Methods for Evaluation of Medical Terminological Systems
A Literature Review and a Case Study

D. G. T. Arts¹, R. Cornet¹, E. de Jonge², N. F. de Keizer¹
¹Academic Medical Center, Department of Medical Informatics, Amsterdam, The Netherlands
²Academic Medical Center, Department of Intensive Care, Amsterdam, The Netherlands

Summary

The usability of terminological systems (TSs) strongly depends on the coverage and correctness of their content. The objective of this study was to create a literature overview of aspects related to the content of TSs and of methods for the evaluation of the content of TSs. The extent to which these methods overlap or complement each other is investigated.

Methods:

Three methods were selected: 1) Concept matching in which two samples of concepts representing a) documentation of reasons for admission in daily care practice and b) aggregation of patient groups for research, are looked up in the TS evaluation in which reasoning on the formally represented content is used to detect inconsistencies; and 3) Expert review in which a random sample of concepts are checked for incorrect and incomplete terms and relations. These evaluation methods were applied in a case study on the locally developed TS DICE (Diagnoses for Intensive Care Evaluation).

Results:

None of the applied methods covered all the aspects of the content of a TS. The results of concept matching differed for the two use cases (63% vs. 52% perfect matches). Expert review revealed many more errors and incompleteness than formal algorithmic evaluation.

Conclusions:

To evaluate the content of a TS, using a combination of evaluation methods is preferable. Different representative samples, reflecting the uses of TSs, lead to different results for concept matching. Expert review appears to be very valuable, but time consuming. Formal algorithmic evaluation has the potential to decrease the workload of human reviewers but detects only logical inconsistencies. Further research is required to exploit the potentials of formal algorithmic evaluation.

Keywords
Terminological systems, evaluation, methods, definitions

1. Introduction

Several developments in health care, such as accountability of care and increased use of electronic patient records, have led to an increased need for accurate, detailed and structured registration of medical data. Many terminological systems (TSs) have been and are still being developed to support this. A TS interrelates concepts of a particular domain and provides their terms and codes [1]. The relations between the concepts within a TS can be hierarchical (e.g. Is-A) or non-hierarchical (e.g. has-location). In addition some TSs hold (formal) rules for the composition of new concepts by combining existing concepts. Examples of medical TSs are the International Classification of Diseases (ICD) [2], the Systemized Nomenclature of Medicine (SNOMED) [3, 4], and the North American Nursing Diagnosis Association (NANDA) terminology [5]. By the direction of the Dutch National Intensive Care Evaluation foundation (NICE) our department is engaged in a continuous effort to develop a TS and corresponding software for the domain of intensive care (IC). This system is called Diagnoses for Intensive Care Evaluation, DICE [6].

For the study described here we distinguish two types of use cases for terminological systems. On the one hand TSs are used by medical staff to document medical data, e.g. patient characteristics or treatment, in the medical record. On the other hand TSs are used to select homogeneous patient groups for research or management purposes. A medical researcher or manager selects from the TS those concepts that define a homogeneous patient group. After selecting the appropriate concepts the researcher/manager can identify patients that fulfill the criteria, by searching their electronic records.

Several authors have specified required characteristics of a TS [7-9]. In 2000 a list of standard requirements for TS was developed and approved by the International Organization for Standardization (ISO) [10]. In this study we will focus on requirements related to the content of a TS, i.e. the concepts, their terms, and the relations between the concepts. The content of a TS is of utmost importance for its acceptance. A physician needs to be able to be complete and sufficiently accurate in depicting the care process, and clinical researchers need to be able to be complete in selecting specific patient groups at any desired level of aggregation. To realize this all concepts, terms and relations belonging to the domain of the TS should be represented and should be correct. For example, we want sufficient terms attached to a concept, and we want the terms to be only the correct ones.

A number of methods to evaluate the content of a TS have been described in literature. A literature study was performed to gain insight into the several types of evaluation methods. The diversity of the terminology used in this context has invited us to compose definitions for the most prominent expressions that are used in this article. In addition, we present three common evaluation methods that focus on (but not restrict to) the coverage and the correctness of a TS’ content. These three methods have been applied in a case study on the TS DICE.

¹ The Dutch National Intensive Care Evaluation (NICE) foundation: www.stichting-nice.nl
The aim of the case study was to analyze the extent to which the results of the three methods overlap or complement each other. For this we compared the results produced by each method.

2. Literature Study

As mentioned in the introduction, in this study we focus on the evaluation of the content of a TS. To gain insight into methods for evaluation of the content of TS that have been applied by others we performed a review of relevant Medline-indexed journal articles by using (combinations of) the following keywords: evaluation, validation, assessment, audit, terminological system, terminology, ontology, classification, thesaurus, nomenclature. In addition, articles were retrieved from reference lists and personal databases. An article was considered relevant if it described the evaluation of the quality of the content of a TS which was developed for a medical domain. Articles were selected from the past ten years.

2.1 Definitions

Our literature study uncovered some inconsistencies in the terminology that is used in this field. We therefore provide this article with some definitions that have been applied in this study.

First of all we have restricted ourselves to the evaluation of TS’s ‘content’. By our definition, the content of a TS includes concepts, the terms attached to these concepts and the relations between these concepts. ‘Concepts’ can be defined as ‘units of thought formed by the characteristics of objects’. Objects might be concrete things such as the heart valve of patient X or abstract things such as the pain of patient Y. ‘Terms’ are used to designate a concept. The ‘relations’ between concepts can be hierarchical (e.g. Is-A) or non-hierarchical [1].

A term that frequently appears in literature considering the evaluation of TS is ‘domain completeness’. ‘Domain completeness’ can be defined as the extent to which the content of a TS covers the intended domain. Domain completeness according to this definition would be hard to measure or to quantify, because the continuous changes in medical knowledge make it impossible to define exactly what comprises a particular medical domain. By lacking of this gold standard, it is impossible to determine to what extent a TS is complete in covering the intended medical domain. Instead, we could take a subset of concepts or terms representative for the intended domain and see to what extent this subset is incorporated in a particular TS. This way we will measure the ‘content coverage’.

Table 1 gives the definitions of ‘content coverage’ and of the ‘occurrence coverage’ and the ‘correctness’ of the separate elements (i.e. concepts, terms and relations) that comprise the content of TS.

The definitions of ‘concept coverage’ and ‘term coverage’ might look straightforward; however the measured coverage can be highly influenced by choices made in the evaluation process. For example one can choose whether or not to consider the occurrence of specific concepts or terms in real practice. Missing a frequently occurring concept or term might be more severe than missing those which are hardly ever used.

We therefore define ‘occurring coverage’ and ‘unique coverage’. For example in Figure 1 we see that the ‘occurring coverage’ is 80% (4/5), whereas the ‘unique coverage’ is 75% (3/4).

Some TSs enable the composition of new concepts by combining two or more existing concepts. We call this feature “post-coordination”. In case a TS enables post-coordination of concepts it is also important to consider whether or not post-coordinated concepts are taken into account when determining the coverage of the concepts. Many concepts might not be present in a TS as a pre-coordinated concept, but can be composed by combining two or more concepts. If only pre-coordinated concepts are considered in the evaluation, then the measured concept coverage will be lower than when post-coordinated concepts are also taken into account.

In this study we also focus on the correctness of the content of TSs. Correctness can only be measured for concepts, terms and relations that are covered by the TS. Definitions of correctness are provided in Table 1.
2.2 Evaluation Methods

Table 2 contains a list of aspects of TSs that have been evaluated by others. These aspects have been categorized according to the aspects of the coverage and correctness of TSs’ content, that were defined in Table 1. Most of the evaluation studies described in literature focus on the coverage of either the concepts or the terms in a TS.

The coverage of concepts or terms is often evaluated through ‘concept matching’ or ‘term matching’ [7, 11-13, 17-19, 22]. Matching implies that a representative sample of concepts or terms is extracted from the domain in which the TS is being used, e.g. the diagnoses of patients at the oncology department. The representative sample can be randomly or non-randomly chosen. The concepts from the sample are then matched with those of a TS. The extent to which concepts or terms in the sample can be matched with concepts in the TS is mostly presented by means of matching categories. For example Chute et al. [12] have applied a scoring scale for the matching of concepts from 0 to 2, where 0 = no match, 1 = fair match, 2 = complete match. A different categorization of matches was used by Warnekar et al. [29]. According to this

<table>
<thead>
<tr>
<th>Concepts in terminological system X:</th>
<th>Concepts in representative sample:</th>
<th>Present in terminological system X?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept A</td>
<td>Concept D</td>
<td>Yes</td>
</tr>
<tr>
<td>Concept B</td>
<td>Concept D</td>
<td>Yes</td>
</tr>
<tr>
<td>Concept C</td>
<td>Concept F</td>
<td>Yes</td>
</tr>
<tr>
<td>Concept D</td>
<td>Concept K</td>
<td>No</td>
</tr>
<tr>
<td>Concept E</td>
<td>Concept A</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Fig. 1 Example of a terminological system and a representative sample containing concepts that are being matched to the terminological system to evaluate the concept coverage

| Table 2 Categorization of quality aspects of the content of TS, as evaluated by others |
|----------------------------------------|-------------------------------------|----------------------------------|
|                                        |                                     | Chute 1996 [12]                |
|                                        |                                     | Campbell 1997 [7]               |
|                                        |                                     | Humphreys 1997 [13]             |
|                                        |                                     | Bodenreider 1998 [14]           |
|                                        |                                     | Cano 1999 [15]                  |
|                                        |                                     | Bowles 2000 [17]                |
|                                        |                                     | Cano 2001 [18]                  |
|                                        |                                     | Bodenreider 2001 [19]           |
|                                        |                                     | Schütz 2001 [20]                |
|                                        |                                     | Bodenreider 2002 [21]           |
|                                        |                                     | Bodenreider 2002 [22]           |
|                                        |                                     | Connet 2002 [24]                |
|                                        |                                     | moss 2003 [25]                  |
|                                        |                                     | Hardiker 2003 [26]              |
|                                        |                                     | Horstle 2003 [27]               |
|                                        |                                     | Warnekar 2003 [29]              |
|                                        |                                     | Wessmann 2003 [30]              |
|                                        |                                     | bowen 2004 [31]                 |
|                                        |                                     | Post 2004 [32]                  |
|                                        |                                     | Ceusters 2004 [33]              |

Methods Inf Med 5/2005
categorization matches could be exact matches, lexical or semantical matches, or no-matches. Wasserman et al. [30] distinguished, apart from the exact matches, synonyms and no-matches that required the addition to the hierarchy of a new ‘leaf’, a new ‘leaf with multiple stems’ or a ‘graft to an existing branch’. The content coverage can be represented, for example, by calculating the percentage of perfect matches. To be representative, the source of the sample of concepts or terms should reflect the intended use of the TS. For example, if a TS will be used by nurses for documentation of nursing information, then the sample of concepts could be well extracted from existing nursing documentation in medical records [11, 17]. While in most cases the matching process is performed by humans, Penz et al. [32] applied two automated mapping tools. Penz et al. were not able to assess their overall value, due to high frequencies of spelling errors and jargon in their sample. In a study of Brown et al. [31] one of the automated mapping tools, the SmartAccess Vocabulary Server (SAVS), has proven to be reliable.

Besides the ‘matching method’, other methods have been applied to evaluate the concept coverage or term coverage. In a study by Bodenreider et al. [14] the system being evaluated had already been in use for some time. The measure of coverage 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 under-representation in the TS. In addition, they evaluated the coverage of hierarchical and other relations within the UMLS. They designed an algorithm to automatically extract the UMLS concepts that are related to procedures in a particular domain. Starting with a few concepts related to the domain, the algorithm recursively selected all their subordinate concepts. This navigation was based on the relations between concepts. Lacking of relations resulted in silence: A concept might seem not to exist in the UMLS only because it is not related to another concept. The amount of relations missing was estimated by comparing the concepts in the subset retrieved by the algorithm to the concepts that are needed for the representation of the procedures within the domain of interest.

Whereas Bodenreider used the relations between concepts to evaluate the content coverage, these relations are more often used for the evaluation of the correctness of the content. Many TSs nowadays consist of more than just a simple list of terms; hierarchical and non-hierarchical relations exist between the concepts. By looking at the relations between concepts, inconsistent, ambiguous or redundant concepts may be revealed. For example, if two individual concepts share the same meaning, they are actually redundant concepts. If ‘polyneuropathy’ and ‘polyneuritis’ were each defined as separate concepts one could find that they share the same relations to other concepts, and that they actually refer to the same disease. In this case one of the concepts ‘polyneuropathy’ and ‘polyneuritis’ is redundant. In case of a hierarchical structuring of the concepts, a concept might be inconsistently classified if it has relations which are in conflict with the relations of its superordinate concept. For example, inconsistency might occur if a superordinate disease is defined to be caused by a bacterium, whereas the subordinate disease is defined to be caused by a virus. The inconsistency here becomes apparent if ‘virus’ and ‘bacterium’ were explicitly made mutually exclusive. Cimino used this kind of method to detect ambiguities, redundancy and inconsistent hierarchical relations within the UMLS [15]. In addition, Bodenreider et al. [14] stated that hierarchical relations between concepts can be used for the evaluation of the categorization of concepts. Their evaluation was based on the idea that concepts inherit properties from their superordinate concept and thus a concept is supposed to belong at least to the same category or categories as its superordinate concept. Evaluation based on relations between concepts has the potential to be automated or semi-automated. For example a computer algorithm could detect concepts that share the exact same definitions or concepts that were assigned to a number of semantic types, of which two are mutually exclusive.

To enable automated evaluation, the TS content (especially the relations between concepts) should be represented in a formal way. Examples of formal representations can be found in the SNOMED-CT [34] and the GALEN terminologies [35]. SNOMED-CT and GALEN both use a Description Logic to represent their knowledge. In a study of Cornet et al. [24] migration of content representation from frame-based to Description-Logic-based has proven to be valuable in determining redundancies in concept definitions and in forcing the knowledge modeler to be aware of ambiguities. Schulz and Hahn [20] have expressed UMLS knowledge in Description Logic. They provide evidence that embedding the knowledge into a formal reasoning framework is effective to identify inconsistencies. Bodenreider et al. [23] applied another approach to the automated detection of inconsistencies, by using the lexical knowledge contained in a terminological system. They assume that all terms are composed of a modifier, such as ‘primary’ or ‘secondary’, and a context, a noun phrase such as ‘adrenocortical insufficiency’. This would result in the terms ‘primary adrenocortical insufficiency’ and ‘secondary adrenocortical insufficiency’. They hypothesize that terms of the form modifier-context and modifier-context are co-hyponyms of the term ‘context’. E.g., ‘primary adrenocortical insufficiency’ and ‘secondary adrenocortical insufficiency’ are hyponyms of ‘adrenocortical insufficiency’. They base their evaluation on the fact that in a consistent terminology the terms modifier-context and modifier-context should be 1) both present and 2) in hierarchical relation with the term ‘context’. The conclusion of this study was that this method alone is not sufficient for ensuring the consistency of a TS.

A completely different method to evaluate the correctness of relations between concepts was applied by Campbell et al. who evaluated the clinical utility of pairs of hierarchically related concepts within three medical terminological systems (SNOMED, READ and UMLS) [7]. Six clinician-informatics specialists manually reviewed random samples of hierarchical pairs. They used a five-point Likert scale (1 = extremely dissatisfied with pairing, 3 = neutral, 5 = extremely satisfied with pairing) to rate the clinical utility of each pair. Review by domain experts was also applied in a study by Hardiker [26].
In summary, we distinguish four evaluation strategies for the content of TSs: 1) concepts matching, 2) evaluation based on relations between concepts, 3) evaluation by domain experts, and 4) evaluation based on lexical knowledge.

Based on this literature review on methods for evaluation of TS we selected three methods for the case study, which will be described below. The three methods reflected the first three of the above mentioned evaluation strategies. The methods were chosen because previous studies deem them promising and because they were applicable to our case study with the DICE TS.

3. Case Study
3.1 Background
A study by de Keizer et al. in 1998 has shown that none of the contemporary TSs met the criteria of a TS for Dutch intensive care [6]. This has been the motivation to develop a new TS, Diagnoses for Intensive Care Evaluation (DICE). The TS DICE comprises reasons for admission to the Intensive Care Unit (ICU), and some of their characteristics, such as the anatomical localization, the dysfunction and the etiology. The DICE TS contains 2373 concepts, of which 1456 are diagnoses that form reasons for admission to the ICU. Other concepts include for example anatomical locations and causes of disease. There are 50 relation types. Thirteen of these, for example “has_anatomical_localization”, may be used for any of the diagnoses. The other 37 relation types are attributes which are specific for certain diagnoses, for example the chronicity (e.g. acute, chronic) of organ dysfunction. Currently a total of 10,425 relations between concepts have been defined. DICE can be incorporated into Patient Data Management Systems to facilitate documentation by physicians, and it is intended to be used for patient selection and aggregation for medical research and management overviews. Two domain experts, i.e. intensive care physicians, and two medical informaticians started seven years ago with a rather simple hierarchy of reasons for ICU admission achieved from the ICNARC Coding Method [36]. 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 homogeneous patient groups we chose to specify the concepts and their characteristics more formally. The DICE content was therefore converged to a frame-based structure. In the development process 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 coverage and correctness, for the intended use of the system.

3.2 Methods
In the case study we will apply three methods: concept matching, formal algorithmic evaluation and expert review, to evaluate the coverage and the correctness of the DICE content and to analyze the extent to which these methods overlap or complement each other. We will compare the overall coverage and correctness measured by the three methods. For methods 2 and 3 the individual missing or incorrect concepts, terms and relations revealed by each method are compared. The methods used are described below.

3.2.1 Method 1: Concept Matching
In this study concept matching was carried out twice with different representative samples of concepts that were matched to the TS. The samples differed in the sources from which they were retrieved. The two sources reflected the two distinct purposes of the system, i.e. 1A) the documentation of and communication about patients’ reason(s) for admission and 1B) the selection of homogeneous (with respect to diagnosis) patient groups for clinical research and management. Concepts from these two representative samples were matched to the content of DICE. Concepts found in DICE could be 1) a perfect match, 2) related (e.g. mitral valve instead of tricuspid valve), 3) too narrow in meaning (e.g. subarachnoid haemorrhage instead of intracranial haemorrhage), 4) too general in meaning (e.g. polyneuropathy instead of infectious polyneuropathy), or 5) a concept could not be coded at all. We applied this categorization because in case of a suboptimal match it enabled us to be specific about why a match was not a ‘perfect match’. The distribution of the concepts among the matching categories was calculated when using all diagnoses occurring in the sets and when using only the unique diagnoses. DICE offers the users the opportunity to compose new (post-coordinated) concepts out of two or more consisting (pre-coordinated) concepts. In this study the diagnoses within DICE that were matched to the diagnoses within the samples could also be post-coordinated diagnoses.

Method 1A: Evaluation for Documentation of Reasons for Admission
DICE was used at the intensive care department of the Academic Medical Center in Amsterdam during March 2001. Attending intensive care physicians used the system in real practice to code actual patients’ reasons for admission. The reasons for admission that the physicians wanted to record in a patient’s medical record comprised the representative sample of concepts. A physician assigned each concept within the sample to a matching category to express the extent to which it was represented in DICE. In case of a non-perfect match the physician entered the actual diagnosis in free text. This enabled checking the correct assignment of concepts to the matching categories.

Method 1B: Evaluation for Aggregation of Patient Groups
We collected all diagnoses that formed (a part of) the in- and exclusion criteria of clinical studies that appeared in two important intensive care journals (Intensive Care Medicine and Critical Care Medicine) between January 1, 2001 and July 1, 2001. These diagnoses comprised the representative sample of concepts that was matched to the TS DICE. The concepts within the sample were assigned to one of the matching categories by two of the authors (DA and EJ), by means of consensus.
3.2.2 Method 2: Formal Algorithmic Evaluation

As an increasing number of medical terminological systems is based on formal representation, it is important to understand the potential of readily available reasoning algorithms exploiting the formal representation. Such algorithms can be instrumental for evaluation of the content of terminological systems. One example of deploying such algorithms is the constraint checking engine based on Protégé Axiom Language (PAL), which has been applied to the Gene Ontology [37]. Whereas in this example a frame-based representation is used, we have used the Description Logic (DL) formalism [38]. The publicly available reasoner RACER [39] was used to perform satisfiability testing, which provides a means for detecting mutually conflicting concept definitions. As the content of DICE has a frame-based representation, it was migrated to a DL-based representation, according to the process described in [40]. In this migration process, assumptions are made on the semantics of definitions, such as for example disjointness of sibling concepts.

For this evaluation we randomly extracted a sample of 80 pre-coordinated diagnoses in DICE. The reasoning process revealed concepts that had inconsistent definitions, which indicated the presence of incorrect (hierarchical or non-hierarchical) relations. For example, if the superordinate concept infectious polyneuropathy (see Table 3) was defined to be caused by a virus and the subordinate leprosy polyneuropathy was defined to be caused by the Mycobacterium leprae, while it was known that Mycobacterium leprae is not a virus, then the subordinate concept would be identified as inconsistent. The inconsistency here could have been caused by the fact that the etiology of the superordinate concept should also include bacterium instead of virus alone, or by the fact that leprosy polyneuropathy should not have been classified as a subordinate concept of infectious polyneuropathy.

3.2.3 Method 3: Expert Review

We printed on paper forms the terms and (hierarchical and non-hierarchical) relations belonging to each of the 80 randomly selected concepts that were also used in the formal algorithmic evaluation (Fig. 2). Six domain experts, all experienced intensive care physicians, manually reviewed the terms and relations belonging to the concepts. If they found a missing or incorrect term or relation they wrote this on the paper forms. One of the authors (DA) collected and analyzed the comments of the domain experts.

3.3 Results

3.3.1 Concept Matching

During the study to evaluate the coverage of DICE for documentation of reasons for admission (1A) 10 ICU physicians registered a total of 164 diagnoses, of which 107 were unique. For the concept matching to evaluate the coverage of DICE for aggregation of patient groups (1B) we retrieved 218 diagnoses, of which 187 were unique. The overlap of the two samples consisted of eight unique diagnoses. The distribution of concepts among the matching categories is displayed in Figure 3.

Concept matching for documentation of diagnoses in daily care practice (1A) resulted in 63% \((n = 67)\) perfect matches when

<table>
<thead>
<tr>
<th>Concept matching for documentation of diagnoses in daily care practice (1A)</th>
<th>% perfect matches</th>
<th>n = 67</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perfect matches</td>
<td>63%</td>
<td></td>
</tr>
<tr>
<td>Matches</td>
<td>97%</td>
<td></td>
</tr>
</tbody>
</table>

Table 3

<table>
<thead>
<tr>
<th>A fictitious example of (inconsistent) concept definitions in a DL-based terminological system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infectious_polyneuropathy</td>
</tr>
<tr>
<td>Leprosy_polyneuropathy</td>
</tr>
<tr>
<td>Mycobacterium_Leprae</td>
</tr>
<tr>
<td>Disjoint (Virus, Bacterium)</td>
</tr>
</tbody>
</table>

Fig. 2 Example of a paper form for expert review

Methods Inf Med 5/2005
only uniquely occurring diagnoses were considered (unique concept coverage) and 74% (n = 121) perfect matches when each single occurrence of a diagnosis in the sample was considered (occurring concept coverage). Concept matching for aggregation of patient groups (1B) resulted in lower frequencies of perfect matches (52% (n = 97) unique, 51% (n = 111) occurring).

The frequency of ‘too general’ matches was higher for the concept matching for aggregation of patient groups for clinical research and management (1B) (25% unique, 24% occurring) than for documentation of diagnoses (1A) (14% unique, 10% occurring). The structure of DICE enables the composition of new diagnoses by specifying their non-hierarchical relations (post-coordination). ‘Too general’ matches indicated that one or more non-hierarchical relations, that are necessary to enable the post-coordination of that specific diagnosis, were missing. DICE enables users to search for a diagnosis based on its characteristics (non-hierarchical relations). For example a user might search DICE for a diagnosis that is an infection that is located in the lungs and retrieve the diagnosis Pneumonia. In one case the physicians appeared to be unable to find a diagnosis based on its characteristics. This indicated that the characteristics that the physician attached to this diagnosis were not consistent with those in DICE. The concept was available in DICE, but some characteristics were missing. The matching category assigned by the physician to this diagnosis was incorrect, i.e. ‘no match’ whereas it actually had to be ‘too general’. The correct category, ‘too general’, was used in the analysis of the results.

3.3.2 Formal Algorithmic Evaluation and Expert Review

The formal algorithmic evaluation of the 80 concepts randomly selected from DICE revealed 28 concepts with inconsistent definitions. Ten inconsistencies were due to erroneous assumptions made during the migration of DICE from the frame-based to the DL-based representation. The remaining 18 inconsistent concepts were caused by 32 errors consisting of ten missing relations, ten incorrect relations and twelve relations that were specified as ‘defining’ instead of ‘qualifying’ relations. An example of the latter is ‘pericarditis’, which was defined as always being an infection (defining relation), whereas in fact this is not always the case, i.e. ‘pericarditis’ can be an infection (qualifying relation). Pericarditis is only an infection in case a bacterium or a virus caused the inflammation.

For the 80 selected concepts the six domain experts found 397 missing relations, 123 incorrect relations, 70 relations that were specified as ‘defining’ in stead of ‘qualifying’ relations, five missing terms, 15 incorrect terms and two diagnoses that should be deleted from the TS content. Of the total of 612 unique discrepancies 173 (28%) were found by two or more domain experts and 83 (14%) were found by three or more domain experts. Twenty-four errors were found both with the formal algorithmic evaluation and by the expert review. Table 4 displays the numbers of the different types of errors and omissions were discovered by each of the three methods applied in this study.

4. Discussion

Over the past 15 years much has been written about the content of TSs. Requirements for the content of TSs and evaluations of the quality of the content of TSs have been described. This paper provides an overview of what aspects determine the quality of a TS’ content and how these aspects can
be evaluated. The aspects most often evaluated were the coverage of terms and of concepts [7, 11-14, 16, 17, 22, 25, 26, 29-32]. Lately the coverage and the correctness of relations between concepts have received increasing attention, especially by Bodenreider et al. [14, 21-23, 28] and Hardiker [19, 26, 27].

The case study described in this paper provides a comparison of three methods that can be applied to evaluate the coverage and/or the correctness of the content of TSs. The first method consisted of two ‘concept matching’ studies which only differed in the origin of the representative samples of concepts to be matched. The overlap of the two samples was relatively small. The matching categories assigned by the physicians for evaluation of DICE for documentation of diagnoses (1A) were checked by the same two people that assigned the categories for the evaluation of DICE for aggregation of patient groups (1B). This makes the assigned matching categories comparable between the two studies. Only once the assigned matching category was found to be incorrect. This indicates that the method, ‘concept matching’ by physicians, produces reliable results.

Differences in the results of the two ‘concept matching’ studies especially concerned the percentage of perfect matches and the percentage of concepts that were found to be too general in meaning. Hales et al. [41] have asserted that the quality of a TS is defined relative to its intended use, which is a major barrier to the evaluation of TSs. The intended use of a TS specifically plays an important role in the evaluation of its content. 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 in the case of DICE we couldn’t rely on a single measure (1A or 1B) to get a complete overview on the coverage of the DICE content.

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. What we measure by concept matching is the ‘concept coverage’ which is merely an approximation of the completeness of all concepts that belong to the domain of interest (‘domain completeness’). In view of the methods for retrieval of the representative samples of concepts (or terms) there is a chance that the sample does not contain concepts that only rarely occur in practice. However, the question of whether rarely occurring concepts are represented in a TS might not be as important as whether concepts that frequently occur are represented. The frequencies of occurrence of concepts in practice can be taken into account by using each single occurrence of a concept within the representative sample for (occurring) concept matching. The increase in concept coverage when measuring the occurring instead of the unique concept coverage (Fig. 4) indicates that, in the case of DICE, the concepts that were not represented were mostly the not frequently occurring concepts.

There appeared to be large differences between the numbers of errors or omissions found by the formal algorithmic evaluation and those found by the domain experts. The physicians identified many more errors and omissions than the formal algorithmic evaluation. The difference stems from the fact that the formal algorithmic evaluation only revealed logically incorrect definitions, whereas the physicians also identified wrong terms and the logically correct, but clinically incorrect definitions. For example, if the definition of encephalopathy stated that it always involves a state of coma, then the formal algorithmic evaluation would render this correct. The physician however would not agree with this definition. Instead (s)he would rather say that a patient suffering from encephalopathy could be in a state of coma.

The formal algorithmic evaluation, as it was performed here, has some other drawbacks. The migration of DICE from a frame-based to a DL-based representation required a number of assumptions that had to be made. In our case these assumptions made some concepts appear inconsistent, whereas they actually were not. Another shortcoming was that by using the formal algorithmic evaluation as presented here the inconsistent definitions could only be identified. The pinpointing of the actual causes of the inconsistencies had to be done manually. We are currently working on ways to automate this identification process [42, 43].

The major drawback of the expert review was that it took the physicians much time to look carefully at all the terms and relations belonging to a concept. Similarly, the analysis of the comments generated by the physicians was a very time-consuming effort. A large number of the comments were given by only one reviewer. Whereas in most cases only comments on which reviewers have reached consensus will be processed, a large part of the comments will not be used for updating the TS. It is arguable how many reviewers have to be involved in the consensus process and how many reviewers have to agree on a comment before it is processed. Automated detection of inconsistent con-

### Table 4 Numbers of missing or incorrect concepts, terms or relations discovered by the three methods

<table>
<thead>
<tr>
<th>Quality aspects of TS’ content</th>
<th>Concept matching</th>
<th>Formal algorithmic evaluation (80 concepts)</th>
<th>Expert review (80 concepts)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1A (164 concepts)</td>
<td>1B (218 concepts)</td>
<td></td>
</tr>
<tr>
<td>Coverage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concepts</td>
<td>24</td>
<td>49</td>
<td>5</td>
</tr>
<tr>
<td>Terms</td>
<td>9</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Relations</td>
<td>19</td>
<td>58</td>
<td>10</td>
</tr>
<tr>
<td>Correctness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concepts</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>Terms</td>
<td>-</td>
<td>-</td>
<td>15</td>
</tr>
<tr>
<td>Relations</td>
<td>-</td>
<td>-</td>
<td>22</td>
</tr>
<tr>
<td>Total number of errors found</td>
<td>52</td>
<td>111</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>612</td>
</tr>
</tbody>
</table>

Methods Inf Med 5/2005
cepts, such as the formal algorithmic evaluation as it was applied in this study, does have the potential to limit and focus the effort of domain experts in the reviewing process. However, this requires that these methods are further explored. We will consider this in our future research. To decrease the efforts for physicians to review the content of a TS we have started the implementation of a system for ‘internet-based terminological knowledge reviewing’, called KEBoRT (Knowledge Editorial Board online Reviewing Tool), which will be used for reviewing the entire DICE content [44].

A shortcoming of the case study is that the evaluation methods were only applied on the TS DICE. There is a chance that the large number of errors and omission identified by the expert reviewers was due to the fact that the current content of DICE is not based on expert consensus. Instead, the developers of DICE consulted two domain experts when building the TS. It might be that expert review as it was applied in this case study would produce fewer comments if all concepts, terms and relations had been approved, by means of consensus between a larger number of domain experts, before they were included in the TS. This should be considered when generalizing the conclusions of this case study, regarding the number of errors found by expert review.

Looking at the results produced by the three methods we see that the ‘concept matching’ methods, that were originally designed to evaluate the coverage of concepts, also revealed some missing terms and (non-hierarchical) relations. The formal algorithmic evaluation revealed only missing and incorrect relations. The expert review revealed mainly missing and incorrect relations, but also a small number of missing and incorrect concepts and terms.

5. Conclusion

Evaluation studies of the content of TS mostly concern ‘term matching’ or ‘concept matching’. More sophisticated methods are being explored. Independent of the method used, it remains important to define exactly what is being measured. The definitions provided in this article could be a starting point in this.

Based on the results of the case study it seems that expert review is most complete in evaluating the quality of the content of TSs. Expert review produces results for all aspects that determine the quality of the content. However the expert review method has some major drawbacks, of which the most important is the fact that it is very time-consuming. In order to get a good overview of the quality of the content of a TS, it is preferable to use a combination of evaluation methods. The ‘concept matching’ method seems to be most useful to determine the coverage of the concepts and terms in the TS. However, it is important to consider the source that was used to retrieve the sample of concepts that are being matched to the TS. The intended purpose of the TS should determine the source of the sample. Different sources can lead to different results regarding the quality of the content of a TS. In addition a clear description of the applied concept matching method is necessary. Formal algorithmic evaluation has the potential to decrease the workload of human reviewers, but further research is required to explore these potentials. From this study it became clear that each method has its strengths and weaknesses. Therefore the three methods should be used in combination with each other.

Acknowledgments

We thank the members of the board of the Dutch National Intensive Care Evaluation (NICE) foundation and Margreeth Vroom, the head of the department of intensive care of the Academic Medical Center in Amsterdam, for their cooperation in this study. We thank the Dutch ministry of Health, Welfare and Sports for their financial support. The preliminary results of this study have been presented in a reduced form in the AMIA 2003 and the Medinfo 2004 proceedings [45, 46].

References


Correspondence to:
D. G. T. Arts
Academic Medical Center
Department of Medical Informatics
P.O. Box 22700
1100 DE Amsterdam
The Netherlands
E-mail: arts@amc.nl