity estimation yielded the clearest clusters in a PCA approach. They did not try a best-match average approach.

No studies are found where SNOMED CT based similarity estimates were compared using a pair-wise technique. Therefore, based on the finding of [24,27] respectively, both all-pair average and best-pair average techniques were explored. The evaluation showed that the aggregation technique affects the result more than the similarity estimation. Looking at the dendrograms the differences in clustering between best-match-average and average can be explained by the fact that the AVG technique gives as much weight to concepts that differentiate two templates as to the concept that are similar. For the AVG dendrograms this means that small templates are likely to be grouped together, just because they do not have so many differences. In addition, the weight on differences means that the AVG technique tends not to group physical examination templates. The reason for this is that the specialised content in specialized physical examination templates differentiates them from the general physical examination templates. In contrast, the BestAVG technique mostly weighs the similarities and groups templates into Swedish and Danish templates and general and specialized ones, and sorts out those which do not have much in common with physical examination templates. This logical grouping is exactly what we hoped to achieve.

The different characteristics of AllAVG and BestAVG methods could maybe have a value in future work; however, for the application in a content analysis context BestAVG will most likely outperform AllAVG.

**Strengths and weaknesses compared to other studies**

The evidence in the field of similarity estimation in the field of CMs, standardization and semantic interoperability is scarce. Actually, only three studies are found in which CMs are compared. In a study by Dugas et al., no semantic similarity estimate is used, it is a simple set-based approach where the number of terms that the templates have in common is used as a metric. The metric is used in a hierarchical clustering approach using dendrograms [41]. In a study by Allones et al., SNOMED CT based semantic search of archetypes is developed. One application of the semantic search is that overlap between archetype content can be detected. The structure of SNOMED CT is used as a resource to enrich the search [42]. In the third study by Gøeg et al., SNOMED CT is used to determine similarities and differences in physical examination templates using both full matches and terminology matches deduced from the structure of SNOMED CT [43]. The contribution of the present study compared to these earlier approaches is that intrinsic similarity estimation is introduced to the field of content analysis which makes semantic similarities quantifiable. This means that the clustering approaches such as the study by Dugas et al. [41] can be expanded with similarity estimation information.
In the evaluation, we chose to include 15 templates, which is comparable to the related studies where the sample size is 4 [43], 7 [41], and 25 [42] respectively. We chose the relatively limited number of templates to make the analysis transparent, which in our opinion is important in this methodological oriented study. Table 1 with the template descriptions serve as a qualitative reference point, so that the value of the dendrograms can be seen in this perspective. Increasing the number of templates significantly would make this methodological transparency impossible. However, in an application study, increasing the number of CMs would be important.

In this study, the degree of automation is more extensive compared to our earlier study [43]. Automation is crucial in content analysis because of the number of similarity estimates calculated for a template comparison equals the product of the SNOMED CT concepts linked to each template, and the number of pair-wise template comparisons needed to perform an analysis raises with the number of templates, see formula (5) which is based on basic combinatorics.

\[ K(n,2) = \frac{n \times (n - 1)}{2} \]  

(5)

With a size comparable to our study i.e. 15 templates with 30 concepts in each template, account for approximately 900 similarity estimates per comparison and 105 comparisons which means approximately 90,000 similarity estimates calculated for the whole study. In a hospital, 15 templates would rarely be enough. Repeating the study with 200 templates would require almost 18,000,000 similarity estimates to be calculated.

Given the scarce evidence, related research is examined. The field of subject clustering based on EHR-information is of special interest. This field is closely related because a patient can be described by a set of clinical terms drawn from ontology much similar to how a template can be described by a set of terms. In addition, the same ontology-systems are typically used to describe patients and templates e.g. ICD, SNOMED CT and the UMLS which combines several terminologies. In [27], patients are described by Read Codes drawn from General practitioners’ records. These were compared using several node-based pair-wise approaches and principal component analysis (PCA). In [44], radiology reports are described using SNOMED CT and compared using an edge-based, group-wise vector approach using k-Nearest Neighbour as clustering approach. Aseervatham et al developed a UMLS-based semantic kernel for categorization of semi-structured documents including clinical observations and radiology notes. The semantic kernel was based on a combination of edge-based and node-based similarity estimates. The categorization was used to automatically assign ICD-9-CM codes [45].
CM analysis methods could draw from the methods proposed in the semantic subject clustering research i.e. apply more sophisticated clustering techniques. However, the hierarchical clustering and the dendrograms have the advantage that they do not presume a defined number of clusters or a certain classifier. The dendrograms make it clear that a template can belong to more than one cluster at the same time which is an important characteristic for CM analysis. For example, a template can both belong to the physical examination cluster and the Swedish physical examination cluster at the same time and both clusters may be important dependent on context.

Future work
Semantic overlap i.e. what is the common content of two or more CMs is one of the themes of the studies by Allones et al.[42] and Gøeg et al.[43]. It would be an interesting follow-up on this study to deduce the common content of user-defined clusters drawn from the dendrograms. For example, a user should be able to choose the cluster with the Danish physical examinations and from that selection get the common content. Common content analysis work has also been done outside the narrow scope of CMs, because common content is related to reaching consensus on the clinical practise in a field. Therefore, common content has been the object of interest of a qualitative content analysis. The qualitative content analysis is characterized by researchers labelling the content that they want to analyze [47]. The study defines a minimum nursing dataset for nutrition based on a qualitative content analysis of different nutrition documentation tools [48]. Analysing semantic overlap is an important process for standardisation purposes and semantic interoperability. Analysis of semantic overlap could be expanded by using both analysis of existing content in EHR systems and guidelines or documentation tools describing the best practise in the clinical field.

Before application, further testing will be needed to establish a solid analysis framework. Testing edge based similarity estimates and applying the methods to a larger number of templates will be logical first steps. Other potential developments could be to improve the template simplification process and develop better similarity estimation techniques for post-coordinated expressions.

Conclusion
This study proposed the use of intrinsic similarity estimation, aggregation and hierarchical clustering for CM comparison. Our evaluation showed that the two similarity estimates, Lin and Sokal & Sneath, did not notably affect the clustering. In terms of aggregation technique, best-pair average techniques outperformed all-pair average. We showed that dendrograms based on intrinsic similarity estimation and best-pair average techniques had the potential of grouping diverse templates in a way that provided
overview of the semantic characteristics of the templates. Developing common content based on the result of the analysis is an important future priority.

Acknowledgements

We would like to thank the EHR units at Odense University Hospital, Regional Hospital Randers, Region Northern Jutland, Östergötland County and Uppsala University Hospital for assisting us with access to their local EHR templates.

Competing interests

This research is part of the first author’s PhD study that is co-financed by Region Northern Jutland and CSC Scandihealth.

References


