Terminological systems and prognostic models as instruments for quality assessment in intensive care

Raiez, F.

Citation for published version (APA):
A Pre-post Usability Evaluation of a Large Compositional Interface Terminology based on SNOMED CT for Intensive Care

Ferishta Bakhshi-Raiez, Nicolette F. de Keizer, Ronald Cornet, Marlijne Dorrepaal, Dave Dongelmans, Monique W.M. Jaspers

Submitted
ABSTRACT

Objective: To evaluate the usability of a large compositional interface terminology based on SNOMED CT. The interface terminology is deployed through a terminology client to register the reasons for intensive care admission in a Patient Data Management System.

Design: Observational study with user-based usability evaluations before and three months after the system had been routinely used.

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

Results: The results of the evaluation study revealed that the usability of the interface terminology fell short (SUS score pretest: 47.2, posttest: 37.5). The qualitative measurements revealed a high number (n=35) of distinct usability problems, resulting in ineffective and inefficient registration of reasons for admissions. This remained steady over time. About 30% of the encountered usability problems was related to the interface terminology based on SNOMED CT, while the remaining 70% was regarded to the terminology client. The problems related to the interface terminology were more severe than the problems related to the terminology client.

Conclusions: This study provides a detailed look at how clinicians interact with a controlled compositional terminology. For any interface terminology, there are tradeoffs between terminology granularity and ease of use. Adjusted terminology content and well designed terminology clients can facilitate the use of a complex compositional interface terminology based on SNOMED CT. Formative and summative evaluations are recommended during the development of a new interface terminology.
4.1 Introduction

SNOMED CT is regarded as the most comprehensive reference terminology for coding clinical information in an electronic patient record. However, its use for actual data entry in a clinical setting is the subject of discussion in many studies (1-3). If all of SNOMED CT is provided to users for systematic data collection, its comprehensiveness forms an impediment, as it contains large amounts of concepts that are irrelevant for most clinical domains. Furthermore, the extensiveness of SNOMED CT, covering all kinds of medical usage, does yet not guarantee that SNOMED CT adequately covers the details for data collection in a specific clinical setting (1-3). Therefore, instead of providing all of SNOMED CT to its users, among others an interface terminology containing a subset of SNOMED CT is proposed to provide easier adjustment to the user requirements in a specific clinical setting (4,5). Interface terminologies are used for actual data entry into electronic medical records, facilitating display and collection of clinical data in a simple way while simultaneously linking users’ own descriptions to structured data elements in a reference terminology or aggregate terminology (1,6). In the interface terminology only the relevant content of SNOMED CT is included. Next, relevant concepts, synonyms, descriptions and terms not included in SNOMED CT but needed for coverage of detailed medical data in a particular clinical setting are added to the subset. Before striving for full implementation, such an interface terminology should be tested on its usability in clinical practice to evaluate whether it meets its intended purpose or to discover areas for improvement. This study reports on the usability evaluation of DICE (Diagnoses for Intensive Care Evaluation), an interface terminology based on a subset of SNOMED CT. DICE is deployed in a Patient Data Management System (PDMS) to register the reasons for Intensive Care (IC) admission (7).

It has been argued that 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 the registration habits and experience of users with the terminological system, and on usability issues concerning the performance and the design of the graphical user interface of the terminology client (8,9). Clinicians will optimally use an interface terminology for structured data entry if the presentation of the interface terminology in its client is intuitive, easy to use, complete, and not time consuming (10). Consequently, for the human computer interaction to be effective, the action sequences that the system users have to perform should be structured in a logical and consistent manner. Furthermore, the system functions should correspond with the goals of the system, and the presentation of the information should build on the users’ system image.

Typically, terminologies are evaluated in terms of content coverage while user interfaces for data entry are evaluated in terms of usability (11-14). Few studies have considered a
combined approach to examine how clinicians interact with the terminological system during data entry using a terminology client (9;15-17). The cause for a user’s inability to find a clinical concept using an interface terminology might be rooted in the misunderstanding of the terminology content or in the terminology client, i.e. its graphical user interface design or its functionalities. Therefore, effective evaluation of an interface terminology should not only concern the interface terminology itself but also the data entry application integrated in Electronic Health Records (HER) (1;6;9). In this study, this combined approach was used to evaluate the usability of the DICE system. Usability, i.e. the extent to which a system can be used by its end users to achieve specified goals, will be measured on five aspects: effectiveness (accuracy and completeness with which users achieve specified goals), efficiency (time spent in relation to the accuracy and completeness with which users achieve goals), overall user satisfaction (users’ attitude to the use of the system), learnability (capability of a system to enable the users to learn it easily and improve their performance over time), and experienced usability problems (6;18-21). To assess these five aspects, we examined how clinicians interacted with the interface terminology during data entry. Besides, we explored areas for improvement of the interface terminology. Different usability evaluation methods were used in a pre-post design. The purpose of the posttest was to evaluate the learnability of the system and the persistency of the problems encountered after the system had been in routine use for three months (20).

4.2 Interface terminology and DICE application

The interface terminology based on SNOMED CT contains an IC specific subset of SNOMED CT with 233,782 concepts (i.e. 83,125 disorder and procedure concepts and 150,657 attribute values to further specify the disorder and procedure concepts) and their English terms (7).

This core SNOMED CT subset was extended with 325 concepts, 1243 relationships for these concepts, and 597 descriptions. This interface terminology is deployed in DICE, a local SOAP-based terminology server, together with a client for terminology browsing (22). The implementation of DICE offers physicians two ways to search for the appropriate reason for admission dependent on admission type (i.e. medical or surgical): (a) a short list containing the most frequently occurring medical or surgical reasons for admission in an Intensive Care Unit (ICU), and (b) entry of (a part of) a preferred or synonymous term for medical concepts, surgical concepts or both. Thereafter, the system returns all terms matching the given free-text query. Clicking on a term, selected by one of the two methods, brings up, a list of the term’s “subordinates” (i.e. child or descendent). DICE is a compositional terminological system. Once a concept is selected, DICE enables the composing of new concepts by qualifying existing concepts with more detail (i.e. post-
coordination). However, such further specification is not mandatory. When a system user is unable to retrieve the 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 provided in Appendix 4.1.1 to 4.1.5.

4.3 Methods

4.3.1 Setting
This study took place in an adult Dutch ICU with 24 beds, with more than 1500 yearly admissions. Since 2002, this ward uses a commercial PDMS, Metavision. 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. As part of the National Intensive Care Evaluation (NICE) project, a national registry on quality assurance of Dutch ICUs, for each patient a minimal dataset among which the reason for admission is extracted from the PDMS 24 hours after admission. In this study, the DICE system is integrated in the PDMS to evaluate its usefulness for structured registration of reasons for ICU admission.

4.3.2 Design
During a pilot from December 2008 to May 2009, we undertook an empirical study on the usability of the DICE system. As shown in Figure 4.1, the basic design of the study was to conduct a pre-post usability evaluation of the system with novice users. This evaluation involved 16 ICU staff members. The participants were representative for the intended end users’ community with respect to parameters 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 introduction session, the use of the DICE system was demonstrated and all basic functionalities of the system were explained using three clinical scenarios. Also the hierarchy of the interface terminology was discussed. The clinical scenarios were representative for data collection in the daily care process and were centered on the core functions of the system.
The pretest was conducted in December 2008 before the system was deployed in the ICU. The posttest was performed in April 2009 when the system had been in routine use for three months. To avoid the source of error that originates from individual differences between the
participants, the same subjects participated in the pretest and the posttest. During routine use, users’ system actions were logged to measure their frequency of use.

Figure 4.1 Overview of the study design.

4.3.3 Clinical Scenarios

We designed 10 tasks (including a number of subtasks) based on clinical scenarios. To prevent recall bias, the task scenarios for the pretest and posttest were slightly different. The clinical scenarios were developed in collaboration with an experienced intensivist who had been involved in the development of the interface terminology. The scenarios were similar to the scenarios used in the training sessions.

The task scenarios were focused on the core functions of the system, namely to search, select and refine the appropriate reason for ICU admission and covered all functionalities of the system. The first eight tasks were easier to perform and had to be executed in a predetermined way in terms of look-up methods. The last two tasks were more complicated and were not preconceived. To prevent sequential bias, the order of the first eight tasks was randomly altered for the different participants. The scenarios for the pretest and posttest are provided in Appendix 4.2.1 and 4.2.2.

4.3.4 Measurement

The qualitative measurements enabled the identification of the usability problems encountered by the participants. These were based on the Think-Aloud method (6;23) where eight participants performed the series of task scenarios while thinking-out loud, verbalizing their actions in interaction with the system.
The sessions took place in the clinical environment of the participants and lasted approximately 40 to 45 minutes. The sessions started with a short introduction of the study. A simple arithmetical problem was used to train the users to 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 in performing the tasks using the DICE system. To consistently apply the Think-Aloud method, 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 experimenter only used the statement ‘keep talking’ to break silences after a fixed interval of 20 seconds (24). 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 task scenarios without thinking-out loud, for measuring the “Time-on-Task”. These quantitative measurements enabled the unbiased assessment of the effectiveness and efficiency of the 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 a modified version of the System Usability Scale (SUS) questionnaire to measure the overall user satisfaction and usability of the system. The SUS questionnaire consists of 10 questions on user satisfactions and system usability with a 5-point Likert scale to calculate an overall usability score between 0 and 100 (25). The modified questionnaire included two additional questions on users’ search preferences in the terminology content.

The Morae Recorder1 3.0 was used to record the 16 pretest and 16 posttest sessions. The recordings contained video recordings and voice recordings of the participant, video recording of the PC screen, mouse click input, keyboard input, and the survey responses.

4.3.5 Data analyses

All data reports were analyzed by two of the authors (FBR and MD) who also served as the observers during the pretest and posttest.

The analysis of the qualitative measurements, Think-Aloud reports, involved the following three steps:

1): Development of a coding scheme by identifying specific occurrences of types of usability problems and aspects of cognitive processes on 10% of the recordings from the pretests (26). For each usability problem type, it was determined whether the encountered problem was related to participants’ misunderstanding of the interface terminology based

on SNOMED CT, or to participants’ misunderstanding of the DICE application, i.e. DICE graphical user interface design characteristics or the DICE system functionalities. This was determined on the basis of the users’ verbalizations and the moment of 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 individual problem types, resulting in a merged list of unique problem types.

2): Independent classification of the remaining 90% of the pretest recordings and all posttest recordings by the two evaluators based on the coding scheme. Interrater agreement was measured as the percentage of corresponding problem classifications by the two evaluators. Disagreements between the two evaluators were then resolved based on consensus resulting in a final list of usability problems.

3): Categorization of the usability problems as violations of the fourteen usability heuristics described by Zhang et al. (27) as presented in Table 4.1, and severity rating of each usability problem based on the severity scales described by Nielsen (28) as presented in Table 4.2. The severity ratings were grounded on the proportion of users who experienced a specific problem, the impact it had on their experience with the system, and the persistency of the problem. Concerning validity of the results, three to five evaluators are needed to independently apply this set of heuristics and severity scales (27). Therefore, three other usability experts were additionally asked to categorize the usability problems as a violation of the heuristics and to provide a severity rating for each type of usability problem.

A single type of usability problem identified by an evaluator could be a violation of multiple heuristics. Once all evaluators had categorized the types of usability problems as violations of at least one of the fourteen heuristics described by Zhang et al. and had assigned a severity score, the results 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 each participant’s time on task and categorization of each participant’s task as being “Completed”, “Partially completed”, or “Failed”. A task was considered “Completed” when 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”). A task was considered “Partially completed” when the DICE-based reason for admission partially covered the reason for admission as described in the scenario (i.e. super-ordinate, subordinate, or a co-ordinate concept such as “Coronary Artery Bypass Graft” instead of “Coronary Artery Bypass Graft with use of Left internal mammary artery”).
A task was considered “Failed” in all other cases (e.g. “Alcohol abuse” instead of “Acute alcoholic pancreatitis”). Disagreements between the two evaluators were resolved through discussion until consensus was reached. Interrater agreement was measured as the percentage of corresponding categorizations of the task results provided by the evaluators individually.

The Think-Aloud recordings and Time-on-Task recordings were analyzed using Morae Manager² 3.0.

Table 4.1 The fourteen usability heuristics described by Zhang et al.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Consistency and standards: users should not have to wonder whether different situations or actions mean the same thing.</td>
</tr>
<tr>
<td>2</td>
<td>Visibility of system state: users should be informed about what is going on with the system through appropriate feedback and display of information.</td>
</tr>
<tr>
<td>3</td>
<td>Match between system and world: the image of the system perceived by users should match the model the users have about the system.</td>
</tr>
<tr>
<td>4</td>
<td>Minimalist: any extraneous information is a distraction and a slow-down.</td>
</tr>
<tr>
<td>5</td>
<td>Minimize memory load: users should not be required to memorize a lot of information to carry out tasks.</td>
</tr>
<tr>
<td>6</td>
<td>Informative feedback: users should be given prompt and informative feedback about their actions.</td>
</tr>
<tr>
<td>7</td>
<td>Flexibility and efficiency: users should be given the flexibility of creating customization and shortcuts to accelerate their performance.</td>
</tr>
<tr>
<td>8</td>
<td>Good error messages: the messages should be informative such that users can understand the nature of errors, learn from errors, and recover from errors.</td>
</tr>
<tr>
<td>9</td>
<td>Prevent errors: it is always better to design interfaces that prevent errors from happening in the first place.</td>
</tr>
<tr>
<td>10</td>
<td>Clear closure: every task has a beginning and an end. Users should be clearly notified about the completion of a task.</td>
</tr>
<tr>
<td>11</td>
<td>Reversible actions: users should be allowed to recover from errors.</td>
</tr>
<tr>
<td>12</td>
<td>Use users’ language: the language should be always presented in a form understandable by the intended users.</td>
</tr>
<tr>
<td>13</td>
<td>Users in control: the users should not have the impression that they are controlled by the systems.</td>
</tr>
<tr>
<td>14</td>
<td>Help and documentation: system should always provide help when needed.</td>
</tr>
</tbody>
</table>

Table 4.2 The severity scales described by Nielsen.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Not a usability problem at all</td>
</tr>
<tr>
<td>1</td>
<td>Cosmetic problem only, need not be fixed unless extra time is available</td>
</tr>
<tr>
<td>2</td>
<td>Minor usability problem, fixing this problem should be given low priority</td>
</tr>
<tr>
<td>3</td>
<td>Major usability problem, important to fix with high priority</td>
</tr>
<tr>
<td>4</td>
<td>Usability catastrophe. Imperative to fix this before product can be released</td>
</tr>
</tbody>
</table>

4.4 Evaluation results

4.4.1 Qualitative measurements: Think-Aloud analysis

Table 4.3 provides a complete overview of the usability problem types revealed in the Think-Aloud sessions in the pretest and posttest. For each usability problem type, the total number of occurrences in the pretest and in the posttest, the user frequency, its categorization based on violations of the heuristics, and its mean severity rating are provided.

The Think-Aloud analysis revealed a total number of 35 distinct usability problem types of which 33 were identified in the pretest and 27 in the posttest. Twenty-five out of the initial 33 problems reoccurred in the posttest. The mean number of distinct usability problem types per user was 1.89 in the pretest and 1.68 in the posttest. The mean number of usability problem encounters per user was 33.63 in the pretest and 24.5 in the posttest. Table 4.4 provides an overview of the percentage of distinct usability problem types per category for the pretest and posttest. The interrater agreement of the classifications of the problems by the two evaluators was 83%.

Most problems were categorized as DICE application problems with a mean severity rating of 2.51, indicating moderate usability problems. The problems related to the interface terminology were more severe with an average of 3.55, indicating major usability problems that, according to Nielsen, should have been fixed before the pilot system release.

4.4.2 Quantitative Measurements: Time-on-task analysis

On average the participants needed 16.1 minutes (Range: 11.75-22.34) to complete all tasks in the pretest and 14.8 minutes (Range: 9.30-30.60) in the posttest.

Table 4.5 provides the effectiveness (as a percentage of the completely solved, partly solved, and failed tasks) and the efficiency (mean time on task in minutes) in relation to the effectiveness in the pretest and posttest. Task 2b and 5b are excluded from these analyses. The interrater agreement concerning the classification of the task results was 91%.

Overall, in the posttest, participants’ effectiveness and efficiency in using the DICE system increased. However, as presented in Table 4.6, this improvement in effectiveness and efficiency, indicating system learnability, varied for the individual participants.

For each task Figure 4.2 shows the average time on task in minutes and the difference in the task results concerning users’ effectiveness in the pretest and posttest. The results enfold important variations and show that the two more complicated tasks, i.e. number 9 and 10 which involved the use of post-coordination, both were poorly solved in the pretest and posttest and that the participants likewise needed more time to complete these tasks.
Table 4.3 An overview of the usability problems from the Think-Aloud analyses in the pretest and posttest. Total Freq.: Total number of occurrence of the problem. User Freq.: Number of users that encountered the problem. + Heuristics and severity score.

<table>
<thead>
<tr>
<th>Nr.</th>
<th>Problem</th>
<th>Pretest</th>
<th>Posttest</th>
<th>Heuristics</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>DICE terminology content</strong></td>
<td>Total Freq</td>
<td>Total Freq</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>The user does not understand the relationships in the DICE model</td>
<td>22</td>
<td>19</td>
<td>Minimalist, Language, Informative feedback, Help and Documentation, Match</td>
<td>3.5</td>
</tr>
<tr>
<td>2</td>
<td>The user does not understand the DICE model hierarchy</td>
<td>18</td>
<td>10</td>
<td>Language, Informative feedback, Help and Documentation, Match</td>
<td>3.7</td>
</tr>
<tr>
<td>3</td>
<td>The user does not understand the DICE post-coordination mechanism</td>
<td>16</td>
<td>17</td>
<td>Informative feedback, Visibility, Match, Help and Documentation</td>
<td>3.7</td>
</tr>
<tr>
<td>4</td>
<td>The level of detail in the DICE model is too extensive for the user</td>
<td>10</td>
<td>3</td>
<td>Informative feedback, Visibility, Minimalist, Flexibility</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Description</td>
<td>Score</td>
<td>Priority</td>
<td>Notes</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>------------------------------------------------------------------------------</td>
<td>-------</td>
<td>----------</td>
<td>---------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>The DICE content is mainly in English and the user has a problem with searching in English</td>
<td>4/8</td>
<td>4/8</td>
<td>DICE model contains a part of the SNOMED CT model which is built using the English language. Some users tried to search with Dutch terms.</td>
<td>Language, Informative feedback, Match, Consistency</td>
</tr>
<tr>
<td>27</td>
<td>The representation of the APACHE IV classification in the DICE model is unknown to the user</td>
<td>1/8</td>
<td>1/8</td>
<td>Besides the SNOMED CT content, DICE also comprises of the APACHE IV classification. If wanted, this classification can be used to register the reason of admission. The APACHE IV classification is marked by numbers in front of the concepts and sometimes with the Dutch preferred terms. The user indicated not to understand the meaning of the numbers.</td>
<td>Language, Informative feedback, Match, Consistency, Help and Documentation</td>
</tr>
</tbody>
</table>

**DICE terminology client: DICE graphical user interface design**

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Score</th>
<th>Priority</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>The changes in the system are not recognizable enough for the user</td>
<td>8/8</td>
<td>1/8</td>
<td>“The list of selected diagnoses” is not visible enough. Font is too small and changes are not noticeable. This list changes by adding, removing or clicking on a diagnosis in the list. The user deleted or added a diagnosis without noticing it.</td>
</tr>
<tr>
<td>7</td>
<td>User mistakenly uses the button &quot;remove diagnosis&quot; as 'back' button</td>
<td>4/8</td>
<td>4/8</td>
<td>The caption for the button &quot;remove diagnoses&quot; (&lt;&lt;) or &quot;empty list of selected diagnoses&quot; (&lt;&lt;&lt;) is often used as back button to undo an action or to return to the previous or main screen (i.e. &quot;List of frequently used diagnoses). However, by pressing this button (a part of) the list of selected diagnoses is removed.</td>
</tr>
<tr>
<td>8</td>
<td>The user is not able to find the button &quot;delete diagnosis&quot; at once</td>
<td>6/8</td>
<td>6/8</td>
<td>In a scenario the user is asked to delete the last added diagnosis from the list of selected diagnoses. However, the user cannot the button &quot;delete diagnosis&quot; right away. The user is clicking all possible buttons until the action is performed. This may have consequences for the list of selected diagnosis. This is time consuming and users are confused which button to press, probably because the button is on the other side on the screen.</td>
</tr>
<tr>
<td>9</td>
<td>User does not recognize the gray specification bar as a button to open the post-coordination window</td>
<td>5/8</td>
<td>6/8</td>
<td>User does not recognize the gray specification bar as a button to open the specification window. In most of the cases, the user kept clicking on a diagnosis in the list to end up at a leaf and automatically get the specification window. This is often not necessary, as the user can open this window easily. It is time consuming and right concepts may not be found.</td>
</tr>
<tr>
<td>10</td>
<td>The font is too small</td>
<td>9</td>
<td>6/8</td>
<td>6</td>
</tr>
<tr>
<td>11</td>
<td>The user is not able to find the button &quot;List of frequently used diagnosis&quot; at once</td>
<td>11</td>
<td>4/8</td>
<td>8</td>
</tr>
<tr>
<td>12</td>
<td>The domain-selection tab for the &quot;List of frequently used diagnosis&quot; is not visible enough</td>
<td>2</td>
<td>1/8</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>The caption of the buttons in the free-text window is small</td>
<td>1</td>
<td>1/8</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>The user is not able to find the button &quot;Add free text diagnosis&quot; at once</td>
<td>1</td>
<td>1/8</td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>The user is not able to find the button &quot;Add relationships value&quot; at once</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>The user is not able to find the button &quot;Empty list the of selected diagnoses&quot; at once</td>
<td>1</td>
<td>1/8</td>
<td>1</td>
</tr>
</tbody>
</table>
Usability Evaluation of an Interface Terminology based on SNOMED CT

<table>
<thead>
<tr>
<th>DICE terminology client: DICE system functionalities</th>
<th>33</th>
<th>16</th>
<th>6/8</th>
<th>Consistency, Visibility, Match, Prevent Error, Help and Documentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>The user is not familiar with the 'Right mouse click' functionality to search in the list of relationships' values</td>
<td>8/8</td>
<td>6/8</td>
<td>The functionality “Right mouse click” to search in the values of the characteristics is not recognized. A user can add a characteristic with a value, for example ‘due to (characteristic) Cholangitis (value)’. When due to is selected a very long list appears. To search through this list using a search string, the user can click on the right mouse button. The user tries to search through the hierarchy by clicking on the tree structure. This is time consuming and results in user frustration.</td>
<td></td>
</tr>
<tr>
<td>The user misses a “back” button or a “undo” button</td>
<td>7/8</td>
<td>8/8</td>
<td>There is no “back” button and user misses the functionality to undo an action. The only option is to start over from the main screen again.</td>
<td></td>
</tr>
<tr>
<td>The user is not familiar with the restrictions for the search string</td>
<td>6/8</td>
<td>4/8</td>
<td>The search functionality in DICE is very limited. The user cannot search with (parts of) two different search terms. These limitations and restrictions of the search string were unknown to the user leading to user frustration. The system feedback was minimum.</td>
<td></td>
</tr>
<tr>
<td>The user is not familiar with the domain specification functionality for the search string</td>
<td>7/8</td>
<td>6/8</td>
<td>The user can limit the search results by searching in a specific domain. This functionality is next to the search bar and four options can be chosen. This functionality is not visible enough to the users. Using the functionality can save time and if users are searching in the wrong domain, wrong concepts might be registered.</td>
<td></td>
</tr>
<tr>
<td>The list of diagnosis is too long</td>
<td>5/8</td>
<td>6/8</td>
<td>The list of search results was sometimes too long (more as 100 items). The user had to search through this long list and got frustrated.</td>
<td></td>
</tr>
<tr>
<td>Wrong assumption that free text or comments must be in English</td>
<td>5/8</td>
<td>-</td>
<td>The user has the wrong assumption that free text or comments must be in English, because DICE is in English. These fields are meant for feedback to the maintainers developers of the DICE and can be in Dutch. The user marked the use of English language as time consuming.</td>
<td></td>
</tr>
</tbody>
</table>

66
23. The user has the wrong assumption that the domain of the search string is determined in the domain-specification-tab of the "List of frequently used diagnoses". The user has the wrong assumption that the domain of the search string is determined in the domain-specification-tab of the "List of frequently used diagnoses". This tab is only applicable for the "List of frequently used diagnoses". The domain of the search string is determined using functionality.

24. The user uses an unspecific search string, resulting in a long list of diagnosis. The user uses a search string that is not specific enough. For instance, the first 3 letters of the search term are used to start a search string. This lead to a long list of results to search through.

25. The user uses a too specific search string resulting in system error. An error is given if the user searches with a too specific search string. The system does not support the user by providing the possible search terms.

26. The user has the wrong assumption that double clicking is necessary to add a diagnosis to the "List of selected diagnoses". Clicking on a diagnosis adds it to the "List of selected diagnoses". Some users thought they had to 'double click' on the items to achieve this goal. However, the second click is seen as a separate system action which opens the specification window for the selected concept. The function 'double click' will work when you add a value of a characteristic to the "List of selected diagnoses". This was confusing for the users.

28. The user starts a search without entering a search string. The user pressed the search button without entering any search string. It was not clear that they had to fill in something. This caused an error telling that the concept was not found. The feedback of the system was minimal.

29. The user has the wrong assumption to start a search string from the main window. The user had the wrong premise that a search action should be started from the main window with the "List of frequently used diagnoses".
Usability Evaluation of an Interface Terminology based on SNOMED CT

<table>
<thead>
<tr>
<th></th>
<th>The user makes a typing error resulting in the search term. This resulted in a system error telling that the concept was not found. The feedback of the system was minimal and confusing.</th>
<th>Informative feedback, Good error message</th>
<th>2.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>The Scroll bar for the “List of selected diagnoses” is unnecessary</td>
<td>Consistency, Visibility, Match, Minimalist, Reversible actions</td>
<td>2.8</td>
</tr>
<tr>
<td>31</td>
<td>The entered comments in the comment field are not visible enough. The comments entered in the comments box are not visible for the user. The comments are saved but the system does not provide feedback on the status of entry. This is confusing for users.</td>
<td>Consistency, Visibility, Match, Informative feedback</td>
<td>2.2</td>
</tr>
<tr>
<td>32</td>
<td>The user has the wrong assumption that “List of selected diagnoses” is editable</td>
<td>Flexibility, match,</td>
<td>1.8</td>
</tr>
<tr>
<td>33</td>
<td>Selection of a diagnosis in the “List of selected diagnoses” is not visible enough</td>
<td>Consistency, Visibility, Match</td>
<td>1.6</td>
</tr>
<tr>
<td>34</td>
<td>The Dutch preferred terms in the “List of frequently used diagnosis” cause user confusion</td>
<td>Consistency, Visibility, Match</td>
<td>2</td>
</tr>
</tbody>
</table>
Table 4.4 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 pretest and posttest in relation to DICE terminology content or DICE terminology client (graphical user interface or system functionalities). A single type of usability problem identified could be a violation of multiple heuristics.

<table>
<thead>
<tr>
<th>DICE terminology content</th>
<th>Pretest %</th>
<th>Posttest %</th>
<th>Mean Severity (SD)</th>
<th>Consistency %</th>
<th>Visibility %</th>
<th>Match %</th>
<th>Minimalist %</th>
<th>Minimize memory %</th>
<th>Informative feedback %</th>
<th>Flexibility %</th>
<th>Good error messages %</th>
<th>Prevent errors %</th>
<th>Clear closure %</th>
<th>Reversible actions %</th>
<th>Language %</th>
<th>Users in control %</th>
<th>Help and documentation %</th>
</tr>
</thead>
<tbody>
<tr>
<td>DICE terminology client</td>
<td>28.0</td>
<td>29.2</td>
<td>3.25 (0.43)</td>
<td>33</td>
<td>50</td>
<td>83</td>
<td>17</td>
<td>-</td>
<td>83</td>
<td>17</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>67</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>User interface design</td>
<td>72.0</td>
<td>70.8</td>
<td>2.51 (0.53)</td>
<td>79</td>
<td>72</td>
<td>76</td>
<td>17</td>
<td>3</td>
<td>28</td>
<td>14</td>
<td>10</td>
<td>48</td>
<td>3</td>
<td>7</td>
<td>-</td>
<td>-</td>
<td>14</td>
</tr>
<tr>
<td>System functionalities</td>
<td>44.4</td>
<td>37.9</td>
<td>2.39 (0.48)</td>
<td>91</td>
<td>82</td>
<td>91</td>
<td>0</td>
<td>9</td>
<td>9</td>
<td>-</td>
<td>-</td>
<td>91</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Total</td>
<td>2.88 (0.51)</td>
<td>71</td>
<td>69</td>
<td>77</td>
<td>17</td>
<td>3</td>
<td>37</td>
<td>14</td>
<td>9</td>
<td>40</td>
<td>3</td>
<td>6</td>
<td>11</td>
<td>11</td>
<td>-</td>
<td>23</td>
<td></td>
</tr>
</tbody>
</table>
4.4.3 Overall user satisfaction: the SUS score

In the pretest, the average SUS score was 47.2 (Range: 32.5-65). In the posttest, this average decreased to 37.5 (Range: 0-65). The results of the SUS questionnaire indicated that, overall, the users were not satisfied and even became less satisfied with the DICE system after three months of use. According to the scale provided by Bangor et al. (29), the acceptability of the DICE system was “OK” in the pretest and decreased to “Poor” in the posttest.

Table 4.5 The distribution of completely solved, partly solved and failed tasks and corresponding mean times on task (+/-SD) in the pretest and posttest.

<table>
<thead>
<tr>
<th></th>
<th>Pretest</th>
<th></th>
<th>Posttest</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Effectiveness %</td>
<td>Time on task</td>
<td>Effectiveness %</td>
<td>Time on task</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (+/-SD)</td>
<td></td>
<td>Mean (+/-SD)</td>
</tr>
<tr>
<td>Completely solved</td>
<td>50</td>
<td>0.95 (0.71)</td>
<td>52.2</td>
<td>0.94 (0.89)</td>
</tr>
<tr>
<td>Partially solved</td>
<td>37.5</td>
<td>2.27 (1.48)</td>
<td>42.0</td>
<td>1.85 (1.24)</td>
</tr>
<tr>
<td>Failed</td>
<td>12.5</td>
<td>1.08 (0.95)</td>
<td>6.8</td>
<td>1.52 (0.92)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1.46 (1.25)</td>
<td></td>
<td>1.37 (1.19)</td>
</tr>
</tbody>
</table>

Figure 4.2 The average user time on task in minutes and the difference in users’ performance per task in the pretest (left bar) and posttest (right bar). A task is “Completed” if the DICE-based reason for admission is semantically equivalent to the reason for admission in the scenario. A task is “Partially completed” if the DICE-based reason for admission partially covered the reason for admission in the scenario. A task was “Failed” in all other cases.
Table 4.6 The (differences in) distribution of completely solved, partly solved and failed tasks and corresponding mean times on task in the pretest and posttest for each participant. A task is “Completed” if the DICE-based reason for admission is semantically equivalent to the reason for admission in the scenario. A task is “Partially completed” if the DICE-based reason for admission partially covered the reason for admission in the scenario. A task was “Failed” in all other cases.

<table>
<thead>
<tr>
<th>User</th>
<th>Frequency of use</th>
<th>Time (min)</th>
<th>Correctly</th>
<th>Partially</th>
<th>Failed</th>
<th>Time (min)</th>
<th>Correctly</th>
<th>Partially</th>
<th>Failed</th>
<th>Time (min)</th>
<th>Correctly</th>
<th>Partially</th>
<th>Failed</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>3</td>
<td>1.49</td>
<td>7</td>
<td>4</td>
<td>0</td>
<td>1.16</td>
<td>5</td>
<td>6</td>
<td>0</td>
<td>-0.33</td>
<td>-2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>J</td>
<td>3</td>
<td>1.44</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>1.30</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td>-0.14</td>
<td>3</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>K</td>
<td>22</td>
<td>2.03</td>
<td>7</td>
<td>4</td>
<td>0</td>
<td>1.40</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>-0.63</td>
<td>-1</td>
<td>-1</td>
<td>2</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>1.96</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>2.78</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>0.82</td>
<td>2</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>M</td>
<td>3</td>
<td>1.15</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>0.84</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>-0.31</td>
<td>-1</td>
<td>3</td>
<td>-2</td>
</tr>
<tr>
<td>N</td>
<td>33</td>
<td>1.44</td>
<td>6</td>
<td>5</td>
<td>0</td>
<td>1.09</td>
<td>7</td>
<td>5</td>
<td>0</td>
<td>-0.36</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>O</td>
<td>5</td>
<td>1.07</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>1.20</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>0.13</td>
<td>-1</td>
<td>2</td>
<td>-1</td>
</tr>
<tr>
<td>P</td>
<td>12</td>
<td>1.10</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>1.14</td>
<td>6</td>
<td>5</td>
<td>0</td>
<td>0.04</td>
<td>1</td>
<td>0</td>
<td>-1</td>
</tr>
</tbody>
</table>

4.4.4 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 pretest, 11 users (69%) answered they would prefer to use the search string; five users (31%) preferred the list of frequent diagnoses. In the posttest, nine users (56%) answered they would prefer to use the search string and seven users (44%) preferred the list of frequently occurring diagnoses.

4.5 Discussion

4.5.1 General findings

In this study we performed a pre-post usability evaluation of a large compositional interface terminology based on SNOMED CT for 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 (SUS score lower than 40). 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, was steady between the pretest and posttest. However, in general, the frequency of individual participants’ use of the DICE system in between the pretest and posttest was low, so no conclusions concerning learnability of the DICE system can be drawn. About 30% of the encountered usability problems concerned the interface terminology based on SNOMED CT, while the remaining 70% was ascribed to the DICE application, of which about 30% concerned the graphical user interface design and about 40% the functionalities. The problems related to the interface terminology were more severe (i.e. average severity rating of 3.6) than the problems related to the DICE application (i.e. average severity rating of 2.5).

IC specific interface terminology based on SNOMED CT
The interface terminology is based on a large, though IC specific subset of SNOMED CT. The extensiveness, complexity of the hierarchy, and the English language usage of this subset were the main causes of the usability problems in the interface terminology. For instance, a medical disorder or procedure may be characterized by ‘due to’, ‘causative agent’, ‘finding site’, or ‘finding site direct’. Most of these characteristics are predefined 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). The users declared that they did not understand this post-coordination principle, i.e. which characteristics they needed to further specify and which were predefined. Furthermore, the users indicated that because of the complex structure and the long lengths of specific lists displaying values (for example the interface terminology contains about 1900 descendants for “virus”), they often were discouraged to search further down these lists and did therefore not make use of the post-coordination possibilities.

The participants’ verbalizations and interaction with the system also revealed that they did not understand SNOMED CT’s hierarchy. A resulting problem was that the users registered two related concepts, for instance “Coronary artery bypass graft x 2” and its descendant “Left internal mammary artery sequential anastomosis. Finally, as no Dutch translation of SNOMED CT exists, searching in the terminology content using English phrases was also an impediment for most of the users.

DICE application
Most of the usability problems concerned the DICE application through which the SNOMED CT subset was displayed to the users. With regard to the DICE user interface, most of the usability problems had a low severity and was related to suboptimal screen layout, wrong positioning of the buttons on the screen and mislabeling of the buttons. The
usability problems concerning the DICE system functionalities were more severe. The DICE system functionalities and their presentation through the DICE graphical user interface do not optimally support the users in searching through a large and complex terminology hierarchy. Furthermore, the DICE application does not fully support the use of post-coordination in SNOMED CT. 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. number 9 and 10 which involved the use of post-coordination, were poorly performed and participants needed more time to complete these tasks.

4.5.2 Findings in relation to the literature and recommendations

Inherent limitations of a large and complex subset from SNOMED CT such as its high granularity or complex hierarchy can reduce efficient use of the interface terminology (1;30) while poor design of the graphical user interface and insufficient functionalities of a terminology client may also hinder the search, navigation and data entry process (8;9). Studying interface terminology and terminology client in parallel offers the possibility to evaluate the causes of these failures, i.e. terminology content or client (9). Yet, few studies have considered a paralleled approach (9;15-17). In this study, different usability methods were used to examine the interactions between users and an IC specific interface terminology.

The usability problems could be linked to either the interface terminology or the terminology application. But the source for the inadequate user performances could be found in the overall effect of a moderate understandability of the interface terminology based on SNOMED CT, poor user interface design and suboptimal support and presentation of the required functionalities by the terminology application. Consequently, to improve the usability of the DICE interface terminology we have to deal with the problems related to the terminology content as well as with those concerning the terminology application.

Interface terminology

The frequently occurring violations of the heuristics described by Zhang et al. (30) show that most of the encountered problems concerning the interface terminology could be attributed to a lack of informative system feedback and inadequate help and documentation functions (27). 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 (2). Ideally, the subset should be as compact as possible while at the same time comprising the level of granularity needed in
the clinical domain of interest (31). Previous research has shown that physicians are not used to think in terms of hierarchies and complex relationships (32), indicating the need for concise subsets drawn from larger terminologies. The large and complex subset with 233,782 concepts as used in our interface terminology is yet far from compact. A previous study showed that a subset of about 2700 concepts from SNOMED CT was sufficient to cover 96% of clinical notes for patients admitted to an IC over a period of 5 years (31). In this earlier study, a large amount of clinical notes was drawn from the clinical information system and was processed by natural language processing (NLP) 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 for an ICU (31). Even under the condition that this subset was extracted in a different setting with a different basic goal, namely retrospective mapping of clinical notes instead of real time data entry by physicians, it might indicate that parts of the concepts included in our interface terminology can be removed from the subset. Shrinking methods can be applied to decrease the number of concepts in our interface terminology (33).

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, for the interface terminology to facilitate sharing and aggregating this kind of data for different purposes, it should enable 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.

The best solution presumably is to focus on the use of value sets and the arrangement and hierarchy of the concepts in the interface terminology rather than on its size. 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 (1;34). 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 on this area emphasizes the need for further research on the contextual adjustments of SNOMED CT subsets for use in interface terminologies.

**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, consistent support for the use of post-coordination, and an adequate search mechanism.

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 (35), as apparent from the frequently occurring violations of the heuristics described by Zhang et al. (30). A majority of these problems were yet of low severity and therefore require low priority in redesign efforts. Once the problems related to the DICE functionalities are dealt with, the problems in the graphical user interface are relatively easy to fix based on the available standards (27;35).

In relation to the DICE system functionalities, constraining how concepts can be put together by using rules for potential concept combinations will enhance their usability (1;36;37). Furthermore, previous studies have demonstrated that usability of an interface terminology is highly related to the efficiency of the search and navigation mechanism and term selection (3;38;39). As most of the DICE 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, the search functionality should be better understandable and should include features to help the users find easily what they are looking for. To this end, the search mechanism in the terminology client could be enhanced with query expansion algorithms, or auto-completion algorithms (40;41).

Graphical presentation of concepts are also a promising way to navigate and select a concept in an interface terminology, the clicking on a body part in a 3D model of the human body for example resulting in a list of related concepts in the interface terminology (42). However, for complex medical domains requiring a large subset of SNOMED CT, 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) can not 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 NLP algorithms are used to encode free-text input (43-46). In general, the use of NLP in clinical practice yields the promise to preserve 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 before. However, the interface clients 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 interface clients would also need components to stimulate clinicians to specify clinical entries that are not detailed
enough through post-coordination. Another drawback for using NLP is that the tools used
in the client interfaces are usually domain and language specific, and, if not available,
costly to build (47;48). In our case, a translation of SNOMED CT terms to Dutch is
required to facilitate the use of NLP in Dutch ICUs.

In conclusion, further research is needed to evaluate the use of NLP and other technologies
for ease of data entry in specific clinical domains. To this end, formative evaluation
methods in which the usability of the system is part of the system’s development cycle
should be used (49). It is widely acknowledged that user-based usability studies throughout
the system development are crucial in designing systems that fulfill end users’ requirements
(50). 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.

4.5.3 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 or hardly use the DICE system
between the pre- and the posttest: 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 posttest 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 pretest and posttest 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 participants’ misunderstanding of the content of the DICE terminology or to the
DICE terminology client. This was determined on the basis of the users’ verbalizations and
the moment of user- system interaction with the DICE terminology content. A limitation of
the study is that the distinction between the problems related to the terminology content and
DICE client was not clear in all cases. However, the high interrater agreement, i.e. 83%,
shows that in most of the cases the evaluators agreed on which problem was related to the
terminology content and which to the DICE client.

Conclusions
This study provides a detailed insight in how clinicians interact with a large controlled
compositional interface terminology to perform data entry. The combination of qualitative
and quantitative analyses in a pre-post approach enabled us to observe and analyze
clinicians’ interactions with a domain specific interface terminology.
For any interface terminology, there are tradeoffs between its terminology granularity and its ease of use. To encourage the clinicians to code the diagnostic information systematically using a large compositional interface terminology based on SNOMED CT, the user interface needs to be (re)designed to make it easy for users to capture their entries with minimal intrusiveness.

Adjusted subsets, enhanced terminology clients with technologies such as NLP present possibilities to easily capture encoded patient data into an electronic patient record. However, further research is needed to evaluate the role of these technologies for data entry in clinical practice.

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.

Bibliography

40. Fung KW, McDonald C, Bray BE. RxTerms a drug interface terminology derived from RxNorm. AMIA Annu Symp Proc 2008;227-31.