A usability evaluation of a SNOMED CT based compositional interface terminology for intensive care

F. Bakhshi-Raiez a,*, N.F. de Keizer a, R. Cornet a, M. Dorrepaal a, D. Dongelmans b, M.W.M. Jaspers a

a Department of Medical Informatics, Academic Medical Center, University of Amsterdam, The Netherlands
b Department of Intensive Care, Academic Medical Center, University of Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 25 May 2011
Received in revised form
30 September 2011
Accepted 30 September 2011

Keywords:
Terminological system
Interface terminology
Usability evaluation
SNOMED CT

ABSTRACT

Objective: To evaluate the usability of a large compositional interface terminology based on SNOMED CT and the terminology application for registration of the reasons for intensive care admission in a Patient Data Management System.

Design: Observational study with user-based usability evaluations before and 3 months after the system was implemented and routinely used.

Measurements: Usability was defined by five aspects: effectiveness, efficiency, learnability, overall user satisfaction, and experienced usability problems. Qualitative (the Think-Aloud user testing method) and quantitative (the System Usability Scale questionnaire and Time-on-Task analyses) methods were used to examine these usability aspects.

Results: The results of the evaluation study revealed that the usability of the interface terminology fell short (SUS scores before and after implementation of 47.2 out of 100 and 37.5 respectively out of 100). The qualitative measurements revealed a high number (n=35) of distinct usability problems, leading to ineffective and inefficient registration of reasons for admission. The effectiveness and efficiency of the system did not change over time. About 14% (n=5) of the revealed usability problems were related to the terminology content based on SNOMED CT, while the remaining 86% (n=30) was related to the terminology application. The problems related to the terminology content were more severe than the problems related to the terminology application.

Conclusions: This study provides a detailed insight into how clinicians interact with a controlled compositional terminology through a terminology application. The extensiveness, complexity of the hierarchy, and the language usage of an interface terminology are defining for its usability. Carefully crafted domain-specific subsets and a well-designed terminology application are needed to facilitate the use of a complex compositional interface terminology based on SNOMED CT.

© 2011 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Interface terminologies are used for data entry into electronic medical records, facilitating collection of clinical data while simultaneously linking users’ own descriptions to structured data elements from a reference terminology such as Systematized Nomenclature of Medicine – Clinical Terms, (SNOMED CT) [1,2]. Before full implementation, the usability of such an interface terminology should be tested in clinical
practice to evaluate whether it meets its intended purpose and to discover areas for improvement. This study reports on the usability evaluation of DICE (Diagnoses for Intensive Care Evaluation), an interface terminology based on SNOMED CT deployed through a terminology application in a Patient Data Management System (PDMS) (Metavision, iMDsoft, Sassenheim, The Netherlands) to register the reasons for admission in Intensive Care (IC) [3].

It has been argued that the correctness and specificity of terminology-based data registration in clinical settings does not only depend on the content of the terminological system, but also on certain user characteristics such as their registration habits and their experience with the terminological system [4,5]. Furthermore, usability issues concerning the design of the graphical user interface of the terminology application impacts the efficacy of terminology-based data registration [4,5]. Clinicians will optimally use an interface terminology for structured data entry if its presentation and browsing of information through the terminology application is intuitive, easy to use and not time consuming [6]. Furthermore, for the human–computer interaction to be effective, the action sequences that the users have to carry out in the application should map to the user’s mental model [7].

Typically, terminologies are evaluated in terms of content coverage while terminology applications for data entry are evaluated in terms of usability [8–11]. Combined approaches to examine how clinicians interact with the terminological system during data entry are gaining interest [5,12–14]. A user’s inability to find a clinical concept using an interface terminology might be caused by a misunderstanding of the terminology content or may be due to the lack of functionalities and the graphical user interface design of the terminology application. Therefore, the evaluation of an interface terminology should not only concern the terminology content, but also the data entry application as integrated in a PDMS [1,2,5]. In this study both aspects of the DICE system were evaluated. Usability, i.e. the extent to which users can achieve specific sets of tasks in a particular environment [15], was measured on five aspects: effectiveness (accuracy and completeness of the recorded reasons for admission), efficiency (time spent to retrieve the correct concept in relation to the effectiveness), overall user satisfaction (users’ attitude with respect to the usability of the DICE system), learnability (whether the users can easily learn to use DICE and improve their performance over time), and usability problems encountered by the DICE users [1,15–18]. To assess the five usability aspects, we examined how clinicians interacted with the interface terminology during data entry. Evaluation was performed both at the baseline and three months after full implementation of the system. The purpose of evaluating the system before and after the implementation was to assess the learnability of DICE and to study whether the usability problems revealed at the baseline persisted after the system had been in routine use for three months [17]. Consequences of usability problems, such as reduced data quality, are out of scope, and hence have not been studied.

2. System description

2.1. Interface terminology and DICE application

The interface terminology based on SNOMED CT contains an IC specific subset of SNOMED CT (referred to as terminology content) with 83,125 disorder and procedure concepts, and 150,657 attribute values to further specify the disorder and procedure concepts and their English terms [3]. This core SNOMED CT subset was extended with 325 concepts, 1243 relationships between these concepts, and 597 descriptions which are not present in SNOMED CT, but are needed to cover reasons of admission in the intensive care unit (ICU). This interface terminology is deployed in DICE (referred to as terminology application), a local Simple Object Access Protocol (SOAP)-based terminology server, together with a client for terminology browsing [19].

For both medical and surgical admissions, DICE offers physicians two ways to search for the appropriate reason for
admission from SNOMED CT: (a) by presenting a short list containing the most frequently occurring medical or surgical reasons for admission in the ICU and (b) by asking the users to enter (a part of) a term for medical concepts, surgical concepts. Thereafter, the system returns all terms matching the given free-text query. When clicking on a specific term, a list of the term’s “subordinates” (i.e. child or descendant) is displayed if available. DICE is a compositional terminological system. Once a concept is selected, DICE allows composition of a new concept by qualifying the existing concept with more detail (i.e. post-coordination). However, such further specification is not mandatory. When a user is unable to retrieve a concept from the interface terminology, he is allowed to enter the diagnostic information in free text. Finally, users can also provide comments on each entry. Screenshots of the DICE user interface illustrating these different functionalities are shown in Fig. 1a and b.

3. Design objectives

3.1. Setting

This study was performed in an adult Dutch ICU with 28 beds, with more than 1500 yearly admissions. Since 2002, this ward uses a commercial PDMS. This PDMS is a point-of-care clinical information system, which runs on a Microsoft Windows platform, uses a SQL server database and includes computerized order entry; automatic data collection from bedside devices; some clinical decision support; and (free-text) documentation of clinical phrases (e.g. reasons for admission and complications) during ICU stay.

In the intensive care, diagnostic information is often recorded in different systems in free text or using a specific classification system resulting in registration insufficiency. For instance, calculation of case-mix-adjusted mortality risks in the Acute Physiology and Chronic Health Evaluation (APACHE) IV prognostic model requires a variety of patient information, such as physiological parameters, comorbidities and reason for ICU admission. The reason for ICU admission is captured using the APACHE IV classification [20]. However, the diagnostic categories in this APACHE IV classification lack detail. Therefore, the same information about the reason for intensive care admission is often separately registered in medical files and discharge letters, but with more detail and in free text. In this study, the DICE system was integrated in the PDMS to evaluate its usefulness for detailed and structured registration of reasons for intensive care admission for calculation of case-mix adjusted mortality rates.

3.2. Design

During a pilot between January 2009 and May 2009, we carried out an empirical study evaluating the usability of the DICE system. This evaluation involved 16 ICU physicians. The participants were representative of the intended end-users’ community with respect to characteristics such as job position, demographic profile, and computer experience.

First, a short demonstration of the DICE interface terminology was given to the participants. During this introductory demonstration session, the use of the DICE system was demonstrated and all basic functionalities of the system were explained using three clinical scenarios. Also the structure of the terminology content was explained. The clinical scenarios were representative for data collection in the daily care process and focused on the core functions of the system.

The pre-implementation test was conducted in January 2009 before the system was deployed in the ICU. The post-implementation test was performed in May 2009 when the system had been in routine use for 3 months (see Fig. 2). To avoid the source of error that originates from individual differences between the participants, the same subjects participated in both tests. During routine use in the ICU, users’ system actions were logged to measure their frequency of use.

3.3. Clinical scenarios

We designed 10 tasks (including a number of subtasks) based on clinical scenarios which are available as electronic supplementary material. To prevent recall bias, the tasks for the pre- and post-implementation tests were slightly different. The clinical scenarios were developed in collaboration with an experienced intensivist who was involved in the development of the interface terminology. The scenarios were similar to the scenarios used in the demonstration sessions.

The task scenarios were focused on the core functionalities of the system, namely to search, select and refine the appropriate reason for ICU admission from SNOMED CT. The first eight tasks were easier to perform and had to be carried out in a predetermined way in terms of look-up method. The last two tasks were more complicated and the look-up method was not predetermined. To prevent sequential bias, the order of the first eight tasks was randomly altered for the different participants.

3.4. Measurements

Usability problems encountered by the participants were determined with the Think-Aloud user test method [1,21]:
eight participants carried out the series of tasks while thinking-out loud, verbalizing their actions during their interaction with the system. The sessions took place in the clinical environment of the participants and lasted approximately 40–45 min. The sessions started with a short introduction of the study. A simple arithmetical problem was used to train the users with the Think–Aloud method, i.e. the participants had to solve a simple equation while thinking out loud. Thereafter, the actual test sessions started in which the users verbalized their thoughts when performing the tasks using the DICE system. The participants were encouraged to speak constantly as if they were alone in the room and they were informed that the observers (FBR and MD) would remind them to keep talking if they fell silent. Interruption by the observers was limited to a minimum; the observer only used the statement ‘keep talking’ to break silences after a fixed interval of 20 s [22]. In case the participants were unable to solve a task, they had to decide themselves when to continue with the next task.

Another eight participants solved the same series of tasks without thinking-out loud, for measuring the “Time-on-Task”. These quantitative measurements enabled the unbiased assessment of the effectiveness and efficiency of the DICE system. The participants were not interrupted during these sessions and when they were unable to solve a task, they had to decide themselves when to continue with the next task.

At the end of each session, the 16 users filled in the System Usability Scale (SUS) questionnaire to measure the overall user satisfaction and usability of the system [23]. The SUS questionnaire, is composed of 10 statements that are scored on a 5-point scale of strength of agreement. Final scores for the SUS range from 0 to 100, where higher scores indicate higher user satisfaction and better usability. The questionnaire included two additional questions on the users’ search preferences. The Morae Recorder1 3.0 was used to record the 16 pre- and post-implementation test sessions. The recordings contained video recordings and voice recordings of the participants, video recording of the PC screen, mouse click input, keyboard input, and the survey responses.

3.5. Data analyses

All data reports were analyzed by two of the authors (FBR and MD) who also served as the observers during the tests. The analysis of the qualitative measurements, based on the Think–Aloud reports, involved the following three steps: (1) Development of a coding scheme by identifying occurrences of usability problem types on 10% of the recordings from the pre-implementation test [14]. For each usability problem type, it was determined whether the encountered problem was related to the participants’ misunderstanding of the interface terminology or to the participants’ misunderstanding of the DICE application, i.e. the DICE graphical user interface design such as screen-layout or the DICE system functionalities such as navigation structure and search engine. This was determined on the basis of the users’ verbalizations and on the moment of the user-system interaction. The two evaluators independently analyzed the users’ performances in terms of outcome for each task. The individual lists with distinct usability problem types provided by the two evaluators were discussed and final consensus was reached on the problem type coding scheme. (2) Independent classification of the remaining 90% of the pre-implementation test recordings and all post-implementation test recordings by the two evaluators using the coding scheme. Interrater agreement was measured as the percentage of problem types that were equally classified by the two evaluators. Disagreements between the two evaluators were then resolved based on consensus resulting in a final list of coded usability problem types. (3) The distinct usability problem types were then classified as violations of the fourteen usability heuristics described by Zhang et al. [24].

Each problem type was also given a severity rating based on the severity scales described by Nielsen and Mack [25]. The severity ratings were based on the proportion of users who experienced a specific problem type and the persistence of the problem. Three to five evaluators are needed to independently apply this set of heuristics and severity scales [24]. Therefore, three other usability experts were additionally asked to categorize the usability problems in terms of violations of the heuristics and to provide a severity ranking for each type of usability problem. The results of the evaluators were summarized into one list. To calculate an overall severity rating for each distinct usability problem type, the ratings from the individual evaluators were averaged.

The analysis of the quantitative measurements, i.e. Time-on-Task recordings, involved measuring the length of the participant’s time on each task and the categorization of each participant’s task as being: (1) “completed”: i.e. the DICE-based reason for admission was semantically equivalent to the reason for admission as described in the scenario (e.g. “Tamponade” for “Cardiac Tamponade”), (2) “partially completed”: i.e. the DICE-based reason for admission (partially) covered the reason for admission as described in the scenario (i.e. a subordinate, subordinate, or co-ordinate concept e.g. “Coronary Artery Bypass Graft” instead of “Coronary Artery Bypass Graft with use of left internal mammary artery”), or (3) “failed”: i.e. all other cases (e.g. “alcohol abuse” instead of “acute alcoholic pancreatitis”, or no solution was provided). Disagreements between the two evaluators were resolved through discussion until consensus was reached. Interrater agreement was measured as the percentage of tasks having the same categorizations as provided by the evaluators before consensus was reached.

The Think–Aloud recordings and Time-on-Task recordings were analyzed using Morae Manager (see footnote 1) 3.0. The Wilcoxon rank test was performed to detect significant differences in the SUS scores between the pre- and post-implementation tests. The significance level was set at a P value of less than 0.05.

4. Evaluation results

4.1. Overall user satisfaction: the SUS score

The results of the SUS questionnaire indicate that the usability of the DICE system fell short of the users’ expectations.

---

In the pre-implementation test, the average SUS score was 47.2 (range: 32.5–65). In the post-implementation test, this average decreased to 37.5 (range: 0–65). So, according to the scale provided by Bangor et al. [26] (score of 0–25: worst, score of 25–39: poor, score of 39–52: OK, score of 52–85: excellent, and score of 85–100: best imaginable), the acceptability of the DICE system was “OK” in the pre-implementation test and decreased to “poor” in the post-implementation test. The change in the SUS score was statistically significant ($P = 0.018$).

4.2. Qualitative measurements: Think–Aloud analysis

The Think–Aloud analysis revealed a total number of 35 distinct usability problem types of which 33 were identified in the pre-implementation test and 27 in the post-implementation test. The mean number of distinct usability problem types per user was 4.1 (33/8) in the pre-implementation test and 3.4 (27/8) in the post-implementation test. The mean number of usability problem encounters per user was 33.6 (268/8) in the pre-implementation test and 24.4 (195/8) in the post-implementation test. The interrater agreement of the classification of the problems was 83%.

Five usability problems (14.3%) concerned the terminology content based on SNOMED CT, while the remaining 30 (85.7%) was caused by the DICE terminology application, of which 11 (31.4%) concerned the graphical user interface design and 19 (54.3%) the DICE system’s functionalities (Fig. 3). The problems concerning the DICE terminology application had a mean severity rating of 2.4. The problems concerning the terminology content were however more severe with an average rating of 3.2. Table 1 provides an overview of the major usability problems with a severity rating of 3.0 and up, indicating that the problem should be fixed with high priority. A complete overview of the usability problems including the total number of occurrences in the pre- and post-implementation tests, the frequencies of users’ encounters, the mean severity ratings, and the categorization of the problem types based on violations of the heuristics is provided in electronic supplementary material.

Table 2 provides the mean usability ratings and the violation of usability heuristics described by Zhang et al. in relation to the DICE terminology content and DICE terminology application (graphical user interface or system functionalities). Most of the problems were related to the (mis)match between the system and users’ mental model (i.e. the representation of the system perceived by users does not match the idea the users have about the system), consistency and use of standards (i.e. different situations or actions in the system mean the same thing), and visibility of the system state (i.e. the users are not informed by proper system feedback about what is going on with the system).

4.3. Quantitative measurements: Time-on-Task analysis

On average the participants needed 16.1 min (range: 11.8–22.3) to complete all ten tasks in the pre-implementation test and 14.8 min (range: 9.3–30.6) in the post-implementation test. Table 3 provides the effectiveness (as the percentage of the tasks that were completed, partially completed or failed) and the efficiency (mean time on task in minutes) in relation to the effectiveness in the pre- and post-implementation tests respectively. Tasks 2b and 5b were excluded from these analyses as for all users the mean time on task was too short to be measured. The interrater agreement concerning the classification of the task results was 91%. Overall, in the post-implementation test, participants’ effectiveness and efficiency increased (Table 4). Fig. 4 shows for each task the average time per task in minutes and the user’s effectiveness as measured in the pre- and post-implementation tests. The results show important variations. The two more complicated tasks, i.e. number 9 and 10 which involved the use of post-coordination, were poorly solved in both tests and the participants needed more time to complete these tasks.
Table 1 – An overview of the usability problem types with a severity of more than 3.0. Total freq., total number of occurrence of the problem; user freq., number of users that encountered the problem.

<table>
<thead>
<tr>
<th>Problem type</th>
<th>Pre-implementation test</th>
<th>Post-implementation test</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>DICE terminology content</td>
<td>22</td>
<td>8(8)</td>
<td>19</td>
</tr>
<tr>
<td>The user does not understand the relationships in the DICE model</td>
<td>18</td>
<td>7(8)</td>
<td>10</td>
</tr>
<tr>
<td>The user does not understand the DICE model hierarchy</td>
<td>10</td>
<td>5(8)</td>
<td>3</td>
</tr>
<tr>
<td>The level of detail in the DICE model is too extensive for the user</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DICE graphical user interface design</td>
<td>24</td>
<td>8(8)</td>
<td>14</td>
</tr>
<tr>
<td>The changes in the system state are not recognizable enough for the user</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DICE system functionalities</td>
<td>16</td>
<td>6(8)</td>
<td>17</td>
</tr>
<tr>
<td>The user does not understand the DICE post-coordination mechanism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The user is not familiar with the ‘right mouse click’ functionality to search in the list of relationships’ values to post-coordinate a concept</td>
<td>33</td>
<td>8(8)</td>
<td>16</td>
</tr>
<tr>
<td>The user misses a “back” button or an “undo” button</td>
<td>19</td>
<td>7(8)</td>
<td>18</td>
</tr>
</tbody>
</table>

4.4. Search method preference

After the test sessions, the participants were asked which search method they preferred; the search string or the list of frequently occurring diagnoses. In the pre-implementation test 11 users (69%) answered they preferred the search string and five users (31%) preferred the list of frequent diagnoses. In the post-implementation test, nine users (56%) answered they preferred the search string and seven users (44%) preferred the list of frequently occurring diagnoses.

5. Discussion and conclusions

In this study we performed a usability evaluation of a large compositional interface terminology based on SNOMED CT for...
Table 2 - Percentage of usability problem types, the related mean usability ratings per category, and the violation of usability heuristics described by Zhang et al. for the pre- and post-implementation tests in relation to the DICE terminology content or DICE terminology application (graphical user interface or system functionalities). A single type of usability problem identified could be a violation of multiple heuristics.

<table>
<thead>
<tr>
<th></th>
<th>Mean Severity (SD)</th>
<th>Match between the system and users’ mental model (%)</th>
<th>Consistency and use of standards in system design (%)</th>
<th>Visibility of system state (%)</th>
<th>Prevention of errors (%)</th>
<th>Informative system feedback (%)</th>
<th>Sufficient help and documentation (%)</th>
<th>Minimum and sufficient amount of information (%)</th>
<th>Flexibility and efficiency of the system (%)</th>
<th>Use or users’ language (%)</th>
<th>Sufficient error messages (%)</th>
<th>Possibility to undo user actions (%)</th>
<th>Minimize user memory load (%)</th>
<th>Clarity of system stages (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DICE terminology content</td>
<td>3.2 (0.43)</td>
<td>83</td>
<td>33</td>
<td>50</td>
<td>–</td>
<td>83</td>
<td>67</td>
<td>17</td>
<td>17</td>
<td>57</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>DICE terminology application</td>
<td>2.4 (0.49)</td>
<td>76</td>
<td>79</td>
<td>72</td>
<td>48</td>
<td>28</td>
<td>14</td>
<td>17</td>
<td>14</td>
<td>–</td>
<td>10</td>
<td>7</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>DICE Graphical user interface design</td>
<td>2.4 (0.41)</td>
<td>91</td>
<td>91</td>
<td>82</td>
<td>91</td>
<td>9</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>9</td>
</tr>
<tr>
<td>DICE System functionalities</td>
<td>2.4 (0.52)</td>
<td>67</td>
<td>72</td>
<td>67</td>
<td>22</td>
<td>39</td>
<td>22</td>
<td>28</td>
<td>22</td>
<td>–</td>
<td>17</td>
<td>11</td>
<td>–</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>2.7 (0.50)</td>
<td>77</td>
<td>71</td>
<td>69</td>
<td>40</td>
<td>37</td>
<td>23</td>
<td>17</td>
<td>14</td>
<td>11</td>
<td>9</td>
<td>6</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
registration of reasons for ICU admission. Overall, the results of the usability evaluation revealed that the usability of the DICE system fell short of the users’ expectations. The use of post-coordination is cumbersome and was influenced by the poor usability of the terminology application as well as the size of the terminology content. The qualitative measurements revealed a high number (n = 35) of distinct usability problems, resulting in ineffective and inefficient registration of reasons for admissions. The efficiency and effectiveness of the system, both not optimal, did not improve after three months of DICE use in practice. Since in this three month between the two tests the frequency of use of the DICE system by the individual participants was low, no firm conclusions concerning the learnability of the DICE system could be drawn. In spite of that, the decrease in the usability problem occurrence showed the users did manage to circumvent the usability problems.

About 14% (n = 5) of the usability problems concerned the terminology content based on SNOMED CT, while the remaining 86% (n = 30) was caused by the DICE terminology application, (31% concerned the graphical user interface design and 55% the DICE functionalities). The problems related to the terminology content were more severe (i.e. 3.2, indicating severe usability problems that according to Nielsen, should have been fixed before the pilot release of DICE system) than the problems related to the DICE terminology application (i.e. 2.4, indicating usability problems of moderate severity).

5.2. Related literature and recommendations for DICE terminology content

The frequently occurring violations of the heuristics described by Zhang et al. [24] show that most of the encountered problems concerning the terminology content can be attributed to a lack of informative system feedback and inadequate help and documentation functions [24]. However, given the high severity rating of and the causes underlying these problems, it is not likely that informative feedback and provision of help and documentation alone would help end-users to overcome these problems.

One of the advantages of using an interface terminology based on SNOMED CT is that users are isolated from the complexity of SNOMED CT [27]. Ideally, the subset should be as compact as possible while at the same time providing the level of granularity needed in the clinical domain of interest [28]. Previous research has shown that physicians are not used to think in terms of hierarchies and complex relationships [29], indicating the need for concise subsets drawn from larger terminologies. The large and complex subset used in our interface terminology is far from compact. A previous study showed that a subset of about 2700 concepts from SNOMED CT was sufficient to cover 96% of the clinical notes of patients admitted to an ICU over a period of 5 years [28]. In this earlier study, a large amount of clinical notes (13 million concept instances comprising about 30,000 unique concept types) was drawn from the clinical information system and was processed by natural language processing procedures including the computation of all SNOMED CT candidate codes. The resulting codes were then processed by a tool which computed the closure of the minimal sub-tree of concept types in the SNOMED hierarchy, thus inferring the complete subset of SNOMED CT that would be necessary to code these concepts [28]. Although this subset was extracted in a different setting with a different basic assumption, namely of retrospective mapping of clinical notes instead of real time data entry by physicians, it might indicate that part of the concepts included in our interface terminology can be removed from the subset. Shrinking methods as described before can be applied to decrease the number of concepts in our interface terminology [30]. However, the domain of IC is rather complex and broad, involving a diversity of clinical problems. A simple appendicitis, for instance, may lead to severe sepsis with a wide range of possible underlying microbiological agents. Furthermore, common microbiological agents in the ICU such as Legionella and Streptococcus pneumoniae are placed at different hierarchical levels in the SNOMED CT. So, to facilitate sharing...
and aggregating this kind of data for different purposes, the interface terminology should support a detailed registration of widely diverging clinical problems and their attributes. It therefore remains hard to determine which parts of our subset can be further restricted.

We believe that to deal with the problems related to the DICE terminology content, the focus should be on the use of value sets and the arment and hierarchy of the concepts in the terminology content rather than on its size [31]. For instance, the 1900 descendants of “virus” in SNOMED CT have a general arrangement purpose as SNOMED CT is a reference terminology. Reference terminologies are meant to provide precise meaning and structure to concepts by formal concept definitions required for consistent and computer-readable coding and storage of clinical data [2,32]. In an interface terminology, the arrangement of concepts could yet be based on the user definitions of these data and routines to ease the data entry process. The lack of specific literature in this area emphasizes the need for further research on the contextual adjustments of SNOMED CT subsets for use in DICE interface terminology and other interface terminologies based on SNOMED CT.

5.3. Problems on the DICE terminology application

Most of the usability problems concerned the DICE terminology application. With regard to the DICE graphical user interface, most of the usability problems had a low severity and were related to suboptimal screen layout, wrong positioning of the buttons on the screen and mislabelling of the buttons. The usability problems identified for the DICE system functionalities and navigation structure were more severe. The DICE system functionalities and their presentation through the DICE graphical user interface are not in line with the user’s mental model and do not optimally support the users in searching through a large and complex terminology hierarchy [7]. Furthermore, the DICE application does not fully support the use of post-coordination in SNOMED CT. For instance, a medical disorder or procedure may be further characterized by the relation ‘due to’, ‘causative agent’, ‘finding site’, or ‘finding site direct’. Most of these characteristics are predefined in the terminology and represent standard values inherited from parent concepts (e.g. “finding site” for viral pneumonia is “lung structure”), but some need to be further specified by post-coordination (e.g. “causative agent” for viral pneumonia is “virus” which can be further specified). Analyses of the cognitive process protocols revealed that the users did not understand the post-coordination principle, i.e. they did not understand which characteristics needed further specification and which were predefined. Consequently, many of the users avoided the use of post-coordination and preferred to search for a pre-coordinated, yet not always available, concept. The quantitative measurements confirm this: the two more complicated tasks, i.e. numbers 9 and 10 which involved the use of post-coordination, were poorly carried out and the participants needed more time to complete these tasks.

5.4. Related literature and recommendations for DICE terminology application

To support optimal recording of clinical data for different purposes, the terminology application should likewise
be improved so that it optimally supports the use of the complex and granular terminology by for instance a more user-friendly graphical user interface design, support for the use of post-coordination, a more adequate search mechanism, and intuitive navigation structure for performing clinical data recording tasks more effectively and efficiently. Intuitiveness of such a navigation structure is obviously to a certain degree correlated with the arrangement and hierarchy of the concepts in the terminology content which should map to the user’s mental model.

Concerning the DICE graphical user interface design, most of the encountered problems could be attributed to a lack of use of standards for good user interface as is apparent from the frequently occurring violations of the heuristics described by Zhang et al. [29]. A majority of these problems were of low severity and therefore require low priority in redesign efforts. The only major problem was that system changes after user actions are often not visible to the user. Once the severe problems related to the DICE functionalities and navigation structure are dealt with, the problems in the graphical user interface are relatively easy to fix using the available standards [24,33].

In relation to the DICE system functionalities, using rules to constrain how concepts can be combined will enhance their usability [2,34,35]. Available standards could also be used to design the search mechanism in the DICE system which was also a major problem. Previous studies have demonstrated that usability of an interface terminology is highly related to the efficiency of the search mechanism and term selection [36–38]. In DICE, the users were not always content with the search mechanism which resulted in a large number of usability problems. However, users indicated to prefer using the search mechanism instead of using the list of frequently occurring diagnoses in selecting a detailed reason for ICU admission. Therefore, the search functionality should be better understandable and should include features to help the users find easily what they are looking for. We suggest to enhance the search mechanism in the DICE terminology application with query expansion algorithms, or auto-completion algorithms which aim to ease the search procedure [39,40].

The graphical presentation of concepts is also a promising way to navigate and select a concept in an interface terminology. Clicking on a body part in a 3D model of the human body for example would result in a list of related concepts [41]. However, for complex medical domains requiring a large subset of SNOMED CT, as is the case in DICE, data entry through graphical representation can be time consuming and especially post-coordination can be cumbersome. Furthermore, some clinical findings such as systemic disorders (e.g. sepsis, or AIDS) are hard to be represented graphically.

The most user friendly way to capture encoded patient data into an electronic patient record may be by automated coding, in which natural language processing (NLP) algorithms are used to encode free-text input [42–45]. In general, the use of NLP in clinical practice yields the promise to shield the users from the complexity of the underlying complex and granular terminologies such as SNOMED CT and to ease the data entry process. Accordingly, the use of NLP could overcome most of the problems related to the DICE application mentioned. However, the terminology applications used for NLP algorithms would also require enhanced components to automatically identify and manage the clinical information in the free text and to suggest relevant SNOMED CT concepts to the users for confirmation. Furthermore, to facilitate a detailed registration of clinical problems, these terminology applications would also need components to stimulate clinicians to specify clinical entries via post-coordination when necessary. Another drawback of using NLP is that the tools used in the terminology applications are usually domain and language specific, and, if not available, costly to build [46,47]. In our case, a translation of SNOMED CT terms into Dutch is required to facilitate the use of NLP in Dutch ICUs. So, further research is needed to evaluate the use of NLP for ease of data entry using a subset of SNOMED CT in the ICU. To this end, formative evaluation methods in which the usability of the system is part of the system’s development cycle should be used [48]. It is widely acknowledged that user-based usability studies throughout the system development are crucial in designing systems that fulfil end users’ requirements [49]. Yet, as was the case in our study, usability evaluation studies mostly aim at gaining insight into system aspects that influenced its adoption and therefore provide summative results.

5.5. Limitations of the study

One weakness of our study is that the learnability of the system could not be measured adequately. More than half of the participants did not use or hardly used the DICE system between the pre- and post-implementation tests; some did not have time because of the nightshifts, whereas others left the registration to resident physicians who were not involved in this study. Nevertheless, those participants who did use the system several times between the pre- and post-implementation tests showed only modest improvement in their task effectiveness and efficiency. In contrast, one participant who never used the system did improve on his task effectiveness while his task efficiency was minimally reduced. Probably, the low frequency of system use and the short time period between the pre- and post-implementation tests impeded an adequate assessment of the learnability of the DICE system.

For each usability problem type, it was determined whether the encountered problem was related to the participants’ misunderstanding of the terminology content or to the terminology application. This was determined on the basis of the users’ verbalizations and on the moment of user–system interaction with the DICE system. A limitation of the study is that the distinction between the problems related to the terminology content and the DICE application was not clear in all cases. However, the high interrater agreement, 83%, shows that in most of the cases the evaluators agreed which problem was related to the terminology content and which to the DICE application.

6. Conclusions and directions for further research

This study provides a detailed insight in how clinicians interact with a large controlled compositional interface
Summary points
What was known before the study?
- Interface terminologies are used for actual data entry into electronic records, facilitating collection of clinical data while simultaneously linking users’ own descriptions to structured data elements from a terminology.
- Clinicians will optimally use an interface terminology for structured data entry if the presentation of the interface terminology by the terminology application is intuitive, easy to use and not time consuming.
- The evaluation of an interface terminology should not only concern the interface terminology itself, but also the data entry application as integrated in an Electronic Health Record system.

What was added by this study?
- The extensiveness, complexity of the hierarchy, and the language usage of an interface terminology is defining for its usability.
- The use of post-coordination in practice is affected by structure and size of the interface terminology.
- The usability problems can be linked to either the interface terminology or the terminology application. Nevertheless the source for the inadequate user performances is found in the joint effect.

data interpretation and critically revised all drafts of the article. MD contributed to the acquisition of data, analyses and interpretation of data and reviewed all drafts of the article. DD contributed to the acquisition of data and the interpretation of the data. He critically revised all drafts of the article. All authors finally approved the final manuscript.

Conflict of interest
None.

Acknowledgements
The authors thank the participants of the usability evaluation sessions for their cooperation in conducting this study. The authors would also like to express their gratitude to the usability experts R. Khajouei and L. Peute for their contribution in rating the severity of the observed usability problems.

Appendix A. Supplementary data

REFERENCES


34. L.R. Horrocks, A comparison of two terminological knowledge representation systems [master’s thesis], Manchester, UK, University of Manchester. 1995. Ref Type Generic.


