TITLE

Usability testing of a respiratory interface using computer screen and facial expressions videos

AUTHORS

Ana Oliveira[a], Cátia Pinho[a,c], Sandra Monteiro[a], Ana Marcos[a], Alda Marques[a,b]

[a] School of Health Sciences, University of Aveiro (ESSUA), Campus Universitário de Santiago, Aveiro, Portugal.

[b] Unidade de Investigação e Formação sobre Adultos e Idosos (UNIFAI), Porto, Portugal.

[c] Institute of Electronics and Telematics Engineering of Aveiro (IEETA), University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal

1st Author 1
Name: Ana Oliveira
email: alao@ua.pt

2nd Author:
Name: Cátia Pinho
email: catiap@ua.pt

3rd Author:
Name: Sandra Monteiro
email: sandramonteiro@ua.pt

4th Author:
Name: Ana Marcos
email: sandramonteiro@ua.pt

Corresponding author
Name: Alda Marques
Address: School of Health Sciences, University of Aveiro (ESSUA), Campus Universitário de Santiago, Aveiro, Portugal
email: amarques@ua.pt
Phone: +351 234372462
Fax: +351 234401597

ABSTRACT

Computer screen videos (CSV) and users' facial expressions videos (FEV) are recommended to evaluate systems performance. However, software combining both methods is often non-accessible in clinical research fields. The Observer-XT software is commonly used for clinical research to assess human behaviours. Thus, this study reports on the combination of CSV and FEV, to evaluate a graphical user interface (GUI).

Eight physiotherapists entered clinical information in the GUI while CSV and FEV were collected. The frequency and duration of a list of behaviours found in FEV were analysed using the Observer-XT-10.5. Simultaneously, the frequency and duration of usability problems of CSV were manually registered. CSV and FEV timelines were also matched to verify combinations.

The analysis of FEV revealed that the category most frequently observed in users behaviour was the eye contact with the screen (ECS, 32±9) whilst verbal communication achieved the highest duration (14.8±6.9minutes). Regarding the CSV, 64 problems, related with the interface (73%) and the user (27%), were found. In total, 135 usability problems were identified by combining both methods. The majority were reported through verbal communication (45.8%) and ECS (40.8%). "False alarms" and "misses" did not cause quantifiable reactions and the facial expressions problems were mainly related with the lack of familiarity (55.4%) felt by users when interacting with the interface.

These findings encourage the use of Observer-XT-10.5 to conduct small usability sessions, as it identifies emergent groups of problems by combining methods. However, to validate final versions of systems further validation should be conducted using specialized software.

Key words: graphical user interface, usability testing, facial videos, screen videos; Observer XT.

1 INTRODUCTION

Healthcare professionals are increasingly challenged to acquire and manage large amounts of information, while still providing high quality health services. Thus, healthcare information systems (HCIS) have become vital to store, organize and share clinical information, which facilitates and improves health professionals' decision making [1]. Although health professionals are the major beneficiaries of these technologies, they often resist to their implementation [2, 3]. This resistance have been attributed to the felling of loss of control expressed by health professionals when interacting with systems [4]. Furthermore, computer systems are often developed by professionals outside the health field who often do not have a full understanding of clinical evaluations and procedures [5]. This may affect the construction of system by being complex and difficult to navigate, contributing to health professionals' resistance to its use. Therefore, systems evaluations performed with the end users are essential, not only in the final version, but throughout the progress cycle to guarantee that the system is develop according to health professionals standards, ensuring its effectiveness, efficiency and usability [6, 7]. To verify and optimise systems, analytical and empirical methods from the area of usability engineering and human-computer-interaction have been applied in HCIS evaluation studies [8]. Kushniruk and Patel [5, 9] have been researching in the field of usability testing and proposed different types of data collection, such as video recordings of the computer screens and users while performing tasks and think-aloud reports.

Computer screen videos (CSV) are one of the most used techniques to develop effective evaluations and assess effectiveness and efficiency of the systems [10, 11].This technique allows researchers to collect observational data of users performance when interacting with a product and capture crucial information, such as the time spent in different tasks and the number of errors occurred [9], during the interaction. The use of CSV have been suggested over qualitative methods, such as interviews and pre-structured questionnaires, as they are more objective and capture the problems in real time [5]. However, it has also been stated that the assessment of these parameters alone, do not guarantee users satisfaction [12]. User satisfaction is influenced by personal experiences with technology, preferred working style, and the aesthetics of systems' design. Such quality aspects seem to be important for users but are not connected to their performance with the system [13]. Furthermore, it is important to assess how people feel when using the system. A variety of methods can be employed to address this aspect, such as i) physiological measures (e.g., electromyography (EMG) and pupil responses), which offer high spatio-temporal resolution, are expensive and require high-level of expertise from the technicians [14]; and ii) various kinds of survey methods (e.g., questionnaires and interview techniques) [12], that are accessible and easy to use but provide limited information, since emotional experiences are not primarily language-based [14]. Thus, recordings of facial expressions emerge as an alternative to these methods.

Facial expressions have been reported as the most visible and distinctive emotion behaviours, reflecting individuals' current emotional state and communicating emotional information [15]. Some studies have been conducted to integrate users' facial expressions response in the usability assessment of graphical user interface (GUI), however they were conducted with expensive software that are not easily accessible in the field of clinical research [16, 17]. The Observer XT is a user-friendly software to collect, analyse and present observational data, often used in social and clinical areas to assess human behaviours [18, 19]. Therefore, this software can be a useful tool to assess users' experience with preliminary GUI in clinical research.

This study aimed to report on the combination of CSV and users' facial expressions videos (FEV) analysed with the Observer XT software to evaluate a respiratory GUI named as LungSounds@UA [20].

2 METHODOLOGY

2.1 GUI description

The LungSounds@UA graphical interface was developed in the scope of a pilot study within a clinical respiratory research project[1]. This GUI aimed to collect and organise respiratory data in a single multimedia database.

A multilayer of windows built with five hierarchy levels, i.e., A, B, C, D and E composes LungSounds@UA interface (figure 1-A). The interface which allows users to record respiratory sounds (figure 1-B) and upload related-respiratory data, such as: clinical parameters; clinical analysis; respiratory physiotherapy monitoring data; functional independence measure (FIM); six minute walk test parameters; spirometry; pain evaluation; imaging reports, e.g., computed tomography (CT) and chest X-ray (Rx); and conventional auscultation.

(insert figure 1 about here)

The organisation of contents in the interface was established according to the physiotherapists' current practice, however alternative navigation controls, such as the vertical buttons displayed on the left side of the computer screen, can be used to easily allow different data entry order. Detail description of the LungSounds@UA graphical interface has been published in Pinho et al. (2012) [20].

2.2 Design

LungSounds@UA was tested in two evaluation sessions conducted on the same day at the University of Aveiro, Portugal. Each session lasted for approximately 70 minutes. The testing room was prepared according to Kushniruk and Patel [5] recommendations, with 4 computers capable of running the software under study and the TipCam Screen Recording Software [21],

[1] Research project ref. PTDC/SAU-BEB/101943/2008.

two desks and two cameras (one camera per desk) to record the participants' facial expressions. Two participants with an individual computer were sited per desk to perform the required tasks (figure 2). Participants were instructed not to interact with each other (i.e. speak, touch or establish eye contact).

(insert figure 2 about here)

2.3 Participants

Eligible participants were selected according to the usability definition of ISO 9241-11, i.e., "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [22, 23] and Nielsen's recommendations on sample sizes [24]. Therefore, eight physiotherapists were recruited to test the GUI, as this class of health professionals were the main target users of the developed application. Physiotherapists were divided in two groups of four each, according to their practice (research or clinical), to maximize the outcomes of the evaluation session [8]. For the same purpose, it was also ensured that all participants had experience in the field of respiratory physiotherapy but never had previous experience with the interface, so maximal reactivity of participants could be observed [7, 25] which would inform necessary improvements. A training session was not applied, as it has been stated that its absence can strengthen the evaluation because full information about the perceived weaknesses are reported, when using the developed applications. Without training session users approach a new system with preconceived ideas based only on their prior experiences, and draw their own conclusions about how it works, which may differ from the designer's intentions [26].

All participants accepted to take part of the study and signed the informed consents prior to any data collection.

2.4 Data Collection

To verify and optimise the interface usability, participants were instructed to enter the same clinical parameters (from a pre-structured case study) in the LungSounds@UA GUI, while their screen and facial expressions were being video recorded [7].

Two facilitators specialized in the interface were in the sessions, however, only intervened to clarify participants' questions. One facilitator read the case study aloud and participants were given enough time to read it by themselves and clarify any doubts before starting the tasks. Then, the facilitators turned on the recorder software and the video cameras.

This methodology (CSV plus FEV) allowed to obtain a more complete evaluation of participants' interaction with the system, when performing the same task. Each camera collected data from two participants (four videos of facial expressions) and CSV were obtained individually (eight CSV), generating twelve video files in total.

3 DATA ANALYSIS

The data were analysed by four researchers. One researcher conducted the analysis of the FEV, one analysed the CSV and two researchers conducted the analysis of the combination of the CSV and FEV.

3.1 Analysis of the facial expressions videos

Facial expressions were studied by analysing the frequency and duration of a list of behaviours (ethogram), derived from: i) the existing literature [27-29]; ii) preliminary observations of the video recordings, regarding engagement aspects with the interface [30] (one trained observer watched all videos and captured the main behaviours of the participants); and iii) the facial acting coding system (FACS) [31]. FACS is a detailed, technical guide that explains how to categorize facial behaviours based on the muscles that produce them. This system follows the premises that basic emotions correspond to facial models [31] and has been proposed and used by many authors to assess their computer systems [14, 17, 32]. The following categories composed the ethogram: i) eye contact with the screen (the user is visibly concentrating on the screen, in order to read, search or understand something in the interface); ii) eyebrows movement; iii) verbal communication; and iv) smile. The first three categories have been reported as indicative of the occurrence of an adverse-event when interacting with the system (e.g., system errors and emotional distress) [27, 33]. Conversely, smile has been associated with agreement and accomplishment [34, 35]. Table I provides a detailed description of each category.

(insert table 1 about here)

One researcher, blinded to the evaluation (that did not participate in the data collection), assessed each of the four FEV and rated facial expressions according to the ethogram, using the specialized software, Noldus The Observer XT 10.5 (Noldus International Technology, Wageningen, the Netherlands). The frequency and duration of the categories were measured [36, 37]. The researcher was trained previously to use the software.

3.2 Analysis of the computer screen videos

Eight CSV were observed and analysed by another researcher, blinded to the evaluation. The frequency and duration of the usability problems found in participants' screens (i.e., warning, error messages and other inconsistencies) were reported. A usability problem was defined as a specific characteristic of the system that hampers task accomplishment, a frustration or lack of understanding by the user [38].

After this analysis, data were coded and grouped into themes and sub-themes, according to previous work conducted by Kushniruk and Patel [5] and Gray and Salzman [39]. Interface (i.e., layout/screen organization, false alarms, time consumption and misses) and user (i.e., unfamiliarity with interface) problems were evaluated through the observation of the CSV. Table II provides a detailed description of each theme and sub-theme.

(insert table 2 about here)

The interface and user problems, were classified when error, warning messages or inconsistencies (other conflicts not reported by these messages) were identified.

3.3 Reliability of the observations

Each FEV was analysed three times by the same researcher to assess the intra-observer reliability of the observations [37]. The intra-observer reliability analysis was conducted for the frequency and duration of each behaviour category with the intraclass correlation coefficient equation ICC (2.1) [40].

Intra-observer reliability was not analysed for the CSV as the findings mainly consist in the objective quantification of messages produced by the graphical interface, and therefore, the intra-observer agreement would have been maximum (ICC=1).

3.4 Combination of the computer screen and facial expressions videos

After the individual analysis of the FEV and CSV, two researchers matched their timelines to relate the coded facial expressions with the usability problems presented by participants in the screens recordings. Disagreements between researchers were resolved by reaching a consensus through discussion. If no consensus could be reached, a third researcher was consulted. After observing all FEV, only facial expressions longer than 20s demonstrated to have significant impact on the participants interaction with the system (a threshold empirically established), and therefore were considered to represent the most important/relevant problems found by participants. Spearman's correlation coefficient was used to correlate each facial expression with each interface and user problems. Correlations were interpreted as weak ($r_s \leq 0.35$), moderate ($0.36 \leq r_s \leq 0.67$) and strong ($r_s \geq 0.68$) [41]. Analysis was performed using PASW® Statistics 18.0 software for Windows (SPSS Inc, Chicago, IL, USA). Significance level was set at $p<0.05$.

4 RESULTS

Each participant took on average 42±6 minutes to complete the proposed tasks.

4.1 Facial expressions

The analysis of the videos took 15 hours to be completed. The behaviour categories analysed in the facial expression are presented in table III and figure 3. Eye contact with the screen was the behaviour category most frequently observed (32±9). The verbal communication was the category with the highest duration (14.8±6.9 minutes). It was also found that eyebrows movement and smile categories occurred less frequently and represented only 2% (2±3) and 1% (2±1) of the users' frequency behaviour, respectively (figure 3).

(insert table 3 about here)

(insert figure 3 about here)

Intra-observer reliability analysis of facial expressions revealed ICC values ranging between 0.91 and 1.00 for all categories except one, indicating an excellent reliability. The lower ICC value represented good reliability and was found for the duration of the smile category (0.54) [42].

4.2 Computer screen videos

The analysis of the videos took 9 hours to be completed. In the eight CSV, 64 problems, both interface (47/64; 73%) and user (17/64; 27%) were found. The major difficulties that emerged from the interaction with the interface were: i) layout/screen organization flaws (26/47; 55%); ii) false alarms *(*9/47; 19%); iii) time consumption (8/47; 17%); and iv) misses (4/47; 9%). The users' problems were all due to unfamiliarity with interface (17/17; 100%).

The majority of the interface and users' problems were reported by error messages (27/64; 42%) however, looking only at the interface problems it is clear that the problems were mainly reported by other inconsistencies (28/47; 60%) (table IV).

(insert table 4 about here)

4.3 Combination of the computer screen and facial expressions videos

After matching the coded facial expressions with the usability problems presented in the screens, it was observed that the same facial expression could be associated with more than one screen problem, and therefore 135 problems were identified. The majority of problems were reported by verbal communication (45.8%) and eye contact with the screen (40.8%). It was also found that the problems identified by facial expressions were mainly related with the participants' lack of familiarity with the interface (55.4%) (figure 4).

(insert figure 4 about here)

Most of the correlations found were moderate (r_s varied from 0.40 to 0.65). Strong correlations were found between the verbal communication and unfamiliarity with the interface (r_s=0.77; p=0.27) and time consumption (r_s=0.69; p=0.59) categories. Smile correlated weakly with the layout (r_s=-0.24; p=0.56) and unfamiliarity with the interface (r_s=0.18; p=0.67) categories. Misses and False alarms did not cause quantifiable reactions in participants and therefore correlations were not found (table 5).

Examples of the combination between the two methods can be found in table 6.

(insert table 5 about here)

(insert table 6 about here)

5 DISCUSSION

To our knowledge, this is the first study reporting on the analysis of FEV and CSV combination using the Observer XT software. The combination of both methods, allowed perceiving and quantifying facial expressions triggered by the “layout” (29.6%), “unfamiliarity with the interface” (55.4%) and “time consumption” (15%) problems. “False alarm” (0%) and “misses” (0%) did not generate relevant facial expressions.

Through the individual analysis of the CSV and FEV it was not clear which were the most relevant problems perceived by users. The analysis of facial expressions alone showed high frequency of eye contact with the screen, which according to Despont-Gros, et al. [33]and Bevan and Macleod [43] indicates that participants experienced difficulties in searching, perceiving, learning, memorising and locating parameters in the interface menu. Long periods of time were found in the verbal communication category , revealing that participants had some difficulties to complete the tasks by themselves, requiring help from the facilitators to proceed [43]. The low percentages identified in smile and in eyebrows movement categories , could denote some displeasure and/or distress, felt by participants when interacting with the interface[27, 35]. These results showed that FEV alone informs about users’ perception and emotions when interacting with the interface however, objective information about the specific interface problems is not provided.

On the other hand, in the CSV analysis, the interface problems represented 73% of the total problems counted in the systems evaluation. This high percentage may be misleading as it suggests a large variety of problems, nonetheless, the same problems were reported by all participants and, in some situations, more than once, by the same participant, overweighing the total of problems counted. The users' problems overestimated the modifications that needed to be performed in the interface, since they were 100% due to unfamiliarity with interface. Thus, through the CSV analysis, the amount of problems and errors reported by the interface can be addressed, but not which ones are truly useful to the user or need to be improved by the interface developers.

The combination of both methods, allowed perceiving and quantifying the facial expressions that were triggered by the "layout", "unfamiliarity with the interface" and "time consumption" problems but not by "false alarm" and "misses". It can be hypothesised that the null values obtained in "false alarm" and "misses" are related to the difficulty in differentiating this two categories from the time consumption problems. The presence of warning and error messages for non-existent problems (false alarm) and/or its absence (misses) in the execution of some tasks may have caused users to be lost in the interface, and therefore spent more time performing the task, which is counted as "time consumption" in the combination of both methods. Nevertheless, these results provide useful data to enhance the interface, mainly in the system "layout" and "time consumption" problems. Different usability methods have been proposed to solve layout problems, such as developing of a "good screen design" by taking in consideration consistency, colour, spatial display and organizational display [44]. Other possibility would be to evaluate users' satisfaction of two different layouts and choose the one which better respond to their requirements [5]. The development of an appropriate layout can significantly reduce the time taken to complete tasks [44], and consequently solves the "time consumption" problems identified in this study.

Major improvements should not in a preliminary assessment be performed on the "unfamiliarity with the interface" (which were the most common problems experienced by users), as it can be justified by the absence of a training session [26, 45] and therefore, users might simply need more time to learn how to interact with the system.

This complementary approach (combination of FEV and CSV) provides valuable information about users' perception regarding the interface problems, which will aid the system developers to establish priorities according to what is crucial to the end users, increasing systems' effectiveness and efficiency.

5.1 Limitations and future research

The present study had some limitations. Firstly, a questionnaire exploring participants' background and familiarity with computers (recommended in some studies [5]) was absent however, as the degree in physiotherapy involves a basic education on computer software, this was not considered a major barrier for the participants' interaction with the system. Secondly, the presence of external observers in the testing room might introduce psychophysiological and emotional changes in test participants [46]. To minimize this effect, only facilitators (whose presence was essential to conduct the evaluation session) were allowed in the present study. Other strategies were also employed to reduce the influence of external factors and enhance participants' performance, such as the organization of the set up room and by following standardized rules in the implemented usability tests [46, 47]. However, due to the complexity of human behaviour it is not possible to guarantee that all variables capable of influence the participants were fully controlled. Thirdly, inter-rater reliability analysis could not be performed as only one researcher observed the FEV. Nevertheless, the inter-rater reliability to detect facial expressions has been found to range from fair to almost perfect agreements (ICC=0.33-0.91) [19]. Fourthly, the use of the Observer XT in this study was very time consuming (15 hours), and may not be appropriate to conduct large validations sessions. To overcome this problem, it would be advisable to perform evaluations with software that automatically match screen videos and facial videos timelines. However, in social and/or clinical sciences Observer XT is commonly available, often used to assess human behaviours, and therefore, researchers are well familiarised with this method which facilitates its implementation and guarantees the reliability and validity of the results found. Fifthly, a high rate of “unfamiliarity with the interface” was observed, mainly, because users did not have experience with the interface prior to the evaluation. A second round of tests would be valuable to confirm this high rate, nevertheless for this evaluations the blindness of the participants to the interface was essential to inform substantially improvements in futures interface versions. Finally, the findings of this study can also be limited by the fact that only groups, and not individual problems, can be identified by the combinations of both methods. Despite the above limitations, this was an evaluation of the first version of the system. The main objective was to have a first feedback of the end users in a controlled environment, therefore, a simple evaluation (as recommended - less expensive and with brief resources [48]) was developed, based on the combination of two usability methods. The combination of different usability methods is reported as a grey research area that requires further investigation to better understand their contributions in the usability field [49]. Therefore, this study constitutes a step towards a better understanding of new usability measures.

6 CONCLUSIONS

The use of CSV or FEV alone does not provide clear information about the most relevant problems perceived by users when interacting with a system, and therefore, these methods alone may not be the most comprehensive measures to assess the interface usability/functionality. The combination of CSV and FEV with the Observer XT leads to a new approach to the traditional techniques for evaluating information systems in medical informatics. However, to validate final versions of software to be use in large organizations, further validation need to be conduct with specialized software.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the funding provided to this project, "Adventitious lung sounds as indicators of severity and recovery of lung pathology and sputum location"- PTDC/SAU-BEB/101943/2008 by Fundação para a Ciência e a Tecnologia (FCT), Portugal.

The authors would also like to thank to all participants of the evaluation session for their contributions, which will improve the LungSounds@UA graphical interface.

REFERENCES

[1] P. Reed, K. Holdaway, S. Isensee, E. Buie, J. Fox, J. Williams, and A. Lund, "User interface guidelines and standards: progress, issues, and prospects.," Interacting with Computers 12, 119-142 (1999).

[2] A. Bhattacherjee and N. Hikmet, "Physicians' Resistance toward Healthcare Information Technologies: A Dual-Factor Model," in *HICSS 2007 40th Annual Hawaii International Conference on System Sciences*, 2007, pp. 141-141.

[3] D. P. Lorence and M. C. Richards, "Adoption of regulatory compliance programmes across United States healthcare organizations: a view of institutional disobedience," Health Serv Manage Res 16, 167-78 Aug (2003).

[4] A. Bhattacherjee and N. Hikmet, "Physicians' Resistance toward Healthcare Information Technologies: A Dual-Factor Model," in *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*, 2007, pp. 141-141.

[5] A. W. Kushniruk and V. L. Patel, "Cognitive and usability engineering methods for the evaluation of clinical information systems," Journal of Biomedical Informatics 37, 56-76 Feb (2004).

[6] A. Kushniruk, V. Patel, J. Cimino, and R. Barrows, "Cognitive evaluation of the user interface and vocabulary of an outpatient information system.," Proc AMIA Annu Fall Symp. : 22-26. (1996).

[7] A. W. Kushniruk, C. Patel, V. L. Patel, and J. J. Cimino, "Televaluation of clinical information systems: an integrative approach to assessing Web-based systems," International Journal of Medical Informatics 61, 45-70 (2001).

[8] L. Peute, R. Spithoven, P. Bakker, and M. Jaspers, "Usability studies on interactive health information systems; where do we stand?," Studies in Health Technology and Informatics 136, 327-332 (2008).

[9] A. W. Kushniruk, V. Patel, and J. Cimino, "Usability Testing in Medical Informatics: Cognitive Approaches to Evaluation of Information Systems and User Interfaces," Proc AMIA Annu Fall Symp 218-222 (1997).

[10] N. M. Diah, M. Ismail, S. Ahmad, and M. K. M. Dahari, "Usability testing for educational computer game using observation method," in *Information Retrieval & Knowledge Management, (CAMP), 2010 International Conference on*, 2010, pp. 157-161.

[11] Ambrose Little and C. B. Kreitzberg. (2009, July, October 10). *Usability in Practice: Usability Testing* Available: http://msdn.microsoft.com/en-us/magazine/dd920305.aspx

[12] M. Thüring and S. Mahlke, "Usability, aesthetics and emotions in human–technology interaction," International Journal of Psychology 42, 253-264 2007/08/01 (2007).

[13] A. Dillon, "Beyond Usability: Process, Outcome and Affect in human computer interactions," Canadian Journal of Information Science 26, 57-69 (2001).

[14] R. L. Hazlett and J. Benedek, "Measuring emotional valence to understand the user's experience of software," International Journal of Human-Computer Studies 65, 306-314 (2007).

[15] C. Darwin, The Expression of Emotions in Man and Animals. London: Murray, 1872.

[16] M. Asselin and M. Moayeri, "New tools for new literacies research: an exploration of usability testing software," International Journal of Research & Method in Education 33, 41-53 2010/04/01 (2010).

[17] J. Staiano, M. Menéndez, A. Battocchi, A. D. Angeli, and N. Sebe, "UX_Mate: From Facial Expressions to UX Evaluation," in *Designing Interactive Systems Conference (DIS '12)*, Newcastle, UK, 2012, pp. 741-750.

[18] B. A. Corbett, C. W. Schupp, D. Simon, N. Ryan, and S. Mendoza, "Elevated cortisol during play is associated with age and social engagement in children with autism," Mol Autism 1, 13 (2010).

[19] J. Cruz, A. Marques, A. Barbosa, D. Figueiredo, and L. X. Sousa, "Making sense(s) in dementia: a multisensory and motor-based group activity program," Am J Alzheimers Dis Other Demen 28, 137-46 Mar (2013).

[20] C. Pinho, D. Oliveira, A. Oliveira, J. Dinis, and A. Marques, "LungSounds@UA Interface and Multimedia Database," Procedia Technology 5, 803-811 (2012).

[21] uTIPu. (nd, 12 April). *TipCam*. Available: http://www.utipu.com

[22] ISO 9241-11, Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11 : Guidance on usability, 1998.

[23] A. Seffah, J. Gulliksen, and M. C. Desmarais, Human-Centered Software Engineering - Integrating Usability in the Development Process, Human-Computer Interaction 8. New York: Springer-Verlag, 2005.

[24] J. Nielsen. (2012, October, 10). *How Many Test Users in a Usability Study?* Available: http://www.useit.com/alertbox/number-of-test-users.html

[25] D. R. Kaufman, V. L. Patel, C. Hilliman, P. C. Morin, J. Pevzner, R. S. Weinstock, R. Goland, S. Shea, and J. Starren, "Usability in the real world: assessing medical information technologies in patients' homes," Journal of Biomedical Informatics 36, 45-60 (2003).

[26] A. F. Rose, J. L. Schnipper, E. R. Park, E. G. Poon, Q. Li, and B. Middleton, "Using qualitative studies to improve the usability of an EMR," Journal of Biomedical Informatics 38, 51-60 Feb (2005).

[27] P. Branco, L. M. E. d. Encarnação, and A. F. Marcos, "It's all in the face : studies on monitoring users' experience," in *Third Iberoamerican Symposium in Computer Graphics*, Santiago de Compostela, 2006.

[28] I. Arapakis, Y. Moshfeghi, H. Joho, R. Ren, D. Hannah, and J. M. Jose, "Integrating facial expressions into user profiling for the improvement of a multimodal recommender system," presented at the Proceedings of the 2009 IEEE international conference on Multi media and Expo, New York, NY, USA, 2009.

[29] R. Ward, "An analysis of facial movement tracking in ordinary human–computer interaction," Interacting with Computers 16, 879-896 (2004).

[30] H. L. O'Brien and E. G. Toms, "The development and evaluation of a survey to measure user engagement," Journal of the American Society for Information Science and Technology 61, 50-69 (2010).

[31] P. Ekman and W. V. Friesen, Facial Coding Action System (FACS): A Technique for the Measurement of Facial Actions. Palo Alto, CA: Consulting Psychologists Press, 1978.

[32] B. Zaman and T. Shrimpton-Smith, "The FaceReader: measuring instant fun of use," presented at the Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles, Oslo, Norway, 2006.

[33] C. Despont-Gros, H. Mueller, and C. Lovis, "Evaluating user interactions with clinical information systems: a model based on human-computer interaction models," Journal of Biomedical Informatics 38, 244-55 Jun (2005).

[34] C. Alvino, C. Kohler, F. Barrett, R. E. Gur, R. C. Gur, and R. Verma, "Computerized measurement of facial expression of emotions in schizophrenia," Journal of Neuroscience Methods 163, 350-361 (2007).

[35] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," IEEE Transactions on Multimedia 8, 500-508 (2006).

[36] J. M. C. Bastien, "Usability testing: a review of some methodological and technical aspects of the method," International Journal of Medical Informatics 79, e18-e23 (2010).

[37] L. Noldus, R. Trienes, A. Hendriksen, H. Jansen, and R. Jansen, "The Observer Video-Pro: new software for the collection, management, and presentation of time-structured data from videotapes and digital media files," Behavior Research Methods Instruments & Computers 32, 197-206 (2000).

[38] J. Nielsen and R. Mack, Usability Inspection Methods. United States of America: John Wiley & Sons, 1994.

[39] W. D. Gray and M. C. Salzman, "Damaged merchandise? a review of experiments that compare usability evaluation methods," Human-Computer Interaction 13, 203-261 (1998).

[40] P. Shrout and J. Fleiss, "Intraclass correlations: uses in assessing rater reliability.," Psychological Bulletin 86, 420-428 (1979).

[41] J. Weber and D. Lamb, Statistics and Research in Physical Education. St. Louis: CV: Mosby Co, 1970.

[42] J. Fleiss, The Design and Analysis of Clinical Experiments. New York: Wiley-Interscience, 1986.

[43] N. Bevan and M. Macleod, "Usability Measurement in Context," Behaviour & Information Technology 13, 132-145 Jan-Apr (1994).

[44] R. G. Saadé and C. A. Otrakji, "First impressions last a lifetime: effect of interface type on disorientation and cognitive load," Computers in Human Behavior 23, 525-535 (2004).

[45] I. Sommerville, Software Engineering, Sixth ed. New York: Addison-Wesley, 2001.

[46] A. Sonderegger and J. Sauer, "The influence of laboratory set-up in usability tests: effects on user performance, subjective ratings and physiological measures," Ergonomics 52, 1350-61 Nov (2009).

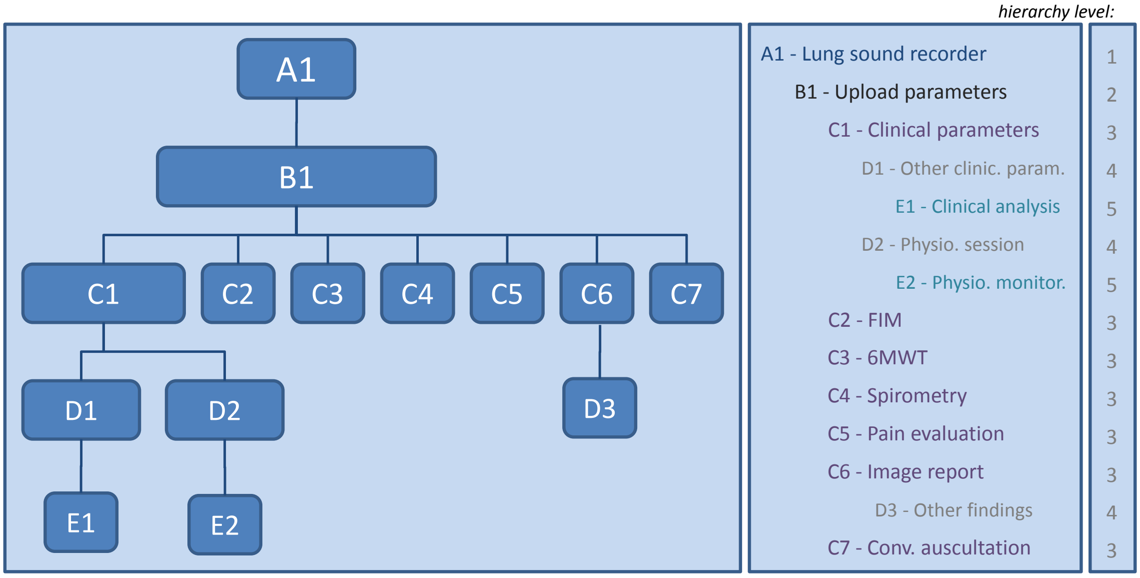
[47] M. Hertzum, K. D. Hansen, and H. H. K. Andersen, "Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload?," Behav. Inf. Technol. 28, 165-181 (2009).

[48] J. Nielsen. (2007, 2 January, 2012). *Fast, Cheap, and Good: Yes, You Can Have It All*. Available: http://www.useit.com/alertbox/fast-methods.html

[49] K. Hornbæk, "Current practice in measuring usability: Challenges to usability studies and research," International Journal of Human-Computer Studies 64, 79-102 (2006).

LIST OF FIGURES (with captions)

[A]



[B]

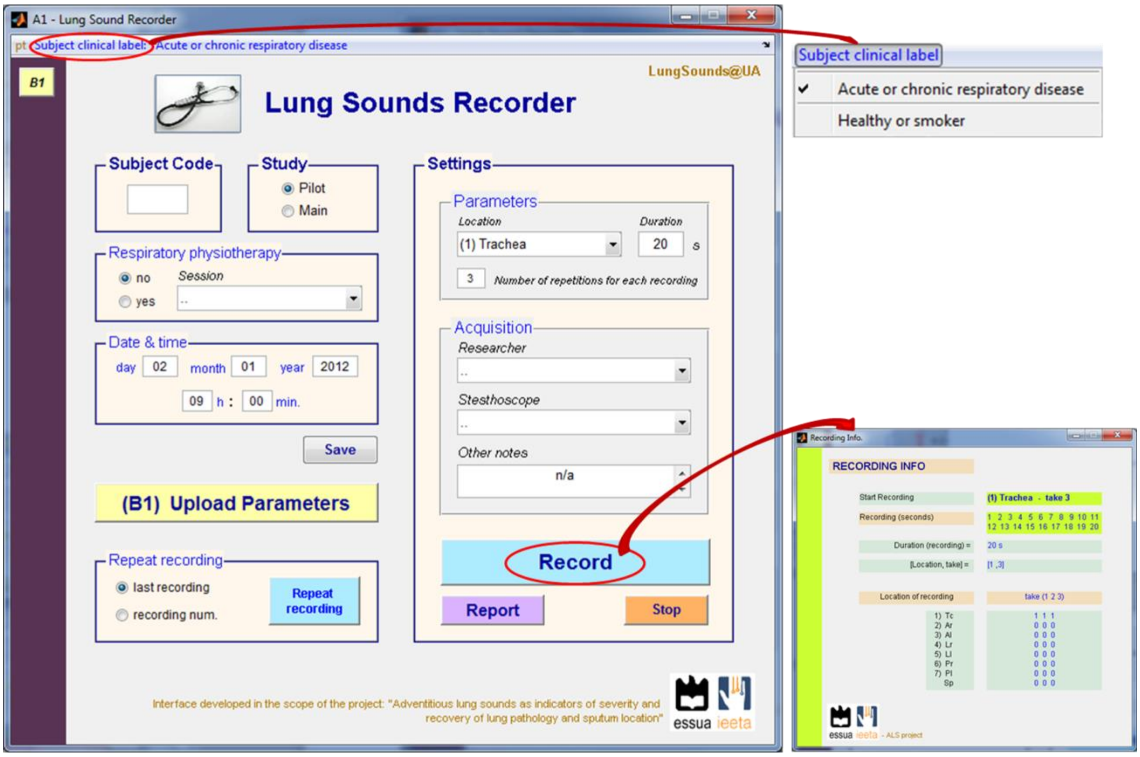


Figure 1: LungSounds@UA interface (A – interface structure – GUI composed by fourteen windows, with a hierarchy of five levels; B - Window A1 - Lung sounds recorder).

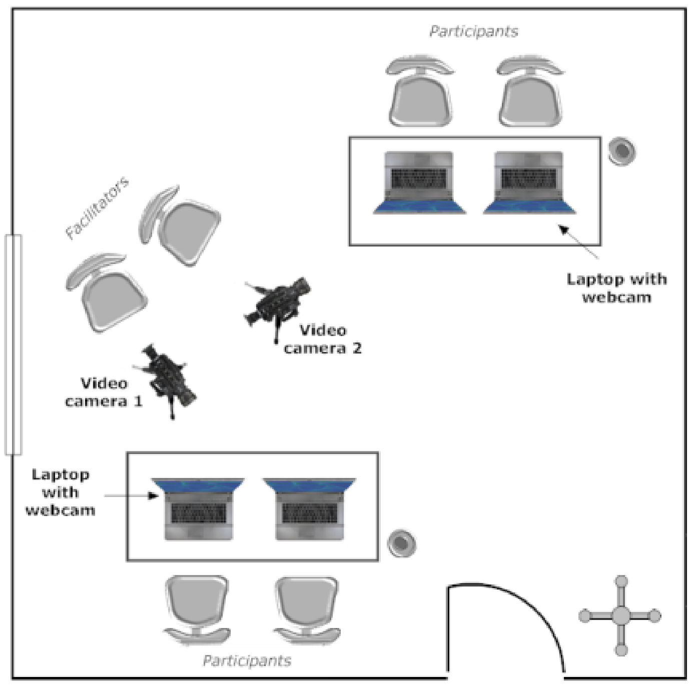


Figure 2: Room setup.

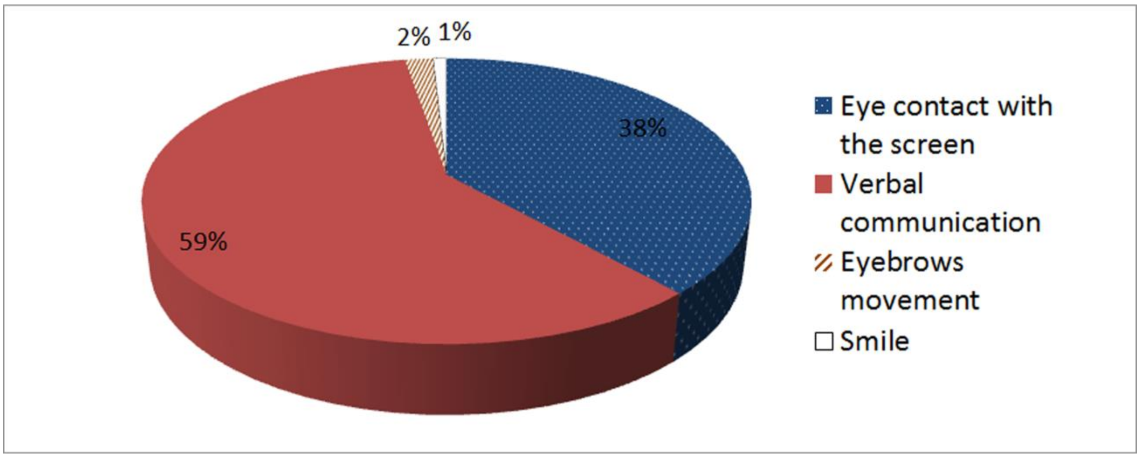


Figure 3: Percentage of users' frequency behaviour during the system interaction.

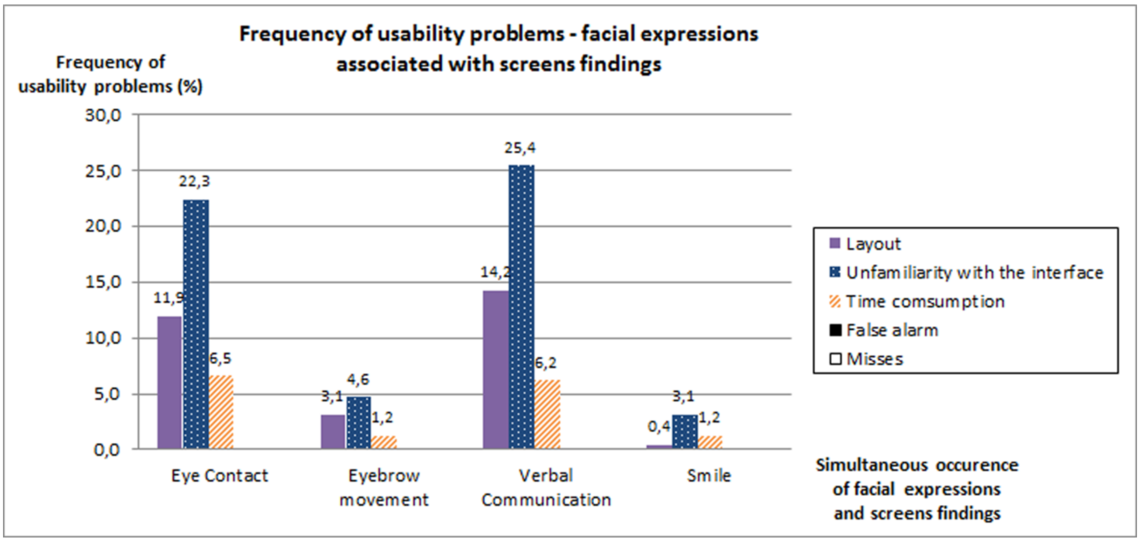


Figure 4: Frequency of usability problems when matching the facial with the computer screen videos.

LIST OF TABLES (with captions)

Table I: Categories of the facial expressions ethogram.

Categories	Description
Eye contact with the screen	The user directs the gaze to the screen, visibly concentrating on the screen, to read, search or understand something in the interface.
Eyebrows movement	The user raises an eyebrow or corrugates both as indicative of frustration or distaste for not understanding the interface or not finding what he/she is looking for.
Verbal communication	The user communicates deliberately and voluntarily with the facilitator using words and/or sentences, to clarify some doubts about the system.
Smile	Facial expression where the lips stretch back or move away slightly (mouth can be half opened) as indicative of agreement, comprehension and accomplishment.

Table II: Themes and sub-themes of the computer screen videos.

Themes & sub-themes	Description
Interface Problems	Problems inherent to the interface.
Layout/screen organization	*Problems related to the layout information in the interface, leading to mistakes or confusion.*
False alarms	*Problems generated when the interface claimed a non-existent problem.*
Time consumption	*Problems related with time-consuming tasks.*
Misses	*Problems generated when the interface did not alert for a specific problem.*
User Problems	Problems originated by users' actions.
Unfamiliarity with the system	*Problems related with the lack of familiarity with the interface.*

Table III: Users' behaviours when interacting with the system – through facial expressions analysis.

Categories	**Type**	**Mean ± SD**	**Minimum**	**Maximum**	**ICC**	**95% CI**
Eye contact with the screen	frequency	32 ± 9	15	45	0.9(9)	[0.99; 1]
	duration (s)	575 ±119	240	1065	0.94	[0.79; 0.99]
Verbal communication	frequency	20 ± 9	4	33	1	[1]
	duration (s)	886 ± 411	435	1458	0.98	[0.94; 1]
Eyebrows movement	frequency	2 ± 3	0	9	0.97	[0.89; 0.99]
	duration (s)	27 ± 42	0	143	0.91	[0.71; 0.98]
Smile	frequency	2 ± 1	0	4	0.98	[0.92; 0.99]
	duration (s)	13 ± 15	0	67	0.54	[0.48; 0.90]

SD – standard deviation
ICC - Intraclass correlation coefficient (2.1) – intra-observer reliability
CI – confidence intervals

Table IV: Interface and users' problems reported in the computer screen videos .

Problems		**Count**	**Problems**	**Count**	**Total**
Interface problems	*Layout/Screen organization*	26	Error messages	10	47
	False alarms	9	Warning messages	9	
	Time consumption	8	Inconsistencies	28	
	Misses	4			
User problems	*Unfamiliarity with interface*	17	Error messages	17	17
			Warning messages	0	
			Inconsistencies	0	
Total		64		64	64

Table V: Correlation between facial expressions and interface and user problems.

Interface and user problems	Facial expressions							
	Verbal Communication		Smile		Eye Contact		Eyebrow Movement	
	r_s	p	r_s	p	r_s	p	r_s	p
Layout	0.58	0.13	-0.24	0.56	0.64	0.09	0.65	0.08
Unfamiliarity with interface	0.77	0.03*	0.18	0.67	0.40	0.32	0.57	0.14
Time consumption	0.69	0.06	0.50	0.21	0.40	0.32	0.52	0.19

False alarms and Misses are not represented as their combination with facial expressions were not observed

* $p < 0.05$

Table VI: Example of matches found between facial expressions and screen problems.

Time	Screen problem	Facial expressions	Example
04:40	Layout/screen organization	Eye contact with the screen	Warning message appears, because the order to enter patient's blood pressure was inverse as it usually appears in clinical documents (i.e., systolic blood pressure/diastolic blood pressure), leading the participant to enter it wrong.
20:41 to 26:07	Time consumption	Eyebrows movement	A participant takes 5.15 minutes to enter the *haemogram, gasometry* and *biochemistry* reference values in the *clinical parameters*.
35:42	Unfamiliarity with interface	Verbal communication	Warning message appears because the participant did not enter the corridor length used for the six-minute walk test, not allowing the interface to calculate the distance walked by the patient.

Time is expressed in (minutes:seconds).

SUMMARY

Propose: Usability testing is essential to optimise information systems and to ensure its functionality to end users. Computer screen videos (CSV) and users' facial expressions videos (FEV) are widely recommended methods to evaluate systems performance. However, software that combines both methods is often expensive and non-accessible in the clinical research field. The Observer XT software is commonly use in this field to assess human behaviours with accuracy. Thus, this study aimed to report on the combination of CSV and FEV (analysed with the Observer XT) to evaluate a graphical user interface.

Methods: Eight physiotherapists with experience in the respiratory field and without any previous contact with the interface entered clinical information in the system while their screens and facial expressions were video recorded. One researcher, blinded to the evaluation, analysed the frequency and duration of a list of behaviours (ethogram) in the FEV using the specialized software, Noldus The Observer XT 10.5. Another researcher, also blinded to the evaluation, analysed the frequency and duration of usability problems found in the CSV. The CSV timelines were also matched with the coded facial expressions to verify possible combinations.

Results: The analysis of the FEV revealed that the category most frequently observed in users behaviour was the eye contact with the screen (32±9) and verbal communication was the one with the highest duration (14.8±6.9 minutes). Regarding the CSV, 64 problems, (47/64; 73%) related with the interface and (17/64; 27%) related with the user, were found. Through the combination of both methods, a total of 135 usability problems were identified. The majority were reported by users' verbal communication (45.8%) and eye contact with the screen (40.8%). The "false alarms" and "misses" did not cause quantifiable reactions in the users and

the facial expressions problems were mainly related with the lack of familiarity (55.4%) felt by users when interacting with the interface.

Conclusions: The findings encourage the combined use of computer screens and facial expressions videos to improve the assessment of users' interaction with the system, as it may increase the systems effectiveness and efficiency. These methods should be further explored with correlational studies and be combined with other usability tests, to increase the sensitivity of usability systems and inform improvements according to users' requirements.