MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-283

Verbal Protocols for Assessing the Usability of Clinical Decision Support: The Retrospective Sense Making Protocol

Panagiotis Balatsoukas, John Ainsworth, Richard Williams, Emma Carruthers, Colin Davies, James McGrath, Artur Akbarov, Claudia Soiland-Reyes, Saurin Badiyani, Iain Buchan

Centre for Health Informatics, Institute of Population Health, University of Manchester and Manchester Academic Health Science Centre, UK

Abstract

We compare the effectiveness of two types of verbal protocol, concurrent think aloud vs. retrospective sense making, for evaluating the usability of a clinical decision support tool. Thirty-five medical and nursing students participated in a usability experiment. Participants were asked to complete seven tasks using the system under evaluation. Eighteen students were allocated to the concurrent think aloud group and the remainder followed the retrospective protocol. The retrospective protocol was significantly more sensitive than the concurrent protocol in recording unique usability problems related to users' cognitive behaviour. These problems concerned the interpretation and comprehension of statistical output, search results and system messages. These findings can be explained by the retrospective protocol's greater ability to detect compound usability problems, capturing the cognitive dimensions of users' interactions with the interface in greater depth. Evaluations of clinical decision support systems should take an evidence-based approach to selecting protocols.

Keywords:

Decision support, Usability, Evaluation Protocol.

Introduction

To date, there are several examples of studies focused on the use of verbal protocols as a means of evaluating the usability of different types of Health Information Systems, such as medical administration work systems [1], electronic patient health record systems [2], clinical decision-support tools [3], clinical terminology interfaces [4], self-care management systems [5] and clinical trial management systems [6]. In the majority of these studies, researchers implemented a concurrent think aloud (TA) protocol as a core data collection technique. This means that participants in the usability study were asked to talk aloud (i.e. to verbalise their actions, thoughts or feelings) while completing a set of tasks using the interface under evaluation. The analysis of this type of concurrent cognitive data can lead to the identification of usability problems.

Although the application of a concurrent TA protocol has been successful in identifying usability problems, there is still a debate about the effectiveness of this method compared to other forms of verbal protocols, especially over the use of retrospective verbalisations [7]. Usually, retrospective protocols follow immediately after a set of tasks is completed by the user using the interface under evaluation. Retrospective verbalisations have been criticised for post hoc rationalisation and reconstruction effects that alter users' cognitive structures after exposure to an interface. However, as opposed to concurrent verbal protocols, retrospective protocols can disclose more information about interaction behaviour. This happens because methods focused on retrospective protocols are based on the vocalisation of the contents of both the short-term and long-term memory, while in the case of the concurrent protocol, the contents of the thought do not require further processing and articulation from long-term memory. There are variations of the retrospective verbal protocol that make use of open-ended questions in order to encourage users to process information from the long-term memory, providing justifications and explanations of specific actions they performed during their interaction with an interface.

Although there were a few attempts made to compare the concurrent TA protocol with different types of usability evaluation methods, like cognitive walkthroughs, surveys, clinical simulations and interviews [e.g. 8, 9], there are quite a few published studies comparing concurrent with retrospective verbal protocols in medical informatics [10]. From a practical perspective, this type of study is important for informing choices about the most effective methods for evaluating the usability of health information systems and provides an evidence-based approach to selecting protocols. In addition, enhanced understanding of this methodology can extend the interpretation of existing usability evaluations.

The aim of this paper is to present the results of a usability experiment comparing two different types of verbal protocols: a concurrent TA protocol and a new retrospective sense making protocol. In particular, the two protocols were compared in terms of the total number and the number of unique usability problems identified, the types and frequency of categories of problems, and how usability problems surfaced (e.g., observation of on-screen behaviour only, analysis of verbalisations only, or a combination of both). Although some researchers outside medical informatics have tried to compare the effectiveness of concurrent and retrospective protocols [e.g. 10, 11], the present study goes a step further by employing a new type of retrospective protocol based on the sense making approach. In addition, we analyse the effectiveness of the two methods in the context of a clinical decision support tool.

This paper is structured as follows: The next section presents the COCPIT tool, a prototype clinical decision making software used as a test-bed for experimentation during the evaluation of the two types of verbal protocols. Then, we describe the research design implemented to compare the two proto-

cols. The results of the comparison follow. Finally, the paper concludes with a summary of the main findings and recommendations for further research.

The system under evaluation

COCPIT (Collaborative Online Care Pathway Investigation Tool) is a tool for visualising, analysing and developing integrated care pathways at the population level. Care pathways define a chronological sequence of steps, most commonly diagnostic or treatment, to be followed in providing care for patients. Care pathways are used to aid health policy making, as they effectively implement clinical guidelines, with customisation possible at a local level to reflect service provision. Interactive care pathways integrate care pathway design with computational intelligence and statistical analysis in a single run-time environment, thus permitting users (typically clinicians and health service managers) to create, edit, re-use and analyse care pathways. This is achieved through a data management framework which provides access to individual-level patient health records. COCPIT's interface has two main components: first, a visual editor for designing care pathways; and second, a data analysis component implementing the methods and techniques for users to analyse and visualize the results for a range of applications.

The visual editor

There are two types of model that can be used and analysed using the visual editor. The first is the care pathway model, while the second is the state model. The care pathway model can contain one or more events which describe a patient's progression through the pathway (Figure 1). An event is defined as a single point in time where something happens. Examples of events include diagnoses, treatments and measurements that can take place at any point. When a patient's health record matches the conditions of an event, the patient is said to have experienced the event. Therefore, a single event from a care pathway counts all patients whose electronic health records indicate that they experienced the event. This matching process is based on the use of standard clinical codes representing conditions, investigations and treatments that a patient may experience. When a clinical code is assigned to a given event in the care pathway, the COCPIT tool's matching algorithm scans the patient record database (accessible through the data management framework of COCPIT) for matches with clinical codes assigned to existing patient records.

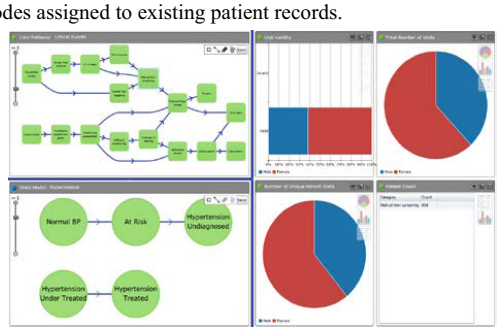

Figure 1- Representative screen-shot of the interface

The state model can contain one or more *states* by which patients can be grouped into discrete categories. Examples of states might be: diabetic, hypertension diagnosed, or undiagnosed hypertension. Similar to events, when patients' health records match the conditions of a state they are said to enter the state. When the conditions stop being met patients leave the state or they get transferred to another state (e.g., from the undiagnosed hypertension state to the hypertension diagnosed state). The role of a state model is to provide a breakdown of patients' characteristics within a specific care pathway, such as the categorisation of patients according to specific states, or risk factors (e.g., diabetes or hypertension), at the time they entered into a specific event (e.g., suspected stroke). Thus, the use of state models can show the reasons why patients entered an event given a particular health state, as well as identify missed opportunities related to health care provision (for instance, the number of patients who by the time they were admitted to a hospital with stroke symptoms [event] had had their hypertension diagnosed but not treated [state]).

The Data Analysis Component

The COCPIT tool provides a space where the statistical output of these two models can be represented visually (Figure 1). In particular, by selecting a specific event in a care pathway, or state in a state model, a user can visualize the number of patients entering that event, or state. Statistical output can be further manipulated, e.g., by clustering the results according to patient demographics or according to selected states. In addition, the tool can calculate the time it takes for patients to transition between events, or the length of time spend in a specific state. Further statistical methods such as analysis of variance, product-limit survival curves, and proportional hazards modelling enable comparisons of the expected and real care outcomes of patients.

Methodology

A total of 35 students in medicine and nursing participated in this study. Participants were divided into two groups: 1. Concurrent TA group (n = 18); and 2. Retrospective sense making protocol group (n = 17). In order to make the groups as comparable as possible, participants were allocated evenly according to their familiarity with clinical decision support tools and other types of clinical information systems as well as their level of knowledge and experience of care pathways and clinical records. This allocation aimed to minimise the effects of between-subject variability with regard to the aforementioned contextual factors vs. participants' behaviours during the data collection process. In order to achieve this level of homogeneity, each volunteer who expressed an interest in participating in the study was asked to fill in a screening background questionnaire in advance. Selected participants were then contacted and meetings were arranged in a usability laboratory. At the beginning of the meeting each participant was introduced to the objectives of this study and filled in a consent form. After the completion of the consent form each participant received five minutes of standard training on the basic functionality of the COCPIT tool.

Following the completion of the training session each participant was asked to complete seven predetermined tasks using COCPIT. The tasks involved creating and editing care pathways and state models, as well as manipulation and visualisation of the statistical output produced from these models. Participants were expected to complete all tasks. Tasks were presented to participants in a different order using a Latin square design. This decision was made in order to counterbalance the effects of learning transfer. All participants interacted with the COCPIT tool using a desktop computer and a 17-inch screen. During task completion a concurrent TA protocol **(condition A)** was applied to 18 of the participants in the study (i.e., par-

ticipants were expected to verbalise any feelings or thoughts that naturally came in their mind during task performance). The remaining 17 participants performed the tasks without been requested to think aloud. However, this latter group of participants, after task completion, was asked to watch a video of their on screen behaviour and comment on it following a retrospective sense making protocol (**condition B**).

Concurrent TA protocol (Condition A)

In the context of this study, a Level-2 TA protocol [12] was implemented because it can be applied concurrently, thus eliminating the bias of post-hoc rationalisation or reconstruction effects. There is evidence that a Level-2 TA protocol, when applied concurrently, does not influence user observed behaviour [13]. Finally, this type of protocol explores the contents of thought. These contents are not always made available in short term memory in verbal form (like in the case of a Level-1 verbal protocol) but may involve the articulation of visual and non-declarative information. This kind of articulation is better suited to common problem-solving tasks.

During the usability test, participants were instructed to talk aloud. If participants remained silent for more than 20 seconds during the task sessions then they were reminded to keep talking. Audio and screen recordings were used for data collection. The transcripts produced both from the audio and screen recordings were analysed using an inductive content analysis.

Retrospective sense making protocol (Condition B)

Sense making is an integral part of any cognitive activity and an efficient way to observe how people think and make decisions when they are trying to perform tasks. Although different versions of sense making protocols exist [14, 15], in the context of this study the gap-bridging approach to sense-making was implemented. This gap-bridging approach has been defined as a series of steps that a person takes across space and time in everyday life in order to solve a problem [15]. Each step of this process consists of gaps and actions taken to bridge them. Sense making can therefore be defined as the cognitive process through which people experience problems (i.e. gaps) and choose to perform certain actions, among alternative ones, in order to solve the problems experienced at a specific point in time. This process is cyclic, and gap-bridging may occur several times as part of small sub-tasks that form larger tasks.

The gap-bridging approach was implemented into the retrospective verbal protocol as a set of questions in order to prompt participants to think aloud. The questions followed the flow of the participants' recorded on-screen behaviour. The aim of these questions was to collect data, in a structured way, about participants' cognitive processes while performing the seven tasks using the COCPIT interface.

The retrospective sense making protocol begins as a common TA protocol where participants are asked to talk aloud while watching a video of their on screen behaviour. Videos were created using a screen recording software that recorded participants' interactions during task performance. Then for each observed click-through activity where an error occurred (e.g., selection of a wrong option) participants were asked a set of three questions by the researcher (as an observed error is defined any action made by the participant which deviates from the optimum working procedure for a task - i.e. any deviation from the set of correct steps or actions needed to complete a task). The three questions were: 1. Why did you choose to make this action? (e.g., why did you choose to click on that link?); 2. What problems, if any, did you encounter after performing this action? (i.e. after clicking on that link); and 3. What actions did you perform in order to solve these problems? These questions followed Dervin's gap-bridging approach and permitted an in-depth but structured analysis of participants' interaction with the COCPIT interface at the individual task level [15].

Data Analysis

An inductive content analysis technique was implemented in order to analyse the data collected from the screen recordings and the concurrent/retrospective verbal protocols. The analysis focused on five variables to compare the effectiveness of the two verbal protocol conditions: 1. The type of observed usability problems; 2. The total number of times each type of usability problem had occurred; 3. The number of unique usability problems (i.e. problems unique to each condition); 4. The usability category within which each type of usability problem had occurred (this was based on six categories of usability problems identified by [9]: Consistency, transparency, control, cognition, context and speed); and 5. The way the usability problems had surfaced (i. through observation of the on-screen behaviour only, ii. through participants' verbalisations only, or iii. through a combination of both). The identification of usability problems and their categorisation into types and categories of usability problems was based on inter-coder agreement [9]. Poisson analyses were performed in order to compare the rates at which usability problems were identified between the two verbal conditions. Results are presented as incidence rate ratios (IRR) with [95% confidence intervals] and two-sided P values. However, due to the small sample size of participants in this study any generalisations should be made with caution.

Results

Number and Type of Usability Problems

The findings showed that more usability problems were identified in the case of the retrospective protocol condition rather than the concurrent condition (IRR = 1.71 [1.40 to 2.10], $P < 0.0001$).

As shown in Table 1, the most frequent types of usability problems were those related to the lack of messages about the validity of the actions performed by users; the interpretation of statistical output; and the misinterpretation of the role of various functions (like the role of a state model). Significantly more usability problems were identified with the retrospective condition for the following types of usability problems: interpreting statistical output (IRR = 4.06 [1.06 to 12.2], P = 0.0012); lack of messages about the validity of the actions performed (IRR = 2.32 [1.43 to 3.90], P = 0.0003); use of ambiguous written instructions and labels (IRR = 4.10 [1.84 to 10.33], P = 0.0001); interpreting search results after a search for clinical codes (IRR = 3.84 [1.71 to 9.71], P = 0.0004); searching for clinical codes (IRR = 2.51 [1.375 to 4.83], P = 0.0015). The concurrent condition appeared to be more sensitive in detecting usability problems concerning the editing of a node in a pathway or the use of edges in a pathway. However, in this case statistically significant differences were observed between the two conditions only for problems concerning the editing of nodes in a pathway (IRR = 0.5 [0.3 to 1.0], P = 0.04).

Table 1 – Occurrence rates (detections per user/session) of types of usability problem

Type of problem	Concurrent (n = 18)	Retrospective (n=17)
Editing a node in a pathway	1.83 (33)	1.0 (17)
Misinterpreting the role of functions	0.89 (16)	1.65 (28)
Data input	0.77 (14)	0.71 (12)
Using an edge in a pathway	1.17 (21)	0.71 (12)
Searching for clinical codes	0.88 (16)	2.23 (38)
Interpreting search results	0.44 (8)	1.71 (29)
Use of ambiguous instructions.	0.44 (8)	1.82 (31)
Takes time to load or save data	0.55 (10)	0.53 (9)
Lack of messages about the validity of the actions performed.	1.39 (25)	3.24 (55)
Interpreting statistical output	0.33 (6)	1.35 (23)

Unique usability problems

The majority of usability problems were identified in both groups of verbal protocols (85%). However, where problems were detected with one protocol only, 12% were uniquely identified by the retrospective protocol, compared with 3% for the concurrent protocol. A closer examination of the problems that were unique to the two protocols showed that in the case of the retrospective protocol this percentage consisted of problems related to the participants' cognitive behaviour, such as problems related to the interpretation of statistical output, the interpretation of the search results after a search for clinical codes, the use of ambiguous written instructions, and the lack of messages or information about whether or not the actions performed were valid. With the concurrent protocol most of the unique usability problems were related to behavioural tasks, such as editing a node in the care pathway, using an edge to connect nodes in the pathway, and searching for clinical codes.

Categories of usability problems

Types of usability problems (Table 1) were classified into six compound categories. These were general categories of usability problems, common in the evaluation of bio-health information systems [9]. As it is shown in Table 2 most types of usability problems were related to the categories of consistency (e.g., standardised use of interface objects, colours, and text as well as predictable behaviour of controls) transparency (e.g., current state is visible, future states can be predicted, action effects are indicated) and cognition (e.g., this refers to the density or ambiguity of information and the amount of cognitive and visual search effort spent). Statistically significant differences between the concurrent and retrospective conditions were seen only in the transparency and cognition categories. In particular, more usability problems related to the transparency of the interface objects were identified by the retrospective protocol (IRR = 2.28 [1.53 to 3.04], P < 0.0001). Similarly, more usability problems related to cognition were identified in the retrospective case (IRR = 1.93 [1.30 to 2.90], P = 0.0005).

Table 2 – Occurrence rates (detection per user/session) of usability problem categories

	Concurrent (n= 18)	Retrospective (n = 17)
Consistency	2.00 (36)	1.94(33)
Transparency	2.17 (37)	4.60 (80)
Control	0.72 (13)	0.59 (10)
Cognition	2.66 (41)	4.18 (75)
Context	1.05 (19)	2.82 (48)
Speed	0.55 (11)	0.47 (8)

Data sources of usability problem identification

Usability problems could have surfaced in one of three ways: 1. Through analysis of verbal transcripts only; 2. Through analysis of on-screen recordings only; and 3. Through a combination of both transcripts. Most usability problems (55%) surfaced from the analysis of verbal transcripts only. Just 10% of usability problems arose from on-screen recordings, and 35% from both transcripts. Finally, the retrospective condition gave rise to significantly more verbally surfaced problems (IRR = 1.70 [1.30 to 2.25], P <0.0001).

Conclusions

The purpose of this paper was to present a retrospective verbal protocol based on the sense-making approach and compare it with the traditional concurrent TA protocol. The findings show that the retrospective protocol identified significantly more usability problems, including more unique problems, than the concurrent protocol. These problems were related to human cognitive behaviour during task completion. Examples of such problems included those related to the interpretation of information (either information represented as statistical output, search results, or ambiguous written instructions and labels). A significant number of these problems were not detected with the concurrent verbal protocol. We suggest that the greater sensitivity of the retrospective sense making protocol is due to the detection of compound usability problems, which recorded in a structured way, more unique usability problems.

Figure 2 illustrates a compound usability problem, where: nodes represent usability problems, the (>>) notation denotes hierarchical relationship between two usability problems, the (|||) notation indicates the presence of usability problems that happen concurrently, and finally the capital letter (*R*) shows the presence of a usability problem that repeats itself. Compound usability problems were typically identified with the three gap-bringing questions integral to the sense making approach of the retrospective protocol. In particular, a compound usability problem contains a root problem. This is documented usually through question Q1 (Figure 2). The analysis of interaction data showed that 73% of root usability problems were related to the consistency category and contained behavioural

errors made by participants while trying to complete the tasks. These errors were associated with various types of behavioural usability problems like the use of an edge to connect nodes in a care pathway, the editing of a node in a pathway, and the search for clinical codes.

A root problem may initiate or lead to additional usability problems through question Q2 (Figure 2). These represented problems that participants experienced after they had encountered the root problem. The majority of these problems (89%) were associated with the following types of usability errors: interpretation of the search results (after searching for clinical codes); interpretation of statistical output; and lack of messages or information about whether the actions performed were valid. These problems were common in the case of the usability categories of transparency and cognition. The majority of problems identified through the analysis of Q2 were uniquely captured by the retrospective protocol and reflected users' cognitive behaviour. In many cases this behaviour could lead to additional errors that formed new or repeated usability problems. These were disclosed using question Q3 (Figure 2).

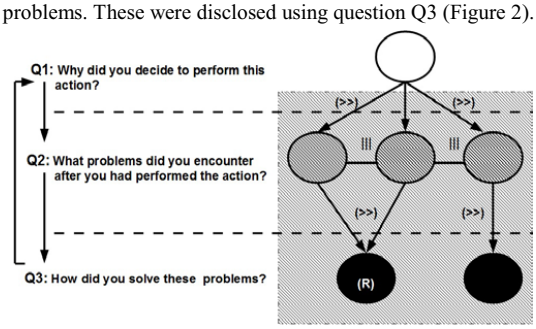

Figure 2- Compound usability problem

Further research is now testing the effectiveness of the retrospective sense making protocol across a variety of medical informatics applications and usability methods. We are also investigating a new technique for calculating the severity of compound usability problems logarithmically, i.e. based on the cumulative severity weights assigned to individual usability problems that form part of a compound and granular problem. Given the increasing prevalence, complexity and importance of clinical decision support systems, the evaluation of their usability needs to be robust. We have shown that a common protocol for testing usability in medical informatics is insensitive, and how to improve on it. There is a need for medical informatics to develop a comprehensive evidence-based approach to inform usability study design.

Acknowledgments

The authors were funded by the UK National Institute for Health Research (NIHR) as part of the Greater Manchester Collaboration for Leadership in Applied Health Research and Care (CLAHRC).

References

[1] Beuscart-Zephir, M, Peloyo, S and Bernonville, S. Example of a Human Factors engineering approach to a medication administration work system: potential impact on patient safety. Int J Med Inform 2010: 79: 43-57.

[2] Walji, M, Katenderum, E, Tron, D, Kookal, K, Nguyen, K, Tokede, O, White, J, Vaderhobli, R, Ramani, R, Stark, P, Kimmes, N, Schoonheim-Klein, M, Patel, V. Detection and characterisation of usability problems in structural data entry interfaces in dentistry. Int J Med Inform 2013: 82(2): 128-38

[3] Carroll, C, Marsden, M, Soden, P, Naylan, E, New, J and Dornam, T. Involving users in the design and usability evaluation of a clinical decision support system. Comput Meth Prog Bio 2002: 69: 123-35.

[4] Backhshi-Raiez, F, de Kaizer, NF, Cornet, R, Dorrepaal, M, Dongelmans, D, Jaspers, MWM. A usability evaluation of a SNOMED CT based compositional interface terminology for intensive care. Int J Med Inform 2012: 81: 351-62.

[5] Lai, TY. Iterative refinement of a tailored system for self-care managmeent of depressive symptoms in people with HIV/AIDS through heuristic evaluation and end user testing. IJIMI 2007: 76: 317-24.

[6] Choi, B, Drozdetski, S, Hackett, M, Lu, C, Rottenberg, C, Yu, L. Usability comparison of three clinical trial management systems. Proceedings of AMIA, 2005: pp.921.

[7] McDonald, S, Zhao, T, and Edwards, H. Dual verbal elicitation. Human Computer Interaction 2013: 1-56.

[8] Li AC, Kannry, JL, Kushniruk A, Chrimes, D, McGinn TJ, Edonyabo, D, Mann DM. Integrating usability testing and think-aloud protocol analysis with near live clinical simulations in evaluating clinical decision support. Int J Med Inform 2012: 81(11): 761-72

[9] Horsky J, McColgan K, Pang JE, Melnikas AJ, Linder JA, Schnipper JL, Middleton B. Complementary methods of system usability evaluation: surveys and observations during software design and development cycles. J Biomed Inform 2010: 43: 782-90.

[10] Jensen, J. Evaluating in a healthcare setting: a comparison between concurrent and retrospective verbalisation. Human Computer Interaction, 2007: 508-16.

[11] van den Haak M., de Jong M, Schellens PJ. Retrospective vs concurrent think aloud protocols: testing the usability of an online library catalogue. Behav Inform Technol 2003: 22: 339-51.

[12] Ericsson K, and Simon, H. Protocol Analysis. 1993: MIT press.

[13] Gerjets, P, Kammerer, Y, & Werner, B. Measuring spontaneous and instructed evaluation processes during web search. Eur Res Int 2011: 21: 220-31.

[14] Pirolli P, Card S. The sense making process and leverage points for analyst technology as identified through cognitive task analysis. Proceedings of the International conference on Intelligence Analysis 2005: pp. 2-4.

[15] Dervin, B. Chaos, order, and Sense-Making: A proposed theory for information design. Information design 1999:: MIT Press. pp. 35-65.

Address for correspondence

Panos Balatsoukas, 1.308 J. McFarlane building, Centre for Health Informatics, Institute of Population Health, University of Manchester, Oxford Road, M13 9PL.