MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Comparison of methods for evaluation of a medical terminological system

Daniëlle Arts[a,b], Nicolette de Keizer[a], Evert de Jonge[b], Ronald Cornet[a]

[a] *Academic Medical Center, Department of Medical Informatics, Amsterdam, The Netherlands*
[b] *Academic Medical Center, Department of Intensive Care, Amsterdam, the Netherlands*

Abstract

The importance of terminological systems (TS) to support standardized and structured documentation of medical data is commonly recognized. The usability of TS in real practice strongly depends on the completeness and the correctness of the content of the TS. We here present four different methods that can be applied to evaluate a TS' content. All four methods were applied in a case study. We make a comparison of 1) the results of two methods that focus on the completeness of the content and that differ in the application of the TS that they focus on and 2) the results of an automated and a manual evaluation of the correctness of the content. Finally we summarize the results of all four methods and analyze whether they overlap or complement each other.

Keywords:

Terminological systems, evaluation methods

Introduction

Several developments in health care have led to an increase in the need for accurate, detailed and structured registration of medical data. Many terminological systems (TS) have been and are still being developed to support this. We define a terminological system as a list of terms that refer to concepts that belong to a specific domain [1]. Some examples of these are the International Classification of Diseases (ICD), the Systemized Nomenclature of Medicine (SNOMED), and the Unified Medical Language System (UMLS) [2]. By the direction of the Dutch National Intensive Care Evaluation foundation our department is engaged in a continuous effort to develop a terminological system and corresponding software for the domain of intensive care (IC), DICE (Diagnoses for Intensive Care Evaluation) [3].

Reasons to evaluate terminological systems might be to provide feedback to software project developers or to justify decisions on adoption, continuation, or termination of an installed system.

Several authors have specified required characteristics of terminological systems which could be evaluated [4]. In this study we will focus on the content of a TS, i.e. the completeness and correctness of concepts, terms and relations between concepts. Physicians need to be able to be complete and sufficiently accurate in depicting the care process, and the clinical researcher needs to be able to be complete in selecting specific patient groups at each desired level of aggregation. Thus, all concepts, terms and relations belonging to the domain of the TS should be represented and the knowledge about the concepts should be correct, i.e. it needs to be compliant with the real world. For example, we do not only want sufficient terms attached to a concept, but we also want the terms to be only the right ones.

The question of how to assess the quality of a TS' content remains. Several authors have applied numerous different methods, each of which focus on either the completenss or the correctness of the TS' content. We present four methods derived from a literature review, of which two focus on (but not restrict to) the completeness and two focus on (but not restrict to) the correctness of the TS' content. These four methods have been applied in a case study to evaluate the content of the TS DICE.

The aim of this study was to analyse the extent to which the results of the four methods overlap or complement each other and the extent to which each of the methods restricts itself to the evaluation of either the completeness or the correctness of the content.Derived from this aim we analysed to what extent the two methods for evaluation of completeness of TS' content, each based on a different purpose of the TS, produce different results and would lead to different conclusions.

For the evaluation of the correctness of TS' content, the use of automated evaluation techniques, e.g. based on the semantics of the TS, is increasing. Considering these developments we compared the results produced by an evaluation method based on manual review to those produced by an evaluation method that automatically detects inconsistencies.

Methods for evaluation of terminological systems

As mentioned in the introduction of this article we focus in this study on the evaluation of the content of a TS, i.e. the concepts, the terms that describe the concepts and the (hierarchical and non-hierarchical) relations between concepts. To gain insight into methods for evaluation of TS' content that have been applied by others we performed a review of relevant journal articles. An article was considered relevant if it described the evaluation of the content of a terminology system which was developed for a medical domain. In the literature review we made a distinction between methods to evaluate completeness and methods to evaluate correctness of a TS' content.

Evaluation of completeness of TS content

The completeness of a TS' content is often evaluated through 'concept matching'. This implies that a representative subset of concepts extracted from the domain of the TS is matched to concepts in the TS [5,6]. The extent to which concepts in the subset can be matched to concepts in the TS is mostly presented as a 'concept-match score'. For example Chute et al. [5] have applied a scoring scale from 0 to 2, where 0 = no match, 1 = fair match, 2 = complete match. To be representative, the source of the subset of concepts should reflect the intended use of the TS. For example if a TS will be used by nurses for documentation of nursing information, than the subset of concepts could be well extracted from existing nursing documentation in medical records [6].

Besides the concept matching method some other methods have been applied to the evaluation of completeness of TS' content. In a study of Bodenreider et al. [7] the system being evaluated had already been in use for some time. The measure of completeness of concepts in the system was based on the number of concepts that had to be added to the system by the users due to underrepresentation in the TS.

Evaluation of the correctness of a TS content

Evaluation of the correctness of a TS' content is often based on its semantics. Many terminologies nowadays consist of more than just a simple list of terms; hierarchical and non-hierarchical relations exist between the concepts, and concepts are described by one or more terms. The relations between concepts form (a part of) the definitions of the concepts and thereby also the semantics of the TS. Analyzing the semantics may reveal inconsistent, ambiguous or redundant concepts. For example Cimino [8] used the semantics of the UMLS to detect ambiguities, redundancy and inconsistent 'parent-child' relationships. Evaluation based on semantics has the potential to be automated or semi-automated. For example a computer algorithm could detect concepts that share the exact same definitions or for concepts that were assigned to a number of semantic types, of which two are mutually exclusive. An example of the latter is that an organism can not be both a bacterium and a virus.

Case-study

Background

DICE is an implementation of the ontology described in [3]. It comprises diagnoses, which form the reasons for admission to IC, and some of their characteristics, such as the anatomical localization, the dysfunction and the etiology. We identify two types of reasons for admission: medical diagnoses, e.g. pneumonia; and surgical procedures, e.g. CABG. DICE can be incorporated in Patient Data Management Systems to facilitate communication between doctors, and it can be used for patient selection and aggregation of patient groups for medical research. Two intensivists and two medical informaticians started five years ago with a rather simple hierarchical list of reasons for ICU admission achieved from the ICNARC Coding Method [9]. Due to the complexity of concepts in the domain, the need for a separation of concepts and terms and the need for a structure to enable aggregation of diagnostic homogeneous patient groups we felt the need to converge to a frame-based structure in which concepts and their characteristics could be specified more formally. In the development proces of DICE we are currently at the stage where we need to evaluate to what extent the current content of DICE meets the requirements, in terms of completeness and correctness, for the intended use of the system.

Methods

For each approach, the evaluation of the completeness and the evaluation of the correctness of the DICE content, we applied two methods.

Evaluation of the completeness of the content

The methods for the evaluation of the completeness of the DICE content are both based on the idea of 'concept matching'. They differ in the way the subsets of concepts that will be matched to the TS are retrieved. The different retrieval methods reflect the two distinct purposes of the system, i.e. the documentation of patients' reason(s) for admission (1A) and the aggregation of patient groups for clinical research (1B).

Evaluation for documentation of reasons for admission

DICE was implemented at the intensive care department of the Academic Medical Center in Amsterdam. Attending intensive care physicians used the system in real practice to code patients' diagnosis. The implemented version of DICE offered the pysicians three ways to find the appropriate diagnosis; (a) in a small list containing the most frequently occurring diagnoses, (b) based on (a part of) its term, or (c) based on (a part of) its definition, e.g. the anatomical localization. The physicians assigned a 'concept match' score to each coded diagnosis. Coded diagnoses could (1) match exactly, (2) be related to the actual diagnosis, (3) be too narrow in meaning, (4) be too general in meaning, (5) have the wrong term, or (6) a concept could not be coded at all. In case of score 2 to 6 the user entered the actual diagnosis in free text. This enabled the checking of the correct assignment of the score by one of the authors (DA) in consensus with another author and IC physician (EdJ).

Evaluation for aggregation of patient groups

During six months we collected all diagnoses that formed (a part of) the in- and exclusion criteria of clinical studies described in two important intensive care journals (Intensive Care Medicine and Critical Care Medicine). The 'concept match' scores for all diagnoses were assigned, by means of consensus, by two of the authors (DA and EdJ). The scores used here were similar to the six scores used in method 1A.

Evaluation of the correctness of the content

The difference between the two methods that evaluate the correctness of the DICE content is that one was performed automatically and the other manually, by domain experts. For these two methods we randomly extracted a 5% (n=80) sample of the basic concepts in DICE.

Automatic evaluation

The DICE content was translated from a frame-based into a Description Logics (DL) based representation. [10]. We used an au-

tomatic reasoning tool (RACER) to reason with the DICE content represented in DL. This reasoning process revealed concepts that had inconsistent definitions, which indicated the presence of incorrect (hierarchical or non-hierarchical) relations. For example, if the 'parent concept' *infectious polyneuropathy* (see figure 1) was defined to be caused by a *virus* and the 'child concept' *leprosy polyneuropathy* was defined to be caused by the *Mycobacterium leprae*, while it was known that mycobacterium leprae is not a virus, than the 'child concept' would be identified as inconsistent. The inconsistency here could have been caused by the fact that the etiology of the 'parent concept' should include *bacterium* in stead of *virus* alone, or by the fact that *leprosy polyneuropathy* should not have been classified as a 'child concept' of *infectious polyneuropathy.*

Polyneuropathy
- Toxic polyneuropathy
- Infectious polyneuropathy
  - Leprosy polyneuropathy

Figure 1 - Example hierarchy of polyneuropathy

Manual evaluation

For each concept (n=80) we printed its terms and (hierarchical and non-hierarchical) relations on paper, each concept on a separate page. Six domain experts, all experienced intensive care physicians, manually reviewed the terms and relations belonging to the concepts and wrote their comments on the paper forms. One of the authors (DA) collected and analyzed all the comments.

Results

Evaluation of the completeness of the content

During the field study 10 IC physicians registered a total of 164 diagnoses. Some diagnoses were registered more than once. The number of unique diagnoses was 107. For the research queries method we selected 91 unique diagnoses. The overlap in the diagnoses selected by both methods consisted of 7 diagnoses. The distribution of the scores for each method is displayed in table 1.

Table 1: Distribution of 'concept match' scores for method 1A and for method 1B

Score	Documentation (1A) (N = 107)	Research (1B) (N = 91)
1. Perfect match	59 (55%)	36 (40%)
2. Related	3 (3%)	2 (2%)
3. Too specific	15 (14%)	7 (8%)
4. Too general	2 (2%)	33 (36%)
5. Wrong term	9 (8%)	2 (2%)
6. No match	19 (18%)	11 (12%)

Score 2 to 4 indicated that the physician did select a diagnosis from DICE, but the selected diagnosis did not completely reflect the actual diagnosis. Score 2, which means that the registered diagnosis was merely related to the actual diagnosis was rarely assigned in both cases. The assignment of scores 3 and 4 differed in both evaluations. During evaluation for the documentation of reasons for admissions (1A) it appeared more frequently that the selected diagnoses was too specific in meaning (score 3), compared to the evaluation for aggregation of patient groups (1B). In contrast, during evaluation 1B it appeared more frequently that the selected diagnosis was too general in meaning (score 4). The structure of DICE enables the forming of new diagnoses by specifying their non-hierarchical relations (post-coordination). In case a diagnosis was assigned a score 4 ('too general'), this indicated that one or more non-hierarchical relations, that were necessary to enable the post-coordination of that specific diagnosis, were missing.

Note that score 5 indicates that the term used for a concept is not the preferred one according to the physician (1A) or is not the one that was used by the authors of an article (1B). It should be noted that in case of score 5 DICE did contain a matching diagnostic concept. In a few cases physicians appeared not to be able to find a diagnosis based on its characteristics (search method c) and gave score 6, while actually it was present in DICE. This indicated that the characteristics that the physician attached to a diagnosis were not consistent with those in DICE. In all these cases it appeared that non-hierarchical relations were missing in DICE.

In one case the assigned score was incorrect, i.e. 'no match' had to be 'too general'. The correct score, 'too general', was used in the analysis of the results.

Evaluation of the correctness of the content

Automatic reasoning with the DICE content represented in DL revealed 28 concepts with inconsistent definitions. Ten of these inconsistencies were due to erroneous assumptions made during the translation from the frame-based to the DL based representation. The remaining 18 inconsistent concepts were caused by 10 missing relations, 10 incorrect relations and 12 relations that were considered defining in stead of qualifying relations.

The six domain experts found a total of 614 unique errors: 397 missing relations, 123 incorrect relations, 70 relations that were considered defining in stead of qualifying relations, 20 incorrect terms and 2 diagnoses that should be deleted from the TS content.

Of the 32 errors found with the automatic evaluation 24 were also found by the manual evaluation. Reversibly, of the 614 errors identified by the manual evaluation, 24 were also found by the automatic evaluation.

Table 2 displays the extent to which different types of errors and omissions were uncovered by each of the four methods applied in this study. A '+' here means that a method revealed a relative large numbers of this type of error or omission. Similarly a '-' or '+/-' indicate that no (-) or a relatively small number (+/-) of these errors or omissions was revealed.

Discussion

This study provides a comparison of different methodologies that can be applied to evaluate the completeness or the correctness of a TS' content. The first two 'concept matching' methods

differed only in the way the subsets of concepts to be matched were retrieved. The number of diagnoses that appeared in the subset of both methods was relatively small. Differences in the results of the two 'concept matching' methods especially concerned the percentage of perfect matches and the percentage of concepts that were found to be too general in meaning. Hales et al. [11] have asserted that the fact that the quality of a TS is defined relative to its intended use is a major barrier to the evaluation of TS. The results of this study endorse the assertions of Hales et al.; the quality of the content of DICE did appear to be relative to the purpose of the system and it appeared that we cannot rely on a single measure.

Table 2 – Extent to which each method evaluated the aspects that determine the quality of a TS content

Types of errors and omissions		1A	1B	2A	2B
Incomplete	Concepts	+	+	-	+/-
	Terms	+/-	+/-	-	+/-
	Non-hierarchical rel.	+/-	+/-	+	+
	Hierarchical rel.	-	-	+	+/-
Incorrect	Terms	+/-	+/-	-	+
	Non-hierarchical rel.	-	-	+	+
	Hierarchical rel.	-	-	+	+

The scores assigned by the physicians for method 1A were checked by the same two people that, in deliberation, assigned the scores for method 1B. This makes the assigned scores comparable between the two methods. Only one score was found to be incorrect. This indicates that the method, 'concept match' scoring by physicians, produces reliable results. However, an actual reliability study with, for example kappa scores, was not performed here. In future studies the reliability of concept match scores should be evaluated.

When interpreting the results of the 'concept matching' evaluation as applied here one needs to keep in mind that it concerns only a sample of concepts. In view of the methods for retrieval of the samples there is a chance that the sample does not contain concepts that only seldom appear in reality. This is an important weakness if one wants to evaluate the completeness of the content.

There appeared to be large differences between the number of errors or omissions found by the automatic evaluation and those found by the manual evaluation. The physicians identified more errors and omissions than the reasoning based on DL did. The difference probably results from the fact that the automatic evaluation only revealed logically incorrect definitions, whereas the physicians also identified the logically correct, but clinically incorrect definitions. For example, if the definition of *encephalopathy* stated that it always involves a state of coma, than the automatic evaluation would find this acceptable. The physician however would not agree with this definition. In stead he would rather say that a patient suffering *encephalopathy* could be in a state of coma.

The automatic evaluation that we performed has some other drawbacks. The migration to DL required a number of assumptions that had to be taken, which in our case made some concepts appear inconsistent, whereas they actually weren't. Another shortcoming of our automatic evaluation was that we were only able to identify concepts with inconsistent definitions. The pinpointing of the actual cause of the inconsistency had to be done manually. We are currently working on a way to automate this identification process.

Bodenreider et al. [12] applied another promising approach to the automatic detection of inconsistencies. They based their assessment of consistency on the lexical knowledge contained in a terminological system. We will consider this approach in our future research.

The major drawback of the manual evaluation was that it is a very time consuming method, both for the domain experts as for the person that analyses the comments generated by the physicians. Automatic detection of inconsistent concepts, as applied here, or as applied by Bodenreider et al. [12], does have the potential to support and focus the effort of domain experts in the reviewing process. However, this still requires that the automatic evaluation method, as applied in our case study, is further explored.

There is a chance that the large number of errors and omission identified by the physicians was due to the fact that the developers of DICE did not structurally consult an editorial board, consisting of domain experts, when building the TS. It might be that the manual method would reveal less results if all concepts and their definitions had been approved by such an editorial board before they were added to the TS' content. This should be considered when generalizing the conclusions of this study.

A difficulty in evaluating TSs' content is the lacking of a gold standard. This makes it impossible to calculate common outcome measures such as sensitivity and specificity for specific evaluation techniques. Consequently we were restricted to a comparison of the frequencies of incomplete or incorrect knowledge identified by each of the three evaluation methods.

If we look at the types of results produced by all four methods it appears that three of the four methods are actually not limited to either the completeness of concepts and terms, or the completeness and correctness of the definitions. The 'concept matching' methods, that were originally designed to evaluate the completeness of concepts and terms in the TS also revealed some missing non-hierarchical relations. The manual evaluation method was originally designed to evaluate the completeness and correctness of the definitions, but also revealed some missing concepts and terms. However, none of the four methods could be used to evaluate all aspects that determine the quality of TS' content.

Conclusion

It appeared that none of the methods used in this study would suffice to evaluate all aspects that determine the quality of a TS' content. In order to get a good overview of the quality of a TS' content, it is preferable to use a combination of different evaluation methods. The 'concept matching' method seems to be most useful to determine the completeness of the concepts and terms

in the TS. However, he intended purpose of the TS should determine the source of the sample. Different sources can lead to different results regarding the quality of the TS' content. Manual evaluation appears to be very worthy, but also very time consuming. Automatic evaluation has the potential to focus human reviewers and decrease their workload. Further research is required to exploit these potentials of automatic evaluation.

Acknowledgements

We thank the members of the board of the Dutch National Intensive Care Evaluation (NICE) foundation, and the head of the department of intensive care of the Academic Medical Center in Amsterdam, Margreeth Vroom, for their cooperation in this study. We thank the Dutch ministry of Health, Welfare and Sport for their financial support.

References

[1] de Keizer NF, Abu-Hanna A, Zwetsloot-Schonk JH. Understanding terminological systems I: Terminology and typology. *Methods Inf Med.* 2000 Mar: 39(1): 16-21

[2] Cimino JJ. Terminology tools: state of the art and practical lessons. *Methods Inf Med.* 2001: 40(4): 298-306

[3] de Keizer NF, Abu-Hanna A, Cornet R, Zwetsloot-Schonk JH, Stoutenbeek CP. Analysis and Design of an ontology for intensive care diagnoses. Methods Inf Med 1998: 38: 102-112

[4] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine* 1998: 37(4-5): 394-403.

[5] Chute CC, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classifications. *JAMIA*1996: 3: 224-233

[6] Bowles KH. Application of the Omaha system in acute care. *Res Nurs health* 2000: 23: 93-105

[7] Bodenreider O, Burgun A, Botti G, Fieschi M, Le Beux P, Kohler F. Evaluation of the Unified Medical Language System as a medical knowledge source. *JAMIA* 1998: 5: 76-87

[8] Cimino JJ. Auditing the Unified Medical Language System with semantic methods. *JAMIA* 1998: 5: 41-51

[9] Young JD, Goldfrad C, Rowan K. Development and testing of a hierarchical method to code the reason for admission to intensive care units: the ICNARC Coding Method. Intensive Care National Audit & Research Centre. Br J Anaesth. 2001 Oct: 87(4): 543-8

[10]Cornet R, Abu-Hanna A. Evaluation of a frame-based Ontology. A formalization-oriented Approach. *Proc MIE* 2002: 90: 488-93

[11]Hales JW, Schoeffler KM. Barriers to Evaluation of Clinical Vocabularies. *Proc MedInfo 1998*: 680-684

[12]Bodenreider O, Burgun A, Rindflesch TC. Assessing the consistency of a biomedical terminology through lexical knowledge. *Int J Med Inform* 2002: 67: 85-95

Address for correspondence

Drs. D.G.T. Arts, Academic Medical Center, Dept. of Medical Informatics J2-257, P.O.Box 22700, 1100 DE Amsterdam, The Netherlands

Email: D.G.Arts@amc.uva.nl