Research Paper ■

Studying the Human–Computer–Terminology Interface

JAMES J. CIMINO, MD, VIMLA L. PATEL, PHD, ANDRE W. KUSHNIRUK, PHD

Abstract **Objective:** To explore the use of an observational, cognitive-based approach for differentiating between successful, suboptimal, and failed entry of coded data by clinicians in actual practice, and to detect whether causes for unsuccessful attempts to capture true intended meaning were due to terminology content, terminology representation, or user interface problems.

Design: Observational study with videotaping and subsequent coding of data entry events in an outpatient clinic at New York Presbyterian Hospital.

Participants: Eight attending physicians, 18 resident physicians, and 1 nurse practitioner, using the Medical Entities Dictionary (MED) to record patient problems, medications, and adverse reactions in an outpatient medical record system.

Measurements: Classification of data entry events as successful, suboptimal, or failed, and estimation of cause; recording of system response time and total event time.

Results: Two hundred thirty-eight data entry events were analyzed; 71.0 percent were successful, 6.3 percent suboptimal, and 22.7 percent failed; unsuccessful entries were due to problems with content in 13.0 percent of events, representation problems in 10.1 percent of events, and usability problems in 5.9 percent of events. Response time averaged 0.74 sec, and total event time averaged 40.4 sec. Of an additional 209 tasks related to drug dose and frequency terms, 94 percent were successful, 0.5 percent were suboptimal, and 6 percent failed, for an overall success rate of 82 percent.

Conclusions: Data entry by clinicians using the outpatient system and the MED was generally successful and efficient. The cognitive-based observational approach permitted detection of false-positive (suboptimal) and false-negative (failed due to user interface) data entry.

■ **J Am Med Inform Assoc.** 2001;8:163–173.

The ability to capture clinical information and represent it by controlled terminology is widely recognized as a necessary aspect of electronic medical record systems.[1,2] The ability to represent information (such as observations and assessments) generated by health practitioners is particularly troublesome, because of the richness and complexity of clinical discourse. Depending on the task at hand, natural language processing can achieve the desired level of encoding: human beings record the data in an unconstrained way, and a computer system generates the coded form. Such approaches can succeed in well-defined, relatively narrow domains, such as mammogram interpretation, but are less applicable to large domains, such as history taking and patient problem lists.[3]

Even when terminologies exist for capturing information in a large domain, the issue of how the information will be transformed from concepts in the clinician's

Affiliation of the authors: Columbia-Presbyterian Medical Center, New York, New York.

Correspondence and reprint requests: James J. Cimino, MD, Department of Medical Informatics, Columbia-Presbyterian Medical Center, 622 West 168th Street, VC-5, New York, NY 10032; e-mail: <jjc7@columbia.edu>.

This work was supported in part by an Electronic Medical Record Cooperative Agreement contract with the National Library of Medicine.

Received for publication: 8/18/00; accepted for publication: 11/16/00.

mind to codes in the computer's database remains. A common approach is to allow clinicians to enter their terms in unconstrained text and then use manual or automated means to code them afterward.[4,5] Such an approach is also used to evaluate the degree of domain coverage provided by clinical terminologies.[6] The obvious disadvantage of this approach is that the clinician is not present to verify that the codes assigned to the text are accurate and represent the best choices available. As a result, the appropriateness of the encoding and the validity of the terminology cannot be guaranteed.[7] An alternative strategy is to have the clinicians interact directly with the terminology to decide for themselves which terms should be used.[8–14]

Typically, terminologies are evaluated in terms of their abilities to represent the concepts at hand, and data entry systems are evaluated in terms of their usability, but the two are rarely examined together. A combined approach offers a critical perspective on how the task is carried out and where it can be improved. Consider, for example, the task of putting a problem on a problem list. If a user wishes to add a problem and fails, the reason might be inadequate completeness (the terminology is not capable of representing the problem adequately), poor usability (the application does not provide adequate access to the terminology), or insufficient representation (some characteristic of the terminology, such as poor organization or inadequate synonymy, interfered with the user's ability to find the proper term). Alternatively, if the user does select a term, the question remains whether the term appropriately captures the intended meaning, and once again, any of these three causes could be to blame. Studying the terminology and the data entry application together offers the possibility of teasing out whether the application has failed (either through lack of data entry or inappropriate data entry) and, if it has, the cause of the failure.

Related work we have conducted over the past six years has focused on issues related to the design of user interfaces, to improve the usability of such systems as computerized patient record systems.[15] We have adopted and modified a number of methodologies from the emerging fields of usability engineering[16] and cognitive science[17] to evaluate systems in terms of both the ease of accessing information and the adequacy of retrieved information. Subjects are typically asked to "think aloud" while interacting with systems to perform representative tasks. In the current work, we are extending this approach to broader issues related to both the design of the user interface and the underlying medical terminology.

Efforts have been made to study the interaction between user interface design and coded data entry. Poon et al.[18] have used timing studies to assess how different user interface features affect data entry speed with a structured-progress note system. They used a variety of methods for presenting lists from which the users selected desired terms. In contrast, Elkin et al.[19] studied the speed and success of users selecting terms by typing words and phrases. Like Poon and colleagues, they used paper-based scenarios from which clinicians were asked to create problem lists. In their scenarios, the desired outcomes (i.e., the specific terms to be entered) were determined in advance. The terminology was known to be complete for the tasks being studied, yet the users entered these terms only 91.1 percent of the time. Because they employed a usability laboratory (which captured detailed video recordings of the user–computer interactions), they were able to determine specific reasons why terms were not entered. This enabled them to differentiate between problems with terminology representation and system usability.

We wanted to examine how clinicians would interact with our controlled terminology, the Medical Entities Dictionary (MED)[20,21] while using an ambulatory record application, the Decision-supported Outpatient Practice (DOP) system for entering real patient data.[22] We chose to study clinicians in the process of actual patient care as they entered a variety of data (regarding problems, allergies, and medications). We employed cognitive-based methods to differentiate between appropriate and suboptimal data capture and to determine the degrees to which problems with completeness, usability, and representation contributed to the unsuccessful data capture.

Methods

Decision-supported Outpatient Practice System

The DOP system was an ambulatory record application that provided a variety of clinical applications, including progress notes, review of reports from ancillary systems, and health maintenance reminders.[22] Users interacted with the controlled terminology to enter problem lists, adverse reactions, and medications[11] using a terminology server that supports stemming and synonym matching and allows users to traverse the terminology's hierarchy.[23] Figures 1 to 3 show various screens from the application.*

*DOP was phased out at the end of 1999 and replaced by a Web-based application called WebCIS.[24]

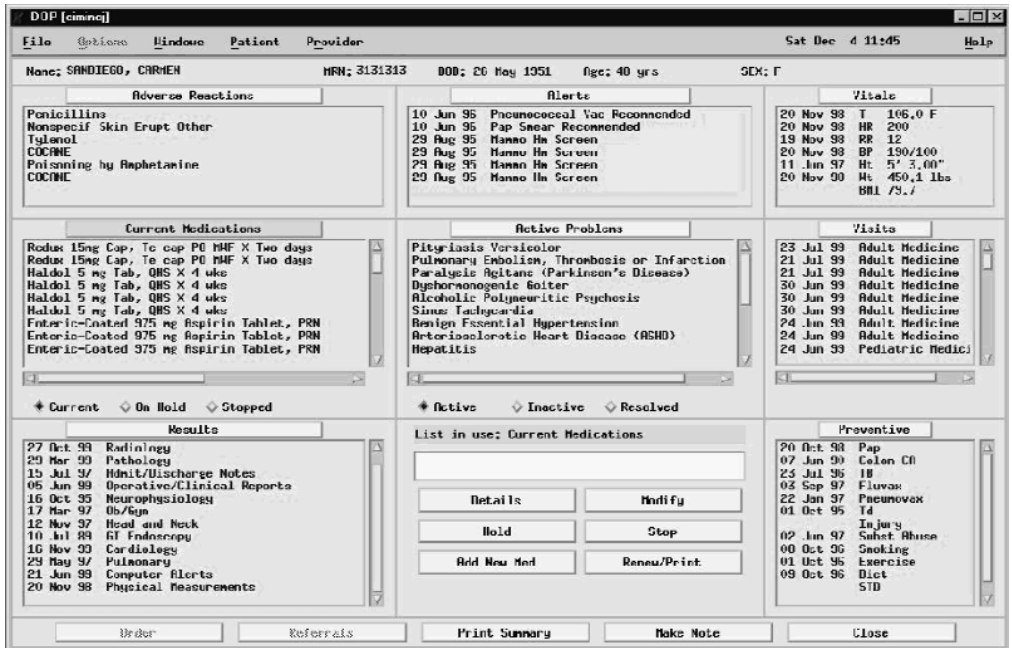

Figure 1 Sample DOP screen. Items in the "Adverse Reactions," "Current Medications," and Active Problems" windows were entered by clinicians using the controlled terminology.

Medical Entities Dictionary (MED)

The MED is a controlled terminology of more than 67,000 terms that is used to encode data in the New York Presbyterian Hospital's clinical repository.[20,25] Terms in the MED each have a unique preferred name and may have one or more synonyms. The terms are organized into a directed acyclic graph of is-a relationships, representing a multiple hierarchy. (Terms are also related through non-hierarchic semantic relationships, but these are not relevant to the current study). The DOP system allowed users to choose terms from selected subsets of the MED by typing in a phrase and providing a list of matching terms selected from the subset (Figure 2). If desired, users can browse the MED hierarchy to find more or less specific terms (not shown). In addition to the problems, adverse reactions, and medications, the MED provides controlled terms related to the route of and dosing frequency for medication orders (Figure 3). Table 1 shows the number of terms in each subset at the time of the study.

Figure 2 Data entry for a patient problem. The user has typed "chf," and the system has returned 13 terms.

Portable Usability Laboratory

We employed a portable usability laboratory to make video and audio recordings of user interactions.[26] A video converter (Scan-DO Ultra, Communication Specialties, Haupaugue, New York) converted the computer display to a video signal (S-video) for capture on videotape using a standard video cassette recorder (VCR). A microphone captured users' statements and questions, along with the keyboard sounds. The video converter and microphone were placed unobtrusively in the work area, and cables were run to the remainder of the equipment (on a movable cart) up to 75 ft away. The audio signal from the microphone was split between the VCR and a cassette recorder. The laboratory operator used a television monitor and headphones to verify the quality of the recordings. A laptop computer running proprietary software was used to control the video convert-

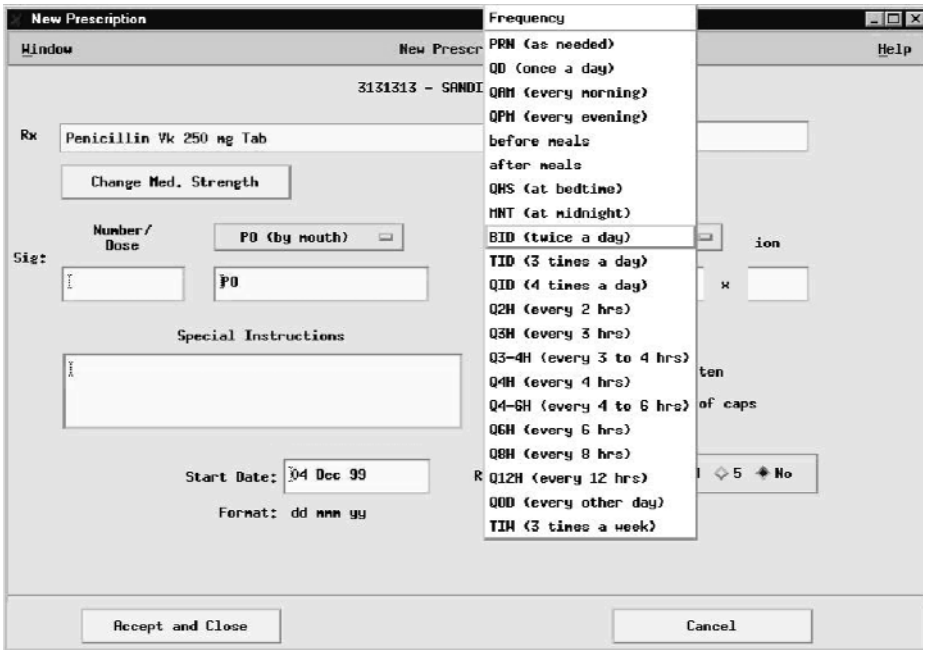

Figure 3 Data entry for a patient prescription. The user has typed "PO" in the dose route field, and the system has recognized it as "by mouth." The user is now selecting the does frequency from the terminology subset using apull-down menu.

er via an RS-232 cable to allow zooming and panning of the video screen to focus on areas of interest. Figure 4 shows the configuration of the system.

Experimental Approach

We recruited—as volunteers—residents, attending physicians, and nurse practitioners who agreed to think aloud and be audio- and videotaped while using the DOP system to record actual patient visits in their offices, in their usual manner. The audiotape of each session was subsequently transcribed and annotated (while reviewing the videotape) to reflect the actions carried out throughout each session. Details of the annotation methodology are presented elsewhere.[26]

Evaluation

In each session, we noted each occurrence of all attempts by users to enter coded data. These included addition of problems, adverse reactions, and medications. For each event, we noted the audible statements from the user (usually regarding the concept of interest), the keystrokes that the user used, the resulting list of terms produced by the system, and the selection made by the user (if any). Medication entries could be further characterized by one of three specific tasks—adding a medication that the patient was currently taking, adding a new medication, and modifying a medication that had been previously added to the list. These tasks were examined separately because the analysis showed them to involve different cognitive activities and different interactions with the data entry application.

On the basis of the users' statements and selections, we coded each event using a limited set of characterizations. When the user selected a term that matched exactly what was typed or said, this was coded as a "success"; if the user did not select a term, this was coded as a "failure"; all other results were considered "suboptimal." Failures and suboptimal results were further characterized into insufficient coverage (the MED did not contain the desired term), inadequate representation (the MED contained the term but it was not returned by the system, due to a missing synonym or unrecognized abbreviation), and usability problems (the system failed to return the term for technical reasons or the user failed to see the term on

Table 1 ■

Size of the Terminology Subsets Used for Data Entry Tasks at the Time of the Study

Terminology Subset	Number of Terms
Patient problems	29,870
Adverse reactions	58,450*
Medications	4,475
Dose routes	16
Dose frequencies	29

*Although some terms in the MED are categorized explicitly as Adverse Reactions, DOP allows users to enter any MED term in the "Adverse Reactions" field; this allows the user to specify, for example, reactions to foods and chemicals.

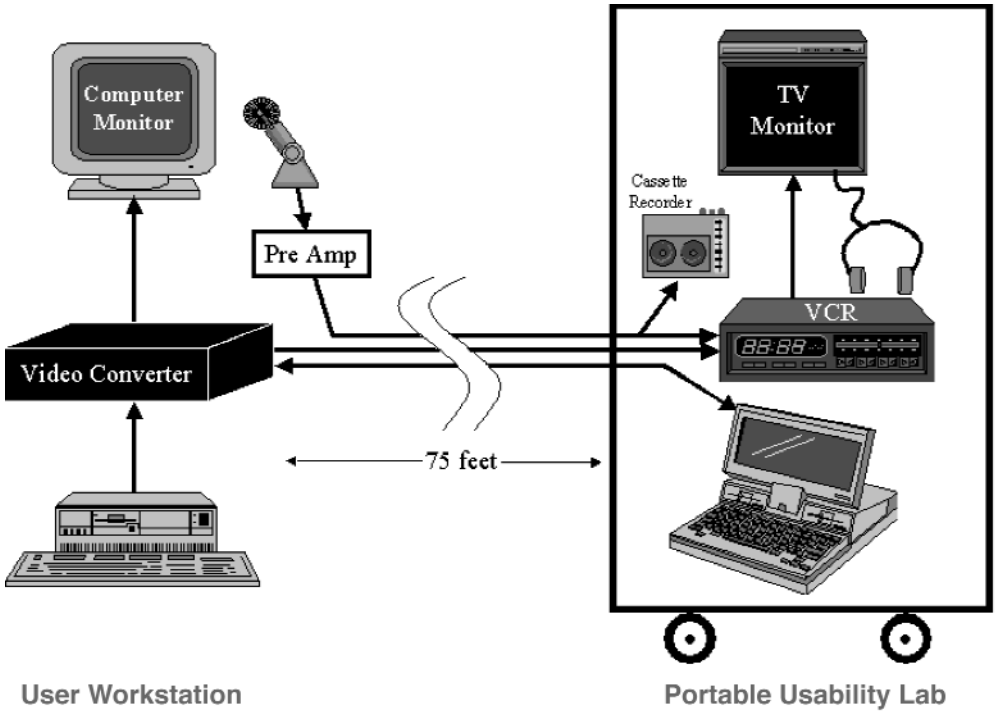

Figure 4 Portable usability laboratory components. The video converter is placed between the computer and the monitor at the user's workstation. The converter sends an NTSC video signal via an S-video cable to the VCR on the portable usability laboratory cart. A microphone is used to provide an audio recording of the user, captured on the VCR and a cassette recorder. A laptop computer on the cart communicates with the video converter via an RS-232 cable. These three wires are bundled together to extend a maximum of 75 feet. The researchser can remain with the cart in the next room, observing the recording with a television monitor and headphones and controlling the panning and zooming functions the video converter.

the returned list). The evaluation method is described more completely elsewhere.[27] The system response time and the total time of each interaction were derived from the timing of the videotape.

All codings were performed by a cognitive scientist and a physician, who reviewed the videotape together. The determination of "failure" was an objective one—if no term was selected, the system was considered to have failed. Distinction between success and suboptimal performance was necessarily subjective; however, we followed very simple rules for classification which would favor down-grading results toward being suboptimal. Discrepancies were resolved through discussion. The distinction between coverage, representation, and usability problems was made by a single terminology expert, familiar with the MED.

Results

We videotaped 27 different users (8 attending physicians, 18 resident physicians, and 1 nurse practitioner) for a total of 32 sessions. Nine of these sessions were previously analyzed for system usability and reported elsewhere.[27] For this study, we re-analyzed these 9 sessions, together with the 23 new sessions, to examine clinicians' interaction with the coded data entry functions. One third of the users (three attending physicians and six resident physicians) considered themselves experienced DOP users.

During the 32 sessions, the users attempted to enter 192 coded terms. In 30 cases, when they did not find a satisfactory term on the list provided by the DOP system, they made one to five additional attempts per instance, providing 46 additional entries, for a total of 238 attempts to add terms. As expected, coding of the data entry events showed that users were adding medications, adverse reactions, and patient problems.

Examination of the videotape showed that the time required to enter, review, and select a term (if one were selected) ranged from 2 to 225 sec (average, 40.4 sec). The number of terms returned by the search ranged from none (in 30 cases) to 1,488. When the 30 search failures and 2 cases of extremely high results (688 when a user entered "arthritis" and 1,488 when a user entered "drug") were excluded, the average

Table 2 ■

Data Entry Events, Categorized by Task

Task	Total Events	Event Time	No. with Results	Response Time	Average No. of Results
Patient problem	68	37.9 (26.8)	56	1.14 (1.34)	23.0 (30.3)
Adverse reaction	19	41.1 (36.8)	16	1.14 (0.99)	19.9 (18.0)
New medication	44	39.5 (31.5)	42	0.45 (0.34)	9.0 (6.3)
Existing medication	69	44.7 (35.1)	60	0.77 (0.93)	14.0 (13.4)
Medication modification	38	37.8 (27.4)	32	0.13 (0.10)	12.2 (12.9)
Total	238	40.4 (31.3)	206	0.74 (0.99)	15.6 (19.7)

NOTE: Event time is measured from start of typing to selection of item from list. Response times and list sizes are shown for searches that produced results. Times are reported in seconds. Numbers in parentheses are standard deviations. The 30 events with no response were excluded from the response time calculations. These events and two events with extremely large lists were excluded from the result counts. All events of less than 0.1 sec were counted as 0.1 sec.

list size was 15.6 terms. Response time of the system varied from instantaneous to 10 sec. Since times below 0.1 sec could not be measured precisely, we treated all response times less than 0.1 sec, including those that appeared to be instantaneous, as 0.1 sec. Using this correction, we found the average response time to be 0.75 sec when results were returned (the 30 searches that returned no results were always instantaneous). Table 2 shows details of the timing data.

Of the 238 attempts to add terms, 151 involved medications. Users attempted the supplementary tasks of adding dose route and dose frequency information for 105 and 104 of these, respectively. These supplementary tasks differed from other attempts to add data in that they offered the user the option of selecting terms from pull-down lists in addition to accepting typed input. Including these supplementary data entry tasks and the 30 search failures, 447 data entry attempts were available for analysis of the success and quality of data entry.

Figure 5 shows some typical events and how they were coded. Excerpts from the transcripts of what users said while interacting with the DOP system, as well as what they did in interacting with the system to retrieve terms, are presented. The numbers in the video transcript refer to the actual counter on the VCR. Using video annotation software, which we refined and modified for supporting video coding, we were able to "time-stamp" sections of the verbal transcript to the actual video footage, allowing for automatic retrieval and review of sections of video recordings. On the right-hand side of the figure, each interaction is coded with regard to the task the user attempted, the term the user entered, the response time of the system, the items returned by the system and the success or failure of the system in providing the user with the desired term. For example, the first interaction in the figure illustrates a successful retrieval of an adequate term in response to the entry of "hand pain". In contrast, for the fourth interaction given in the figure (starting at time 00:51:55), the user enters "manic-depression" but does not select a resulting term (presented by the DOP system) because the system failed to provide the user with an appropriate synonym.

Table 3 shows the results of coding the 447 attempts by users to enter coded data. Users chose some term from the list 381 times (86 percent), although in 16 cases (4 percent), the analysis of the videotape showed the result to be suboptimal; that is, it did not appear to match the users' spoken intentions. Of the 82 cases in which the user either selected no result or a suboptimal result, the cause was missing MED concepts in 33 cases (40 percent of the 82 cases, 7 percent of all cases).

Typically, the user typed a specific term and got no results or typed a general term (such as "constipation") and spoke about a more specific term (such as "chronic constipation"). Less-specific data entry generally produced a list of more-general terms. Sometimes, the user selected a more general term (such as "constipation") and sometimes typed a modifier (such as "chronic") in the comment field. Usually, however, the user did not select a less specific term. The cause of the MED's deficiency was roughly divided between missing a more specific problem term (such as "chronic constipation") and missing a specific medication (5 were repeated attempts by one user to enter "oxygen" as a medication) or the desired strength of a medication.

Video Transcript	Coding
00:15:50 Adds a problem "two active problems, and now I've got to create a new problem, lets see what is it, I guess it is a kind of hand pain" **00:15:57 Enters term "hand pain"** "And it come up with pain involving hand I guess that will do, it is possible arthritis but I suspect diabetic neuropathy" **00:16:25 Adds text to the problem description** "Ok, I'll go back to my note, lets see I noticed..." **00:16:30 Answers phone call**	Task: Enter current problem Said: hand pain Typed: hand pain RESPONSE TIME: 1.4 sec Found: 1 item Picked: Pain Involving Hand TOTAL TIME FOR EVENT: 28 sec Interpretation: Success
00:21:40 Adds a problem "She has history of peptic ulcer disease. She has a, she was admitted in the past with peptic ulcer disease, that was when I first saw her..." **00:21:51 Enters term "Peptic Ulcer disease"** "I'll just check this. So I have to see this, change the start date of her. Um, OK, We have, not sure what the date was."	Task: Enter current problem Said: history of peptic ulcer disease RESPONSE TIME: 1 sec Found: 2 items Picked: Personal History of Peptic Ulcer Disease TOTAL TIME FOR EVENT: 28 sec Interpretation: Success
00:30:18 Adds a problem "Uh, I'm sure it will be a long list. Back pain, there is a lot of secondary pain involved with it, she has been on disability for a very long time... **00:30:32 ENTERS SEARCH WORDS "back pain"** "Back ache, I like that one. Started one year ago..." **00:30:48 Selects term from list**	Task: Enter current problem Said: chronic back pain Typed: back pain RESPONSE TIME: 0.5 sec Found: 3 items Picked: Backache, Unspecified TOTAL TIME FOR EVENT: 30 sec Interpretation: Suboptimal due to missing concept
00:51:55 Adds a problem, types: manic-depression **User scans list, doesn't select anything** **00:52:08 User tries again** (next time, user finds "Bipolar Affective Disorder, Unspecified")	Task: Enter current problem Said: nothing Typed: manic-depression Found: 7 items RESPONSE TIME: 0.5 sec Picked: nothing TOTAL TIME FOR EVENT: 13 sec Interpretation: Failure, system did not recognize synonym
01:15:11 Enters plan, types: dc medformin, start glynase at 2 mg; pt had been having episodes of hyoglycemia on 2.5 milligrams dose. See me, follow up 1 month **01:16:27 to 01:16:28 Adds a medication** "Make that 1.5 mg." **01:16:11 Selects term from list**	Task: Modify existing medication Said: glynase 2mg Typed: glyburide RESPONSE TIME: 0.5 sec Found: 5 items Picked: Glyburide 1.5 mg TOTAL TIME FOR EVENT: 60 sec Interpretation: Successful, system helped user find correct dose

Figure 5 Examples of video transcription and coding. The left column shows transcription of audio portion of videotape, with timing marks and some statement about user interaction with the computer. The right column shows the coding of the interaction while reviewing the videotape.

Table 3 ■

Analysis of Data Entry Events by Task

Task	Total	Success	Suboptimal			Failure		
			C	R	U	C	R	U
Main task:								
Patient problem	68	45 (66)	5 (10)	1 (1)	0	7 (10)	5 (7)	5 (7)
Adverse reaction	19	9 (47)	1 (5)	5 (26)	0	0	3 (16)	1 (5)
New medication	44	36 (82)	0	0	0	7 (16)	0	1 (2)
Existing medication	69	50 (72)	1 (1)	0	2 (3)	4 (6)	7 (10)	5 (7)
Medication modification	38	29 (76)	0	0	0	6 (16)	3 (8)	0
Subtotal	238	169 (71)	7 (3)	6 (3)	2 (1)	24 (10)	18 (8)	12 (5)
Supplementary task:								
Dose route	105	99 (94)	0	0	0	0	6 (6)	0
Dose frequency	104	97 (92)	0	0	1 (1)	2 (2)	4 (4)	0
Subtotal	209	196 (94)	0	0	1 (0.5)	2 (1)	10 (5)	0
TOTAL	447	365 (82)	7 (2)	6 (1)	3 (1)	26 (6)	28 (6)	12 (3)

NOTE: Successful codings were those in which the term selected by the subject matched his or her intent, based on analysis of his or her spoken comments. Suboptimal codings were those in which a term was selected that did not match the subject's intent. Failed codings were those in which no term was selected. Reasons for suboptimal or failed codings included insufficient coverage (C), inadequate representation (R), and usability problems (U). Numbers in parentheses are percentage of total.

Review of the videotape and analysis of the MED showed that in 49 cases (60 percent of the 82 cases, 11 percent of all cases), the MED actually did contain the desired term but the user did not find it (sometimes choosing a suboptimal term but generally choosing no term). In general, the reason for failure was that the MED lacked a synonym or abbreviation (such as "htn" for "hypertension"). In 14 cases, we attributed the failure to problems with the user interface, including the lack of phonetic spell-checking and failure to hit the "Enter" key, which was required inconsistently in an early version of the DOP system.[27]

The success rates for different tasks, and the reasons for failure, are comparable for all the tasks except the drug route and frequency data entry. These two tasks differ from the others in that they allow only a very limited terminology and provide optional pull-down lists. Users appeared to be generally familiar with these restricted terminologies, and most problems were due to lack of recognition (for example, "p.o." was not recognized as "po" and "qDay" was not recognized as "qd"). The MED was missing only two concepts that users attempted to enter for drug frequency: "pain" and "1/2 H AC and QS."

The tool for navigating the hierarchy of the MED was used only twice (by the user searching for "oxygen"), neither time successfully. Although this "tree walking" capability could have been used to identify more-specific terms, it was never invoked for this purpose.

Discussion

This study provides a detailed look at how clinicians interact with a controlled terminology to carry out data entry. To our knowledge, the only previous study to observe and analyze clinicians in this manner was carried out by Elkin et al.[19] Their study was carried out in a realistic setting in which they provided physicians with preconstructed scenarios. The behavior they studied was whether the users could find in the terminology the terms suggested by the scenarios. They successfully demonstrated that their system was easy to use and that their users were comfortable with their terminology. Their report did not include any analysis of how their system and terminology performed when users wanted to enter additional terms.

The evaluation method used in this study was a combination of objective and subjective techniques. The

differentiation between success, suboptimal results, and failure was readily determined from review of the videotape, with a high degree of agreement between the cognitive scientist and the physician. Differentiation between content, representation, and usability problems was a task that basically determined whether the MED contained the desired term, either as entered by the user or in some other form. This required an intimate knowledge of the MED; as is often the case,[5] only one evaluator was available who had the necessary expertise for this task.

We believe that the use of our evaluation technique provided valuable insight into the users' interactions with the MED. A study that was limited to review only of DOP logs (including entry of dose terms) would suggest that users succeeded in finding MED terms 86 percent of the time. The closer inspection permitted by the observational approach, however, showed that users chose a suboptimal term because the MED lacked the desired term 2 percent of the time (and for other reasons an additional 2 percent of the time). Likewise, analysis of the 14 percent data entry failures shows that 8 percent of the time, the MED actually contained the desired term. Thus, the degree of MED coverage in this experiment was found to be 92 percent, a rate that compares favorably with other studies of terminology coverage.[6]

The user interface used in the DOP system takes a common approach to coded data entry—allow the user to type in a phrase and attempt to match it to a known phrase. The observational approach showed that the user interface was responsible for 3 of the 16 suboptimal results and 12 of the 66 failures, or 15 of 82 (18 percent) of the problems. A spell-checking feature would have mitigated this failure rate, suggesting that the application was quite usable with respect to data entry. Some of the remaining causes of user interface problems were navigation problems and were noticed in the first set of subjects. The system developers addressed these, and the navigation problems subsided in subsequent sessions.[27] When these cases were excluded from analysis, the rates of successful, suboptimal, and failed codings were almost identical before and after the changes in the system.

Although we postulated that terminology representation might affect data entry, we found few instances of this, although the need for richer synonymy was evident in some areas. The system allowed the user to navigate the MED hierarchy, but the users did not avail themselves of this option. The hierarchy of the MED therefore did not interfere with data entry, but we cannot comment on whether it could have helped in cases where the user could not find an appropriate extant term.

The DOP system was phased out at the end of 1999, but the MED and the terminology server that it used remain in use and are being integrated with DOP's replacement. The lessons learned in this experiment therefore have implications for us as we move forward; we believe that there are also implications for others interested in clinician data entry.

First, our approach allowed us to pinpoint and distinguish problems with both user interface design and terminology coverage. We could see where the user interface needed improvement (spell-checking and navigational consistency) and where the terminology was particularly lacking (adverse reactions).

Second, we observed that when users type into a coded data entry system, they often type less than what they are intending to record (based on what they say aloud and what they select from the proffered lists). This has implications for studies of terminology completeness, in which an evaluation might conclude that a terminology does have a term (when a user might have wanted a more detailed term) or does not have a term (because the user did not provide it).

Third, we documented that, in some situations, the direct interaction with the terminology allows the user to select a better term than would otherwise be chosen. For example, in several instances a user corrected a mistaken medication dosage because the desired dose was found not to exist (as with the glyburide example in Figure 5).

This study permitted an evaluation of the whole coded data entry experience, end to end, in actual clinical situations. Using this approach, we have been able to evaluate user interfaces and underlying terminologies, providing a basis for the iterative refinement and improvement of both. This type of study seems, to us, to be critical to the success of clinician data entry, because it identifies problem areas that are not detected with simple exit questionnaires.[27] The results of this study show us, in quantifiable terms, what problems exist and allow us to address them directly.

Studying 27 volunteer users for an hour or so each is a start, but the obstacles to clinician data entry are likely to be multifactorial. Additional work is needed to learn how the heterogeneous population of all clinicians reacts to such tasks. Although the think-aloud approach allows us to see whether users are finding

terms they deem appropriate, further studies are needed to determine whether their intentions are appropriate. Do clinicians do a "better" job at recording their data when using a computer-based approach? Do they make additional effort? Do the computer-based resources help them carry out a task in a more informed manner? We can start to get answers to these questions when we notice user behaviors, such as repeated attempts to say what they mean or efforts to correct their attempts to prescribe nonexistent drugs. But formal studies of these behaviors, along with comparison of paper-based approaches, remain to be done.

The importance of such studies should not be underestimated. If we are to use clinicians' data to drive automated decision support, help with case management, reduce errors, and facilitate clinical research, we must pay careful attention to the quality of the data they are generating. Such data must be at least as good as those available in paper records if we are to justify the expense and effort of collecting them. Clinicians must convey their meanings as precisely as possible if computer systems are to assist them in making effective, informed decisions. Thus, data capture is as crucial as data manipulation. Just as human beings can fail to express themselves for physical or cognitive reasons (such as motor aphasia vs. expressive aphasia), so too can our systems fail to capture data for technical or semantic reasons. Our study demonstrates that it is possible to tell the difference. The distinction is important if we are to fix the problems, since ultimately our systems will be easier to change than our users.

Conclusion

Our methodology allowed us to distinguish deficiencies of terminology content from shortcomings of a clinician data entry interface. In our particular case, we learned that the addition of terminology for adverse reactions, the correction of navigational difficulties, and the addition of spell-checking to the term look-up function will be particularly beneficial. Our approach is generalizable to help other terminology developers and system designers better focus their quality improvement efforts.

This work was supported in part by an Electronic Medical Record Cooperative Agreement contract with the National Library of Medicine. The authors thank Peter Elkin, Brian Kaihoi and Chris Chute for demonstrating the value of a usability laboratory for terminology evaluation. The authors also thank Randy Barrows for his support in studying the DOP system, the study participants for their cooperation, and Andria Brummitt for editorial assistance.

References ■

1. United States General Accounting Office. Automated Medical Records: Leadership Needed to Expedite Standards Development: Report to the Chairman/Committee on Governmental Affairs, U.S. Senate. Washington, DC: USGAO/IMTEC-93-17, Apr 1993.
2. Sittig DF. Grand challenges in medical informatics? J Am Med Inform Assoc. 1994;1:412–3.
3. Rector AL. Clinical terminology: why is it so hard? Methods Inf Med. 1999;38:239–52.
4. Campbell JR, Givner N, Seelig CB, et al. Computerized medical records and clinic function. MD Comput. 1989;6:282–7.
5. Brown SH, Miller RA, Camp HN, Giuse DA, Walker HK. Empirical derivation of ane electronic clinically useful problem statement system. Ann Intern Med. 1999;131:117–26.
6. Campbell JR, Carpenter P, Sneiderman C, et al. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions and clarity. J Am Med Inform Assoc. 1997;4:238–51.
7. Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems. Implications for outcomes research. Ann Intern Med. 1993;119:844–50.
8. van der Lei J, Duisterhout JS, Westerhof HP, et al. The introduction of computer-based patient records in The Netherlands. Ann Intern Med. 1993;119:1036–41.
9. Huff SM, Pryor A, Tebbs RD. Pick from thousands: a collaborative processing model for coded data entry. Proc 17th Annu Symp Comput Appl Med Care. 1993:104–8.
10. Scherpbier HJ, Abrams RS, Roth DH, Hail JJ. A simple approach to physician entry of patient problem list. Proc 18th Annu Symp Comput Appl Med Care. 1994:206–10.
11. Barrows RC, Johnson SB. A data model that captures clinical reasoning about patient problems. Proc 19th Annu Symp Comput Appl Med Care. 1995:402–5.
12. Gundersen ML, Haug PJ, Pryor TA, et al. Development and evaluation of a computerized admission diagnoses encoding system. Comput Biomed Res. 1996;29:351–72.
13. Campbell JR. Strategies for problem list implementation in a complex clinical enterprise. Proc AMIA Annu Symp. 1998:285–9.
14. Elkin PL, Bailey KR, Chute CG. A randomized controlled trial of automated term composition. Proc AMIA Annu Symp. 1998:765–9.
15. Kushniruk AW, Patel V. Cognitive computer-based video analysis. In: Greenes R, et al. (eds). Proc 8th World Conference on Medical Informatics. 1995:1566–9.
16. Nielsen J. Usability Engineering. New York: Academic Press, 1993.
17. Ericsson KA, Simon HA. Protocol Analysis: Verbal Reports as Data. Cambridge, Mass: MIT Press, 1993.
18. Poon AD, Fagan LM, Shortliffe EH. The PEN–Ivory project: exploring user-interface design for the selection of items from large controlled vocabularies of medicine. J Am Med Inform Assoc. 1996;3:168–83.
19. Elkin PL, Mohr DN, Tuttle MS, et al. Standardized problem list generation, utilizing the Mayo Canonical Vocabulary embedded within the Unified Medical Language System. Proc AMIA Annu Fall Symp. 1997:500–4.
20. Cimino JJ, Hripcsak G, Johnson SB, Clayton PD. Designing an introspective, controlled medical vocabulary. Proc 13th Annu Symp Comput Appl Med Care. 1989:513–8.

21. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Inform Assoc. 1994;1:35–50.
22. Barrows RC, Allen BA, Sherman E, Smith K. A decision-supported outpatient practice system. Proc AMIA Annu Fall Symp. 1996:792–6.
23. Cimino JJ. Terminology tools: state of the art and practical lessons. Presented at: IMIA Working Group 6 Conference; Dec 16–19, 1999; Phoenix, Arizona.
24. Hripcsak G, Cimino JJ, Sengupta S. WebCIS: large scale deployment of a Web-based clinical information system. Proc AMIA Annu Symp. 1999:804–8.
25. Cimino JJ. From data to knowledge through concept-oriented terminologies: experience with the Medical Entities Dictionary. J Am Med Inform Assoc. 2000;7:288–97.
26. Kushniruk AW, Patel VL, Cimino JJ. Usability testing in medical informatics: cognitive approaches to evaluation of information systems and user interfaces. Proc AMIA Annu Fall Symp. 1997:218–22.
27. Kushniruk AW, Patel VL, Barrows RC, Cimino JJ. Cognitive evaluation of the user interface and vocabulary of an outpatient information system. Proc AMIA Annu Fall Symp. 1996:22–6.