Human factors methods in health information systems' design and evaluation: The road to success?
Peute, L.W.P.

Citation for published version (APA):
Peute, L. W. P. (2013). Human factors methods in health information systems’ design and evaluation: The road to success?
Human Factors Methods in Health Information Systems' Design and Evaluation

The Road to Success?

by Larry Wall & Kathleen Russell
from the book: The Road to Success is always under construction

A quitter never wins, a winner never quits.
Action speaks louder than words but not nearly as often. Failing to prepare, we prepare to fail.
Past failures are guideposts for future success.
There is no right way to do a wrong thing.
There can be no rainbow without a cloud and a storm.
If your dreams turn to dust… vacuum.
Seek joy in what you give not in what you get.
Procrastination is the thief of time.
Success comes in cans.
Failure comes in can'ts.
Anger is one letter short of danger,
greatest remedy for anger is delay.
2/3 of promotion is motion.
Of all the things you wear,
your expression is the most important.
Human Factors Methods in Health Information Systems’ Design and Evaluation

The Road to Success?
The research described in this PhD dissertation was carried out at the Department of Medical Informatics, Academic Medical Center, Amsterdam, the Netherlands.

Human Factors Methods in Health Information Systems’
Design and Evaluation
The Road to Success?
PhD dissertation, University of Amsterdam, Amsterdam, the Netherlands

Cover design by Linda W.P. Peute, Edited by Robin Deckert, Print service Ede
ISBN/EAN: 978-94-91602-12-2
Printed by Print Service Ede

Copyright © 2013, L.W.P. Peute, Amsterdam, the Netherlands.
All rights reserved. No part of this publication may be reproduced, stored, or transmitted in any way or by any means, without prior permission of the author.
Human Factors Methods in Health Information Systems’ Design and Evaluation

The Road to Success?

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit van Amsterdam op gezag van de Rector Magnificus prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel op woensdag 10 april, te 10:00 uur

door

Linda Wilma Petronella Peute

geboren te Haarlem
Promotiecommissie

Promotores: Prof. dr. M.W.M. Jaspers
            Prof. dr. P.J.M. Bakker

Overige leden: Prof. dr. C. Nøhr
               Prof. dr. D. Svanæs
               Prof. dr. A. Sturk
               Prof. dr. M.B. Vroom
               Prof. dr. J.H. Ravesloot
               Dr. J.E.W.C. van Gemert-Pijnen

Faculteit der Geneeskunde
The strongest oak of the forest is not the one that is protected from the storm and hidden from the sun. It's the one that stands in the open where it is compelled to struggle for its existence against the winds and rains and the scorching sun.

Napoleon Hill, 1883-1970

This thesis is dedicated to my sister Selma, for her courage and strength in undertaking the great challenges life has set upon her and her commitment and love for her children.
Contents

Chapter 1. Introduction 9

Chapter 2. The Significance of a Usability Evaluation for an Emerging Laboratory Physician Order Entry System 23


Chapter 4. Effectiveness of Retrospective and Concurrent Think Aloud in Formative Usability Testing; Which Method Performs Best? 65

Chapter 5. Classification and Prioritization of Usability Problems Using an Augmented Classification Scheme 85


Chapter 7. Reducing Clinicians’ Cognitive Workload by System Redesign; a Pre-post Think Aloud Usability Study. 125

Chapter 8. From Expert-driven Guidelines to a User-oriented Decision Support System; a Usability Comparison Study. 133

Chapter 9. Discussion - Summary of findings and Conclusion 153

Samenvatting - Dutch Summary 167

Acknowledgements 174

Curriculum Vitae and Portfolio 176
Chapter 1

General Introduction

Introduction

Decades of research on Health Information Systems’ (HIS) success has taught us numerous lessons and revealed factors associated with why these systems have failed or succeeded [1]. Nevertheless HIS success remains an on-going topic of discussion. Even though new publications keep shedding light on the complexity of HIS design and implementation, lessons learned have had only limited effect so far on accelerating the adoption of HIS [2, 3]. Estimations from 2010 showed that at least 40% of HIS fail and are abandoned in practice [4]. Considering that the publication of the IOM report ‘to err is human’ from 2001 promoted with urgency the need for incentives on implementing safe and effective HIS, this failure rate is alarming [5].

Despite these failures, implementation of HIS is considered to provide the healthcare community with promising tools for enhancing the efficiency, effectiveness and safety of (delivered) healthcare processes. Several studies have indeed shown positive impacts of HIS on health practices [6], but likewise case-studies continue to be published where HIS introduces unintended consequences and adverse effects to the health care practices [7-10]. These unintended consequences include unforeseen changes in work and communication patterns, changes in roles and responsibilities in the organizational structure and technology-induced errors due to poor usability of HIS (11-14). Usability in this context refers to the extent to which a system or system functionality can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use [15].

Rather than improving the quality of care, inadequately designed HIS may thus only marginally support their users, lead to additional work, potentially hamper (clinical) health practices, and even endanger patient safety [8, 10, 12]. As a result, ‘Usability’ has become a key factor influencing HIS implementation and adoption by healthcare professionals. How to design usable (easy to use) system
functionalities supporting effective, efficient and safe (clinical) task performance now is one of the primary aims in development of HIS. This focus on usability directly brings to view the challenges underlying adequate elicitation of users’ requirements, system development, and evaluation to ensure efficient, effective and safe use of HIS.

Understanding and optimizing how people interact with technology is the subject of the research field of Human Factors (HF) and usability engineering. Studies on information systems design and implementation in other safety-critical industries, such as aviation, have indicated that by applying methods from the HF/usability field in User-Centered Design (UCD) approaches, significant benefits and advances for the design of effective, efficient and safe HIS may likewise be gained [16]. UCD is a structured system development methodology that concentrates around prospective users of a system in order to create a system that meets their needs while considering their limitations, and preferences [17]. In a UCD approach, HF/usability engineering is employed throughout a system’s development process.

In this thesis, various case studies are described that provide sight on key lessons learned in applying HF/usability methods in UCD approach of HIS. The studies concern various categories of HIS, among which computerized physician laboratory and medication order entry systems, a clinical query tool to support analysis and benchmarking of Intensive Care registry data, a patient oriented web-based HIS and a computerized clinical decision support system. In these case studies, various HF/usability methods were applied in various phases of UCD of these healthcare information systems. In this way a broad perspective of the application and value of HF/usability methods in UCD of HIS is given, with the final aim to advance the knowledge of how to achieve implementation and adoption of HIS more successfully.

In summary, the aim of this thesis is twofold:

- To advance knowledge on the value of application of HF/usability methods in UCD of HIS and on the performance of usability evaluation methods on identification of usability problems of HIS.
- To extract key lessons learned to consider in UCD of HIS based on each of the case studies presented in this thesis.

In this chapter first the principles of the UCD approach of system development are described. Thereafter, the HF/usability methods applied in the case studies of this thesis are presented. Subsequently, the research questions addressed in this thesis, followed by an outline of the thesis are given.
The User-Centered Design process

UCD is both a philosophy as well as a system design approach that is grounded in continuous and central involvement of end-users, in which their needs and limitations are given extensive attention at each stage of the design process. The term ‘user-centered design’ originated in Donald Norman’s research laboratory at the University of California San Diego (UCSD) in the 1980s and became widely used after the publication of a co-authored book entitled: User-Centered System Design: New Perspectives on Human-Computer Interaction [18].

A UCD process therefore focuses on users through the planning, design, development and evaluation of a product. On the basis of the UCD, the international standards ISO 13407: Human-centred design process and its successor ISO 9241-210, 2010 define a general process for including user-centered activities throughout a system development life-cycle (figure 1). However, these standards do not specify exact methods supporting the conduct of the various phases of the defined multi-stage UCD process model. Nevertheless, UCD is considered as the basis for designing usable systems that increase the chances of successful system implementation and adoption [19].

In a way, UCD can be characterized as a problem solving process that not only requires designers to analyse and foresee how users are likely to use a product, but also to test the validity of their assumptions with regards to user behaviour in real world tests with actual users. Such testing is necessary as it is often very difficult for the system designers to understand intuitively how a first-time user would experience their product design, and what users’ learning curves may look like. The chief difference with other system design philosophies is that the UCD philosophy is focused on optimizing the system around how users can, want, or need to use the system, rather than forcing the users to change their behaviour to accommodate it.
The UCD process adopted from the ISO 9241-210:2010 standard incorporates the following four main activities:

1) To understand and specify the context of use; the analysis of user characteristics, user tasks and organizational, technical and physical environment of system implementation is at the primary focus of this phase.

2) To specify the user requirements; the identification of the user needs and specification of the functional and non-functional (among which usability) requirements for the system in relation to the intended context of use are at focus.

3) To produce design solutions; in this phase, the user interface of a system is defined and adapted to meet the user requirements, thereby taking into consideration the overall user experience. The focus is also on altering the system design solutions in response to user-centered evaluation and feedback and communicating recommendations for (re)design solutions to those persons responsible for their implementation. Formative usability evaluation can be performed by usability evaluation experts inspecting a user-interface of a system, or based on user testing that provide insight into the usability with or user perception on the produced prototype system. Formative evaluation may be iteratively applied in this phase of system development.

4) To evaluate design against requirements; in this phase it is determined if the requirements have been satisfied through summative end-user evaluations of a system. Summative end-user evaluation focuses on evaluation of a system through predefined measures, rather than on the diagnosis and correction of specific design problems, as in formative usability evaluation of systems.
Human Factors and Usability methods in Health Informatics

The Human factors (HF) domain in Health Informatics (HI) is defined as ‘the scientific discipline concerned with the understanding of interactions among humans ‘healthcare professionals, or patients’ and other elements of a system, and the profession that applies theories, principles, and methods to HIS design in order to optimize user well-being and overall system performance’ (definition adopted by the International Ergonomics Association in 2000). HF/usability engineering is the discipline focusing on the interactions between people and devices. The critical element in these interactions is the machine user-interface as depicted in figure 2.

![Human-machine interaction model adapted from Redmill and Rajan (1997).](image)

The human-machine interaction comprises how humans perceive information from a machine, and subsequently process this information. Based on the information processed, humans can decide on ensuing (control) actions (and translate them) to performance of potential physical actions through the user-interface of the machine. This is followed by a system’s processing and reaction on this input, after which the machine provides feedback to the user about the effect of their actions via the display in the user interface. The cycle then starts again, so that humans are able to perceive (and assess) the
changes in system state through the user interface in to allow them to plan new information processing actions and thus react accordingly.

Principles and methods from Human factors/usability engineering are applied to design the machine-user interfaces. These user-interfaces include all components with which users interact while performing their tasks in the system.

In this thesis, the focus is on the methods that HF scientists and engineers may apply during UCD of HIS with a main focus on usability methods for evaluating HIS and determining potential sources of usability problems in the user-system work practices. Usability problems are defined as ‘obstacles’ to the effective and efficient accomplishment of a specified task by a specified user while he interacts with a system through its interface. HF/usability engineers have a great number of usability methods available which they can apply for formative (to provide input in new (re)design efforts) and summative (to test the usability of a final system) system evaluation during UCD of HIS. Since the 1990s, these methods have been steadily introduced into the HF/usability community and since then undergone a variety of modifications to their approach and objectives [20-23]. Two general classes of formal usability evaluation methods can be distinguished: usability inspection methods and usability testing methods. During usability inspection, usability evaluators assess a (prototype) system design on potential usability problems present in its user-interface. Inspection methods applied within the case studies of this thesis are the Heuristic evaluation (HE) and Cognitive walkthrough (CW) [24, 25]. In contrast to inspection methods, usability testing involves representative end-users of a system who are (systematically) observed under controlled conditions to determine the user-experience in using the (prototype) system. The most often used method in usability testing is the Think Aloud method [26-28].

Both classes of evaluation methods can be performed for revealing usability problems and inform the design of a system from an early UCD phase on. Usability evaluators have to choose methods best suited for producing informative results in relation to the system functionalities, user population, user tasks and context of use. Research comparing and contrasting the results of usability inspection and usability testing methods in uncovering usability problems of interactive computer systems has mainly be performed in domains other than healthcare. These studies have focused on comparison of usability evaluation methods on their detection rate in numbers of usability problems.

Yet, each method’s value in detecting different types of usability problems in a system’s design may differ depending on its context of use. Since health care is characterized by its intricate organisational environment, variety of HIS (clinical) users, and safety-critical work processes, the ‘context of use’ may impact the utility and performance of usability methods [29]. Since the resources and time for conducting system usability evaluations are often limited, it is worthwhile to know whether think
aloud usability testing provides better sight on the types of usability issues leading to medication ordering errors in CPOE system, than a cognitive walkthrough. Knowledge on which usability evaluation methods are best suited for revealing different types of usability issues in a certain type of system used in a certain healthcare context is relevant to choose among methods. Therefore, in this thesis, comparison studies of usability evaluation methods focus on their performance in identification of different types of usability problems concerning different types of HIS.

Complete and consistent classification of human-computer interaction problems revealed through usability evaluation is essential for high quality problem reporting. Classification not only supports consistent and clear communication of usability problems to system designers but aids in identifying trends and patterns across problem sets of usability evaluation case studies. These studies have classified usability problems in different ways, including bottom up classification, based on severity of problems or on high level heuristic principles. Bottom-up classifications focus on analysing the content of verbal data from usability tests and groups categories of usability issues emerging from this raw data. These categories mostly reflect user-interface flaws concerning the layout and visibility of information on a screen. Severity classifications were introduced by Nielsen and focus on assessing usability problems severity commonly based on their frequency, persistence and their effect on users’ task performance while they test the system [30]. Heuristics evaluation principles are used for high level aggregation and classification of usability problems encountered by usability experts in inspection of a system interface. However, above mentioned classification methods all lack a formal approach for clustering usability problems and may become biased by subjective judgement of evaluators. Bottom-up classification may produce different grouping of usability data by different evaluators and lead to inconsistent and incomplete classification of problems. Heuristic principles are used for identification of usability problems on an aggregated level rather than for their classification. As these high-level principles are not mutually exclusive, they hamper consistent classification of usability problems as the same problem may be ‘classified’ by multiple heuristics. Though classification of usability problems based on their severity does support prioritization of problems in a system (re)design process it does not address the characteristics of the usability issue nor its potential cause. A comprehensive classification of usability problems to guide complete, consistent and accurate problem reporting seems needed to support characterization of HIS usability problems and properly guide system redesign efforts. The User Action classification framework from the field of Human Computer Interaction, may address these issues [31]. The application of this framework will result in more meaningful problem clusters since it groups usability problems according to different phases of user interaction in which they occur (i.e. the planning, translation, physical action, and assessment phase, see figure 2. This will aid the identification of trends and patterns across problem sets of the usability evaluation case studies and facilitate the identification of usability problems that share common characteristics and that may impact user effectiveness, efficiency or patient safety.
The main HF/usability methods that are addressed in this thesis are:

Ethnographic observation – Ethnography is a form of research focused on the qualitative properties of observation. Ethnographic studies focus on compiling data about human activities. Researchers are allowed to closely observe and study a particular group in order to better understand the customs, work processes and habits of the group under study.

Interviews/ Focus groups – These methods focus on involving people while asking them about their perceptions, opinions, beliefs, and attitudes towards a mock-up or (prototype) HIS. In focus groups questions are asked in an interactive group setting where participants are free to talk with other group members. During interviews secluded sessions are being held providing either open ended or closed ended questions or these combined to a participant.

(Historical) document analysis – Document analysis is a form of qualitative research in which documents are interpreted by the researcher to give voice and meaning around an assessment topic. Analyzing documents incorporates analyzing and coding content into themes similar to how focus group or interview transcripts are analyzed. Evidence obtained form multiple sources is weighted and correlated. Another important aspect of document analysis is evaluating the provenance, purpose, motivation and constraints of the information contained in the documents within a particular (historical) context.

Usability Heuristics – a usability inspection method by which two or more usability specialists/evaluators inspect the interface of a system by reviewing its compliance to a set of established usability principles, the so-called heuristics [25].

Cognitive walkthrough – a usability inspection method focused on learning by exploration known as system learnability. One or more evaluators work through a series of tasks while they use a mock up or a prototype system and ask a set of questions from the perspective of the user population and context of use. It is based on an explicit analysis of all user goals and corresponding task action-sequences that need be performed by end-users in order to assess potential risks to end-user task performance based on pre-defined task scenarios [24].

Think Aloud – The Think Aloud method stems from cognitive psychology and is considered to be a golden standard in usability user-testing. The think-aloud method was introduced in the usability field by Clayton Lewis, and further developed based on the techniques of protocol analysis by Ericsson and Simon [28]. The Think Aloud method involves a sample of approximately 5 to 8 participants thinking aloud (verbalizing their thoughts on what actions to perform and corresponding feelings) as they are performing a set of specified tasks (scenarios). Test sessions are audio and video recorded to support accurate analysis of the usability problems experienced by the participants as portrayed in the acquired verbal protocols. The purpose of this method is to gain sight on the implicit cognitive user actions.
during task performance. There exist two different experimental procedures for obtaining think-aloud protocols. The first one is a procedure by which think aloud protocols are obtained during task performance: the concurrent *think-aloud protocol*. The second procedure gathers think aloud protocols after task performance: the *retrospective think-aloud protocol*.

*Mock-up* - Mock-ups are used by system designers mainly to acquire feedback from users about designs and design ideas early in the design process. Mock-ups are ‘very early prototypes’ made of cardboard or otherwise low-fidelity materials. The user, aided by the designer, may test the mock-up (imagining that it works) and thus provide valuable feedback about functionality and understanding of the basic design idea.

## Research questions and outline of this thesis

The *research questions* that are addressed in this thesis are:

**Question 1.** How can HF methods contribute to investigating the human, social and organizational issues that influence HIS implementation?

**Question 2.** How do usability methods compare regarding their potential to identify different types of usability problems concerning (prototype) HIS and what are ensuing merits of applying HF/usability methods in UCD of HIS?

**Question 3.** How should usability problem descriptions be classified so that they can be effectively reported on?

**Question 4.** How does Think Aloud usability testing contribute to redesign of HIS in terms of usability problem types detected, efficiency and effectiveness of users’ task performance before and after redesign?

**Question 5.** What are the key lessons learned of application of HF/usability methods concerning the UCD phases of HIS?

The research questions above are discussed in the following case-studies presented in chapters 2 through 8 in this thesis, as outlined in the following paragraphs.

**Chapter 2** describes a *longitudinal analysis of an implementation process* of a Computerized Physician Order Entry system for Laboratory ordering (CPOE-L) in an academic hospital setting. In
Chapter 1

This study the contextual human, social and organizational issues and their interrelations are analyzed in terms of their effect on the system failure after its implementation by use of HF methods. A contextual model to support the interpretation of the qualitative data from focus groups, interviews and historical document analysis was developed. Based on this model, recommendations for other CPOE-L implementations are provided.

Addressed research question: question 1

Key lessons learned - Question 5 - extracted for UCD phase: 1. understanding and specifying context of use.

Chapter 3 describes the results of a formative usability evaluation study of the prototype CPOE-L system, as part of the producing design solutions process of the final system studied after its implementation (chapter 2). The usability problems of the CPOE-L system are assessed in terms of their type and their effect on healthcare providers’ task ordering efficiency and effectiveness (errors and omissions in ordering or cancelled orders). In addition comparison is made between the results of the applied HF/ usability methods: the Cognitive Walkthrough and Think Aloud method and their merits in evaluating prototype systems are discussed.

Addressed research question(s): question 2

Key lessons learned - Question 5 - extracted for UCD phase: 3. producing design solutions.

In Chapter 4 an existing web-based Physician Data Query Tool to provide clinicians the opportunity to query National Intensive Care Evaluation (NICE) data from the NICE Registry, is evaluated on its usability [31]. Two variants of the Think Aloud HF/ usability testing method, the Concurrent (CTA) and the Retrospective (RTA) Think Aloud are applied. Focus of the study is on the comparison of their performance in detecting (different types of) usability problems. Also, their potential merits in system (re)design are discussed in this chapter. The formative usability evaluation study provided input to the query tool’s redesign addressed in chapter 7 of this thesis.

Addressed research question(s): questions 2

Key lessons learned - Question 5 - extracted for UCD phase: 3. producing design solutions.

In Chapter 5 an existing Computerized Physician Order Entry system (CPOE) for medication ordering was evaluated on its usability by use of the Cognitive walkthrough and Think aloud methods. The study proposes a framework for the classification and prioritization of usability problems to support informed (re)design of evaluated (prototype) systems. An adapted version of framework is subsequently applied to classify the usability problems revealed by the Cognitive Walkthrough and Think aloud usability evaluation of the CPOE system. The proposed framework augments to an
existing classification framework, the User-Action Framework, with usability problem severity rating and a qualifyable effect rating of the usability problems on healthcare professionals’ task performance and task outcome.

Addressed research question(s): question 3

Key lessons learned - Question 5 - extracted for UCD phase: 3. producing design solutions.

In Chapter 6 a website for education of Dutch childhood cancer survivors is developed based on a user-centered design approach specific for health websites (the Website Developmental Model for the Healthcare Consumer). In this study the applicability of a concise subset of the WDMHC is explored. The human factor methods applied in this study are information needs analysis to inform mock-up creation of the website and focus groups in order to develop a website prototype. In the usability evaluation of the website, the results of a Heuristic evaluation and the Think Aloud, method are compared.

Addressed research question(s): questions 2 and 4

Key lessons learned - Question 5 - extracted for UCD phase: 2. Specifying the user requirements, 3. Producing design solutions, and 4. Evaluating the design against requirements.

In Chapter 7 A pre-post usability evaluation study of the intensive care registry Physician Data Query Tool is described. Usability problems revealed by the concurrent Think Aloud method before and after redesign of the Query Tool are classified according to the User-Action Framework (also described in chapter 5) and compared. Further, users’ task effectiveness and efficiency is compared before and after redesign of the Query Tool. These measures give insight into physicians cognitive work load in terms of user task efficiency and task effectiveness and provide sight on the merits of a UCD approach in redesign of the tool.

Addressed research question(s): question 3 and 4

Key lessons learned - Question 5 - extracted for UCD phase: 3. Producing design solutions, and 4. Evaluating the design against requirements.

In chapter 8 a Clinical Decision Support System (CDSS) supporting the retrieval of childhood cancer survivor’s follow-up screening procedures by health professionals is developed based on a user-centered design approach. The usability of a paper-based guideline defining childhood cancer survivor’s follow-up screening procedures was first analysed by having healthcare professionals think aloud while they retrieved the screening procedure for simulated patient scenarios. The think aloud verbal protocol analysis and usability problems revealed during use of the paper-based guideline informed the design of the CDSS. The usability evaluation results of the designed CDSS were compared to those of the paper-based guideline in terms of the type of usability problems experienced.
and the efficiency and effectiveness of retrieval of childhood cancer survivor’s follow-up screening procedures by health professionals participating in the study.

Addressed research question(s): question 4

Key lessons learned - Question 5 - extracted for UCD phase: 4. evaluating the design against requirements.

This thesis ends with chapter 9, in which the results of the case-studies presented in this thesis are summarized and key lessons learned in the application of HF/usability methods are presented.

REFERENCES

15. ISO 9241-11 1998, 'Guidance on usability', International Organization for Standardization,


Chapter 2

The Significance of a Usability Evaluation of an Emerging Laboratory Order Entry System

Linda W.P. Peute, Monique W.M. Jaspers


Abstract

Objectives: To assess the usability of an emerging POE system, OM/Lab, for the computer-supported ordering of laboratory tests. We were more specifically interested in the relation of the usability problems detected in the user testing sessions with the order behaviour in terms of efficiency and errors in ordering. Methods: A cognitive walkthrough of the OM/Lab system was conducted by two analysts using four real-life scenarios for ordering laboratory tests, which were reviewed for comprehensiveness by an expert clinician. Thereafter, the OM/Lab system was evaluated on its usability in testing sessions with seven potential end users of the system performing these same four scenarios. The results of these end-user testing sessions were used to analyze the effect of usability flaws on the quality of ordering in terms of omissions, errors in orders and cancelled orders. Results: The analyses revealed a total of 33 usability problems, of which 25 problems were revealed both by the cognitive walkthrough and in the end-user sessions. These 25 usability problems indeed led to inefficient order behaviour, omissions and errors in orders and even to cancelled orders. Discussion: Our results revealed that the OM/Lab system suffered from a high number of usability flaws. The interface design flaws were, among other things, related to misallocation of buttons on the screen, incomprehensibility of button labels and feedback containing no relevant information to the user about the cause of errors made and consequences of a user’s action. Additionally, our user test session results indicated that the OM/Lab system also suffered from user interaction problems of a more socio-technical nature. These sessions revealed, among other things, that the more specific action sequences to be executed within the ordering model of the OM/Lab system did not correspond to the daily working routines of end-users and that the grouping of laboratory tests within clusters did not match the paper-based order forms. The seemingly negative effects of these usability flaws on the quality of ordering and the inefficiency of work processes during the pilot implementation finally lead to withdrawal of the OM/Lab system from clinical practice. Though the system implementation failed, our usability study brought research on (re)designing and evaluating clinical computer applications at a higher status in our institution. It is now recognized that usability evaluation studies that will support good quality of clinical practice are highly important.

Keywords: Physician Order Entry Systems; Usability Evaluation; Information Systems.
Introduction

The potential benefits of Computerized Physician Order Entry (CPOE) systems in enhancing the quality and efficiency of health care delivery are well recognized. The recognition of these benefits by health care organizations has led to numerous attempts of CPOE implementations of which however only few have succeeded [1-6]. CPOE implementations that failed their expectations have nevertheless shed light on the causes underlying these failures, among which clinicians’ discontentment with the CPOE system has been proven to be a main factor [7]. Clinicians’ adoption and usage of CPOE systems has been hampered by their poor design. A thorough understanding of the fundamental principles behind human activities is required to design systems that are usable in supporting these activities [8, 9]. Usability is thus critical to successful CPOE implementation and adoption by clinicians. As defined by the International Standards Organization, usability refers to the effectiveness, efficiency and satisfaction with which users can achieve specific sets of tasks in a particular environment [10]. User effectiveness relates to accuracy and completeness with which specified users can achieve specified goals in particular environments. User efficiency refers to the resources expended in relation to the accuracy and completeness of goals achieved. User satisfaction relates to the comfort and acceptability of the system to its users. Usability studies on CPOE have however found these factors to be highly interrelated. These studies have shown that user satisfaction with CPOE system is highly correlated with ease of use of CPOE systems (effectiveness) and with productivity and system response time (efficiency) [11, 12]. Hence, physicians would be more willing to accept CPOE if the system is fast and easy to use, that is, if the user interface behaves consistently and meets the requirements of the physicians [11, 13]. Ease of CPOE system use is influenced by the ease with which its functions can be learned and memorized by users. Since the functionality of a CPOE system is made available through its user interface, its design has a huge influence on its usability [14-18]. A cumbersome user interface may not only slow down users but introduce a new class of cognitively based errors into the process [19]. Inconveniently displayed, not easily noticeable or obscured data may cause users to make erroneous decisions or to fail to take corrective actions [20]. For the human computer interaction being effective, the tasks and procedures that the user may perform with the system need to be structured in a logical and consistent manner. This means that the system functions should correspond one to one with the goals that the user sets him self in performing his tasks and the order in which the user wants to attain these goals. Furthermore, the number of actions that a user has to perform in accomplishing a task should be minimized. Also, the presentation order of information should match the order in which a user processes this information [21]. The representational structure of displayed information indeed determines the level of cognitive complexity any given task will require [22-26] and thus information should be displayed in such a way...
that human perception of it is promoted. Otherwise, inefficient order behaviour, omissions and errors in orders may occur [27].

A recent trend in evaluating health care applications on their usability is the application of cognitive analysis methods of human-computer interaction with the aim to enhance a system’s user interface. Besides, in these studies, the relation between hypothesized, actual usability problems and the occurrence and frequency of errors that result from suboptimal user interfaces is explored [19, 26]. These studies have shown that cognitive analysis methods are not only suited to characterize user-interaction flaws in a system but also to identify (the opportunities for) errors attributable to these usability problems. In evaluating CPOE systems, use of these methods may focus on analyzing the ease with which physicians can order tests, on analyzing to what extent the system functions are structured in a, for the end-user, logical and consistent manner, on analyzing whether system terminology follows terminology used in daily practice and on analyzing whether the system provides adequate feedback in case of errors [19, 20].

A variety of cognitive analysis methods and techniques are used to evaluate a system’s design, among which the cognitive walkthrough and the think aloud technique [28, 29]. Cognitive walkthrough is a usability inspection method, based on evaluation of computer systems by an analyst, whereas the think aloud technique has been applied in usability testing sessions, based on system evaluations by true or potential end-users of a system. Cognitive walkthrough applies general principles from cognitive psychology and is to be conducted by an analyst who simulates the cognitive processes and user actions needed to accomplish a certain computer-supported task [30, 31]. The cognitive walkthrough method is conducted by use of predefined scenarios and focuses on the ease with which action-sequences can be executed in the system. The think aloud method formally belongs to the verbal report methods and stems from the field of cognitive psychology. It was specifically developed to gather information on the cognitive behaviour of humans performing tasks. This technique has proven to be of high value for evaluating a system’s design on usability flaws and is therefore frequently used to gather information about a system’s usability in testing computer systems with potential end-users. Users ‘interact’ with a system or prototype interface during recorded usability sessions according to a predetermined set of scenarios while verbalizing their thoughts [32].

Several studies have found that approximately 1/3 of a system’s usability problems are identified by all methods. Different methods have a unique complementary value in discovering usability flaws [33, 34]. Usability inspection methods focus on how intuitive an application is from a general human cognitive perspective, whereas usability testing methods focus on whether the application is compatible with user’s activity and with the cognitive aspects of that activity in particular [35, 36]. It is therefore highly recommended to combine both usability inspection and testing methods in a system usability study.
The study presented in this paper is concerned with the usability evaluation of a working prototype CPOE system for ordering laboratory tests (OM/Lab) that in a pilot was to be implemented at the outpatient neurology clinic and the outpatient laboratory of the Academic Medical Center (AMC) – Amsterdam. If the OM/Lab pilot implementation would prove to be a success, the system would be implemented at other clinical departments of the AMC.

We were asked to evaluate the prototype OM/Lab system. The main objective of this evaluation was to assess whether the OM/Lab system complied with the user requirements and to reveal potential usability flaws in the system. Besides these aims we were also interested in the relation of the usability problems detected in the user testing sessions with end-users’ order behaviour in terms of omissions, factual errors in orders and cancelled orders.

OM/Lab system background

The planning for the development and the implementation of the OM/Lab system followed a standard software design cycle. From both departments to be concerned in the pilot implementation of OM/Lab (the outpatient laboratory and the outpatient neurology clinic) the head of the department was involved in the project as representative and as OM/Lab acceptant. During the software prototype phase the request to evaluate the OM/Lab system was brought in for the choice of the 'one-page' human computer interaction model implemented in OM/Lab. Use of this 'one-page' model, offering users a total overview of the entire tasks to be executed in the system, proved to lead to better understanding by users of their progress in the system as opposed to a ‘page browsing interaction’ model [37]. This ‘one-page’ model had therefore been much discussed by the OM/Lab project team as an interaction model not only for ordering laboratory tests, but also for ordering of other tests such as radiology images. It was only after the prototype phase that the project implementation team asked us to evaluate the OM/Lab prototype system. So we were not involved in any earlier phase of OM/Lab development.

The ‘one-page’ user interface of the OM/Lab system (see Figure 1) consists of a six-step model with six bars containing drop down panes, one for each of the following: (1) Entering information identifying the requester, (2) Selecting laboratory protocol based order sets, (3) Ordering individual laboratory tests, (4) Entering patient data, (5) Entering a research question and (6) Sending the order. In step 1 the requesting clinician enters personal information, such as his name, and telephone number. Step 2 ‘Selecting a protocol’ offers the user the opportunity to request laboratory tests by choosing a protocol based order set for a specific neurological disorder. For example the protocol 'TIA' Trans Ischemic Attack defines 14 laboratory tests that should be ordered by the physician to assure if the patient has had a TIA. By choosing the TIA protocol the user does not have to select all these 14 tests individually by searching for each of these tests in step 3 and selecting these tests one by one by clicking on it with the mouse. Instead, after the selection of a specific protocol, tests are automatically selected as a set. After the user has chosen a protocol, he may add more tests or deselect tests from the
protocol pre-selected test set. As described earlier, ordering tests by predefined order sets (protocols) may save time and may improve the quality of ordering. The OM/Lab order entry prototype offered the following 5 protocols: Cardio Vascular Accident (CVA), Young Stroke (YS), Trans Ischemic Attack (TIA), Poly Neuropathy (PN) and Carpal Tunnel Syndrome (CTS).

In Step 3 ‘Ordering laboratory tests’ users may select the required laboratory tests one by one (see Figure 1). Users may otherwise choose between different kinds of materials; blood, urine portion, collected urine and liquor, after which the system displays the relevant tests accordingly. These sheets are further subdivided in several tab sheets: general, hormones, vitamins, toxicology. Within a sheet the tests are alphabetically ordered according to types of lab tests such as ‘chemistry’, ‘haematology’, ‘coagulation’, ‘blood bank’ or ‘purchase tube’. Tests which are not commonly requested are only made visible after the user has clicked on the ‘extra tests’ button. In step 3 a clinician may specify a specific time and date for the order to be carried out and may indicate whether the order is/should be carried out by the clinician himself/the department nurse or by the laboratory personnel. Step 4 'Enter clinical information' is implemented as a type-in field in which a clinician may enter additional information concerning the patient. A user is however not allowed to skip step 4 if the laboratory personnel needs additional patient information concerning the analysis of a specific test ordered by the physician in step 3. For example, laboratory personnel may need to know whether a patient is a diabetic or not in analyzing a specific test. Step 5 'Enter research question' will only be activated by the system after a specific test is ordered in step 3 for which laboratory personnel needs to know the physician’s reason for selecting that particular test, such as requests for not frequently ordered and expensive tests. Step 6 'Sending the order' provides an overview of all tests as selected in step 3. In this step, physicians may review their selection of lab tests after which they can accept and send the order to the laboratory. Once the order is sent, a form is printed with a code, to be scanned at the laboratory. The patient receives this form, and brings the form to the laboratory.

**Methods**

In evaluating the “one-page” human computer interaction model of the OM/Lab system two analysts conducted a cognitive walkthrough. We specifically chose to apply this usability inspection method as it focuses on the identification of goals and of subsequent actions to be performed to accomplish tasks within the system, thereby analysing whether the users’ background knowledge would enable completion of these tasks. As the focus of our study was on evaluating whether the one-page model was intuitive in use from a human cognitive perception, we felt that applying the cognitive walkthrough method would be more informative than the application of a heuristic evaluation method, in which the focus is on evaluating a system’s design using a set of general heuristics or principles.
The Significance of a Usability Evaluation

Figure 1. Screen dump of the OM/Lab system's 'one-page' order entry screen (in Dutch).

The cognitive walkthrough consists of a preparation phase and an execution phase. In the preparation phase, the analyst constructs a correct action sequence for the task to be computer-supported. The user-background of the end-users is then defined on the basis of their computer experience. Thereafter, the execution phase is started. For each action in the action sequence the analyst answers four questions: 1) Will the user try to achieve the right effect (goal setting), 2) Will the user notice that the correct action is available, 3) Will the user associate the correct action with the effect trying to be achieved, and 4) If the correct action is performed, will the user notice that progress is being made towards completion of the task? With the description of the user’s computer experience in mind, the evaluator analyses whether each question leads to success or failure by stepping through the system. After all actions have been evaluated, the usability problems are merged into a list.

In this study, the cognitive walkthrough was likewise used to define a framework of sub-goals and related action-sequences per sub-goal to be performed in the system. An illustration of a goal and subsequent actions to complete the task of selecting non-frequently ordered laboratory tests in the OM/Lab system is shown in Table 1. The analysts independently used this framework to identify potential usability problems related to the action-sequences to be executed in the OM/Lab system. Each of these potential usability problems was then coded according to ‘Norman’s theory of action’ into ‘goal problems’ (the user tries to accomplish the wrong thing) or ‘action problems’ (the user
would like to perform the correct action but does not know how) [38]. Subsequently all potential usability problems were discussed and given a severity classification, ranging from 1 (minor problem) to 5 (usability catastrophe). Our purpose of rating potential usability problems in terms of severity was to help us communicate prioritized usability flaws that would need fixing by the system designers.

Table 1. Example of a goal-action sequence in the OM/Lab system.

<table>
<thead>
<tr>
<th>System’s state: Subtask (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Goal</strong></td>
</tr>
<tr>
<td><strong>Goal Action</strong></td>
</tr>
<tr>
<td><strong>Action</strong></td>
</tr>
<tr>
<td><strong>System response</strong></td>
</tr>
</tbody>
</table>

**Coded Potential Disparities:**

| Goal problem 1) | The function of the ‘extra tests’ button is not clearly defined in the system |
| Action problem 2) | End users might not notice the button ‘extra tests’ due to its invisibility on the screen. |
| Action problem 3) | After a user has clicked on the ‘extra tests’ button, the tests added are not visible on the screen. |

A think aloud usability test session starts with a preparation phase to familiarize the user with thinking aloud. Thereafter, the actual user test sessions start in which users then verbalize their thoughts in performing the task supported by the system. These sessions are preferably videotaped [31]. After all user sessions have been completed, the evaluator produces a complete list of the usability problems detected through review of the video-tapes and verbal protocols. In this study, these user sessions took place in the clinical environment of the potential end users and lasted 40 to 50 minutes. Sessions were videotaped; screen sequences of the OM/Lab system were captured for subsequent analysis. The verbal protocols of end users were transcribed and physicians’ utterances were analysed for usability problems encountered in these testing sessions. Besides, the user sessions were also analyzed to characterize the ways in which these usability problems negatively influenced the clinicians’ task behaviour in terms of omissions in orders, errors in orders or cancelled orders.
Subjects and materials

Four real life clinical scenarios, which were reviewed by an expert clinician for comprehensiveness, were constructed and were used in both the cognitive walkthrough and end-user testing sessions to assess the usability of the OM/Lab system. The scenarios were designed in such a manner that they overall covered all action-sequences to potentially be followed by end-users ordering tests by use of OM/Lab. Standard practice laboratory protocols were used as input for three of the four scenarios. Seven potential end users of the OM/Lab order entry system, three neurologists and four neurologists in training, were asked to perform these four simulated laboratory-ordering tasks supported by the OM/Lab system. The test users differed in age and computer experience. They all worked with the paper-based laboratory ordering forms in use at the outpatient neurology clinic. None of the end users had prior experience with a system for ordering laboratory tests; they had not been involved in the OM/Lab system requirement analysis phase, nor in its development phase.

Results

Cognitive walkthrough: potential usability problems

The cognitive walkthrough identified six subtasks and 29 associated actions that needed to be executed by a user in order to enter and send an order. The six subtasks included: (1) Check name and phone number of requester, (2) Select a protocol, (3) Select laboratory tests, (4) Enter patient clinical data, (5) Enter reason for ordering laboratory tests, and (6) Send the order to the outpatient clinical laboratory. These subtasks represent the six states of the 6-step model of the OM/Lab system (see Figure 1). For example, an action associated to the subtask of selecting laboratory tests is: click on the drop down menu ‘material’. In executing a relatively simple order, that is, ordering one laboratory test with the help of the system, only four of these six subtasks are to be performed and 13 actions. In executing a relatively complex order all six subtasks of the system need to be performed and all 29 associated actions, of which some are to be performed several times.

The in-depth cognitive walkthrough analysis of the OM/Lab user interface revealed a total of 25 potential usability problems associated with actions to be performed in executing the 6-step model. Table 2 provides the results if the cognitive walkthrough analysis, the severity of problems detected and the number of encounters by end-users in the Think Aloud user testing.
Table 2. Results of the cognitive walkthrough (verified by the user testing sessions).

<table>
<thead>
<tr>
<th>CW Code</th>
<th>Characterization potential usability problem</th>
<th>Severity</th>
<th># Users</th>
<th># Encounters.</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. A.202</td>
<td>Colour and location of the term ‘change’ on a bar may confuse users</td>
<td>1</td>
<td>4 / 7 *</td>
<td>Mentioned 5 times</td>
<td>Inefficient order behaviour</td>
</tr>
<tr>
<td>1. A.203</td>
<td>The required action sequence for opening step 1 might be difficult for end-users to comprehend.</td>
<td>2</td>
<td>3 / 7</td>
<td>3 / 28 **</td>
<td>Inefficient order behaviour</td>
</tr>
<tr>
<td>1. C.203</td>
<td>An inexperienced user may not know if nor how he can change the telephone number.</td>
<td>1</td>
<td>2 / 7</td>
<td>3 / 7</td>
<td>Errors of omission</td>
</tr>
<tr>
<td>2. A.203</td>
<td>Inflexibility of selecting more than one protocol. Cancellation of a selected protocol set in the system is not possible. This may lead to inefficient order behaviour.</td>
<td>3</td>
<td>3 / 7</td>
<td>4 / 7</td>
<td>Errors of omission Cancelled orders</td>
</tr>
<tr>
<td>3. A.201</td>
<td>Visibility and location of dropdown pane for selecting specific laboratory material.</td>
<td>2</td>
<td>2 / 7</td>
<td>3 / 28</td>
<td>Errors of omission</td>
</tr>
<tr>
<td>3. B.102</td>
<td>The understandability and naming of labels of laboratory tests may result in inefficient search strategies.</td>
<td>4</td>
<td>4 / 7</td>
<td>Mentioned 6 times</td>
<td>Factual errors Errors of omission</td>
</tr>
<tr>
<td>3. B.201</td>
<td>The chosen subgroups for laboratory tests on the screen may lead to difficulties in locating tests as it is in a substantially different format as the paper laboratory form.</td>
<td>2</td>
<td>2 / 7</td>
<td>13 / 28</td>
<td>Inefficient order behaviour Factual errors</td>
</tr>
<tr>
<td>3. B.202</td>
<td>Locating and memorizing particular lab tests might be difficult for end-users due to their invisibility, by use of the scrollbar.</td>
<td>4</td>
<td>7 / 7</td>
<td>298 / 28</td>
<td>Inefficient order behaviour</td>
</tr>
<tr>
<td>3. D.202</td>
<td>The location of labels on tabloids and different naming of tabloids might be difficult for end-users to notice when searching for tests in the system.</td>
<td>2</td>
<td>3 / 7</td>
<td>12 / 28</td>
<td>Inefficient order behaviour</td>
</tr>
<tr>
<td>3. E.103</td>
<td>The function of ‘Extra Tests’ button is not clear and might not be understood by end-users.</td>
<td>3</td>
<td>3 / 7</td>
<td>3 / 7</td>
<td>Errors of omission</td>
</tr>
<tr>
<td>3. E.202</td>
<td>The button ‘Extra Tests’ on the screen might be difficult for end-users to notice due to its invisibility and unclear label.</td>
<td>4</td>
<td>7 / 7</td>
<td>4 / 7</td>
<td>Inefficient order behaviour Errors of omission</td>
</tr>
<tr>
<td>3. G.101</td>
<td>End-users might deselect tests both in cito as in the general sheet due to the display of the filled in selection boxes in both sheets after selection of a test only in the general sheet.</td>
<td>5</td>
<td>7 / 7</td>
<td>13 / 14</td>
<td>Inefficient order behaviour Factual errors</td>
</tr>
<tr>
<td>3. G.103</td>
<td>End-users might unsuccessfully try to order a test in cito as the action sequence required by the end-user is tedious.</td>
<td>4</td>
<td>7 / 7</td>
<td>Mentioned 4 times</td>
<td>Factual errors</td>
</tr>
<tr>
<td>3. J.203</td>
<td>End-users might be confused due to unclear error messages providing no feedback</td>
<td>4</td>
<td>6 / 7</td>
<td>14 / 28</td>
<td>Inefficient order</td>
</tr>
<tr>
<td></td>
<td>Concerning the User Action</td>
<td>Frequency</td>
<td>Notes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>------------------------------------------------------------------------------------------</td>
<td>-----------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. K.203</td>
<td>The fill-in box for entering a date for the execution of the order might not be noticed by end-users because of its screen allocation and missing error message if no information is entered.</td>
<td>3</td>
<td>‘All 7 end-users did not notice nor filled in the date’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. L.103</td>
<td>The fill-in box for entering a time to execute the order might not be noticed by end-users because of its screen allocation and missing error message if no information is entered.</td>
<td>3</td>
<td>‘All 7 end-users did not notice nor filled in the time to execute the order’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. M.103</td>
<td>End-users might not indicate who is to perform the order, for this is no standard option on the known paper ordering form and the system does not signal the user to make a selection.</td>
<td>2</td>
<td>2 / 7 2 / 7 Errors of omission</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. A.103</td>
<td>The action sequence required by an end-user to fill in clinical information is hard to understand and might lead to inefficient order behaviour.</td>
<td>2</td>
<td>7 / 7 26 / 28 Inefficient order behaviour</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. B.103</td>
<td>End-users may not go on to step 5 in the system if the end-user does not fill in information in step 4 when the system request additional patient information for a specific test.</td>
<td>3</td>
<td>2 / 7 3 / 28 Inefficient order behaviour Errors of omission</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. A.102</td>
<td>Inefficient action sequence in step 5 leads to too many actions to be performed by end-users.</td>
<td>5</td>
<td>5 / 7 16 / 21 Inefficient order behaviour Errors of omission</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. A.103</td>
<td>The user does not associate the goal ‘filling in research question’ with the correct action to complete the task in the system.</td>
<td>5</td>
<td>4 / 7 11 / 21 Errors of omission</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. C.104</td>
<td>No consistent system reaction to go to next step in the system, if required information has not yet been entered by the user.</td>
<td>5</td>
<td>7 / 7 16 / 21 Inefficient order behaviour Factual errors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. D.101</td>
<td>The go-on button might be clicked by end-users before finishing the ordering step, due to its invisibility and confusing label terminology.</td>
<td>2</td>
<td>4 / 7 11 / 21 Inefficient order behaviour</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. A.202</td>
<td>User might not be able to change selected tests in the system, for the system correct action sequence to fulfil this goal might be difficult to see for end-users.</td>
<td>2</td>
<td>3 / 7 5 / 28 Factual errors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. C.103</td>
<td>The button ‘start over’ might be incorrectly associated with the goal to add another order for the same patient, due to its location and label terminology.</td>
<td>2</td>
<td>1 / 7 1 / 28 Cancelled orders</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* = 4 out of 7 end-users

** = encountered errors 3 times in 28 scenario’s
Cognitive walkthrough: results verified by think aloud sessions

All potential problems detected by the cognitive walkthrough method were indeed encountered by end-users during the think aloud usability tests. In the next section, examples of the most severe and most frequently encountered usability problems, revealed by the cognitive walkthrough and verified by the think aloud analysis, will be described per subtask.

**Subtask two ‘Select a protocol’**

*Inflexibility of protocol selection:* The cognitive walkthrough analysis characterized this usability problem as the inflexibility of reversing a protocol selection in the system. The think aloud usability test showed that two end users choose a wrong protocol and after discovering their mistake were not able to undo their selection. Instead of restarting the system, one end user deselected all of the system’s protocol based pre-selected set of 16 laboratory tests, before selecting the set of laboratory tests he had decided on. Thus, the inability to undo the selection of a protocol within a laboratory order forced users to inefficient and time-consuming ordering behaviour.

**Subtask three ‘select laboratory tests’**

*Inability to navigate to laboratory tests in the system:* The cognitive walkthrough detected a potential usability problem referred to the invisibility of laboratory tests on the screen and the time spent in selecting laboratory tests by use of the scroll-bar (see Figure 1). In performing the four scenarios, the seven end users clicked the scrollbar 298 times. Because searching of tests with use of the scrollbar arrows took additional time, all end users were dissatisfied and complained about system performance. Moreover, though all tests were alphabetically organized the users on average still needed two and a half minutes to find the test they were searching for. Five of the seven users verbalized more than once, while searching for a test in the OM/Lab system, that the laboratory tests in the program were not organized as they were used to with the paper ordering forms. They all expressed that the system was needlessly inefficient and that they needed a search function by which they could find tests more easily, for example by entering the first letter of a test or a test synonym.

*Invisibility of button labels and incomprehension of button function:* For ordering infrequently ordered laboratory tests a button ‘extra tests’ had to be selected in the system. After activation of this button additional laboratory tests appeared on the OM/Lab screen. The cognitive walkthrough analysis revealed that due to the location of this button on the screen and its name label, users might not notice nor comprehend the function of this button, and might therefore not find the tests additionally supplied by the system after selection of this button. During the think aloud sessions end users indeed experienced severe difficulties in ordering these laboratory tests; they did not notice this button on the screen and if they did, they did not comprehend what its function was. The button was only found in
three of the seven times it should have been selected. Most end users complained that many laboratory tests were difficult to find and that selecting a test required too many actions by the user. Because of this problem in ordering tests, several errors of omissions were made by the users in the usability testing sessions.

**Appearance and information content of error messages in the system:** The cognitive walkthrough revealed that most error messages in the system provided the user no information concerning the cause and effects of his preceding action. In the think aloud testing sessions, in 14 of the in total 28 scenarios that were to be performed by the seven end users, system error messages were presented to them. All users appeared to be distressed and confused by these messages, leading to inefficient ordering behaviour by the end-users.

**Understandability of required sequence of actions:** The cognitive walkthrough revealed that the sequence of actions to be performed by end-users to order tests cito (tests that require results in a shorter time frame) was inefficient and could be confusing to them. After an end-user had selected a laboratory test on the general sheet, this test would also be given a mark and grey label on the (pink coloured) cito sheet. However, this mark on the cito sheet was only to indicate to an end-user that he or she had previously selected the test on the general sheet, though not in cito. If a user would however like to indicate that this specific test needed to be performed cito, he or she would first have to deselect the test by clicking on its box on the cito sheet, and then select the test again by clicking on the same box once more, after which action the colour of this specific test label was changed from grey into pink. The cognitive walkthrough analysis found this action sequence to be highly inefficient. Indeed, during the user test-sessions none of the seven end users executed the required sequence of actions in the OM/Lab system to order tests cito. The think aloud analysis provided insight into a different problem though related to the understandability of required system actions. Users misinterpreted the marked and grey labelled tests on the cito sheet and believed that all these tests were already ordered cito, though in fact they were ordered but not yet cito. Though the system designers had deliberately decided to use a highly noticeable colour for the cito sheet (namely pink), the mixing of colours for labels (grey for tests ordered in the general sheet, pink for tests ordered cito) disoriented all users as they interpreted grey marked tests on the cito sheet as if these tests were to be performed cito. They therefore deselected those tests for which it was not their intention to have them performed cito. This resulted in missing laboratory tests on each of the seven occasions it could have occurred.

**Subtask 4 ‘Enter clinical patient data’**

**System flexibility:** To analyze certain lab tests, laboratory personnel needed additional patient information. Clinicians had to indicate, for example, whether a patient was suffering from diabetes. The cognitive walkthrough revealed however that if an end-user did not supply the additional patient information required by the system in step four, the end-user could not go on to the next step in the
system. As the system offered no user feedback in the form of error messages on the reason for the need of entering this patient information, users indeed tried unsuccessfully to skip this task during the think aloud sessions. The end-users remarked several times “I don’t need to give additional information, I want to go on with sending the order”. Finally, they filled in “not necessary”, after which they could go-on with step five in the system. However, in the analysis of the test sessions these orders were regarded as incomplete orders, as in these cases the information the laboratory personnel (would have) needed to adequately perform tests had not been supplied by the ordering physician.

Subtask 5 ‘Enter reason for ordering laboratory tests’

Understandability of required sequence of actions/ User workload:

So that laboratory personnel could adequately process certain expensive or infrequently ordered laboratory tests, physicians needed to enter their reason for selecting these kinds of tests in the OM/Lab system. The cognitive walkthrough revealed that the system did not provide any feedback in support of the actions to be performed to successfully complete this step. Five of the seven users were indeed disoriented about the correct sequence of actions they needed to execute in entering a reason for ordering a specific laboratory test in the OM/Lab system. Users ineffectively tried to skip this step to accomplish sending the order. After on average executing the wrong action three times, four end-users finished this subtask correctly. However it seemed that this system feature was not easily learned, for an end user performed the same ‘ineffective’ action sequence several times in different task scenarios.

Think aloud analyzed usability problems

During the think aloud analysis eight specific additional usability issues arose which were not identified by the cognitive walkthrough. These problems were all revealed in the context of actions to be performed in the system that did not directly match the standard paper ordering routines. Five of these usability problems are discussed in the next paragraph.
**Table 3- Usability problems additionally revealed by the think aloud analysis.**

<table>
<thead>
<tr>
<th>TA per system state</th>
<th>Characterization TA analysis</th>
<th>Usability problem</th>
<th>Type of resulting error</th>
<th># users in usability testing</th>
<th># encounters in usability testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 2: TA 1</td>
<td>Locating required laboratory tests after selection of the ‘extra test’ button because of reorganization of laboratory tests.</td>
<td>Does not adhere to working practices</td>
<td>Errors of omission</td>
<td>7/7*</td>
<td>12/14**</td>
</tr>
<tr>
<td>Step 3: TA 2</td>
<td>The end-user’s goal to send an order in the system is prematurely as opposed to the additional tasks to be performed in the system. The text on button ‘start over’, confused end-users’.</td>
<td>Does not adhere to working practices</td>
<td>Cancelled orders</td>
<td>2/7</td>
<td>3/28</td>
</tr>
<tr>
<td>Step 3: TA 3</td>
<td>Disconformities in ordering protocol based laboratory order sets. Users order additional tests caused by interpretation differences of the paper based protocols in clinical practice.</td>
<td>Does not adhere to working practices</td>
<td>Factual errors in orders by adding tests</td>
<td>5/7</td>
<td>12/21</td>
</tr>
<tr>
<td>Step 3: TA 4</td>
<td>System content mismatched laboratory test terminology as applied in clinical working practice thereby leading to the belief that some test were not included in the system though in fact they were.</td>
<td>System terminology</td>
<td>Factual errors</td>
<td>6/7</td>
<td>10/14</td>
</tr>
<tr>
<td>Step 4: TA 5</td>
<td>The need to enter clinical information about the patient was not in accordance with the paper-based laboratory ordering process. In practice requesters did not supply patient clinical information on the paper order forms, this additional task in the system was not a goal the users sets himself in ordering tests in the system.</td>
<td>Does not adhere to working practices</td>
<td>Inefficient order behaviour</td>
<td>7/7</td>
<td>10/24</td>
</tr>
<tr>
<td>Step 4: TA 6</td>
<td>The system did not support the users’ goal to enter additional information concerning a specific test, for example a patient’s weight or length. In practice this was implicitly solved by adding written text on the paper laboratory form next to a test to be ordered.</td>
<td>Does not adhere to working practices</td>
<td>Potential errors in practice:</td>
<td>4/7</td>
<td>4/-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Inability to correctly analyze lab tests</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Double orders</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 5: TA 7</td>
<td>Forced entering of additional information concerning specific tests in step 5, after selection of a protocol based order-set, was considered as redundancy of information entering. End-users expressed that they had supplied this information by selecting a protocol.</td>
<td>Does not adhere to working practices</td>
<td>Factual errors</td>
<td>7/7</td>
<td>21/21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Requested information was not supplied by end-users</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 5: TA 8</td>
<td>Though the system was immature, users expected that in selecting a protocol based order set in step 2, all test according to the protocol would be ordered by the system, though some tests could not be ordered by use of the system.</td>
<td>Does not adhere to working practices</td>
<td>Errors of omission</td>
<td>3/7</td>
<td>9/21</td>
</tr>
</tbody>
</table>

* = 7 out of 7 end-users, ** = encountered errors 12 times in 14 scenario’s
Subtask three ‘select laboratory tests’

Difficulty in finding required laboratory tests:

After a user had clicked on the ‘extra tests’ button in step 3 of the OM/Lab system the lists of tests on the screen was extended with additional laboratory tests. However, the list of tests was then reorganized and regrouped leading to end-users experiencing difficulties in locating the correct test they had planned to order in the system. The cognitive walkthrough analysis did not reveal this reorganization of laboratory tests as a usability flaw as the analysts found the list of tests logically organized according to their alphabetic order, and after regrouping were still in alphabetic order. All seven end-users however experienced this reorganization of laboratory tests as a problem, as the list of tests was very large and some tests were after reorganization not directly visible on the screen. In reaction to this usability flaw, users skipped the selection of several tests as they claimed they could not find the test they needed. This resulted in errors of omission in 12 of the 14 orders this could have occurred during the users testing sessions. The users also commented on the tests arranged under the subgroup “blood tube” on the general sheet. They explained that, in order to parallel the paper based order forms, these tests should in fact have been grouped under a completely different tabloid that is under the tabloid ‘extensive coagulation’. This likewise contributed to the users’ experiencing severe difficulties in finding tests in the system.

Premature user goal setting of sending order in the system:

The system offered the possibility in step 3 to completely start over with the ordering process by clicking on the single button ‘start over’. Two end-users made the error to click the ‘start over’ button. The button ‘start over’ was interpreted by these users as ‘this order is ready, send the order and start a new order’. The two end-users who made this error also verbalized that they had selected the required tests and didn’t want to provide clinical information nor a research question. They were convinced that this was the next right action to send the order to the laboratory. In their clinical practice it was not customary to supply additional patient information. Also an error handling message in the system related to the button ‘start over’ in this case might have prevented 3 cancelled orders.

System terminology versus terminology used on the paper-based forms:

The names of several laboratory tests differed from the names used on the paper based laboratory forms. This was caused by the fact that the OM/Lab system automatically abstracted terms and names for lab tests from the ‘resource catalogue’ of the Laboratory Hospital Information System. Six of the seven end-users remarked that a test was not included in the system and could therefore not be ordered, though in fact the test could be ordered in the system but yet under a different name. This resulted in 10 omissions in orders.
Subtask 4 ‘Enter clinical patient data’

Self-descriptiveness of tasks in the system vs. users’ unawareness of additional system goal:

In the think aloud analysis all seven end-users verbalized that they did not understand the necessity of filling in clinical patient information in a laboratory order, as only they themselves would review the order results. Since physicians did not understand the reason for entering this extra information, they started searching for different routes in the system to send the order. However, this information was needed by the laboratory personnel for adequately analyzing tests. Since the end-users of the OM/Lab system in daily practice do not supply patient clinical information on the paper order forms, they were not willing to fill in any such information. This led to 10 occasions in which the required patient information was not provided by end-users.

Subtask 5 ‘Enter reason for ordering laboratory tests’

Redundancy of information vs. users’ unawareness of additional system goal:

Users, after having selected a specific protocol in step two of the system, were forced to add information for each of the tests ordered as a set by the protocol concerning the reason for its ordering. All seven end-users felt that they had supplied this information in step two just by selecting a protocol. Therefore, on nine occasions during the user sessions, three of the seven end-users deliberately omitted the requested information.

Discussion

In this study, we performed a usability evaluation of the OM/Lab system, an emerging CPOE laboratory system. Overall, the results of the usability evaluation revealed that the OM/Lab’s usability fell short of the designers expectations. Despite the “one-page” human computer interaction model implemented in the OM/Lab system, which seemed an intuitive model to follow in ordering laboratory test sets, our evaluation revealed a total of 33 usability flaws in the system, which all resulted in insufficient or erroneous order behaviour in the user test sessions. Though the system designers had felt that the six-step model could facilitate completion of the task of test ordering in a logical manner, many of the action-effect relations of the OM/Lab system were not transparent to the evaluators and the actual end users; in fact many of the system action sequences impeded rather than facilitated the ordering process.

As to an extent the design of the system influenced its low usability. Our usability inspection results, all verified in the user testing sessions, revealed a large number of human interaction problems in the OM/Lab interface that were of a general nature. These ‘surface’ human interaction problems of the OM/Lab system were related to inconveniently displayed buttons, incomprehensibility of button
labels, and feedback containing no information to the user about the cause of the error made and the consequences of a user’s action. Likewise the inflexibility of protocol selection and the cumbersome OM/Lab navigation structure for searching tests to be ordered were brought forth by conceptual design aspects deliberately chosen by the system designers. The poor perceptual cues of the interface and error messages without clarifying information content resulted in the users making faulty inferences regarding next actions required and hence triggered the users in applying inappropriate task strategies. These “surface” problems indicate that their existed mismatches between the designers’ and the users’ conceptualizations of the general task hierarchy (implemented as the six-step model in the OM/Lab system) of ordering lab tests. Yet, these problems could have been easily fixed before the OM/Lab pilot implementation effort.

Moreover, the analysis of the test sessions with end-users showed that the more specific action sequences to be performed within the general task hierarchy eq. six-step model of the OM/Lab system did not correspond to the daily working routines of end-users. These usability problems, of a more socio-technical nature, dealt with the order in which our end-users normally process information in their daily working practice in contrast to the way they had to process this information in the OM/Lab system. Further, in addition to the problem of the invisibility of laboratory tests on the OM/Lab screen, revealed in the usability inspection, certain laboratory tests were clustered in the system in such a way that these categories did not correspond with the way our clinicians tend to cluster tests in clinical practice whereas the paper-based order forms follow these clusters. This classification of tests in the system and within each of these classifications the alphabetic ordering of tests, considered logical by the system designers, did not facilitate but instead hampered end-users in searching for particular laboratory tests. A related usability problem of a socio-technical nature was the fact that users had to provide specific clinical patient information on system’s demand which did not correspond to their goal setting. End-users did not consider this information relevant for the ordering of laboratory tests.

As explained earlier, during the requirement analysis phase of OM/Lab only one representative of the outpatient neurology clinic and one representative of the outpatient laboratory had been consulted about the design of the user interface of the OM/Lab system. The total lack of communication between these representatives might explain why the system lacked in its design as it caused the designers to fulfil user requirements expressed by the representative of the outpatient laboratory which demanded actions of the neurologists which thus far were not required in their daily practice of ordering laboratory tests, that is in ordering tests on paper-based forms. Presumably, the representative of the outpatient laboratory had thought to take full advantage of the OM/lab system development efforts by expressing additional information needs of laboratory personnel so as to enable them to perform tests more efficiently than in the paper-based situation. Yet, these information needs of laboratory personnel were not discussed with the neurologists who were to provide this patient information. As a
consequence the neurologists felt that they did not need to provide this information which in their opinion was redundant and irrelevant for their (own) working practices.

These more socio-technical usability aspects were not only the basic cause for omissions and errors in orders and cancelled orders in our evaluation study but also in the pilot implementation of OM/Lab. Due to time pressures these flaws were not repaired in the OM/Lab system before its pilot implementation. However, these flaws indeed brought about a disruption of work patterns during the pilot implementation, such as delays in the time it took to process orders by the laboratory personnel and the necessitated return of patients to the hospital in order to finish incomplete laboratory orders. During the implementation of OM/lab it became clear that clinicians still experienced difficulties in ordering tests. Among other things, they could not easily localize tests in the system and they did not provide the patient clinical information required by the laboratory personnel to perform laboratory tests adequately. The clinicians seemed not aware that they had to provide this information which led to a vast amount of unfinished and incomplete orders. This, in the end, contributed to rejection of the system by the neurologists as well as the outpatient laboratory personnel. Yet, most of these usability problems in OM/Lab could not have been fixed effortlessly as they were part of the basic database structure of the system, which had not been developed in-house but by a commercially based vendor.

It is well known that the domain knowledge and daily working experiences users bring to the interface may determine its success or failure [39, 40]. Ideally a user’s domain knowledge and working behaviour would be considered in an early stage of system design to analyze in more detail whether the use of certain design features, such as the order of action-sequences to be performed and related menu structures, the terminology used for labelling categories, menus, buttons, and system feedback contents do bear a resemblance to the task domain and related working routines under consideration. However due to the fact that we had not been involved in the requirement analysis phase of the OM/Lab system we did not have had sight on the neurologists’ and laboratory personnel’s daily working routines in ordering or performing laboratory tests, nor on their additional information needs or other requirements to be fulfilled by the OM/Lab system. In the context of the relatively high number and severity of the usability problems encountered in our usability evaluation, we can only but assume that if the requirements capture activity for OM/Lab would have been performed more profoundly, this would have resulted in a higher usability of the OM/Lab system.

Usability evaluation studies require the construction of task scenarios. The manner of how to construct these task scenarios however seems to be a somewhat neglected aspect in the description of how to apply usability methods; task scenarios are a fundamental prerequisite for performing usability evaluations in a structured manner but explanations of usability methods provide no explicit guidance on how to construct these scenarios [40]. Variables such as task selection and task coverage have yet been found to affect the success of usability evaluation methods; detailed task descriptions can significantly change the number and type of, both severe and less severe, usability problems found in a
system’s design [41]. In our study, we were able to construct task scenarios covering all action-sequences to potentially be followed by end-users ordering tests by using OM/Lab. This may have enhanced the discovery of usability flaws in our evaluation study. Usability evaluations of applications more complex than OM/Lab will however be more demanding in time and human resources and can probably not be evaluated on all their aspects in one test. As it is with most systems not practically possible to include all a system’s aspects in one evaluation test, it is important to thoroughly analyse and set the goals for usability evaluations and focus the construction of task scenarios the system is to be evaluated against accordingly. Usability evaluators should verify these task scenarios in a systematic way to ensure that all system aspects relevant in the context of the goal are covered by these task scenarios and will be explored in the usability assessment. The ultimate goal is to develop end versions of health information systems that map on users tasks and strategies in performing these tasks and trigger the cognitive dynamics of the user in such a way that the intended tasks can be accomplished with minimal cognitive effort [42].

This study aimed at evaluating whether the OM/Lab system fulfilled end-users’ requirements and in so doing providing the OM/Lab system developers with information on how to redesign the system. However, due to management implementation pressures and unavailability of additional resources, the AMC system implementation project leader decided that the major usability problems brought forward by our usability evaluation would not be fixed before the OM/Lab system would be implemented in a pilot phase at the outpatient neurology clinic at the AMC. After a six weeks pilot implementation period, OM/Lab was withdrawn by the project leader due to the inadmissible problems the system caused in clinical practice such as delays in the patient waiting time at the outpatient laboratory and the inability of physicians to work with the system. Its pilot-evaluation results showed that OM/Lab’s low usability directly instigated clinicians’ rejection of it. In the end, the ‘one-page’ model was aborted, and a new setting was chosen for the sequel of order management at the AMC.

The failure of the OM/Lab system was nevertheless exactly the reason for the AMC to assemble a new medical informatics research team on (re)designing and evaluating clinical computer applications in an early stage of system development. With this new approach to system development and implementation, our aim is to bring forth clinical systems that support clinicians to the optimum in their daily working routines and consequently enhance good clinical practice.

Acknowledgements

The authors thank M. Smeulers, J. de Gans, P. F. Mooijweer-Groen, T.L.F. Urbanus and P. J. M. Bakker for their cooperation in conducting this study.
REFERENCES

Chapter 3

Anatomy of a failure: A Sociotechnical evaluation of a Laboratory Physician Order Entry System Implementation

Best paper selection – IMIA Yearbook of Medical informatics 2010: page 24-25

Linda W. Peute, Jos Aarts, Piet J.M. Bakker, Monique W.M. Jaspers


*Published in special issue: Human Factors Engineering for Healthcare Applications Special Issue. Edited By Marie-Catherine Beuscart-Zéphir, Jos Aarts and Peter Elkin*
Abstract

Objective: To investigate the human, social and organizational issues surrounding a Computerized Physician Order Entry system for Laboratory ordering (CPOE-L) implementation process and to analyze their interrelated effects on the system implementation failure in an academic medical setting. Second, to provide lessons learned and recommendations on to how to manage challenges of human, social and organizational nature surrounding CPOE-L implementations.

Methods: The themes surrounding CPOE introduction were identified by a heuristic analysis of literature on CPOE implementations. The resulting set of themes was applied as a reference model for 20 semi-structured interviews conducted during the CPOE-L implementation process with 11 persons involved in the CPOE-L project and in reviewing all CPOE-L related project documentation. Data was additionally gathered by user questionnaires, by user discussion rounds and through an ethnographical study performed at the involved clinical and laboratory departments. In analyzing the interview transcripts, project documentation and data from user questionnaires and discussion rounds a grounded theory approach was applied by the evaluation team to identify problem areas or issues deserving further analysis.

Results: Outlined central problem areas concerning the CPOE-L implementation and their mutual relations were depicted in a conceptual interpretative model. Understanding of clinical workflow was identified as a key theme pressured by organizational, human and social issues ultimately influencing the entire implementation process in a negative way. Vast delays in CPOE introduction, system immaturity and under-functionality could all be directly attributed to a superficial understanding of workflow. Consequently, final CPOE integration into clinical and laboratory workflows was inhibited by both end-users as well as department managers and withdrawal of the CPOE-L system became inevitable.

Conclusion: This case study demonstrates which human, social and organizational issues relevant to CPOE implementation cumulatively led to a failure outcome of the CPOE-L pilot introduction. The experiences and considerations described in this paper show important issues for CPOE systems to be successfully introduced and to be taken into account in future CPOE implementations. Understanding and consideration of (clinical) workflow aspects by project managers and the involved clinical organization is of extreme importance from the very start of a CPOE implementation process.
Introduction

Though potential benefits abound [1–3], successful implementation of Computerized Physician Order Entry (CPOE) systems for electronic entering and retrieving of medical orders is known to be difficult and expensive [4,5]. Failure of CPOE introduction seems to be tied to the current lack in understanding of current clinical practices surrounding order creation and of how to integrate CPOE into the apparent complexity of order workflow [6]. Sociotechnical approaches that focus on this interrelation of organizational environment and technology are considered valuable in enhancing the understanding of the CPOE implementation process [7]. This paper describes the evaluation of a failed implementation of a CPOE system for ordering laboratory tests (CPOE-L) at a large university hospital, the Academic Medical Center (AMC) in Amsterdam. For a good understanding of the issues influencing the CPOE-L project’s failure, the aim of this study is to characterize the nature of involved key issues, explore their interrelations, and use these insights to support new implementation efforts in our institution and possible other institutions. Frameworks or models that focus on analyzing factors that influence the adoption of Information and Communication Technology (ICT) have been applied in research for decades. Models such as those of Delone and McLean (IS Success model [8]), Seddon (variance model of Information System success [9]), Mirani and Lederer (framework to measure benefits derived from IS projects [10]), Grover (framework for measuring IS effectiveness [11]) and Smithson and Hirscheim (conceptual framework for IS evaluation [12]) analyze the impact, success and effectiveness of a system based on different classes of measures such as, among others, system quality, information quality, service quality, organizational impact, productivity and system usage. However, these models of ICT success cannot be applied in analyzing the variables and interpret their mutual interrelations influencing a complete system implementation process, from feasibility study to a first pilot phase. Delone and McLean themselves for example argue that ‘top management support’ and ‘user involvement’ may cause success rather than being a part of success and that their model does not take these variables into account. Other issues that might be relevant to an implementation study, such as the culture and organizational characteristics are also not included in these IT success models. Sociotechnical based frameworks could yet offer more insight into the sociotechnical transfusion of Information Technology (IT) in healthcare and its adoption by its intended users. However, existing frameworks such as the IT Adoption Model (ITAM) focuses on the individual users’ perspective and attributes of users’ system adoption but fails to operationalize the organizational, human and social aspects surrounding a system development and (first) implementation effort [13]. The review by Van Der Meijden et al. [14] was likewise aimed at identifying attributes to assess the success of electronic health care systems after implementation, not at revealing variables and their mutual interdependencies that might have influenced the software development life cycle and in the end may explain the nature of the cause of systems’ successes or...
failures. We were specifically interested in analyzing and interpreting the variables and problem areas that may cause CPOE success or failure in order to explain what went wrong during the CPOE-L implementation process. Our research focus is on interpreting human, social and organizational problems encountered in the entire CPOE-L implementation process, from feasibility to pilot implementation. We interpret our results by developing a three layer conceptual model based on the analysis of factors influencing CPOE implementation described by Ash et al., Kuperman, Sittig and Massaro [4,15–20]. Research of these factors dates back to 1970 and so the knowledge base of CPOE implementations has been built on over more than 25 years of experience [21,22]. Recent studies on CPOE implementation underscored the importance of understanding clinical workflow as a key issue for effective CPOE adoption in clinical practice [6]. But although success and failure factors for CPOE implementation have been thoroughly explored in studies on CPOE adoption, the influence of the human, social and organizational issues surrounding an implementation process has only limited been examined in the literature [23]. The main objective of this study is to investigate the impact and interrelation of experienced implementation problems on the whole CPOE-L implementation process from feasibility study to pilot introduction in the Academic Medical Center of Amsterdam. In doing so, we provide lessons learned and recommendations for acting upon challenges concerning human, social and organizational issues during a complete CPOE implementation process.

Research methodology

A longitudinal study design was chosen to analyze the implementation process of the CPOE-L system; from January 2004 until December 2004 [24]. From the restart of the project in June 2004 until the pilot implementation in September 2004 we conducted 20 semi-structured interviews of approximately 2 h, with 11 persons involved in the CPOE-L implementation process. All interviews were taped and transcribed with consent of the interviewees. The interview script was developed on the basis of insights from a heuristic review of CPOE literature. We examined the CPOE literature by systematically searching for issues or principles mentioned or described concerning CPOE introduction. In “heuristically” analyzing these principles with multiple reviewers we were able to define 36 CPOE ‘implementation themes’ to be covered in the interviews. Table 1 summarizes these themes in four categories: ‘Policy and strategy of the implementation project’, ‘Project organization, resource management and education’, ‘CPOE Technology and system development’, and ‘Social, cultural and organizational context’. This categorization is based on system evaluation domains known in information technology literature [25] enriched with CPOE domains of consideration formulated by Ash et al. [18]. In representing themes in these four domains, we aim to evaluate human, social and organizational issues that influenced the CPOE-L implementation process. Interview questions were prefaced by stating that the study concerned ‘what
<table>
<thead>
<tr>
<th>Interviewed person</th>
<th>Role or function in CPOE-L implementation process</th>
<th>Dates of interview</th>
<th>Main focus of interviewee CPOE implementation theme(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principle CPOE-L project</td>
<td>Head responsible for IT implementations at the AMC, physician champion</td>
<td>May '04</td>
<td>CPOE vision/Scope and objectives, role of physician champion, power/control and politics, organizational readiness, top level commitment, implementation costs</td>
</tr>
<tr>
<td>Project leader</td>
<td>Manager CPOE-L implementation team from October '03</td>
<td>November '03</td>
<td>Staging of the implementation, implementation strategy, involvement of contacts and essential people, organizational readiness, implementation motivators, Staging of the implementation follow up, internal communication, collaboration and trust, teamwork, social relations, organizational readiness (change of opinion)</td>
</tr>
<tr>
<td>Medical Information analyst</td>
<td>End-user representative, part of CPOE-L implementation team</td>
<td>June '04</td>
<td>System value to users, collaboration and trust, implementation strategy, user centered design, culture of involved departments</td>
</tr>
<tr>
<td>Software developer</td>
<td>Developer user interfaces CPOE-L system, part of CPOE-L implementation team</td>
<td>June '04</td>
<td>System customization, order sets, views and attitude, teamwork, Internal communication</td>
</tr>
<tr>
<td>System tester and implementer</td>
<td>Implementation coordinator IT at the AMC, responsible person for CPOE-L implementation (linking purchased database to developed UIs but also responsible for training and maintenance CPOE-L system)</td>
<td>June '04</td>
<td>System testing and evaluation, workflow mismatches, complexity of working practices, Training, user support, teamwork, internal communication and feedback, complexity of working practices, collaboration and trust</td>
</tr>
<tr>
<td>Hospital (Laboratory) Information systems controller</td>
<td>IT administrator, laboratory IS</td>
<td>May '04</td>
<td>Collaboration and trust, working practices and social relations, Internal communication</td>
</tr>
<tr>
<td>Outpatient laboratory department head</td>
<td>Head of the AMC laboratories</td>
<td>January '04</td>
<td>Power/control and politics, organizational readiness, top level commitment, workflow complexity</td>
</tr>
<tr>
<td>Outpatient laboratory department manager</td>
<td>Outpatient laboratory coordinator, direct end-user of the system, project contact person</td>
<td>January '04</td>
<td>Workflow, views and attitude, power/control and politics, involvement of end users, system value to end users</td>
</tr>
<tr>
<td>Outpatient laboratory end-user</td>
<td>From 1997 till 1999 laboratory end-user, function shift in 2000 and became laboratory quality manager, was however still contacted for information by the implementation team as end-user project team contact person</td>
<td>September '04</td>
<td>Internal communication and feedback, social relations, collaboration and trust</td>
</tr>
<tr>
<td>Outpatient neurology manager</td>
<td>Outpatient Neurology coordinator and planner, involved in CPOE-L development from 1997</td>
<td>January '04</td>
<td>Involvement in CPOE project, workflow changes, views and attitude, social relations, internal communication and feedback, Involvement in CPOE project, workflow integration, views and attitudes, internal communication and feedback, organizational readiness</td>
</tr>
<tr>
<td>Outpatient neurology physician end-user</td>
<td>End-user CPOE-L system, project team contact person</td>
<td>August '04</td>
<td>Development order sets, collaboration and trust, internal communication, workflow, views and attitude, views to end users</td>
</tr>
<tr>
<td></td>
<td></td>
<td>September '04</td>
<td>Views and attitude towards CPOE-L, system value to end users, collaboration and trust, end-user involvement</td>
</tr>
</tbody>
</table>
you have learned during the CPOE implementation’ and ‘things that happened that you didn’t expect or could not have foreseen’. Hence, the focus was both on the insights gained from literature as on unforeseen or unanticipated circumstances, both desirable and undesirable. The questions were designed to be as neutral as possible to avoid bias. Interviewees were first presented with a theme and its description and then asked to openly describe their view on that theme in the CPOE-L implementation and its’ (un)foreseen consequences as they had experienced themselves. Interviewees were continually stimulated by further probing when discussing a new topic that they related to the implementation theme under discussion. In doing so we aimed to gain insight into interrelations between implementation issues and their effect on the duration and course of the implementation process as perceived by the different interviewees. Table 2 provides an overview of these interviews and the themes the interviewees focused on. Apart from interviews, field and documentary historical methods were applied. All project documentation was analyzed, both before and after the pilot introduction of the system. Observations were made at the sites involved in the project (laboratory sites and neurology department) and after the pilot roll-out an ethnographic study was performed that observed the end-users during their daily use of the CPOE system. The end-users were asked to keep diaries to report on human, social and organizational as well as workflow issues, and to write down a list on technical, or other unspecified problems they encountered in practice (self-reported problem lists) on a daily basis [26]. This CPOE-L pilot implementation, conducted as a first attempt for system roll-out in clinical practice, was subsequently evaluated by means of end-user questionnaires during the fourth week of introduction. The questionnaire administered was based on the validated Questionnaire for User Interaction Satisfaction (QUIS), complemented with questions on usefulness, training and provided user support. Six weeks into its pilot stage, the CPOE-L implementation was completely aborted. Meetings with all departments involved were organized to discuss their CPOE-L implementation experiences from the perspective of main problem areas and interrelations as identified in the CPOE-L implementation analysis. Fig. 1 presents the applied study flow outlined by the CPOE-L project implementation process. Researchers from the department of Medical Informatics of the Academic Medical Center-University of Amsterdam and an affiliated hospital performed the study. Their professional background is in cognitive psychology, medical informatics and social and organizational science; their background experience in medical informatics varies between 5 and 25 years. Interviews were held by the same interviewer, experienced in interview techniques for over 5 years. Interviews were transcribed and coded bottom-up in conformance with the grounded theory approach [27]. The ethnographic methods, observations and analyses of self-reported problem lists by the first author during the implementation process provided insight into the current working practices and views of potential end-users on the CPOE-L system introduction. Analysis was done iteratively: documents and field notes were first analyzed separately. Then, after each set of interviews was completed, the first and last author coded them and met to agree on final coding and discussed if root patterns could be classified in one of the themes uncovered in literature or if new themes should be
added. The second author assisted and supported interpretation and comparing of the analysis results on basis of sociological and sociotechnical experience in CPOE system implementation. This triangulation of methods and blending of data is known to enhance the scope of detection and the validation of results, in this study the human, social and organizational issues influencing the CPOE-L implementation [16,28].

Table 1 – Implementation themes derived from a heuristic analysis of CPOE implementation literature.

<table>
<thead>
<tr>
<th>CPOE implementation category</th>
<th>Implementation themes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Policy &amp; strategy (political, strategy)</strong></td>
<td>CPOE motivation, communicated</td>
</tr>
<tr>
<td></td>
<td>CPOE vision</td>
</tr>
<tr>
<td></td>
<td>Scope and objectives, transparency</td>
</tr>
<tr>
<td></td>
<td>Top level commitment, higher level support</td>
</tr>
<tr>
<td></td>
<td>Physician leadership/champions</td>
</tr>
<tr>
<td></td>
<td>Contacts, personnel, essential/key people</td>
</tr>
<tr>
<td></td>
<td>Costs, sufficient funds</td>
</tr>
<tr>
<td></td>
<td>Internal communication and feedback, transparency</td>
</tr>
<tr>
<td></td>
<td>Staging of the implementation, flexible planning and strategy</td>
</tr>
<tr>
<td></td>
<td>Requirements analysis, user needs analysis</td>
</tr>
<tr>
<td></td>
<td>Multidisciplinary teamwork</td>
</tr>
<tr>
<td></td>
<td>End-user involvement</td>
</tr>
<tr>
<td></td>
<td>Project evaluation</td>
</tr>
<tr>
<td></td>
<td>User-support during introduction</td>
</tr>
<tr>
<td></td>
<td>Sufficient training before and during introduction</td>
</tr>
<tr>
<td><strong>Project organization, resource management</strong></td>
<td>Workflow analysis</td>
</tr>
<tr>
<td><strong>Project organization, resource management</strong></td>
<td>Personal order sets</td>
</tr>
<tr>
<td>and education (managerial, economy, education)</td>
<td>Support clinical protocols</td>
</tr>
<tr>
<td></td>
<td>User centered design, usability</td>
</tr>
<tr>
<td></td>
<td>Consistent, intuitive, user friendly interface</td>
</tr>
<tr>
<td></td>
<td>Decision support</td>
</tr>
<tr>
<td></td>
<td>Customization, flexibility, adaptability</td>
</tr>
<tr>
<td></td>
<td>System speed</td>
</tr>
<tr>
<td></td>
<td>Available functionality</td>
</tr>
<tr>
<td></td>
<td>System maturity</td>
</tr>
<tr>
<td></td>
<td>System testing and evaluation</td>
</tr>
<tr>
<td></td>
<td>Multi-dimensional integration</td>
</tr>
<tr>
<td><strong>CPOE Technology and system development (technical</strong></td>
<td>Complexity of work practices</td>
</tr>
<tr>
<td><strong>CPOE Technology and system development (technical</strong></td>
<td>Integration into order workflow</td>
</tr>
<tr>
<td><strong>Social, cultural and organizational context</strong></td>
<td>Value to users</td>
</tr>
<tr>
<td><strong>Social, cultural and organizational context</strong></td>
<td>Collaboration and trust</td>
</tr>
<tr>
<td><strong>Social, cultural and organizational context</strong></td>
<td>Social relations, open attitude</td>
</tr>
<tr>
<td><strong>Social, cultural and organizational context</strong></td>
<td>Culture of involved department</td>
</tr>
<tr>
<td><strong>Social, cultural and organizational context</strong></td>
<td>Power, control and politics</td>
</tr>
<tr>
<td><strong>Social, cultural and organizational context</strong></td>
<td>Organizational readiness</td>
</tr>
<tr>
<td><strong>Social, cultural and organizational context</strong></td>
<td>Involvement of end-users/contacts</td>
</tr>
</tbody>
</table>
Setting and background of the CPOE-L implementation process

The CPOE-L project dated back to 1997. The main purposes of CPOE introduction at that time were to improve quality and efficiency in ordering and to offer clinicians the possibility to order diagnostic or therapeutic tests electronically. The decision to commence with the laboratory functionality was made from the early start of the project. The outpatient laboratory seemed ideal to start with CPOE implementation for several reasons: the large volume of laboratory orders requested, the apparent large number of incorrectly filled-in orders (33% in 1997), and the fact that the outpatient laboratory had a full information technology infrastructure in place. The choice to involve the neurology outpatient department in a pilot implementation was also made in the first stage of the project. The neurology outpatient department was known for its well-organized working practices including the use of clinical protocols. A CPOE database system was purchased, while the clients’ user interfaces of the system would be developed in-house. The integration of the CPOE database into the existing ICT infrastructure required software changes of the Hospital Information System (HIS). Due to staffing changes a new project team was assembled in November 2003 to realize a product version of CPOE-L with the target to implement it in the first quarter of 2004. Two information analysts, with a background in medical informatics, performed the requirements analyses and developed the system specifications. Their experience as clinical information analyst ranged from 5 to 8 years, respectively. The system specifications were drawn up in meetings with the outpatient neurology clinical contacts, potential system end-users who contributed to the CPOE-L project, and the outpatient laboratory contacts of the CPOE-L project, but were also partly based on requirement analyses previously performed in 2002. The software engineers, experienced in system development for over 10 years, were to develop the user interfaces for the CPOE-L system based on these specifications.

Fig. 1 – Timeline CPOE Implementation process and graphical display of the evaluation study flow.
Analysis of human, social and organizational issues of the CPOE-L implementation and their impact on workflow understanding and vice-versa

Table 3 provides an overview of the frequency in which implementation problem themes were mentioned and discussed in relation to each other during the interviews. Eight out of the 36 implementation themes covered in the interviews and document analysis appeared to be associated with direct problems in the system implementation process finding its central problem in the ‘understanding of clinical workflow’. The observation study at the outpatient neurology department and the outpatient laboratory validated these results and provided additional insight into the issues influencing this understanding of clinical workflow.

The conceptual model in Fig. 2 depicts these eight main problem themes that influenced the CPOE-L implementation process, in circles. The arrows between the circles represent patterns of influence with bi-directionality. Texts provided in the figure, either on top of or in the middle of an arrow or on top of the circles depict syntax patterns derived from the analyses of field notes, documents and interviews. The observational-ethnographical study, conducted during the CPOE-L system’s pilot implementation provided insight into the impact of the unresolved issues, more specific the lack of understanding of clinical workflow, on problems experienced during the CPOE-L system pilot introduction. Problems in system use and disruptions in clinical work that end-users were confronted with on a daily basis could be attributed to the issues revealed through the interviews, document analysis and observations before pilot introduction of the CPOE-L system. It appeared that, from the beginning of the restart of the project in 2004 the project team(s) had tried to reduce the complexity of the CPOE-L implementation by dividing it up in separate implementation aspects to be tackled. By working on sub-aspects the project team(s) assumed that the overarching implementation of the system was being handled. However, as departments involved in the CPOE implementation process were contacted separately, gaining an integrated view on ordering procedures and ordering workflow was severely hampered. As a result, the CPOE-L project lacked a mechanism to act on dynamical changes that occurred in both project and department organizations. Needed innovations in organization and order workflow of the clinical and laboratory departments were continuously carried through while the pilot introduction was being prolonged by the implementation team. For example, the ordering of laboratory tests on paper was continuously innovated; at a certain moment in time neurologists only had to pick a single paper test ordering form and sign it to order clinical protocol related laboratory tests. The initially stated benefits of CPOE implementation to the end users were already partly achieved by these relatively (simple) changes in work procedures. At the start of the CPOE project, the system implementation was seen as an essential addendum to simplify laboratory test ordering. But as a result of the achieved benefits of relatively simple workflow innovations, this need for CPOE
decreased and, the interest and collaboration of the departments involved in the CPOE-L project diminished. Consequently, the project team only realized far along the course of the implementation process that certain AMC laboratories, such as the bacteriology and virology, still worked with paper sheets and that several other third party laboratories worked with lab systems incompatible to the CPOE-L database.

Figure 2: Three layer conceptual model depicting influencing factor interrelations of the CPOE project and their perceived effect on workflow.

It became clear that the project’s management grossly underestimated the CPOE-L implementation project. The technical infrastructure of some laboratories, though in place, proved to be enormously complex and the large number of specialized laboratory services added to this complexity. This severely hindered the linkage of the purchased CPOEL database to the Laboratory-Hospital
<table>
<thead>
<tr>
<th>Problem theme in interview</th>
<th>Short description</th>
<th>Discussed problem theme interrelations in interviews</th>
<th>Frequency of problem theme interrelation in interview transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Involvement of project contacts</td>
<td>Key personnel in each service responsible for negotiating and supplying information regarding POE development. Ash et al. [16,18,19], Singh and Stod [20]</td>
<td>Internal communication and feedback</td>
<td>21 times mentioned in 20 interviews; experienced as important to project contacts while this importance was unrecognized by the implementation team</td>
</tr>
<tr>
<td>Internal communication and feedback</td>
<td>Communication to and from all involved persons in the implementation of the POE system. Ash et al. [16,18,19], Massaro [4]</td>
<td>View and attitudes, Collaboration and trust</td>
<td>Workflow 10 times mentioned in 17 interviews; mostly experienced by the project implementation team. Project leader commented on this more than 8 times in 2 interviews</td>
</tr>
<tr>
<td>Collaboration and trust</td>
<td>Collaboration of involved clinical departments in system development, trust in delivering a good system that supports clinical workflow. Ahmad et al. [41], Massaro [8]</td>
<td>Collaboration and trust</td>
<td>Workflow 14 times mentioned in 20 interviews; described in interviews with both the implementation team as well as the project contacts</td>
</tr>
<tr>
<td>Complexity of working practices</td>
<td>Complex structure of work practices of medical personnel. Kupperman and Gibson [15], Massaro [4]</td>
<td>Workflow</td>
<td>18 times mentioned in 16 interviews; discussed by system tester and implementer</td>
</tr>
<tr>
<td>Value to end-users</td>
<td>Stated benefits for end-users that follow from POE use in practice. Ahmad et al. [41]</td>
<td>Collaboration and trust</td>
<td>Value to end-users 14 times mentioned in 17 interviews; discussed by end-user contacts</td>
</tr>
<tr>
<td>Organizational readiness</td>
<td>Technical infrastructure and the social system. Do they offer the right background to implement a POE system. Ash et al. [16,18,19], Shulkin et al. [18]</td>
<td>Workflow</td>
<td>24 times mentioned in 14 interviews; discussed by the project implementation team members and the laboratory contacts</td>
</tr>
</tbody>
</table>
Information System and prolonged the implementation process by 3 months. Linkage of the CPOE-L database to specific blood test codes of the laboratory system required updating and supplementing of laboratory codes in the CPOE-L database. When these problems were encountered, the project team members stated that these implementation issues were instigated by a lack of organizational readiness of the laboratory environment to implement CPOE-L. Contrasting to this view is the project team’s limited insight into the complicated order workflow of the laboratories that intrinsically contains concealed information needs to be covered by a CPOE-L system. As the CPOE-L project team initially focused on the technical properties of the system, project actions were design driven and based on the CPOE system’s constraints and opportunities. The first result of this approach was that project decisions were based on the database layout and potential issues surrounding the CPOE’s infrastructure. Second, the project’s communication channeling was unidirectional; project contacts delivered input required for defining CPOE-L system contents. Information relevant to the CPOE-L project, such as the intermediate changes in order workflow, was not specifically requested by the project team and thus not brought to their attention. The project team commissioned department contacts to deliver particular information needed for system development without in turn informing them on the project status. This cultivated a strong negative attitude among the department contacts towards and also loss of trust in system development. Also, the fact that no arrangements had been made for time or financial compensation for the department contacts on efforts spent on the CPOE-L project intensified these negative attitudes and stimulated diverging interests to CPOE implementation between the departments and the CPOE-L project team. Instead of a gradually evolving synergy between the departments’ workflows and the CPOE-L system under development, the negative attitudes towards the CPOE-L system even promoted changes in project management and continuous innovations in the departments’ workflow. The resulting mismatches between the neurology department’s order workflow and the CPOE-L system design were brought to attention by a usability study performed on a prototype version of the system [29]. Project time restraints, apparent lack in resources and aforementioned database problems were held accountable by the project team for the inability to deal with the revealed usability flaws in the system. However, the analyses of interview transcripts revealed that these mismatches between the CPOE system design and the neurology department’s workflow resulted from the earliest lack in understanding and consideration of workflow issues relevant to the design of the CPOE-L system. Resolving these system flaws would however have required too many system changes, for which a large time investment could not be granted. The project team likewise failed to anticipate that their limited laboratory order workflow analysis would be insufficient to commence with CPOE-L design. The poor understanding of order workflow in the earliest development phase of the system subsequently led to severe delays in system development and also to the fact that these usability issues could not be solved within an appropriate time span.
The CPOE-L pilot implementation

After a prolonged system implementation process of 8 months the project leader nevertheless consulted the department heads and decided to introduce a yet immature, but considered ‘ready’, CPOE-L system in a pilot phase at the outpatient neurology clinic. The system was considered immature as some technical aspects (for example the CPOE linking with the external laboratory systems) were not yet taken care of but planned to be realized in the near future. The ethnographical study performed during the CPOE-L pilot phase aiming at analyzing the extent of CPOE-L system adoption and integration into order workflow, revealed that end-users indeed missed certain system functionalities and experienced usability problems in their daily use of the system. In total, 96 problems were
described in the self-reported problem lists by end-users at the outpatient laboratory and over 47 by the clinical neurology department during the 6-week introduction of the CPOE-L system. These problems ranged from direct problems in system use by end-users to problems in system stability and printer problems. Also, a 15 min increase in patients’ waiting time at the laboratories was reported. Problems in system use were said to directly obstruct the laboratory workflow. As the problems encountered could not be solved easily, 33 out of the total of 206 laboratory orders (16%) entered by the neurologists during this 6 weeks of use of the CPOE system were ordered double and 25 (12%) were not ordered at all. While extracting this information from the CPOE database it became clear that 58 entered laboratory orders could not be realized by the laboratory. Six weeks after the start of the CPOE-L pilot, all department heads abandoned the CPOE-L system from their units. The CPOE-L project leader thus decided to put the system out of use. Several project meetings were organized in order to discuss unresolved issues surrounding the CPOE-L implementation. CPOE-L project members as well as project contacts were in agreement that problems in usability, performance, system immaturity and poor functionality were the direct cause of system failure. The discussions stressed that the root of these problems could indeed be found in pitfalls of the implementation strategy and organization of communication in the CPOE-L project. Table 4 recapitulates the analysis of the implementation process as described above as lessons learned on the themes of project and organization, human and social, and their interrelation with workflow understanding. Because the nature of a case study such as this one is not to provide general lessons, the lessons learned from this analysis are labeled as ‘hazardous issues’ and ‘prerequisites’ for future implementations. ‘Hazardous issues’ are these scenarios described in this case study which pose a potential risk to system introduction or how its management is organized. ‘Prerequisites’ are these conditions revealed in this case study that can be viewed upon as conceivable prerequisites for new CPOE implementation efforts by other institutions.

Discussion and conclusions

Studies that describe CPOE implementation efforts have shown that implementing these kind of systems is challenging and not only in need of organizational transformation. Abroad variety of implementation preconditions should also be fulfilled, including resource commitment, good leadership and organizational readiness [30–33]. However, as failure stories continue to arise, it seems that issues influencing the course of an implementation process and finally inducing implementation failures have not been analyzed fully. Though it is often stated that the issues responsible for CPOE implementation failure are mostly unanticipated, elucidating these issues might be helpful to enhance new implementation efforts. This study combines a longitudinal analysis of an implementation process of a CPOE system with an in-depth analysis of problem themes identified from CPOE literature. It revealed the human, social and organizational implementation problems that were attributed to a
superficial understanding of (clinical) workflow, and subsequently led to the CPOE-L implementation failure. In our research strategy we focused on factors or issues described in CPOE implementation literature to be of (potential) influence on a CPOE implementation outcome. This approach limits the identification of themes outside of the CPOE literature. Recent publications offer more insight into factors possibly influencing health informatics systems’ implementations in general. Brender et al. identified 110 success factors and 27 failure criteria distributed on several in essence sociotechnical categories [34]. The success and failure factors we identified from the CPOE literature appear to be quite similar to the factors identified by Brender et al. This shows that CPOE implementation literature covers most factors for success and failure identified in health informatics systems’ implementation studies. There are however some CPOE specific technical and functional aspects to be taken into account in a CPOE design process such as the need for specific order sets, high system flexibility and multi-dimensional integration of the system with the electronic patient record (EPR) and other systems. At first sight, a CPOE implementation could therefore be viewed upon as a regular health informatics system implementation project. New research on CPOE implementation indeed supports this statement. Outcome research on user acceptance, though often reported the major cause for CPOE failures, has gradually made way for research into the more intrinsic organizational and workflow problems primarily experienced during CPOE adoption in clinical practice [6]. In our CPOE-L study all 36 issues identified in the CPOE literature were considered relevant, however our analyses made clear that the failure of the CPOE implementation team to adequately guide and coordinate the project led to very severe implementation problems. The mismatch between the clinical workflows at the laboratory units and the CPOE system design, provoked by the chosen implementation strategy of the project team, accounted for the inhibition of the system’s adoption in clinical neurology care. Literature elaborates on the importance of an extensive workflow analysis as it is indeed seen as a key issue in designing successful systems [6,33,35] without such an analysis, major problems with the adaptation to a CPOE system by clinicians after its introduction into clinical practice may occur [36]. However, Pratt et al. note that analyzing medical order workflow is a fairly complex and elaborate task as work flow is a dynamic process filled with events involving many intricate levels in the organization [37]. In managing a system implementation process, focus should not be so much on (technical) sub-problems surrounding the implementation, but rather on the overarching relation between the development of the system and the clinical workflow the system is to be implemented in. Moreover, adjustments in workflows often have to be realized and therefore foreseen to optimally profit from a CPOE system implementation. In contrast, literature shows that CPOE systems have mostly been designed on very simplistic ordering models. Without understanding the very intricate and complex activities of clinical workflow, CPOE systems are unable to cope with ordering complexity [38–40]. Then, CPOE success largely depends on the organizational readiness to receive CPOE and the readiness of the AMC should therefore have been evaluated and discussed by the project team before starting CPOE development [33,41]. Though an organization’s readiness may be
determined by technical developments and acknowledgement to changes, its angles such as the IT infrastructure, staff expertise, organization of care management processes, and organizational culture must be explored fully before starting an implementation process [40]. The fact that the CPOE-L implementation was not considered a necessary ‘change process’ from the beginning of the project already points at a too simplified view within the AMC on a CPOE implementation. When implementing CPOE it must be clear which processes are being targeted for change and how these processes relate to other parts of the organization, from order-generating departments to order-executing departments. In line with this, usability studies can reach their true value to a system’s implementation process only if based on extensive insights into clinical workflow and only when those results are integrated in system design. The problems users experienced in working with the user interfaces of the CPOE-L system during the pilot implementation were an apparent consequence of the superficial analysis of workflows. Even the involvement of medical informatics specialists in the CPOE-L implementation process did not suffice to reveal the end-users’ tacit knowledge of the clinical and laboratory workflow to a level needed for a satisfying CPOE-L design. The most prominent solution to bridge this gap between the developers and the users today is to introduce the actual engineers of the software directly into the client’s environment [42]. We did conduct an ethnographic study but this activity mainly served to investigate the final consequences of the encountered implementation problems during the system implementation process on the outcome of the CPOE pilot implementation. Yet, amore interventionist approach sensitizes that fieldwork and informing of system design should be closely interwoven [43]. An ethnographic analysis on workflow should therefore be incorporated into a system implementation process and be performed at predefined phases in the software development cycle continuously informing system design in order to build systems that could potentially fit (changing) workflows. Our approach was primarily focused on gaining understanding into the issues influencing the CPOE-L implementation, with a focus on its implementation from the restart in 2003. From this case study, we have learned important lessons relating to management and organization of the overall system development process and how to integrate these insights into CPOE system design. These lessons are considered extremely valuable in the AMC and will continually be incorporated in new implementation projects, thereby promoting the AMC organization’s readiness for a new implementation effort of the CPOE-L system. The three layer conceptual framework that evolved during this case study helped to understand the cumulative effect of those issues that dominated our CPOE-L implementation failure. Case study research can yet serve the building of theories [44]. In developing a conceptual model, connecting main problem implementation themes, we have shaped a basis for a theory on how to approach and interpret human, social and organizational interactions between clinical workflow and CPOE system design [44, 45]. In applying these insights to other sites, such a theory may increase and may thereby enhance the success of future CPOE implementations.
Acknowledgements

The authors thank all members of the CPOE-L project team and all project contacts involved for their cooperation in conducting this study.

REFERENCES


Chapter 4

Effectiveness of Retrospective and Concurrent Think Aloud in Formative Usability Testing; Which Method Performs Best?

Linda W.P. Peute, Nicolette F. de Keizer, Monique W.M. Jaspers


Abstract

Objective: To compare the performance of two usability testing methods, the Concurrent (CTA) and Retrospective (RTA) Think Aloud, in a formative usability study of an Intensive Care Registry Physician Data Query Tool.

Methods: Sixteen representative target intensivists participated in the usability evaluation study. Subjects were allocated to one of the method CTA/RTA condition by a matched randomized design. Each subject performed six usability testing tasks of varying query development complexity in the Query Tool. The methods’ performance was compared in terms of overall effectiveness in detecting usability problems within the average time subjects spent per method. Subsequently, the thoroughness of the methods in detecting usability problems weighted by problem severity level and problem type was analyzed.

Results: The usability evaluation of the Data Query tool revealed a total of 43 unique usability problems. CTA performed significantly better than RTA in detecting usability problems. Overall, CTA usability problem detection effectiveness was 0.80 vs. 0.62 (p<0.05) respectively with an average difference of 42% less time spent per subject compared to RTA. In addition CTA was more thorough in detecting usability problems of a moderate (0.85 vs. 0.7) and severe nature (0.71 vs. 0.57). Regarding problem type detection, CTA detected unique usability problems concerning graphics/symbols, navigation issues, error messages and the organization of information on the Query Tool’s screens. RTA detected unique issues concerning system match with users’ language and applied terminology. Qualitative analysis of the CTA and RTA verbal protocols showed that RTA verbal protocols contained significantly more explanatory comments regarding the cause of a usability problem and comments concerning additional system requirements.

Conclusion: CTA is more effective in usability problem detection but does not outperform RTA. RTA additionally provides sight on unique usability problems and new user requirements for specific user groups. Based on the results of this study we recommend the use of CTA in formative usability evaluation studies of health information technology. However, we recommend further research on the use of RTA in usability studies focusing on user profile customized (re)design.

Keywords: Usability Evaluation, Think Aloud, Intensive Care Information System
Introduction

In the current era, clinicians are increasingly becoming dependent on interactive healthcare applications to provide them access to the information they require [1-3]. Easy navigation and high understandability of the application’s interface have therefore become imperative to clinicians for efficient and effective system use [4]. In other words, interactive healthcare applications need to be designed with explicit regard to their usability; where usability is defined as the effectiveness, efficiency and satisfaction with which specific users can achieve a specific set of tasks in a particular environment [5].

Formative usability studies provide means to improve on a system design by uncovering those interface design flaws that clinicians might encounter when interacting with a system in a clinical setting [6]. To perform a formative usability study a broad spectrum of Usability Evaluation Methods (UEMs) is available and these methods are increasingly used in interactive healthcare applications design and evaluation [7-8]. However, selection of a UEM in a specific healthcare setting is often limited by practicality and by accessibility of required human resources and time to perform the evaluation study. The spare number of publications on UEM performance in formative studies on interactive healthcare applications furthermore limits usability practitioners in motivated selection of a UEM [9]. UEMs detection scopes may however differ significantly and may prove to be more or less informative to system (re)design in different study settings. To support justified UEM selection this study compares the performance of two known UEMs; the Concurrent Think Aloud (CTA) and Retrospective Think Aloud (RTA), in a formative usability study of a web-based Intensive Care (ICU) Registry Data Query Tool.

CTA finds its foundation in the cognitive and behavioral sciences [13-14]. Its application in usability testing studies yet has shown its value in providing rich and valid data on the usability of a system’s design [15]. In CTA usability testing subjects are instructed to verbalize their thoughts while conducting predefined tasks in the system concurrently. The RTA is a variation on the CTA and emerged to bypass certain limitations of the CTA in usability testing studies [16]. Concurrent verbalization for instance might interfere or slow down subjects’ task performance and in doing so may influence the reliability of CTA usability testing measures such as task efficiency and efficacy [17]. In contrast to CTA, the RTA method instructs users to recall their thoughts and actions after they have finished the computer supported predefined task(s). So, during RTA subjects verbalize their thoughts while reviewing a screen recording of their performance while interacting with the system under study. In this way, no direct interference of a subjects’ task performance occurs. However, linguistic studies and studies from psychology have shown that, when compared, the methods do appear to collect verbalized data of a different quantity and differing kind [18,19]. A basic assumption for this difference in methods’ output can be found in the workings of the human memory. In
concurrent verbalizations thought processes are expressed during task performance while retrospective verbalizations rely on the retrieval (recall) of these thought processes on tasks already completed. However, if these methods differ in their collection of data they might also differ in their detection scope of usability problems. It is of importance to understand the methods’ differences in performance given that (re)design efforts of interactive health application are primarily based on a usability method’s results.

In this paper the following research questions are addressed: Are the CTA and RTA method comparable in terms of:

*Overall effectiveness* in detecting usability problems with regard to *time subjects spent* per method?

*Thoroughness* in detecting usability problems of a minor, moderate or high severity?

*Types* of usability problem detected?

### Background Test Object; ICU Data Query tool

In 1996 the Dutch National Intensive Care Evaluation (NICE) foundation started a registry collecting data on patients admitted to Dutch Intensive Care Units (ICUs). A data set of about 100 items on demographic, physiological and clinical variables is collected for each individual patient. Based upon this data several prediction models such as APACHE can be calculated to correct measures of crude hospital mortality for illness severity at admission [20]. The registry aims to detect differences and trends in quality and efficiency of ICU care and provides quality reports and benchmarking information to its participating hospitals on a quarterly basis. In 2004 the request was made by participating ICUs to develop a tool to support ICU management and scientific reporting. To accommodate their request, a web-based application for clinical data querying was developed called ‘NICE Online’. NICE Online provided NICE subjects the opportunity to query the NICE database while protecting the privacy of patients and hospitals included in the Registry.

A standard software design cycle was applied in 2005 in the development project of NICE Online. A graphical interface was build providing the functionality (utility) of querying collected ICU data and the utility to benchmark the ICU data collected in the NICE registry. The user’s view on query commands was reduced to a custom designed webpage showing a structured query model (figure 1) to support clinicians in developing queries themselves.
Effectiveness of RTA and CTA

Figure 1. Screenshot of the Physician Clinical Data Query Tool: NICE Online. N.B. Text is translated from Dutch.

Nice Online use is limited to subjects having a user account. After logging into the system, ‘standard queries’ are presented to the user. The user can decide to select one of these ‘standard’ queries and choose to either directly view the query’s resulting table or graph or change the query by adding (data) elements to the query model. Another possibility is to start a new query, also called ‘custom query’ in the system. A user is then presented with a blank query model, in which he/she must add all (data) elements needed to generate the query. The query model consists of four components: functions, benchmarking/mirror, splitting/intersection, and selection of subpopulation. For each component the user can select from a large list of elements that present either statistical models or data elements in the NICE Registry. Functions allow the user to select the statistics of interest: for example the hospital or ICUs Standard Mortality Rate (SMR). With ‘benchmarking’ the user’s ‘own ICU data are compared to ‘national data’. The splitting/intersection and selection of subpopulation components offer the user the possibility to split the data in gender categories or to create a subpopulation with regard to for example a certain time period, as a few examples. When a user is finished with the query model, he/she can select the ‘graph/table’ button to create the resulting graph or table. The resulting graph or table has to be interpreted by the user himself. Thus background knowledge on statistics is required by users to accurately interpret the resulting data.
At the time of the study a limited number of NICE subjects had requested a user account for NICE Online. In July 2008, NICE Online registered 80 users of 26 of the 79 ICUs participating to NICE at that time. A log file analysis was subsequently performed to gain insight into NICE Online usage patterns. It showed that only 17% of the registered user accounts actually utilized the query functionality on a regular basis; more than 5 times by the quarter of a year. Next, a telephonic information needs analysis provided insight into reasons for the low adoption based on end-user experience with NICE Online. It indicated that users were willing to use the tool but the structured model for query development was considered difficult to use. However it did not become clear in what way the cognitive burden of a query development by use of the Tool was influenced by a potential lack in the Tool’s usability. In addition, planned expansions to the NICE Database required a high level of user-friendliness of NICE Online. This made it necessary to assess its usability and if necessary improve on its functionality. More information on the NICE online system can be found in [21].

**Materials and Methods**

**Subjects**

To select a number of representative study subjects, target users were categorized based on a log file analysis of NICE Online usage patterns. Eight experienced users having developed at least 30 individual unique queries in the Tool in the 12 preceding months, and eight new users holding a user account but without use experience with the Tool were randomly selected. For testing of the Physician Data Query Tool formal permission to contact the users was obtained from the NICE Foundation. The selected subjects were then contacted by email with an attached formally signed letter with the NICE foundation approval and the request to participate. All users agreed to participate to the NICE Online evaluation study. In total 16 subjects were contacted.

**Tasks**

Six usability testing tasks were developed, each composed of two to six subtasks. Input into the development of these tasks was given by two data managers of the NICE Foundation. The NICE data managers are skilled in ICU clinical query development and were able to provide relevant tasks of varying complexity with a gold standard for how to perform and finalize each task supported by NICE Online. The tasks were preceded by a query description, or short clinical question, which could be answered in NICE Online. The formal usability test started with two standard query tasks, randomly given to each subject. Then four tasks with a varying degree of complexity were randomly presented to the subject. These four tasks consisted of two custom query tasks and two tasks consisting of a
clinical question to be translated into a query in NICE Online. In the custom query tasks the subject had to enter query statements in line with the structured query model of the Query tool. In the clinical question tasks a descriptive scenario was given which the subject had to translate to a query in the Query Tool. An example of a custom query task and a clinical question query task is given in Table 1.

Table 1- Examples of the usability testing tasks

<table>
<thead>
<tr>
<th>Examples Main question</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Custom query task</strong></td>
</tr>
<tr>
<td>‘Please select’</td>
</tr>
<tr>
<td>1) the percentage of patient admissions</td>
</tr>
<tr>
<td>2) split by admission type (non surgical, elective surgical, urgent surgical)</td>
</tr>
<tr>
<td>3) for the data of your own hospital,</td>
</tr>
<tr>
<td>4) within a sub selection of the last two years.</td>
</tr>
<tr>
<td><strong>Translating a clinical question task</strong></td>
</tr>
</tbody>
</table>

Usability testing experiment

To characterize the subjects’ user profiles a questionnaire was handed out to each subject before the start of the usability experiment. This questionnaire contained eight questions concerning the subjects’ age, gender, and subjective measurements of their experience levels with computer usage, statistical data processing, and database query demand development (e.g. use of SPSS or Microsoft Access applications). Also the place (at home, at work, both) and manner of use (for research or management purposes) regarding computers, statistical data processing, and development of database queries was inquired. Users with similar profiles were evenly assigned to CTA and RTA in a matched randomized way.

The experiments took place in the actual clinical working area of the subjects. A portable usability laptop with TechSmith’s Morae software (Morae recording tool) allowed recording of all the subjects’ verbalizations in combination with a screen recording of their (mouse) actions in the system and a video recording of the subjects performing the actual tasks in NICE Online on the usability testing laptop. Subjects received the tasks in paper form and were given oral instructions on how to carry them out on the laptop. Before testing started, users were asked if they were right or left handed (for mouse configuration).
**CTA condition:** In the CTA condition, subjects were instructed to think aloud while performing the usability testing tasks. In line with the Think Aloud procedure described in [13], it was made clear that the accompanying researcher, the facilitator, would not interfere with the session by giving assistance. The facilitator would remind the subject to keep thinking aloud if the subject would fall silent. Prior to starting the test each subject in the CTA condition received think aloud training tasks to practice verbalizing their thoughts. The training tasks consisted of verbalizing the performing of simple computations.

**RTA condition:** In the RTA condition, the subjects received the same usability testing tasks in paper form. They were instructed to carry out the tasks in support of the Tool, without assistance of the facilitator (silent condition). Directly after subjects were finished, the silent test condition was stopped. Subjects then received the think aloud training task to practice thinking aloud. Subsequently, subjects were asked to verbalize their thoughts retrospectively while viewing the video recording of their actions in the system. In doing so they constructed the retrospective verbal reports. In the RTA condition the Morae recording Tool recorded two files: the user task performance while performing the tasks in the Query tool and the retrospective verbalizations of subjects reviewing their task performance.

**Data analyses**

All audio of the 16 subjects recorded with the Morae software was analyzed and transcribed to verbal protocols. Per subject two usability researchers independently analyzed the time spent in minutes by logging on the time-line of the Morae software. Time spent per subject was calculated as the time between starting a first task in the Tool until the moment of ending a final task in the Tool. The final task was considered completed when the subject mentioned to be finished performing a task (as instructed in the task description). For the RTA condition the time spent on retrospective verbalizations was additionally calculated in the same manner.

Based on the verbal protocols of the CTA and retrospective verbal protocols of the RTA a coding scheme for analysis of usability and task relevant verbal statements was developed. Six randomly chosen verbal protocol transcripts per test condition were first selected and coded. Subsequently all transcripts were coded based on the coding scheme and matched to corresponding video recordings analyzed by two usability evaluators. The same two usability researchers individually categorized the usability problem descriptions into problem types based on the work of Kushniruk et al [22]. Evaluators’ usability coding and categorization consistency was assessed using the Cohen’s kappa statistic. For each method condition a list of detected usability problems agreed upon was then generated. By union of these lists a total set (without redundancy) of usability problems found in the Query Tool was constituted. This overall list is considered to contain all ‘real’ usability problems.
existent in the query tool. Before further analysis could proceed, the average detection rate of usability problems per method was assessed. This measure indicates whether the small sample sizes assessed in the study were sufficient to cover the number of usability problems that potentially could exist in the Query Tool. Next, the frequency (the number of times a usability issue is mentioned) and persistency (the recurrence of usability problems in different tasks) of user experienced usability problems was assessed. Based on these analyses and the potential impact of a problem on task performance a severity rating of 1 (minor) to 4 (high severe) was given to each usability problem, as defined by Nielsen [23]. The severity of usability problems was then rated by two usability researchers and the head software engineer of the NICE Query Tool. To assess the inter rater reliability of three evaluators Fleiss’ kappa was measured [24]. To compare the methods’ performance first their ability to detect usability problems is assessed. Hartsen et al. researched how to balance the validity and reliability of different usability methods in a standardized manner. They assessed specific criteria which UEMs should be able to fulfill in order to be suitable for a specific usability study approach. Some of these criteria are combined in the measure of a methods’ usability problem detection effectiveness, where Effectiveness = thoroughness of the method * validity of the method [25]. In this study Effectiveness measurements are analyzed with reference to the time spent per subject in the testing.

Thorroughness refers to the ability of a UEM to find as many usability problems as possible when the user performs tasks with the evaluated system. To measure the thoroughness of CTA or RTA the proportion of the usability problems found by either CTA or RTA to the total of usability problems found by both methods in the Query Tool is assessed. Validity measures whether different UEMs, when compared, find problems that are real problems in system use. Validity of the CTA or RTA method is then measured by the proportion of real usability problems found by a method to the total number of potential usability issues that were found (in first instance) by use of the method. Pearson chi-square test is performed to assess whether a difference in detection effectiveness between methods is significant. In addition to the overall comparison of the methods’ effectiveness their thoroughness was weighted by problem severity levels (1 low to 4 high severe).

Then, the methods’ output in terms of type and severity of problems detected were compared qualitatively. In addition, the verbal protocols were qualitatively analyzed in-depth to explore further reasons for potential differences in verbal protocol output.

Results

General results
In Table 2 the user characteristics of the 16 subjects are given. Subjects in both method conditions were similar in age, profession and gender. The CTA analysis resulted in 38 potential usability problems of which two were excluded after discussion between the usability evaluators, leading to a total of 36 unique usability problems detected. The RTA found 36 usability issues of which five were not considered to impact the current usability of the Tool but were considered informative to redesign. After discussion between the evaluators a total of 31 unique usability problems were found to be detected by the RTA. After union of the CTA and RTA results in total 43 unique real usability problems were detected in the Query Tool. Inter rater reliability of the usability problem detection consistency was measured $\kappa = 0.93$, $p<0.001$. The 43 usability problems were categorized in seven usability types, $\kappa = 0.83$, $p<0.001$, showing almost perfect agreement among evaluators. Of the 43 usability problems, 20 were classified with a severity 3 or higher. Inter rater reliability of the severity classification of usability problems was $\kappa = 0.69$, $p<0.001$; showing substantial agreement. Table 3 provides an overview of the number of usability problems detected per method.

Based on these first results we additionally assessed whether the small sample sizes were sufficient to cover the potential number of usability problems that exist in the Query Tool at the moment of evaluation. We therefore computed the average detection rate of usability problems per method [26]. The average detection rate of usability problems in the CTA condition was 0.61 (sd 0.07); in the RTA 0.53 (sd 0.06). Usability literature states that for average detection rates higher than 0.45, 8 subjects per method condition provide at least 95% of the existing usability problems in a system [25, 26]. Based on these measured detection rates further analyses of the usability data output was considered reliable.

### Table 2. User characteristics of study participants

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>CTA (n=8)</th>
<th>RTA (n=8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean age (sd) in years</td>
<td>38.6 (sd 4.27)</td>
<td>42.3 (sd 6.96)</td>
</tr>
<tr>
<td><strong>Profession</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICU Physician</td>
<td>7 (87.5%)</td>
<td>8 (100%)</td>
</tr>
<tr>
<td>ICU manager</td>
<td>1 (12.5%)</td>
<td>-</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male, n(%)</td>
<td>7 (87.5%)</td>
<td>7 (87.5%)</td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>1 (12.5%)</td>
<td>1 (12.5%)</td>
</tr>
</tbody>
</table>

### Table 3. Overview of distinct usability problems detected per method and severity level.
Effectiveness of CTA versus RTA

Table 4 shows effectiveness of the two method conditions calculated by their thoroughness and validity. The validity of the RTA was considerably lower than CTA. The five usability issues detected but which did not affect subject task performance mainly caused this difference. In this study the CTA mounted up to 18% higher effectiveness in detecting usability issues compared to the RTA method. Additional Pearson chi-square test showed that the effectiveness of the CTA was indeed significantly higher than the effectiveness of the RTA in revealing unique usability problems ($\chi^2 = 4.54 \ (1) \ P=0.03$).

In both conditions the verbal protocols (concurrent or retrospective) were matched with the video recordings of subject task performance to detect and assess usability problems in NICE Online. To assess the effectiveness of the methods performance the average time spent per subjects in both method conditions was calculated. Table 5 gives an overview of the average time spent per subject on task performance and retrospective verbalization. During Task performance the RTA amounted to significant less time effort (22%) for subjects to perform the tasks in NICE Online compared to CTA. The Retrospective verbalization approximately doubled the time subjects spent on the usability testing of NICE Online. This leads to respectively 42% time saving of the CTA pertaining to the 18% higher effectiveness.

With regard to CTA and RTA thoroughness in detecting moderate and severe usability problems a variation in methods performance is also visible (See Figure 2). The difference in thoroughness was most prominent in usability problems of a severity 2 (15% higher effectiveness of CTA) and severity 4 (respectively $0.71$ vs. $0.57$; a 14% difference in higher effectiveness of CTA). The thoroughness of the RTA weighted per severity level was overall lower than the thoroughness of CTA.
Table 4. Thoroughness, validity and effectiveness per method condition

<table>
<thead>
<tr>
<th>Method condition</th>
<th>Thoroughness</th>
<th>Validity</th>
<th>Effectiveness (Th*Va)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTA (n=8)</td>
<td>(36/43) 0.84</td>
<td>(36/38) 0.95</td>
<td>0.80</td>
</tr>
<tr>
<td>RTA (n=8)</td>
<td>(31/43) 0.72</td>
<td>(31/36) 0.86</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Table 5. Average time spent per subject per method condition

<table>
<thead>
<tr>
<th>Time spent</th>
<th>RTA n=8</th>
<th>CTA n=8</th>
<th>CTA vs RTA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td>(min, sec)</td>
<td></td>
<td>(Min, sec)</td>
</tr>
<tr>
<td>- Task performance</td>
<td>41.07</td>
<td>7.49</td>
<td>50.16*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>+ 22%</td>
</tr>
<tr>
<td>- Retrospective verbalization*</td>
<td>46.05</td>
<td>5.62</td>
<td>-</td>
</tr>
<tr>
<td>Overall time spend per participant</td>
<td>87.12</td>
<td>-</td>
<td>50.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- 42%</td>
</tr>
</tbody>
</table>

* During retrospective verbalization a subject was able to stop the recording to verbalize his thoughts. Morae recorded the total time of reviewing the video potentially leading to a longer time on the retrospective verbalization than the actual task performance.

Figure 2. Thoroughness of the CTA and RTA method as a function of the severity
Type of usability problem detected by CTA versus RTA

Figure 3 gives an overview of the distribution in problem types of overlapping and unique usability problems found in the CTA and RTA analyses. Unique usability problems detected per method signify that CTA and RTA usability problem detection scope differed. CTA showed higher sensitivity to usability issues related to the Query Tool’s graphics/symbols, navigation issues, error messages and the layout and organization of the information on the Query Tool’s screens. Of these usability issues detected only by the CTA, three problems were of a severity 4 and three of severity 3 (see Table 6). RTA revealed two unique problems related to understandability of graphics/symbols used in the query tool to generate a query (severity 3). Furthermore, in the RTA condition 4 unique usability problems on terminology and meaning of labels (of which two of severity 4) were revealed. These problems first signified a mismatch between the user (intensivist’s) language and the terminology implemented in the Query Tool. For example, to generate the desired query specific data elements had to be selected by use of search functionality or by browsing through a list of labeled elements. Naming of these labels, such as ‘gender, admission type’ was partly derived from the NICE data-dictionary which contains all data-elements collected by the NICE registry. In both CTA and RTA testing, subjects selected incorrect data elements in several scenarios to generate the queries. During CTA all subjects mentioned the usability problem of incorrect ordering of elements and the organization of the elements on the screen. However, in RTA subjects mentioned that they were unable to deduce the proper use of an element in the query design by means of their label. It appeared that data elements in the Query tool were labeled on differing levels of cognitive complexity; from single elements such as ‘gender’ to more complex elements such as ’age group per ten years’ (which supports data analysis of patients over periods of ten years). This added to the cognitive complexity of developing a query in the Tool and was remarked upon by seven subjects in the RTA condition. In addition, subjects could select the same type of data elements under multiple components of the query model in the Tool; ‘age group per ten years’ could be selected under ‘splitting/intersection as well as under ‘mirror’. To comprehend how either selection impacts the resulting figure of a query, direct result review becomes a requirement, which was not supported by the Query Tool.
Figure 3. Graphical overview of Type of problems detected per method.

Table 6. Distribution of usability problems per method by problem type and severity

<table>
<thead>
<tr>
<th>Problem Type</th>
<th>Total</th>
<th>CTA/ RTA overlapping* (n=24)</th>
<th>CTA Unique (n=12)</th>
<th>RTA Unique (n=7)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Severity</td>
<td>Severity</td>
<td>Severity</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>CTA</td>
<td>RTA</td>
<td>1</td>
</tr>
<tr>
<td>Visibility of system status</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Overall ease of use</td>
<td>7</td>
<td>7</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Error messages/ help instructions</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Meaning of labels/terminology</td>
<td>4</td>
<td>8</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Layout/screen organization</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Graphics/ symbols</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Navigation</td>
<td>5</td>
<td>3</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>36</td>
<td>31</td>
<td>3</td>
<td>11</td>
</tr>
</tbody>
</table>

*overlapping = problems detected in both methods of similar nature
In-depth qualitative analysis of CTA and RTA verbal protocols

Furthermore CTA and RTA verbal data showed additional differences in usability problem detection. First of all, though CTA consisted of a higher number of verbal protocols, RTA verbal protocols proved more explanatory towards the more severe usability problems. More precisely, the RTA verbal comments provided additional insight in underlying causes of the occurrence of these severe problems. Subjects in the CTA condition who did not fully comprehend what the term splitting/intersection meant in the query model, showed irritation and commented upon their lack of understandability of the Query Tool terminology. RTA subjects however commented upon the terminology to be ambiguous and provided synonyms to splitting/intersection they perceived would describe it more clearly. Also, subjects who were already familiar with the use of Query Tool prior to usability testing provided in the RTA condition highly detailed information when confronted with a problem they were already familiar with. Usability issues they had experienced before in using the Tool had thus been already thought over by them. During the RTA testing they would specifically address the issue showing how important they thought the problem actually was. In doing so they provided solutions to improve on detected usability issues. Next to solutions the more experienced subjects in the RTA also provided sight on five usability issues which were classified as additional user recommendations for extended system functionalities (utility) of the Tool. For example, one of these recommendations was related to use of the Tool for a deeper scientific analysis of patient cohorts.

The analysis of the number and type of usability problems detected in the CTA and RTA showed that the mental model of subjects, defined by Nielsen as “what the user believes about the system at hand”, did not match the query design model in the Tool [27]. In both CTA and RTA subjects’ expectations and believes on query design in the tool, resulted in severe problems in task performance. Subjects who were not familiar with the Tool as well as experienced subjects encountered the same type of usability issues. This indicated that experienced users had not changed their mental model of query development in the Tool based on their learning experience with the Tool. To extract the mental model of subjects in querying data in the Query Tool their coded articulations on how they intended to solve the usability testing task in the system was further analyzed. The CTA verbal protocols proved more informative for this task. They contained a higher number of verbalizations, provided more insight into usability problems detected and provided more verbalizations on direct task performance compared to the RTA protocols. Therefore the coded intensivist’s verbalizations on query formulation were first modeled for the CTA condition in accordance with their information search behaviour. The RTA coded verbal transcripts were subsequently compared to the resulting query design model to assess differences between subjects information search behaviour in both methods. Results of this analysis showed that overall subjects approached the solving of the Query in the testing tasks in the same convergent manner, for both simple to more complex queries.
The Query model in the Query Tool did not support the query solving approach of subjects. Instead of ‘function’, ‘splitting’, ‘intersection’ and ‘mirror’, subjects started from the basis of what they were supposed to measure, to which data elements they needed to compare or include in the query. Only when they had sight on the data they were interested in, the last phase of the query design model was approached: benchmarking of the data. The CTA results, presented in this study, provided sight on the system requirements and were used as input in redesign of the Query Tool to match its users’ mental model.

**Discussion**

This study compares the performance of the CTA and RTA method in a usability case study of an ICU Query Tool. The analysis revealed that the effectiveness of the CTA in general and its thoroughness for detecting both moderate and severe usability problems was higher than that of RTA. The methods also differed in their detection scope of usability problem types. The CTA condition in this study in general detected 12 more unique usability issues on graphics/symbols, navigation issues, error messages, layout and organization of information on the Query Tool’s screen, overall ease of use and visibility of system status. The RTA in this study detected more usability problems related to unclear system terminology for users. In addition, RTA offered insight into new system functionalities.

From the field of psychology certain studies provide consistent results to ground this research. Firstly, Kuusela and Paul compared verbal protocol segments of the CTA and RTA approach. They showed that the CTA condition produced a higher number of protocol segments than the RTA condition and elucidated on this difference to be caused by the cognitive nature of the CTA [28]. In contrast to RTA the CTA method appears to evoke subjects to verbalize their short-term memory (STM) task-related thoughts. As verbal protocols in CTA are collected during task performance this leads to a high number of protocol segments. Our study adds to this, by indicating that the difference in cognitive focus of the methods leads to a difference in usability focus. During CTA the verbal protocols reflect STM thoughts on task performance thereby leading to a higher effectiveness of the CTA in detecting usability issues related to the direct user task performance. As a result, usability evaluation studies using only the RTA method may be less able to detect usability problems that make instant requests of the STM of subjects. However, the in-depth analysis of the verbal protocols of the RTA showed a merit of using RTA in formative usability evaluation for the purpose of system redesign. RTA verbal statements consist of a higher rate of user recommendations and are of a more explanatory nature than the CTA. This is in line with prior research of Bowers and Snyder in which they state that RTA verbal protocols consist of more explanations and design statements [16]. However, our research also indicates that subject experience levels could be of influence on this result. In this study, the subjects experienced with NICE Online, and with expertise in clinical data analysis, prior to TA testing were
more informative in providing design solutions and in stating new requirements for additional functionalities of the Tool. Research on expert performance in many domains such as chess and computer programming; have revealed maximal adaptations of experts to domain-specific constraints [29]. Experts input into usability tests might thus differ considerably compared to novice users. However, because of the small number of experienced subjects in the RTA condition a conclusion on this cannot be drawn. Further research into subject computer and system experience on usability testing results of RTA and CTA is therefore required. In this study, even distribution of new and more experienced users in the methods’ conditions prohibited potential bias of subject expertise level on Effectiveness measurements.

Compared to CTA, the retrospective verbalization of RTA may thus evoke a more rational and reflective report on task information processing behaviour as the computer supported task has already been completed and can no longer be changed. These results seem to indicate that though RTA gives sight on unique issues regarding user terminology it might be better equipped for revealing the underlying cause of a usability problem or new user requirements for specific user groups or customized design.

Next to the results of the CTA and RTA comparison, this research also addresses the complexity of clinician querying of large computerized repositories of clinical data variables. This is as far as we know the first usability test on the cognitive complexity of query design by clinicians themselves in an online Query Tool. The foremost problems of the NICE Online tool concerned the comprehensibility of Query model in the Tool and its applied terminology. This mismatch between the Query Tool’s Query model and the mental model of end-users data querying inhibited the system’s convenient use. From the results of the CTA and RTA test a new query model was developed. This new query model could be extracted from the verbal protocols. Additionally, several user requirements were revealed by this study. Among these was the need to provide clinicians the ability of an overall view on all data items in complex data queries combined with direct review of the query result. Following the results of this study the Physician Data Query Tool has been redesigned. A subsequent pre-post study assesses the effect of the Think Aloud testing on the usability of the Query tool’s redesign. In this study the differences in the cognitive complexity of data querying in the old versus the new tool and its effect on the mental workload of data querying by clinicians in the new NICE Online are assessed.

To conclude; though this study indicates that CTA is more effective in usability problem detection the method does not outperform RTA. RTA additionally provides sight on unique usability problems and is of a more explanatory nature than CTA. Based on the results of this study we recommend the use of CTA in formative usability evaluation studies of health information technology. CTA enables the researcher to assess the cognitive processes of subjects during task performance, thereby providing insight into the mental model of users interacting with the healthcare application. The RTA in this study seems to lack the potential to assess the cognitive processes in direct task performance; signified
by less usability output. However, based on the high explanatory value of RTA verbal protocols we recommend RTA in usability studies focusing on user customized (re)design.

Acknowledgements

The authors thank the NICE Foundation for their assistance and support in performing this study. Also all NICE participants who acted as subjects in this study are thanked for their input. We are specifically grateful to the chief software engineer of the NICE registration, Eric van der Zwan, for his suggestions and continuous support in performing this study. The authors also acknowledge Bas de Veer for his contribution as usability evaluator in the performance of the usability testing and subsequent analyses. We thank Joost Dusseljee and Sylvia Brinkman for their input in conducting this study as NICE data managers.

REFERENCES


Classification and Prioritization of Usability Problems Using an Augmented Classification Scheme

L.W.P. Peute, R. Khajouei, A. Hasman, M.W.M. Jaspers

*Journal Biomedical Informatics. 2011 Dec;44(6):948-57*
Abstract

Usability evaluation studies in health care typically classify usability problems in terms of their type, number, and severity. These classes are usually devised by the evaluator for the purpose at hand and the used problem types often are not mutually exclusive, complete and distinct. We investigated whether the UAF classification framework augmented with severity ratings of usability problems and assessments of their potential impact on final task outcomes could improve the classification, analysis, and prioritization of usability problems. To investigate the merits of such an augmented scheme, a usability case study of a computerized physician order entry system (CPOE) was performed using the Cognitive Walkthrough and Think Aloud usability evaluation methods. All 57 usability problems detected in the case study could be classified with the help of the augmented classification scheme into exclusive paths through the hierarchy of the UAF belonging to different phases of user interaction: planning, translation, physical action and assessment. We could show that this classification scheme enables usability evaluators to cluster problem descriptions that share common characteristics and that it addresses the underlying cause of the usability problems. It helps differentiate problems that look similar but yet affect the user-system interaction and the task results differently. Moreover the evaluators classified the majority of the usability problems identically. This work is of value to other researchers who intend to report the results of their usability evaluation studies.

**Keywords:** Classification, User-Computer Interface, usability, Cognitive Walkthrough, Think Aloud, Medical Order Entry Systems
Introduction

Health care information systems have the potential to support clinicians in making clinical decisions and hence to improve patient safety [1] but clinicians’ reluctance to use these systems in daily clinical care has become a well known dilemma [2]. Studies have shown that usability problems are among the factors negatively affecting system’s acceptance [3;4] and limiting their effectiveness in supporting and streamlining clinical care [5-8]. Usability is a quality attribute that assesses how easy user interfaces are to use. Usability problems can be defined as aspects of a user interface that may cause the corresponding system to have a decreased usability for the end user.

So far, research on health care information system usability evaluations has focused on the identification of specific problems that compromise effective, efficient and safe use of the systems [7;9-11]. Evaluators use usability evaluation methods (UEMs) to detect usability problems. Whenever a problem is observed the evaluator usually writes down a usability problem description in an ad hoc manner, expressed in whatever terms seem appropriate to the evaluator. It is not unusual for two observers to write substantially different descriptions of the same problem. Therefore the evaluators have to cluster problem descriptions into unique problem types (bottom-up analysis). The determination whether different usability problem descriptions are referring to the same underlying usability problem is usually done by expert judgment, and is subject to much variability.

An alternative is to classify problems in terms of recognized usability principles, the so-called heuristics [12] (top-down approach). However, it has been shown that various difficulties arise when these heuristics are used as a classification scheme due to their incompleteness, lack of mutual exclusiveness, and lack of specificity [13]. Heuristics are therefore not sufficiently comprehensive to be used for the systematic classification of usability problems. Moreover, this heuristic classification does not provide insight in the underlying causes of the usability problem types reported. As a consequence, detected usability problems might be typified as similar, although the underlying causes could be fundamentally different. It is clear that there is a need for a standard way to describe usability problems and a framework with which usability problem descriptions can be more easily and more directly compared. Severity ratings of usability problem types are commonly based on Nielsen’s classification [14] grounded on the proportion of users who (will) experience a specific problem, the impact it (will) have on them, and whether the usability problem will be a problem only the first time they encounter it, or whether it will persistently bother them [15;16].

Usability problems not only bother users during interaction with the system but also have an impact on their task performance. Therefore, these problems can be a common source of errors [8], potentially compromising patient safety. Hence, in prioritizing redesign efforts the potential of usability problems to evoke user errors, for example leading to wrong medication orders, should be taken into account.
However, usability problems are typically reported in terms of type, number, and severity without considering an evaluation of the impact on the task performance.

Consistent and detailed reporting of usability problems, their underlying causes, their impact on the user and their potential effect on the final outcomes are not only needed for guiding and prioritizing redesign efforts but in addition are prerequisites for classifying usability studies. This allows researchers and system designers to learn from the errors made by their colleagues, because they are able to retrieve the relevant knowledge using the classification of the problem type they are interested in.

To minimize individual differences in reporting a framework that supports describing usability problems is needed. Such a framework should contain a complete set of the attributes applicable to a problem type, and distinguish a problem of one type from a problem of another type. The User Action Framework (UAF) is such a framework developed by the usability engineering field. The UAF is a four level classification. This framework only supports the coding of problem types. It does not include a severity rating or an assessment of the potential impact of usability flaws on the final task outcomes.

Therefore the aim of this study is to explore whether the UAF classification framework augmented with severity ratings of usability problems and an assessment of their potential impact on final task outcomes could be of value for classifying, analyzing, and prioritizing these problems in a generic way. To explore this research question, usability problems identified in a usability evaluation of a working CPOE system were used as input. It was investigated whether the framework in the hands of evaluators leads to the same problem classifications. Also the advantages of the four level UAF classification over existing shallower classifications were studied.

The paper is organized as follows. Section 2 provides an overview of the methodology that we followed to develop the augmented classification scheme. Section 3 presents the results of applying this scheme to a CPOE system. Section 4 discusses the findings and the strong and weak points of the study. Section 5 concludes the paper.

Methods

Augmented classification scheme

We propose that a classification scheme for usability problems in the domain of health care should satisfy the following requirements. The scheme should

- rely on theory from the domain of usability engineering: a classification scheme should include a robust taxonomy that allows the assignment of the problems to the different stages
of user interaction and describes the trend by which a design flaw of a system resulted in a usability problem (cause of the problem). This scheme should allow consistent problem reporting and the tracking the usability problems through successive stages of iterative design.

- rank the effects of the problems on the extent of physical and cognitive effort they require from the user: Users in the health care domain are very busy and often have to take care of different issues simultaneously. Problems that severely hinder users during interaction should be distinguished from problems that have a small effect on the interaction.
- address the potential effect of the problem on task outcome: As the impact of usability problems may threaten the health of people the classification scheme should be customized to the health care setting and address the impact of the problems on clinical task outcome.

The classification should therefore help to prioritize the problems based on the extent of physical and cognitive effort required from the users and on the impact of the problems on task outcome.

Figure 1 shows the proposed augmented scheme for classifying and prioritizing usability problems in the domain of health care. In the following section the development of this scheme is described. We applied this scheme to investigate whether two evaluators classified and prioritized usability problems identified by a cognitive walkthrough (CW) [10] and a think aloud (TA) usability testing of a CPOE system [17] identically.

![Figure 1. The augmented scheme for classifying and prioritizing usability problems](image-url)
The UAF classification

In the field of Usability Engineering a standardized classification of usability problems having various dimensions is considered essential for accurate, complete and consistent problem reporting and for identifying the underlying cause of the problem. One classification scheme which is developed and applied in usability research is the User Action Framework (UAF) [15]. The UAF was built by adapting and extending Norman's theory of action model [18], a model that highlights issues about the way people interact with machines in terms of cognitive and physical user actions. The UAF classification provides insight into what the users think, perceive and do throughout each cycle of interaction with a computer system. The purpose of the interaction cycle is to model the flow of user interaction in any interactive system. The UAF interaction cycle contains four phases; planning (high level), planning (translation), physical actions and assessment.

Planning is the part of the interaction cycle that contains all cognitive actions by users to determine what to do. Supporting users in planning involves helping them understand the system model and helping them keep track of where they are within a task. High-level planning focuses on the system model and metaphors and the user’s knowledge of the system states and modalities. Planning includes user work goal decomposition across a hierarchy of plan entities: the user establishes a goal, decomposes the goal into tasks to be performed on computer and establishes an intention of what to do to accomplish the task.

Translation is the part of the interaction cycle that contains all cognitive actions by users to determine how to carry out the intentions that arise during planning. In the translation phase, a user specifies the action sequence and determines which physical actions have to be executed in order to accomplish an intention, translating intentions into plans for physical actions. Cognitive affordances (e.g. visual clues) support the users' ability to plan physical actions. Therefore, usability issues concerning translation include those that pertain to the system’s cognitive affordance presentation, and their content or meaning.

The physical action part of the User Action Framework is about executing the actions by manipulating user interface objects, and includes issues of interaction complexity and styles, manual dexterity, and layout. In graphical user interfaces, physical action mainly involves clicking, dragging, and selecting.

The assessment part includes a user perceiving, interpreting and evaluating the resulting system state. It concerns issues about the existence, presentation, contents and understandability of feedback and how it supports the user's ability to assess whether the outcome of physical action was desirable or effective. [15;16].
The UAF provides a quasi-hierarchical tree of usability concepts and issues organized around the users’ cognitive and physical actions, structured into four levels of abstraction, making it possible to classify and report usability problems from general towards more specific perspectives [15]. To each phase of the UAF interaction cycle one or more usability issues are associated with mutually exclusive standardized usability attributes or sub-categories below each issue. A usability problem can be classified using two or more levels of the UAF hierarchy. The UAF is viewed as a standard way to normalize usability problem descriptions [19] and is known to provide a highly reliable means for a detailed classification of usability problems.

Figure 2 presents the path for classifying a usability problem by usability evaluators using the framework. This problem concerns a user who does not notice the buttons (“m2” or “kg”) on the main screen, provided for calculating the medication dosage (Figure 3).

Usability evaluators first investigate in which phase of the interaction the problem occurred. In this example the problem occurs after the user planned to calculate the dosage that has to be entered in the dosage entry field. Therefore, the evaluators conclude that the problem does not occur in the planning phase but in the translation phase of the interaction. To translate his intention of calculating the dosage into a physical action, the user needs physical cues. In this example the user does not see the existing buttons for this purpose. Next, the evaluators try to map the problem to the different subcategories of the translation phase, first for level 2, then for level 3 and 4. The framework provides the standardized (sub)categories that the problem could belong to. For instance, one of the subcategories of translation phase at the second level is “existence” (Figure 2). The evaluators check whether the buttons needed to perform the physical actions exist on the screen. Since in our example the buttons exist on the screen, the problem does not belong to this subcategory. Evaluators continue comparing the problem with the other attributes at this level. In this example the problem is not concerning the “meaning” of the buttons because as soon as a user sees the buttons he can understand the meaning of the labels. The problem is that the buttons are presented in such a way that they may be overlooked by a user. After finding the relevant category at level 2 (presentation) the evaluators try to map the problem to the subcategories of this category. This process continues until the problem cannot be mapped any further. In this example it is difficult to observe the buttons, which is a “perceptual issue”. And finally the reason that the user may overlook the buttons (underlying cause of the problem) is that the user may not notice them because of their small size, color close to the background, and the relatively ‘large’ distance from the dosage entry field.
Figure 2. The path for classifying the usability problem presented by Figure 3 (the UAF (sub) categories belonging to this path are shown in gray with bold arrows)

* only categories belonging to the “translation” phase of the UAF are presented

Figure 3. Screenshot of a CPOE showing hardly noticeable buttons for the calculation of the medication dosage
Two usability evaluators with more than 5 years of usability expertise subsequently classified all identified usability problem descriptions that resulted from two previous usability evaluations of a CPOE system by traversing the UAF decision structure and selecting the most appropriate classification category and sub-category at each level of the hierarchy. Each usability evaluator analyzed the problem description using the classification instruction [16] and the clinical scenario used for conducting the usability evaluations. Usability problems were then coded via the UAF hierarchy to the most detailed level. The resulting sets of usability problem classifications were reviewed in a meeting and disagreements were discussed. Any remaining disagreements were resolved through discussion with a third evaluator.

Severity rating of the usability problems.

The severity of usability problems were rated based on Nielsen’s classification [14] (Figure 2). In one case study, Nielsen showed that the probability of getting a severity rating within ±0.5 rating unit from the true severity of a problem on a 5-point rating scale was only 55% with a single usability evaluator, but 95% for the mean of ratings of four independent evaluators [20]. He recommended collecting ratings from at least three evaluators. Therefore, in this study the severity rating of each problem type was determined by consensus of three usability evaluators. For determining the severity rating, the frequency with which a problem (might) occur(red), the (potential) impact of the problem on the users and the (potential) persistence of the problem are taken into account [21].

Determining the potential effect of the problems on task outcome

The two usability evaluators reviewed all usability problem descriptions and the corresponding system states at each moment in the user system interaction to determine what the potential effect of a usability problem could be on the ordering outcome. The potential effect of usability problems on ordering outcome was classified (for this computer application) as wrong medication name, dosage, frequency, duration, and route of administration.

Results

In total 57 usability problems identified in previous usability evaluation studies were classified by the evaluators. The UAF classification of the usability problems together with each problem’s severity and the potential effect of each problem on the ordering outcome were determined (Tables 1-4). Two (4%) of these problems could be classified up to the second, 16 (28%) up to the third and 39 (68%) up to the fourth level of the UAF hierarchy. At the first level of the UAF, 55 out of 57 usability problems were classified identically by the two evaluators. One of the faulty classified problems classifications concerned the lack of an undo button to be able to see or change the result of previous user actions.
Since this usability problem occurred when a user wanted to translate her intention of undoing a step into a physical action, this problem should be classified under translation. Since an undo button or an active link were missing, one of the evaluators classified this problem under the assessment phase as (lack of) “existence” of information. The other faulty classified problem concerned the situation where a user did not recognize the possibility to enter a specific dosage in the drop down-menu of an alert screen. This problem also occurred when the user searched for a way to change the dosage. It should therefore have been classified under translation. As the user did not notice the function of an existing field, one evaluator wrongly considered this problem as a faulty presentation of information related to the assessment phase of the UAF. At the other levels of the UAF three problems (two at level 3 and one at level 4) were classified differently. These disagreements were mostly caused by different interpretations of the details of a problem by one of the evaluators. This evaluator had not participated in the usability evaluation and was not aware of the context in which the usability problems occurred. There was no disagreement at these levels when the evaluators looked at the interface of the system for clarification. All usability problems could be classified by the evaluators using several levels of the UAF hierarchy and the evaluators agreed in more than 90% of the cases.

Planning

Of the total number of identified usability problems six (10%) problems were found relating to this phase of user-system interaction. Classification of the usability problems showed that usability problems in this phase of interaction were caused by the user’s difficulties to understand the overall CPOE system model, system state or modalities and the users’ inability to keep track of the ordering steps that were completed and to determine the next goal to be accomplished. Several users had for example difficulties in remembering which steps they had completed, and what next steps they had to carry out to enter a medication prescription.

The severity of problems concerning the planning phase was low except for one problem that could potentially result in wrong medication durations. Table 1 shows that problems with a similar severity and description can have different underlying causes when they are categorized with the UAF. For example one of the problems was that it took a while before the users entered a start date for a medication, and another problem was that it took a while before the users started the ordering process. These two problems look similar based on their description and severity, but when they were classified using the UAF framework the underlying cause of these problems appeared to be different. In the first problem when the users decided to order medication using order sets, they could not decompose their goal into the tasks required to proceed with ordering. In the other one the users already determined the task (start ordering process) but as the model of the system differed from what they expected they could not decide which action to take to accomplish the task.
### Table 1. Example of a usability problem classification in the UAF\(^a\) planning phase with its severity and potential effect on task outcome

<table>
<thead>
<tr>
<th>UAF categories</th>
<th>Example</th>
<th>No. (Severity)</th>
<th>Potential effects on outcome (^b) (No. problems)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User’s ability to keep track of how much is done</td>
<td><em>CW and TA:</em> Users could not infer from the system state how much of the ordering task was achieved.</td>
<td>1 (2)</td>
<td></td>
</tr>
<tr>
<td><strong>Level 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Users knowledge of system state, modalities</td>
<td><em>TA:</em> User did not know that system does not allow the opening of “Patient” menu during ordering and tried to open it.</td>
<td>1 (1)</td>
<td></td>
</tr>
<tr>
<td><strong>Level 3</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goal decomposition</td>
<td><em>TA:</em> Users could not immediately infer from the screen ‘Order set’ that they had to enter a start date for a medication in order to proceed with the ordering process.</td>
<td>1 (1)</td>
<td></td>
</tr>
<tr>
<td>Supporting human memory limitations</td>
<td><em>TA:</em> Users forgot to enter start and stop times. The layout of the start and stop time entry field did not attract the attention of the users.</td>
<td>1 (3)</td>
<td>Wrong medication duration (1)</td>
</tr>
<tr>
<td><strong>Level 4</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Users model of the system</td>
<td><em>CW:</em> From the layout of the main screen, Users can not instantaneously find out how to start the order process.</td>
<td>2 (1,1)</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)User Action Framework

\(^b\)Potential effect(s) of given example on outcome(s) are shown in italic bold

### Translation

Sixteen (28%) of the usability problems concerned this phase of the interaction. The UAF classification revealed that usability problems in this phase were mainly caused by the lack of certain functionality so that certain tasks cannot be performed or, if the functionality is available, the lack of cognitive affordances showing how to successfully perform a task. Also poor presentation and design, poor content and meaning of available cognitive affordances may cause problems. Mismatches between the users’ terminology and the terminology, abbreviations and labeling of buttons used in
Mediator also were a source of usability problems. Seventy five percent of the problems concerning this phase of interaction had a minor severity (severity 2). Although their severity was similar, their underlying cause was different and these problems were categorized under different subcategories of the translation phase of the UAF. Classification of the problems with the augmented scheme showed that some of the problems would get a low priority if they were classified based on their severity only while these problems will be given a high priority if also their impact on the task outcome is considered for prioritization. For example, of the five problems leading to a medication error, the severity of three problems was rated as minor.

Table 2. Example of a usability problem classification in the UAF\(^a\) translation phase with its severity and potential effect on task outcome

<table>
<thead>
<tr>
<th>UAF categories</th>
<th>Example</th>
<th>No. (Severity)</th>
<th>Potential effects on outcome(^b) (No. problems)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 2</td>
<td>Content and meaning Clarity, precision and predictability Completeness and sufficiency of meaning</td>
<td>2 (2,2)</td>
<td>Wrong medication selection (2)</td>
</tr>
<tr>
<td></td>
<td>CW and TA: Users did not understand the meaning of some abbreviations used in the medication list.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 3</td>
<td>Naming of labels Cognitive directness</td>
<td>2 (2,4)</td>
<td>Wrong medication dosage (1)</td>
</tr>
<tr>
<td></td>
<td>CW and TA: Users could not infer the functions of two buttons &quot;change&quot; and &quot;record&quot; based on their labels.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 4</td>
<td>Existence of a way</td>
<td>1 (2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CW and TA: Users could not undo an action in the system and change a previously made selection in the ordering process.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Existence</td>
<td>Cognitive affordances</td>
<td>3 (2,2,2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TA: Users expressed that they needed help information to know which consolidation (1, 2 and 3) is suitable for the patient.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presentation</td>
<td>Preferences and efficiency issues</td>
<td>2 (2,3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CW and TA: Users preferred to select from an alphabetically organized list of order sets rather than checking all items in a non</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
organized list.

<table>
<thead>
<tr>
<th>Perceptual issues</th>
<th>Noticeability</th>
<th>CW and TA: Users did not notice the buttons provided for calculating medication dosage on the main screen.</th>
<th>Wrong medication dosage (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>4 (2,2,3,3)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Discernability</th>
<th>CW: Users cannot comprehend the function of the button “new order” from its shape.</th>
<th>1 (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task structure and interaction control</td>
<td>Consistency and compliance of task structure</td>
<td>CW: The labeling of menus and submenus provided for initiation of an order is not consistent with the labeling of buttons provided for the same purpose.</td>
</tr>
</tbody>
</table>

\(^a\) User Action Framework

\(^b\) Potential effect(s) of given example on outcome(s) are shown in italic bold

### Physical actions

Nine (16%) of the usability problems were encountered in this phase of the interaction. The UAF classification showed that these usability problems were caused by faulty presentation and lay out of the physical objects to be manipulated on the screen, lack of user control over screen objects as they were being manipulated and failure of the system to meet specific preferences of users for doing physical actions. Based on severity 78% of the problems in this category were similar (rated as severity 2) while the causes of these problems were different as was apparent from the assignment to different subcategories of the UAF. One of the problems, which was rated as severity 2 based on its frequency, persistence and impact on the user, could lead to a wrong medication selection and, thus, should get a high priority for fixing.
### Table 3. Example of a usability problem classification in the UAF<sup>a</sup> physical action phase with its severity and potential effect on task outcome

<table>
<thead>
<tr>
<th>UAF categories</th>
<th>Example</th>
<th>No. (Severity)</th>
<th>Potential effects on outcome&lt;sup&gt;b&lt;/sup&gt; (No. problems)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 1</strong></td>
<td>Perceiving objects as they are being manipulated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceiving physical</td>
<td><em>CW and TA: Different behavior of the button “select” in the patient selection window than that expected by users.</em></td>
<td>1 (2)</td>
<td></td>
</tr>
<tr>
<td>objects</td>
<td>Discernability</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Level 2</strong></td>
<td><em>CW: Users cannot perceive in which format they should enter patient’s date of birth.</em></td>
<td>1 (2)</td>
<td></td>
</tr>
<tr>
<td><strong>Level 3</strong></td>
<td>Manipulating objects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical control</td>
<td><em>CW and TA: User could not open tab motivation text to enter the motivation for an order.</em></td>
<td>2 (2,3)</td>
<td></td>
</tr>
<tr>
<td><strong>Level 4</strong></td>
<td>Physical layout</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proximity issues</td>
<td><em>TA: User by accident clicked on a wrong option in the medication order set list.</em></td>
<td>2 (2,2)</td>
<td>Wrong medication selection (1)</td>
</tr>
<tr>
<td>Preferences and</td>
<td><em>CW: User cannot review the system’s calendar year by year to set date of birth</em></td>
<td>1 (2)</td>
<td></td>
</tr>
<tr>
<td>efficiency</td>
<td>Preferences and efficiency</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>CW and TA: Users could not directly enter the number of days for medication duration. Instead, Medicator requires the entering of a start and stop date.</em></td>
<td>1 (4)</td>
<td>Wrong medication duration (1)</td>
</tr>
</tbody>
</table>

<sup>a</sup> User Action Framework

<sup>b</sup> Potential effect(s) of given example on outcome(s) are shown in italic bold
Assessment

In total 26 (47%) of the 57 identified usability problems were classified in the assessment phase. These problems concerned the existence, presentation, content and meaning of system feedback about the course of the user-interaction and the display of information resulting from users’ actions. Around half (12 out of 26) of the problems concerning the assessment phase were severe problems (severity 3 and 4) and they could influence the task outcome. Not all the problems influencing the outcome were highly severe problems since three of the problems potentially resulting in ordering errors were assigned severity 2. The UAF classification showed that 19 (73%) of the problems concerning the assessment phase of interaction were caused by a lack or suboptimal design of system feedback or by unclear feedback contents. The remaining seven (27%) problems in this phase were caused by the absence, unclear contents and meaning, and poor presentation of information displayed after the users’ actions.

Table 4. Example of a usability problem classification in the UAF assessment phase with its severity and potential effect on task outcome

| UAF categories | Example | No. (Severity) | Potential effects on outcome
<table>
<thead>
<tr>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
<th>(No. problems)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feedback</td>
<td>Content and meaning</td>
<td>Completeness and sufficiency of meaning</td>
<td>CW and TA: Users could not understand the function of column “dosage percentage” in the medications table.</td>
</tr>
<tr>
<td></td>
<td>Cognitive directness</td>
<td></td>
<td>TA: Users did not understand the recommendation provided by the alert “medication dosage-unit control” correctly.</td>
</tr>
<tr>
<td></td>
<td>Error avoidance</td>
<td></td>
<td>CW: User can enter frequency of medication in letters, but system alerts “no dosage [not frequency] is entered”. User should only enter numbers.</td>
</tr>
<tr>
<td></td>
<td>Existence</td>
<td>Existence of a cognitive affordance</td>
<td>CW and TA: Users do not receive a feedback or warning when they forget to enter stop date for a medication and can proceed to the next step.</td>
</tr>
<tr>
<td></td>
<td>Presentation</td>
<td>Perceptual issues&gt;</td>
<td>CW and TA: Alert screen “medication dosage-unit control”</td>
</tr>
</tbody>
</table>

Wrong medication selection (1), Wrong medication duration (4), wrong medication dosage (3), wrong medication frequency (1)
Timing shows up too late in the ordering process. 

<table>
<thead>
<tr>
<th>Information display</th>
<th>Human memory aids</th>
<th>CW: Users are not provided with the unit of calculated dosage in “Dosage calculation” windows.</th>
<th>2 (3,3) Wrong medication dosage (1), Wrong medication selection (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content and meaning</strong></td>
<td>Error avoidance</td>
<td>TA: User was confused when the system retrieved an alternative name of the medication than the one typed by the user.</td>
<td>1 (2)</td>
</tr>
<tr>
<td><strong>Layout and grouping</strong></td>
<td>-</td>
<td>CW: Independent of the administration time entered by users a different administration time is shown in a different section of main screen.</td>
<td>2 (2, 4) Wrong medication duration (1)</td>
</tr>
<tr>
<td><strong>Presentation</strong></td>
<td>Perceptual issues &gt; noticeability</td>
<td>TA: User did not notice that the administration time was already recorded and re-entered it.</td>
<td>2 (2,3) wrong medication duration (1)</td>
</tr>
</tbody>
</table>

*a* User Action Framework  

*b* Potential effect(s) of given example on outcome(s) are shown in italic bold

## Discussion

This study assessed whether an augmented classification scheme from the usability engineering field could be of value for classifying and prioritizing usability problems revealed in a CPOE usability case study. Examination of problems at the various levels of the UAF enabled us to identify meaningful problem clusters containing problems that shared common characteristics and hence revealed main usability issues at both more global and more specific levels. Moreover we could determine the underlying cause of the usability problems and the manner in which they influenced the user interaction by classifying problems across different levels of the UAF hierarchy. In addition the scheme in the hand of both evaluators led to the same classification in more than 90% of the 57 problems descriptions.

The augmentation of the classification scheme with severity ratings of usability problems and an assessment of their potential impact on ordering outcomes enabled us to prioritize usability problems from two perspectives: the (potential) impact on the system-user interaction and on patient safety. As such, the applied methodology advanced our understanding of the usability problem set of the CPOE usability case study.
Many studies have evaluated the usability of a variety of interactive healthcare applications and reported on the usability problems and their impact on the user interaction. Apart from the application of usability heuristic classifications in some studies, none of these studies used a systematic framework for guiding and structuring the evaluation and reporting of these usability flaws.

The augmented scheme provides better classifications than the current usability problem classification strategies such as heuristic [22] and bottom up classification [23;24] in the sense that classifications of usability problems by this new framework based on the UAF are distinguishable, mutually exclusive, complete and specific. Heuristic taxonomies lack these characteristics leading to usability problems of a very different nature being classified to the same heuristic, a single problem classified to more than one heuristic or to one heuristic not adequately capturing its essence and to impossibilities to classify some problems at all. Moreover, the utility of a classification that is essentially a by-product of an evaluation technique such as heuristic analysis can be limited because classification is not the intended purpose of the technique [13].

In bottom up classifications of usability problems, the frameworks of reference evolve with the analyses and as such not only become highly dependent upon the expertise of the evaluators, but more importantly lack a common foundation for future usage: the consistent, complete and accurate reporting of usability problems and understanding of their underlying causes. The need for standardized classification and reporting of usability issues is clear when the aim is to share the usability experiences by the development of a knowledge base concerning usability problem characteristics and their impact on user-system interactions.

Severity classifications of identified usability problems, providing insight in their frequency, persistence and impact on users, are commonly used in prioritizing system redesign efforts. Severity classifications applied to usability problem datasets represent somewhat isolated characteristics of these sets. This classification only address the criticality of the problem without providing more information about what the problem is, how it occurred and what could be its effect on task outcome. Furthermore, these classifications do not account for the potential impact of usability problems on the final outcomes of the user-system interaction.

We augmented the UAF with a severity classification of problems and appraisals of their potential impact on final task outcomes. This total scheme can be useful for analyzing and classifying usability problem sets, for revealing the core problems and hence for prioritizing the order in which to address problems in a system redesign. Some of the problems assigned a low severity in this study could potentially result in ordering errors. This means that prioritizing problems only on frequency, persistence and effect on users does not reflect their potential effect on patient safety. The problems with a low severity would then be given higher priorities if their potential effect on the ordering outcome is also taken into account. Therefore, when severity ratings of usability problems are used for
prioritizing redesign efforts of CPOE, these should preferably be accompanied by assessments of the possible influence of identified usability problems on final task results.

Going deeper into the hierarchy of the UAF, the number of the problems that were classified under different categories increased. This highlights the fact that problems seeming similar at the surface (the first levels of the UAF) can have different underlying causes [19]. A one level classification such as severity rating or bottom up approaches treats all problems that are clustered in the same category similarly without paying attention to the potential differences in their cause. In order to be able to tackle the usability problems fundamentally a redesign effort should target the underlying cause of each problem. Therefore, there is no unique design solution applicable to all the problems when classified at a high level of the categorization. This indicates that the one-level classifications such as heuristic analysis used in other usability evaluation studies might suffer from this critique. In order to provide details of each problem, one level classifications should be expanded horizontally. This can not be done for classifications such as severity rating and classifications based on heuristics because the problem classes are pre-defined.

Furthermore, our scheme could be of value for studies comparing the strengths and weaknesses of usability evaluation methods for user interface evaluation. Common practice is to compare the proportion of minor and major problems found by these methods. While such a criterion may be useful in examining which method has a higher detection rate of major problems, it does not provide insight in the capability of each method to detect specific classes of usability problems (e.g. problems concerning a specific phase of interaction). Though our problem set was too small to compare the CW and TA on these aspects, larger problem datasets may provide the opportunity to examine these issues.

This study has some limitations. The two usability experts were comfortable with classifying problems according to the UAF and showed a low number of disagreements among their ratings. Andre et al. [15] has shown that the reliability of the UAF for categorization of problems is higher than the heuristic classification and than the Usability Problem Taxonomy [13], a multi level usability problem classification. These results support the notion that the UAF provides a reliable classification system that is helpful in developing a common understanding of the different usability problem attributes. We acknowledge the fact that these results are based on the use of the UAF by two usability experts only and based on one usability problem set concerning a single CPOE system. Moreover, the two evaluators were familiar with the CPOE system, one of them conducted the usability tests with end users and thus was aware of the context in which usability problems occurred. Evaluators not familiar with the type of system under study, not having participated in the usability end user tests may experience more difficulties in the UAF classification of usability problems. But in principle in practice the evaluators will have to use the UAF classification directly after the usability evaluation and therefore will be in the same position as the authors. Furthermore, the usability problem dataset used for examining the value of the augmented classification scheme was based on one CPOE
usability case study at one academic hospital. While it is likely that similar usability problems occur in the 15 other Dutch hospitals that use the same CPOE system, we acknowledge that the use of one scenario and usability testing of the CPOE system by 10 clinicians from two clinical departments limits the generalizability of the results. Despite these limitations, we believe that this study highlights an important issue and serves as a model for other researchers seeking to enhance their insights into the impact of CPOE designs on usability on both more global and specific levels.

In a general sense, usage of this augmented scheme will minimize subjective analysis and inconsistent classification of usability evaluation. Widespread application could help in a complete and consistent classification of usability problems based on their underlying causes and as a result produce problem reports of higher quality. But most important, adoption of the scheme in practice would aid in more easily revealing trends and patterns across problem sets of usability evaluations case studies. This would assist in building comprehensive usability knowledge bases and advancing human computer interaction science in the healthcare domain. The usefulness and value of the approach exemplified in this study to classify and prioritize usability problems of a CPOE medication system should be further researched in usability evaluations of other health care applications. Future studies could examine how the consideration of the potential task outcome of usability problems in severity rating enhances the prioritization of usability problems.

Conclusions

The augmented classification scheme used in this study enables usability evaluators to analyze and classify usability problem sets in relation to the phase of the user-system interaction and addresses system users’ cognitive, physical and assessment efforts triggered by each of these problems. This classification differentiates problems that on the basis of their description and severity seem similar but could affect the user–system interaction and the task results differently. Evaluators using the scheme independently arrived at the same problem classification in more than 90% of the cases.

REFERENCES

2. Boonstra A, Broekhuis M. Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. BMC Health Serv Res 2010;10:231.


Chapter 6

A Concise and Practical Framework for the Development and Usability Evaluation of Patient Information Websites

Sebastiaan L. Knijnenburg, Linda W.P. Peute, Leontien C. Kremer, Monique W.M. Jaspers

submitted
Abstract

Background: The application of user-centered design principles is an essential part of modern website development and evaluation. The Website Developmental Model for the Healthcare Consumer (WDMHC) is an extensive and successfully evaluated framework that incorporates these design principles. However, due to its extensiveness the application in daily practice may be limited. In the current study we apply a limited subset of the WDMHC framework in a case study concerning the development and evaluation of a website aimed at childhood cancer survivors (CCS).

Objective: To assess whether the implementation of a limited subset of the WDMHC-framework is sufficient to deliver a high-quality website with few usability problems, aimed at a specific patient population.

Methods: The website was developed using a six-step approach divided into three phases, all of which were derived from the WDMHC: 1) information needs analysis, mock-up creation and focus group discussion; 2) website prototype development; 3) heuristic evaluation (HE) and think aloud analysis (TA). The HE was performed by three double experts (knowledgeable both in usability engineering and childhood cancer survivorship), who assessed the site using the Nielsen heuristics. For the think aloud analysis eight end-users were invited to complete three scenarios covering all functionality of the website.

Results: For the information needs analysis, a survey was completed by 145 potential end-users, which led to a structured and prioritized requirements document. This document served as the input for the mock-up creation and both this mock-up and the requirements document were used as an agenda for the focus group discussion during which refinements for both were proposed. After the prototype website was built, the HE and TA were performed concurrently. The HE resulted in 29 unique usability issues; the end-users performing the TA encountered 67 usability problem occurrences which could be classified into eleven unique problems. All the major usability issues discovered in the heuristic evaluation were likewise revealed by the think aloud user test sessions and vice versa. The four additional issues revealed by heuristic evaluation concerned cosmetic design flaws, whereas the two additional problems revealed by think aloud were related to website content.

Conclusion: Due to the involvement of both end-users and double experts throughout the project, we were able to deliver a prototype website that closely matched the expectancy of the end-users and resulted in relatively few usability problems during end-user testing. With the successful application of this limited subset of the WDMHC, we provide developers with a small, clear and easily applicable framework for the development of healthcare websites with high usability aimed at specific medical populations.

Key words: usability evaluation, website development, user centered design
Introduction

In the Netherlands, almost 40 per 100,000 children and adolescents develop cancer each year[1]. Great improvements in treatment for malignant disease in childhood have led to a major increase in survival rates, up to 85%[1]. With the rise in survival, a new and quickly growing population of long term survivors of childhood cancer emerged. In a recent study, Geenen et al. found that 75% of all survivors suffer from at least one adverse event at a median follow-up of seventeen years, emphasizing the need for adequate life-long follow-up care [2]. In order to improve the quality of care and to successfully conduct clinical research, the DCOG (Dutch Childhood Oncology Group) LATER (Long term effects after childhood cancer) project pursues to follow-up all Dutch 5-year survivors and screen them based on evidence-based guidelines[3-5].

Previous research showed that survivors of childhood malignancies are not well informed about their past disease and the potential late effects [6-8]. Due to the long intervals between and the limited timeframe within follow-up visits, childhood cancer survivors may have a need for a more continuous source providing this information. As for instance Lewis showed, well informed patients tend to be more compliant to screening and therapy, facilitating earlier diagnosis and treatment of late toxicities of treatment[9]. Patient information websites allow patients to educate themselves using web-based, personalized information at the time and place of their preference [10, 11].

In 2007, Johnson and Turley presented the Website Developmental Model for the Healthcare Consumer (WDMHC), a framework for the user-centered, iterative design and development of healthcare websites [12]. This framework incorporates well-documented user-centered design principles aimed at assisting the development of websites for healthcare consumers in the broadest sense. Although the framework was successfully applied in practice during the redesign of a consumer health information website, it is too extensive for projects with a limited scope and with limited resources [13]. Specifically, for the current study we had a well-defined end-user population with a specific medical background and only limited resources available for website development and evaluation. We hypothesized that the deployment of a limited subset of the WDMHC principles would be sufficient for delivering a patient information website aimed at a small and well-defined population that meets the end-users’ requirements and expectations by providing high-quality and trustable content and does so with a sufficient usability. For example, one of the usability methodologies used in the WDMHC was the keystroke level model [14]. This method focuses on efficiency of a given system, which was not relevant for this project due to its limited scope.

In most studies reporting on website development, different usability testing methods are employed in different phases of the software development process [15-18]. Often, usability inspection methods, like the heuristic evaluation, are used early in the development process, where system mock-ups and system prototypes are evaluated by usability experts, whereas end-user testing methods like the think
aloud method are more often used later in the system’s development process [15, 17]. A literature review by Jaspers showed that the think aloud method gains the best insights in the cognitive processes of end users, but at a high cost [19]. On the other hand the heuristic evaluation can also give good results at a much lower cost. However, it is known that the heuristic evaluation can find a broad range of cosmetic and minor problems which may disguise more severe problems [19].

In the present study we aimed to test the hypothesis that implementation of a subset of the WDHMC principles is sufficient for the development and evaluation of a patient information website aimed at a specific medical patient population, in our case long term childhood cancer survivors. Additionally we aimed to compare the surplus value of two widespread usability evaluation methods, the heuristic evaluation and the think aloud method, in the context of the deployed user-centered development approach.

**Methods**

The original WDMHC framework consists of more than 15 steps conducted in four phases: 1) user-, task- and environmental analysis; 2) functional- and representational analysis and their relation to visual and content representation criteria; comparative analysis to similar websites; 3) cognitive walkthrough, keystroke level model, heuristic evaluation and incorporation of metrics and guidelines and 4) content-based testing, expert-based testing and user-based testing.

The multifaceted approach we undertook for the development of our website consisted of six steps in 3 phases: 1) information needs analysis, mock-up creation and focus group; 2) website prototype development; 3) heuristic evaluation and think aloud analysis. The whole process is outlined in Figure 1. The following sections will shortly describe the methods undertaken for each step.
Figure 1. The Website development model for the healthcare consumer (above X1), and the adapted Subset (X2) for patient centered information websites.
Chapter 6

Information needs analysis

Understanding the end-users’ information needs as well as the required functionalities are prerequisites for delivering a website that end-users intend to visit. Requirements for the website were elucidated using a structured questionnaire comprising 24 questions, sent out to 160 survivors of childhood cancer or their parents. We asked our respondents to score a 5-point Likert-scale on 22 items involving different kinds of (health) information and functionalities which might be of interest for survivors of childhood cancer or their relatives. The information needs analysis resulted in a prioritized list of information requirements for the website.

Mock-up, focus group and prototype development

The visual representation or design of the website and its content should be optimized to allow for an optimal match between functionality and structure on the one hand and the goals of its end-users on the other hand. Based on the prioritized list of information needs, a simple mock-up website was created depicting its main functionality, content and navigational structures. Subsequently, we organized a focus group to assess whether the prioritized requirements list gathered from the survivors’ survey and the mockup website corresponded with end-users’ expectations. To allow for a friendly and open discussion environment, we decided to limit the amount of participants. We invited two childhood cancer survivors and two parents of survivors who previously had responded to the questionnaire as well as a pediatrician, a pediatric oncologist and two human factors engineering and usability experts. A trained focus group facilitator chaired the session using a structured agenda in combination with a brainstorm approach. All participants were free to discuss their opinions and make suggestions to our initial mock-up. Minutes of the focus group in combination with the initial information needs assessment were used as input for the development of a working prototype of the website. The development of the prototype was outsourced to an external web design company, who incorporated established website usability guidelines during the development process [20].

Usability evaluation

To evaluate the prototype website on its usability we aimed to compare two well-known methods from the field of usability engineering: the heuristic evaluation and the think aloud analysis [19, 21]. Although both methods are usually deployed in different stages of the software development life cycle, we were interested in the (additional) value in terms of differences in output of both methods from the viewpoint of the complete user-centered development approach.

Heuristic evaluation
In a heuristic evaluation (HE), usability evaluators inspect a system according to a predefined set of rules of thumb: heuristics. The most widespread set of heuristics was developed in the early nineties by Nielsen and Molich and consists of ten heuristics. [22] Each evaluator first assesses the system under investigation in a global way, to get an idea of structure and navigation of the system and then makes a second iteration to assess user interface elements in detail. All possible usability issues are written down and the severity of each issue is rated on scale from 0 (cosmetic) to 4 (catastrophic) [23]. For our heuristic evaluation three evaluators independently inspected the prototype website according to Nielsen’s heuristics. All three were trained in usability evaluation and usability research and had domain expertise in the field of childhood cancer survivorship care.

**Think aloud method**

The think aloud method (TA) is a verbal report method that stems from the field of cognitive psychology. During a think aloud session, participants interact with a system according to predefined protocols/task scenarios while they verbalize their thoughts aloud. These verbalizations get recorded and in combination with the screen recordings of the users’ website navigation patterns and their task performances, a deep insight can be gained into the system’s usability and the related usability problems.

For this study we created four scenarios that covered all the functionality the website offered: using the late effects search structure, finding information about diagnoses and anti-cancer therapy, looking up outpatient clinic information, etcetera (see Appendix 1). We selected a representative sample of eight potential end-users, as this number would be sufficient to reveal over 90% of real usability problems, as empirically shown by Nielsen. [24] We visited the participants in their home situation as this was the intended environment where the website would be used. We used a mobile ‘usability laboratory’ consisting of a laptop with Morae™ software (TechSmith Corp., Okemos, MI, USA) to capture the screen, mouse gestures and mouse clicks through screen recording and the participant’s face and voice through an integrated webcam. After all task scenarios were completed, users were automatically presented a 10-item System Usability Scale questionnaire to get a quick insight in the user’s perception of the prototype website usability [25].

All recorded data were analyzed with the Morae™ software. Firstly, all occurrences of usability problems were marked by reviewing the recordings. From the list of all occurrences a coding scheme was developed bottom-up by one author by grouping all the unique usability problem types into categories. A second evaluator (SK) individually reviewed all recordings, marked all usability problem occurrences and used the coding scheme to categorize them. Results of both reviewers were compared and discrepancies were resolved by discussion.
Assessment of the surplus value of both usability evaluation methods

From the literature it is clear that the HE and the TA both have their proper uses in the software developmental lifecycle.[16, 26-31] However, the costs associated with the TA are much higher than those of the HE. We aimed to compare the surplus value of each method in the perspective of our multi-faceted and user-centered approach. To assess possible differences in the detected problems by both methods, we qualitatively mapped the problems found with both methods to each other.

### Table 1. Information- and functional requirements based on the information needs analysis

<table>
<thead>
<tr>
<th>Requirements for the website</th>
<th>Mean rating</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address of the specialized late effects outpatient clinic</td>
<td>4.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Phone number of the outpatient clinic</td>
<td>4.2</td>
<td>0.9</td>
</tr>
<tr>
<td>Information about late effects tailored to your personal situation</td>
<td>4.2</td>
<td>0.7</td>
</tr>
<tr>
<td>Information about all possible late effects</td>
<td>4.1</td>
<td>0.8</td>
</tr>
<tr>
<td>Information about recognizing symptoms of possible late effects</td>
<td>4.5</td>
<td>0.6</td>
</tr>
<tr>
<td>Information on possible therapies for the late effects tailored to your specific situation</td>
<td>4.3</td>
<td>0.7</td>
</tr>
<tr>
<td>Information on possible therapies for all possible late effects</td>
<td>4.1</td>
<td>0.8</td>
</tr>
<tr>
<td>Information about self-care to minimize the impact of possible late effect</td>
<td>4.4</td>
<td>0.7</td>
</tr>
<tr>
<td>Up-to-date news about late effects of childhood cancer</td>
<td>4.2</td>
<td>0.8</td>
</tr>
<tr>
<td>Information about the rights you have as a patient</td>
<td>4.1</td>
<td>0.7</td>
</tr>
<tr>
<td>Asking health-related/medical questions through email and receiving an answer via email or telephone</td>
<td>4.0</td>
<td>0.8</td>
</tr>
<tr>
<td>A search engine to search the site</td>
<td>4.0</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Rating were scored on a 5 point Likert scale (1: very unimportant to 5: very important).

Abbreviations: SD: standard deviation.

### Results

#### Information needs assessment

One-hundred forty-five survivors or parents of survivors returned the questionnaire on information requirements (90.6%). Of these 145, only 15 respondents (10.3%) used the internet to look for information about late effects of childhood cancer and only four (27%) of them managed to find the information they were looking for. On the other hand, only 4% of all respondents stated they would not visit a late effects information site when available. Table 1 presents all items concerning
survivors’ information needs that were received a mean rating of ‘important’ or higher. The complete results of the questionnaire have been presented elsewhere [32].

Mock-up, focus group and prototype development

Based on items listed in Table 1, we envisioned a website structure divided in six sections: Home, News, LATER Information, Outpatient Clinics, Links and Contact. As the need for tailored information supply among survivors is high whilst their prior knowledge of disease, treatment and risks of late effects is low, we devised a search structure where survivors could search information in a linked network of diagnoses, their associated treatments and the potential late effects caused by these treatments. This search structure was envisioned through three columns (diagnosis, treatment, late effects) with clickable data labels, where the selection of an item in one of the three columns would automatically filter the related items in the other two columns.

During the focus group the results of the information needs assessment were discussed and the mock-up of the website was presented. The proposed division into six sections was used as an interview guideline and each section was discussed separately. The survivors’ response on the mock-up was very positive, even though the design was still very basic. Input from the survivors was useful in prioritizing the requirements. There were three main comments regarding the mock-up website: firstly, there was a need for additional information on diagnoses and treatments apart from information on late effects; secondly, functionality to seek contact with a health care provider was not necessary, as the goal of the website was to inform survivors, not to facilitate contact between survivors and physicians. Also, the proposed functionality of a discussion board was discarded as there was an existing and well-visited discussion board on the website of the Dutch Childhood Cancer Parent Organization (DCCPO). Survivors also expressed the opinion that discussion boards should be separated from websites visited by health care professionals, to keep a sense of privacy. The third comment was that the communication of all information should be in a conservative manner, to focus on reassurance instead of giving rise to anxiety.

Based on the original requirements and the feedback from the focus group, an external web design company developed the first working prototype of the website, incorporating the suggested search structure for finding targeted information on late effects. All content for the website was generated and peer-reviewed by members of the Information Service of the DCOG LATER project in cooperation with the Dutch Childhood Cancer Parent Organization (DCCPO). Screenshots of the prototype can be found in Figure 2 and Figure 3.
Figure 2. Screenshot of the start page of the prototype information website for childhood cancer survivors

Figure 3. Screenshot of the information search model of the prototype information website for childhood cancer survivors

Usability evaluation

Heuristic evaluation

Three double experts (usability experts with domain knowledge on childhood cancer survivorship) listed a total of 40 occurrences of usability issues according to the 10 heuristics described by
After removing duplicates, 29 issues represented unique problem types (average pairwise inter-rater agreement: 33.3%). The usability problems with a severity rating higher than 2 and those problems found by more than 1 evaluator are listed in Table 2. The problems rated as most severe were the fact that hyperlinks to external websites were not depicted as such and that using the back-button while browsing the search structure for finding targeted information on late effects and its underlying information pages did not function correctly. These issues could severely hamper the end user’s website interaction experience. Other usability issues were related to design consistency over different pages (different colors for hyperlinks on different pages, the search box was placed on different positions), deviations from standard web design guidelines (the website’s logo should be clickable to return the homepage) and the website’s content (the explanation of the search structure’s mechanism was regarded as too complicated for the target audience).

Table 2. Results of the heuristic evaluation

<table>
<thead>
<tr>
<th>Description of usability problem</th>
<th>Mean severity rating</th>
<th>Evaluators (found / total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>External links are not clearly depicted as such</td>
<td>3.5</td>
<td>2/3</td>
</tr>
<tr>
<td>Search structure ‘back’ doesn’t function correctly; results and selections in tables change after ‘back’</td>
<td>3.5</td>
<td>2/3</td>
</tr>
<tr>
<td>Automatic change of results when using search structure is not visible because the user needs to scroll down to see those changes</td>
<td>3</td>
<td>1/3</td>
</tr>
<tr>
<td>Deeplink of diagnosis results do not lead to relevant information directly</td>
<td>3</td>
<td>1/3</td>
</tr>
<tr>
<td>Search bar only matches full strings</td>
<td>2.5</td>
<td>2/3</td>
</tr>
<tr>
<td>‘all diagnosis/all…’ selectable items in the search structure all seem to be active and selected, but only ‘all diagnosis’ is really active and selected.</td>
<td>2</td>
<td>2/3</td>
</tr>
<tr>
<td>No ‘back’ at search structure explanation</td>
<td>2</td>
<td>2/3</td>
</tr>
<tr>
<td>Coloured blox on homepage not clearly clickable</td>
<td>2</td>
<td>1/3</td>
</tr>
<tr>
<td>Inconsistent use of hyperlink colours</td>
<td>2</td>
<td>1/3</td>
</tr>
<tr>
<td>Inconsistent searchbox placement</td>
<td>1.5</td>
<td>2/3</td>
</tr>
<tr>
<td>Submenu is not clearly visible</td>
<td>1.5</td>
<td>2/3</td>
</tr>
<tr>
<td>2 hyperlink lines in results of treatment, diagnosis which link to the same site</td>
<td>1</td>
<td>2/3</td>
</tr>
<tr>
<td>Logo’s on homepage not clickable</td>
<td>1</td>
<td>2/3</td>
</tr>
</tbody>
</table>

**Think aloud method**

The think aloud population consisted of five survivors and three parents of childhood cancer survivors. Average age of the testers was 35 years and mean time since initial cancer diagnosis was 16 years. All participants had broadband connections available at home and used the internet for at least two hours a
week. Testing sessions lasted for approximately 30 minutes per participant. Testers expressed little difficulty with thinking out loud, although some of the testers had to be reminded to keep talking on multiple occasions during the session.

### Table 3. Results of the think aloud evaluation

<table>
<thead>
<tr>
<th>Division</th>
<th>Coding category</th>
<th>TA ID</th>
<th>Usability problem</th>
<th>Total occurrences</th>
<th>N of testers encountering problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Menu or submenu</td>
<td>C</td>
<td>M1</td>
<td>Users think information about late effects is found under ‘information’</td>
<td>5</td>
<td>5/8</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>M2</td>
<td>Outpatient clinic abbreviations are not known to users</td>
<td>2</td>
<td>2/8</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>M3</td>
<td>Users cannot find/do not notice submenu</td>
<td>4</td>
<td>4/8</td>
</tr>
<tr>
<td>Search structure</td>
<td>N</td>
<td>Z1</td>
<td>User cannot return to previous page</td>
<td>5</td>
<td>3/8</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Z2</td>
<td>User cannot proceed to full information (read more)</td>
<td>5</td>
<td>4/8</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>Z3</td>
<td>Selection is gone after browser ‘back’, has to be performed again</td>
<td>10</td>
<td>6/8</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>Z4</td>
<td>User wants to display results by clicking</td>
<td>12</td>
<td>6/8</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>Z5</td>
<td>Users see a selection displayed, but this is not actual active selection</td>
<td>8</td>
<td>6/8</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>Z6</td>
<td>Users cannot distinguish which hyperlink line leads to correct information</td>
<td>3</td>
<td>3/8</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>Z7</td>
<td>Users do not perceive links to external websites as such</td>
<td>4</td>
<td>4/8</td>
</tr>
<tr>
<td>Search bar</td>
<td>S</td>
<td>B1</td>
<td>Users do not get results with legitimate search query</td>
<td>9</td>
<td>7/8</td>
</tr>
</tbody>
</table>

Interrater agreement for categorization was 72%, after which all discrepancies were resolved by discussion. Coding category abbreviations: S: System response; N: Navigational options; C: Representation of content; M: mismatch between system and user’s expectations.

In total the eight testers encountered 67 usability issue occurrences which could be classified into eleven unique problems (Table 3) and four categories. System response concerned usability problems related to how the prototype website responded to actions taken by the participants. Problems that arose during navigation through the website were categorized as navigational problems and the category ‘representation of content’ comprised all usability problems relating to participants not
understanding the presentation of certain website content. The fourth category comprised all usability problems concerning ‘mismatches between system and user’s expectations’.

The majority of the issues were encountered in the late effects information search structure. The most occurring problem, encountered 12 times by 6/8 users, was that users clicked on the search structure’s list entries, assuming that that would take them directly to information of interest, while clicking on these entries only updated the selection of relevant links displayed below the search structure itself (Z4). Another major problem was that the selected entries in the diagnosis/treatment/late effects list boxes were not persistent when using the browser’s back button: depending on the used internet browser the search structure would either forget the previous user selection (Z3) or, if it did show, the results of that selection were not updated correspondingly (Z5). Also, the search bar that was incorporated to search on the overall website did not behave as the test users expected from an online search engine. Legitimate search queries, for instance for ‘leukemia’, did not retrieve relevant information, as the leukemia information on the site was actually a link to an external page on the site of the DCOG.

Assessment of the surplus value of both usability evaluation methods

Table 4 lists a comparison of problems found in both the TA and the HE. All the major usability issues discovered in the heuristic evaluation were likewise revealed by the think aloud user test sessions and vice versa. The four additional issues revealed by heuristic evaluation concerned cosmetic design flaws, whereas the two additional problems revealed by think aloud were related to website content.

The time spent on both analyses differed a lot. The heuristic evaluation took each evaluator approximately one hour, after which it took another hour to aggregate the results, resulting in a total of four hours spent on the HE. The think aloud analysis cost around 30 minutes to perform, but took a lot more time to analyze: two evaluators both went through all the participants audio-visual recordings to transcribe the usability problems the participants uncovered. This took at least another 30 minutes per participant per evaluator, even though we used a sophisticated software suite specialized at this kind of tasks. We estimate the total time spent on the think aloud analysis at least 30 hours (excluding travel time to/from the participants).
Table 4. Comparison of the results of the think aloud analysis and the heuristic evaluation.

<table>
<thead>
<tr>
<th>Think aloud usability problem descriptions</th>
<th>N</th>
<th>Associated findings of the heuristic evaluation</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users think information about late effects is under general ‘Information’ heading.</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of unknown abbreviations</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Users cannot find/do not notice submenu</td>
<td>4</td>
<td>Submenu is not clearly visible</td>
<td>1.5</td>
</tr>
<tr>
<td>User cannot return to previous page</td>
<td>5</td>
<td>No ‘Back’ at search structure explanation</td>
<td>2</td>
</tr>
<tr>
<td>User cannot proceed to full information (read more link)</td>
<td>5</td>
<td>Automatic change of results not clear when using search structure</td>
<td>3</td>
</tr>
<tr>
<td>User wants to display results by clicking</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selection is gone after browser ‘back’, reselection needed</td>
<td>10</td>
<td>Search structure ‘Back’ doesn’t function correctly</td>
<td>3.5</td>
</tr>
<tr>
<td>Users see a selection displayed, which is not the actual active selection</td>
<td>8</td>
<td>‘all diagnosis/all…’ selectable items in the search structure seem to be all active, but only ‘all diagnosis’ is really active and selected</td>
<td>2</td>
</tr>
<tr>
<td>Users cannot distinguish which hyperlink line leads to correct information</td>
<td>3</td>
<td>2 hyperlink lines linking to the same site</td>
<td>1</td>
</tr>
<tr>
<td>Users do not perceive external links as such (and get lost)</td>
<td>4</td>
<td>External links are not depicted as such Deeplink of diagnosis results do not lead to relevant information directly</td>
<td>3.5</td>
</tr>
<tr>
<td>Users do not get results with legitimate search query</td>
<td>9</td>
<td>Search bar only matches full strings</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Coloured blocks on homepage not clearly clickable</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inconsistent use of hyperlink colours</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inconsistent placement of search box</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Logo’s on homepage not clickable</td>
<td>1</td>
</tr>
</tbody>
</table>

N: the amount of occurrences of the usability problem found by think aloud analysis. S: Severity score of the usability problem found by heuristic evaluation

Discussion

In this study we successfully implemented a practical multi-faceted, user-centered and iterative development approach for the development of a patient information website aimed at childhood cancer survivors and their relatives. Due to the involvement of end-users from the start of the project on and continuing in every phase of the development, we were able to deliver a prototype website that closely matched the expectancy of the end-users and resulted in relatively few usability problems during end-user testing.

User-centered design is one of the critical factors for the success of a website development project. When end-users are involved throughout the development process, the chance that the implemented solutions fulfill the needs of the end-users increases. This will subsequently lead to user adoption and user goal realization.[33] The most comprehensive framework for website development that
The Development and Evaluation of Patient Information Websites

incorporates user-centered design is the Website Developmental Model for the Healthcare Consumer (WDMHC).[12] This framework consists of more than fifteen methodological steps aimed to assist the development of websites for healthcare consumers in the broadest sense. Although these steps were shown to be successfully applicable in a case study on the redesign of consumer health information website,[13] the model is too extensive for projects with a limited scope and with limited resources. By testing whether the implementation of the limited subset of the WDMHC resulted in a website with good usability, we provide developers with a small, clear and easily applicable framework for healthcare websites aimed at specific medical populations.

In the current study we showed that the deployment of a minimalist subset of the methods described in the WDMHC can also be sufficient to build a website revealing relatively few problems during end-user testing. For our subset of the WDMHC we decided a priori that we had to include at least one method for the requirements phase, the development phase and for the evaluation phase. We then selected those methods that we reckoned to give the highest additional value with the lowest associated costs. The only exception was the think aloud analysis, which is known to have high costs, as we wanted to include at least one end user evaluation method to test whether end-users could read, understand the website and would trust the website content. Of course the methods we used in our subset can be interchanged with similar methodologies. For example, Kinzie et al. earlier developed a website development framework that can be mapped to the WDMHC as well.[33] They also started with a needs assessment including interviews, surveys and focus groups, followed by solution identification and task analysis, and finally iterative cycles of website design (guided by website design guidelines), development, and evaluation using cognitive walkthroughs and think aloud end-user testing. We recommend that the choice of the methodologies used for the specific phases of the development process should be adapted to the characteristics and needs of an individual project.

Childhood cancer survivors are at risk for developing late adverse effects of their treatment.[2] However, they have little knowledge about these late effects, as well as about their initial diagnosis and the treatment they received.[8, 34] These knowledge deficits may hamper survivors’ abilities to engage in a healthy lifestyle, to participate in screening programs and to adequately be able to engage in self-management and self-care. The results of the information needs assessment proved that survivors did indeed have unmet information needs regarding late effects of childhood cancer.[32] Respondents rated an extensive list of possible information items and website functionalities, resulting in a detailed prioritization of requirements for the website.

Focus groups, if used at all in the development process of health information websites, are often employed near the end of the process, after usability testing or during redesign iteration.[16, 35] We reckoned that a live meeting with all the stakeholders early in the development process would result in a more detailed specification of requirements, resulting in a closer match between website content, functionality and structure and end-users expectations and needs. The focus group proved worthwhile.
at this stage of the development: we gained additional insights regarding the requirements of the potential end-users. The participating survivors knew what information and functionality was already available online and which resources were still missing, enabling us to focus on the information and functionality with the highest priority in designing the website. Additionally, based upon the mockup, survivors were able to pinpoint possible (usability) problems with both the navigational structure of the site and search structure for late effects information before we even started building a prototype. The information gathered during these two initial phases was valuable for the development of the first version of the website.

Several studies have been performed comparing the performance of heuristic evaluation and think aloud analysis applied either concurrently or at different time points in the software development life. [26-28] In comparison to the think aloud analysis, heuristic evaluation is known to find a higher number of usability issues, with both low- and high severity, at much lower costs. Our results are in line with these common findings: the heuristic evaluation revealed many more, especially low-severity or cosmetic problems in comparison to the think aloud analysis at much lower cost. However, several studies state that end-user usability testing is imperative to uncover the remaining problems not found in heuristic evaluation. [13, 16, 29-31] This was not the case in our study. All but two minor issues encountered by end-users during testing were also covered by the heuristic evaluation. The two items solely found by participants of the TA were “Users think information about specific late effects is under general ‘Information’ header instead of under ‘Late Effects’ header” and “Use of unknown (medical) abbreviations”. The domain knowledge and the involvement in the development process of the three double experts who performed the HE may explain that these two issues were overlooked in the heuristic evaluation. It should be noted that these two issues were the only ones found that related to readability, credibility, accuracy and understandability of the website’s content. We think that the deployed methodology, in combination with the editing and peer-reviewing of the content by a team of medical professionals and members of the DCCPO, resulted in a website that end-user can read, understand and trust. Formal evaluation using validated measurements may be performed in future studies to confirm this.

The good results of the heuristic evaluation in comparison to the think aloud evaluation may have had several reasons. The user-centered approach, from the initiation of the project and onwards, may have led to a close match between the design and the end-users’ strategies in approaching the website, resulting in less (severe) usability problems during testing. The inclusion of three evaluators knowledgeable both in usability evaluation studies and childhood cancer survivorship may also have contributed to the relatively better performance of the HE. Nielsen and Landauer found that, in a series of six heuristic evaluations of different user interfaces, the use of 5 evaluators resulted in a detection coverage of 75% of the usability problems [36]. Detection rates improve strongly though with the expertise of the evaluators: one skilled usability expert with additional medical knowledge regarding
the application’s content domain may account for up to 60% of all usability issues.[36-38] In situations that skilled usability evaluators are not familiar with the website’s medical domain, a work-domain expert could assist the evaluator in tackling domain specific problems. These work domain experts in these so-called ‘participatory heuristic evaluations’ may help usability evaluators in considering how the system contributes in reaching certain goals in that particular area of skills[19].

A possible limitation of this study was the scope of the website under study. Use of the website was uncomplicated as the system mainly consisted of static information content with some dynamic behavior implemented through JavaScript. Of course appropriate usability testing is also essential for smaller websites, but future studies should investigate the reliability of our methodology when implemented for the development of more advanced healthcare websites.

In conclusion, we successfully implemented a practical, user-centered and iterative development methodology for the development of a patient information website for childhood cancer survivors. If time and resources are limited, heuristic evaluation with double experts may be sufficient to discover all usability problems on a patient information website. In our case all encountered usability problems have been translated into recommendations for redesign and were given as feedback to the website developers. The redesigned website was launched in 2009 and in the final evaluation respondents were satisfied with its usability and contents [39]. As patient information websites should be designed to match end users’ expectations, our approach can serve as a low-cost and practical framework for their development and evaluation. In the future, we aim to extend the current website to a patient portal where survivors can login to exchange medical record data with healthcare professionals and where they can receive truly tailored health information based on their electronic medical record.

Acknowledgments

This study was financially supported by the Dutch Childhood Oncology Group. The sponsor had no role in the design and conduct of the study, the collection, analysis and interpretation of data or in the preparation, review or approval of the manuscript.

Reference List


29. Ostergren MJ, Karras BT. ActiveOptions: leveraging existing knowledge and usability testing to develop a physical activity program website for older adults. AMIA Annu Symp Proc 2007:578-582. PM:18693902


Chapter 7

Reducing Clinicians’ Cognitive Workload by System Redesign; a Pre-Post Think Aloud Usability Study

Linda.W.P. Peute, Nicolette F. de Keizer, Eric P.A. van der Zwan, Monique W.M. Jaspers

Abstract

Interactive Health Information systems are often considered cognitively complex by their users, leading to high cognitive burden and increased workload. This paper explores if Think Aloud usability testing provides valuable input to effectively redesign a web-based Data Query Tool in Intensive Care and to reduce physicians’ cognitive workload during system interaction. Pre and post redesign usability testing demonstrated a major reduction in the cognitive task workload after redesign of the tool. Classification of revealed usability problems by means of the User Action Framework pointed out that usability problems related to the cognitively planning of actions by system users foremost affected cognitive task workload. This result may support Health Information system (re)design efforts on how to tackle the system’s cognitive complexity and in so doing improve on its usability.

**Keywords:** Assessment-Evaluation, User Interfaces, Health Professional Workstation, Design aspects, Usability


Introduction

Usability evaluation is an essential but complex part of Health Information (HI) system development. Its purpose is to identify usability problems to improve the system interface design so that it can be used efficiently, effectively, satisfactorily and foremost safely by clinicians (1). In contrast to conventional usability evaluation methods, usability methods that emerged from the field of cognitive psychology are ever more viewed upon as essential in HI system (re)design. They are considered to transcend the level of revealing interface design flaws and advance to the stage of providing insight into clinicians’ cognitive processing in achieving system tasks (2). In doing so, these methods eventually aim to contribute to designing intuitive HI systems that support and facilitate clinical care by keeping the cognitive task workload of its users to a minimum (3). The classic Thinking Aloud (TA) method is one of the usability evaluation methods that stem from cognitive science and is generally considered the “gold standard” in usability testing (4). Scientific research on the validity of the method has shown that subjects’ verbalized information in TA usability testing accurately reflect users thought processes when interacting with a system (5). Inverting these insights however to successful system (re)designs and thereby minimizing the cognitive task workload of system usage is still challenging.

This paper investigates if users’ cognitive task workload in interaction with a web-based Data Query Tool of a National ICU Quality Registry is reduced in terms of improved efficacy (correctly performed tasks) and efficiency (task completion time) after redesign of the system with direct input from TA usability testing. Additionally, usability problems revealed in the pre and post TA test were classified and compared by use of the User Action Framework (UAF) (6). UAF classification is based on Norman’s theory of action and categorizes usability problems in the sequence of a user’s cognitive and physical actions in performing a task in a system. We hypothesize that the earlier a user is obstructed in this sequence, due to usability problems in a system, the higher the cognitive task workload of a user is. The potential of TA usability testing in supporting HI system (re)design with the aim to reduce users’ cognitive task workload in a system and the beneficial effect of applying the UAF classification in this perspective are furthermore discussed in this paper.

Methods

System background: NICE online

The Dutch National Intensive Care Evaluation (NICE) registry collects demographic, physiological and clinical data on patients admitted to Dutch ICUs to detect differences and trends in quality of ICU delivered care. To provide participating ICUs with the possibility to query their own data and compare their performance with their peers or with national averages a web-based Data Query Tool was
developed in a standard software development cycle. In the NICE Query Tool users define a ‘query’ themselves to compose a graph or a table depicting the selected information. An example of such a query in NICE Online is: ‘compare an ICU’s standardized mortality ratio (SMR) of medical patients to the national mean SMR of medical patients in the year 2009’. Figure 1 provides screen shots of both the first and redesigned NICE Query Tool. Additional information on the development of the Tool and its functionality is published in (7).

Pre-Post study design

In October 2008, a pre-TA usability evaluation was performed (pre-test) to assess the overall usability and cognitive complexity of developing queries in the, at that moment available, Query Tool. Eight end-users were contacted to participate in a Think Aloud (TA) study, with an equal representation of new and more skilled users. A portable usability laptop with Morae software was used to document subjects’ TA verbalizations and video and record their (mouse) actions in the system. Sessions took place in the clinical workspace of the subjects and six predefined germane tasks, consisting of several subtasks and varying in complexity, were all given to subjects during the TA test in random order. Revealed usability problems were input to redesign the Query Tool interface. In the beginning of 2010 post redesign TA testing was performed on a beta-test version of the redesigned tool to measure the effectiveness of the redesign efforts. Again, eight end-users were contacted of which four new test users participated who were comparable to the pre TA user test group in terms of computers skills and previous experience with the Query Tool, and four users who had also participated in the pre TA study. Bias of pre-defined task learnability for these four users was negligible, since the time between the pre and post study was around one year. In the post TA testing, similar circumstances were upheld as in the pre TA test, including the tasks to be performed in the system.

To compare overall task efficacy between the pre-and post TA sessions, the shortest routes to correctly perform the tasks and the corresponding end-results in both the old and redesigned system were determined by highly experienced data managers of the NICE registry. The correct task end-results were then applied as the ‘golden standard’ to measure the percentage of tasks correctly completed by subjects in the TA sessions.

To compare overall task efficiency, time on task measurements had to be adjusted for optimization of system response for the display of query results (e.g. users had to wait over one minute for display of the query result in the old system in contrast to 3 seconds in the new system). Pre-post overall task efficiency measurements were therefore compared in terms of the additional time it took users to complete tasks in the system both pre and post as opposed to the time it takes to complete the tasks by the shortest route in both system designs.
Usability problem classification; the User Action Framework

UAF classification places detected system usability problems in the context of four subsequent phases of the user interaction cycle; Planning- high level, Planning- translation, Physical actions and Assessment (6). Usability problems relating to the planning phase concern users’ cognitive actions for planning how to perform a task; e.g. the inability to track where you are in a system. The translation phase is about cognitive actions to determine how to carry out the intentions; related usability problems are incorrect button labeling or vague symbols. Physical action pertains to executing the actions by manipulating user interface objects; usability problems are e.g. button proximity or small size. The assessment phase is about perceiving, interpreting and evaluating the resulting system state to decide whether the action was indeed accurately performed; related usability problems concern users’ misunderstandings of system feedback. UAF classification starts with the four user interaction phases on the first level; accurate classification can go up to six levels. This paper limits its results to the first level to provide a general insight into the relation between performance measures on the cognitive task workload and the UAF classified usability problems found.

Results

Overall, 12 subjects were included in the pre and post TA study (4 subjects were included in both pre and post TA). In total, 36 usability problems were revealed by two usability analysts in the pre TA test and 35 usability problems were revealed in the post TA test. UAF categorization of usability problems was performed by both usability analysts separately (κ=0.91). Of the usability problems detected in the pre-test 34 (94%) were resolved by redesigning the Query Tool, 5 (14%) were considered overlapping
with usability problems found in the post-test and thirty new usability problems were revealed in the post TA Test.

Table 1. Pre and post redesign measurements of task efficiency and efficacy.

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Pre TA Test (8)</th>
<th>Post TA Test (8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Task Efficiency (min)</td>
<td>50.16 (sd 7.62)</td>
<td>19.40 (sd 6.01)</td>
</tr>
<tr>
<td>- Optimal route (min)</td>
<td>20.05</td>
<td>12.51</td>
</tr>
<tr>
<td>- Adjusted Overall Task Efficiency (min)*</td>
<td>+ 30.11</td>
<td>+ 6.49</td>
</tr>
<tr>
<td>Overall Task Efficacy</td>
<td>24 (50%)</td>
<td>46 (96%)</td>
</tr>
</tbody>
</table>

* Efficacy: Total number of tasks (completed) (8x6=48), * deviation in min from optimal route

Overall task efficiency in the pre TA test was extremely low; users took on average 30 minutes longer to perform the tasks in the system compared to the time it would have taken them when they would have known what to do and how to act in the system (shortest route) (Table 1). This extra time was reduced to less than 7 minutes in the post test. The fact that subjects were able to complete only 50% of the tasks during the pre TA test confirmed that usage of the Tool for developing queries was cognitively complex. This percentage increased to 96% after redesign.

Table 2 shows the usability problems categorized by ‘first level’ UAF classification in both the pre and post TA tests. The majority of usability problems in the pre-test concerned the ‘planning’ phase (64%). It appears that these problems were accountable for the high cognitive task workload associated with the low values measured for task efficacy and efficiency in the pre test. After redesign the majority of the 30 new usability problems detected in the post TA test concerned the ‘assessment’ phase (63%), showing an evident shift in the phase of interaction in which the new revealed usability problems occurred compared to the pre-test. However, these post TA usability problems did not or only minimally seem to affect users’ cognitive task workload. Apparently, usability problems related to the assessment phase did not have a great impact on task efficacy or efficiency. Analysis of the verbal protocols and video recording of users’ actions showed that usability problems in the Assessment phase were mostly related to users’ preferences in interface layout, such as graph colour and display of system feedback related to information on the screen.
Table 2. Usability problems in the Pre and Post TA test classified by first level UAF

<table>
<thead>
<tr>
<th>UAF Phase</th>
<th>Pre TA Test (36)</th>
<th>Post TA Test (35)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Planning</td>
<td>8 (22%)</td>
<td></td>
</tr>
<tr>
<td>2. Planning (translation)</td>
<td>15 (42%)</td>
<td>10 (29%)</td>
</tr>
<tr>
<td>3. Physical Action</td>
<td>13 (36%)</td>
<td>3 (8%)</td>
</tr>
<tr>
<td>4. Assessment</td>
<td>22 (63%)</td>
<td></td>
</tr>
</tbody>
</table>

Discussion/Conclusion

In this study the input of TA usability testing in redesigning a web-based Data Query Tool of a National ICU Quality Registry led to a clear reduction in its complexity and hence the cognitive task workload of its users. Optimization of the task accuracy was obtained from 50% of tasks completed in the pre-test to 96% of similar tasks completed in the post test. However, redesign of the Tool also caused thirty new usability problems to occur. This is not surprising as it is well known that usability evaluation is an iterative process; subsequent changes to a user interface design might reveal other problems that again need user testing (8). The fact that users’ efficiency was highly improved after redesign of the Tool indicates that the new usability problems detected in the post TA test minimally affected their cognitive task workload in use of the redesigned system.

Applying UAF classification in our study was particularly useful to compare the nature of the pre and post detected usability problems and their effect on users’ cognitive task workload. UAF classification revealed a potential cause-effect relation between the occurrence of usability problems in the planning phase of the users’–system interactions and their apparent negative effect on their cognitive task workload in terms of task efficiency and efficacy. Indeed, input from the pre TA test to the Query Tool redesign efforts offered insight on how to tackle the usability problems in the planning phase and in so doing furthered the development of the Query Tool to better support users’ cognitive processes in Data Querying. The new usability problems detected in the post test were mostly related to the Assessment phase, indicating that these problems were more or less of a cosmetic nature. As such, they did not provoke additional cognitive burden. Those usability problems that placed a high cognitive burden on system use were thus successfully reduced in just one redesign iteration of the Query Tool.
Future studies that apply TA testing in a redesign cycle should focus redesign efforts on those aspects of the system that affect the planning of tasks by end-users, especially when high cognitive task workload of complex HI system tasks is seen as a major barrier for system use.

REFERENCES

1. STANDARDIZATION SIOO. ISO 9241-11 Ergonomic requirements for office work with visual display terminals (VDTs) – part 11: guidance on usability. 1998.
8. B. Kaplan, K.D. Harris-Salamone, White paper: Health IT project success and failure: recommendations from literature and an AMIA workshop, K AM Med Infom Assoc 16 (2009), 291-
Chapter 8

From an Expert-driven Paper Guideline to a User-centered Decision Support System; a Usability Comparison Study

Ellen Kilsdonk, Linda W.P. Peute, Rinke J. Riezebos, Leonie C. Kremer, Monique W.M. Jaspers

*Conditional Acceptance special issue paper "CDSS evaluation". In Artificial Intelligence in Medicine*
Abstract.

Objective: To assess whether a user-centered prototype Clinical Decision Support System (CDSS), providing patient-specific advice, better supports healthcare practitioners in terms of types of usability problems detected and efficient and effective retrieval of childhood cancer survivor’s follow-up screening procedures compared to an expert-driven paper-based guideline.

Methods and material: A user-centered design (UCD) process was employed to design a working prototype CDSS. Usability problems in information retrieval with the paper-based guideline were assessed by Think Aloud analysis with 13 participants. Both simple as well as more complex tasks were applied. The analysis provided input into the user-centered design of the prototype. The usability of the prototype CDSS design was subsequently evaluated by Think Aloud analysis with the same participants. Usability problems of the paper-based guideline and the prototype CDSS were compared by classification of usability problems using the Classification Usability Problems (CUP) classification scheme. In addition, efficiency (time to complete task) and effectiveness (completeness of retrieved screening procedures) of information retrieval of participants in the expert-driven paper-based guideline and the user-centered prototype CDSS were compared.

Results: Usability problems in both the paper-based guideline as well as the CDSS prototype were mainly classified as ‘incongruent with participants’ mental model’. The CDSS design reduced this type of problem from 17 to 6 problems. Time to perform simple information retrieval tasks in the CDSS increased with 58 seconds when using the prototype CDSS, however resulting in a 58% improvement in task completeness compared to the paper-based guideline. Time to perform complex scenarios decreased by 3:50 minutes with the prototype CDSS, with 17% higher completeness compared to the paper-based guideline.

Conclusion: Analysis showed that usability problems experienced by healthcare practitioners when using a paper-based guideline could be overcome by implementing the guideline in a user-centered CDSS design. Different types of usability problems were experienced with the CDSS; however, these problems did not inhibit effective and efficient performance of tasks in the system. Usability analysis of the paper-based guideline effectively supported comparison of usability problems with both information retrieval systems and supported the UCD of the CDSS.

Keywords. Clinical Decision Support System, Screening guidelines, Childhood Cancer Survivors, Usability, Software design
Introduction

In clinical practice, evidence-based guidelines are developed and implemented to guide and support healthcare practitioners’ performance in making more efficient and effective clinical decisions in patient care [1]. However, though the underlying principles of the guidelines are generally acknowledged, uptake of and compliance to paper-based guidelines in clinical practice is generally low [2]. Research into the causes of this low uptake shows that next to management and workload related issues, paper-based guidelines appear to not support healthcare practitioners’ practical information needs in patient care. This might be caused by an expert-driven design process of paper-based guidelines [3].

An expert-driven design process is characterized by a focus on scientific validation of the guideline. In the design process of guidelines, experts give higher priority to reaching consensus than to content related issues. A lack of focus on the presentation of this content in the guideline leads to complex and intricate content structures. As a result, clinicians experience difficulties in using the resulting paper-based guideline in daily practice [3], leading to a lack of familiarity with the guideline and low guideline adherence. Consequently, evidence-based clinical practice based on paper-based guidelines is hampered [4].

A possible solution for improving the use of guidelines into clinical practice is to communicate them through Clinical Decision Support Systems (CDSS), preferably linked to an Electronic Medical Record (EMR) [5]. CDSS may overcome problems in use of paper-based guidelines in clinical practice by offering health care practitioners patient-specific advice based on the guideline recommendations. CDSS indeed have shown to improve guideline adherence by increasing healthcare providers’ knowledge of preferred practice, by reducing inertia to previous practice, and by reducing guideline complexity [6]. However, it is important to realize that the usability of a CDSS, to what extent it accommodates clinicians’ cognitive workflow, is critical for the effective and efficient use of that CDSS [7]. A computer system’s usability is strongly tied to the extent to which a user's mental model, a set of beliefs of how the system works, matches and predicts the actual actions of the system [8]. When a system design does not match the user mental model of the system, users will have a hard time in learning how to use the system. This is a problem commonly found with CDSS automation [9]. Designing a CDSS that optimally supports healthcare practitioners in clinical practice is thus challenging. One approach to achieve these challenging goals is referred to as User Centered design (UCD) [10]. UCD is a system development methodology that explicitly focuses on analysing end-users needs, mental processes, limitations and preferences in order to design a system that meets end-users requirements. Usability testing is a fundamental part of UCD for evaluating whether end-users
are supported in achieving their tasks in the system in an effective and efficient manner [10]. Since research has shown that many CDSS are not usable [11] UCD has been strongly recommended for CDSS development [11;12]. Insight into the context of use and usability problems that healthcare practitioners encounter while processing information in a paper-based guideline can be useful in revealing the requirements for the design of a CDSS. Such a user-centered design approach may enhance CDSS acceptance among healthcare practitioners because the system can then be designed to support end-users’ needs which were not met or provided for in the paper-based guideline [3].

We followed a UCD approach in the development of a CDSS prototype implementing an expert-driven paper-based guideline. The guideline is used in daily practice by healthcare practitioners to define follow-up screening procedures for childhood cancer survivors. The UCD approach focused on the usability problems healthcare practitioners encountered with the paper-based guideline as input into a CDSS design implementing the guideline. This study explores the type of usability problems clinicians experience in using the expert-driven paper-based guideline when compared to a user-centered CDSS prototype and assesses the efficiency and effectiveness of clinician information retrieval in both information systems.

**Methods**

**Research context**

Over the last decades, advances in diagnosis and treatment of childhood cancer have dramatically increased long-term survival [13]. As a result, the number of childhood cancer survivors is growing rapidly and it has become increasingly clear that the former disease and the received treatments can significantly impair long-term health and quality of life by causing late-effects. For this reason, the LATER (LATe Effect Registry) taskforce was started in 2004 by the Dutch Child Oncology Group (DCOG-LATER). Within LATER a nationwide cohort of over 6000 survivors has been identified and patient data such as: general patient information, medical history, tumor diagnosis and detailed treatment information, are collected in paper-based and electronic medical records. In a collaborative effort of the DCOG-LATER, an evidence-based guideline was developed for screening childhood cancer survivors for these possible treatment-related late effects. The goal of the LATER guideline is to promote uniformed and high-quality follow-up care across 7 academic hospitals in the Netherlands involved in follow-up care of childhood cancer survivors. [14].

The LATER guideline was developed by the task-force consisting of 16 multidisciplinary teams, each covering a specific clinical domain (e.g. cardiology or nephrology). Teams were instructed to follow a uniform procedure in developing the guideline. The expert-based approach of designing the guideline
resulted in a paper-based guideline document of 53 pages structured according to the clinical domains covered by the multidisciplinary teams. Subdivision of the 16 clinical domains resulted in a total of 24 domains in the paper-based guideline. Each domain, for instance cardiology, first describes domain specific inclusion criteria (e.g. survivors that received radiation to the thorax). It then continues to describe the potential resulting long term effects (e.g. arrhythmia) related to the inclusion criteria. Next, it describes the screening procedures to detect these long term effects (e.g. an echocardiogram) and the frequency with which to perform the screening procedures. At last it describes the advices that need to be given to the patient and specific points of interest for the anamnesis and physical examination.

The LATER guideline is currently communicated through a paper-based format and provides information on screening protocols based on the treatment that a cancer survivor received in childhood. To prepare a survivor’s follow-up screening visit, the healthcare practitioner retrieves a patient’s historical diagnosis and treatment data from the paper-based and electronic medical records at their site. Based on the childhood cancer history, they determine the patient-specific screening procedures, points of interests for the anamnesis and physical examination, and the advices to be given to the patient by reviewing the contents of the paper-based guideline. Because of multi-modal cancer treatments and the large variety of related possible late effects, healthcare practitioners experience difficulties in defining the overall follow-up screening procedure for individual patients. They have to go through more than 700 clinical rules to define the screening procedure for each individual cancer survivor. Consequently, defining the overall follow-up screening procedure for individual childhood cancer survivors by use of the paper-based guideline proves to be challenging for healthcare practitioners in daily clinical practice. The healthcare practitioners involved in follow-up care of childhood cancer survivors expressed a need for a CDSS providing them with patient-tailored advice to better support them in defining childhood cancer survivor’s follow-up screening procedures than the paper-based guideline.

**Research design**

This study consisted of 2 phases: a usability evaluation of the paper-based guideline and a formative usability study of the prototype CDSS. Thirteen healthcare practitioners from the follow-up clinics participated in both study phases. The usability problems that practitioners encountered in defining the follow-up screening procedures for individual patients by use of the paper-based guideline were analyzed by use of the Think Aloud method [15]. The results from this usability study provided input into the requirements and design of a working prototype CDSS. A formative usability study was performed on the CDSS design to assess if healthcare practitioners were more effective and efficient in defining the overall follow up screening procedures for individual patients by use of the prototype
CDSS than by use of the paper-based guideline. The paper-based guideline evaluation was conducted in April 2011; the formative usability study of the prototype CDSS in June 2011.

In the paper-based guideline evaluation, healthcare practitioners were asked to prepare two patient visits to the follow-up clinic on the basis of two fictitious patient scenarios, one simple and one complex. For each scenario, healthcare practitioners were asked to define the relevant follow-up screening procedures by use of the paper-based guideline. The patient scenarios were constructed from real-life patient cases. An information sheet described the patient’s general information (age, sex etc.), information about his/her cancer history, and information about the treatment the survivor had received in childhood. An expert physician on late effects of treatments for childhood cancer reviewed and validated the scenario descriptions. For each scenario, a list of all relevant screening procedures was constructed by reviewing the paper-based guideline contents. These lists were verified by the expert physician. Participants were randomly assigned to either start with the simple scenario or to start with the complex scenario. They received instructions about the think aloud method and a training task to practice the verbalization of thoughts prior to starting their first scenario. During task performance, a microphone recorded participants’ verbal utterances and a video camera recorded participants’ activities with the paper-based guideline.

In the formative CDSS usability study, the healthcare practitioners were again asked to prepare two clinic visits on the basis of fictitious patient scenarios and to define the overall follow-up screening procedures by use of the prototype CDSS. Two new patient scenarios, one simple and one complex, were developed which were similar to the ones used in the first study phase in terms of the number of screening procedures that applied to the fictitious patient. The expert physician once again validated the scenario descriptions and verified the list of relevant screening procedures constructed for each scenario.

A microphone recorded participants’ verbal utterances. Screen and mouse clicks, face expressions and voice and time records of the participant working with the prototype CDSS were captured with Morae software (TechSmith Corporation, Okemos, MI, USA).

Study population

Study participants were recruited through the coordinators of the long-term follow-up clinics by email. Participation was on a voluntary basis but participants were only included if they bear responsibility for preparing follow-up visits and defining the overall follow-up screening procedures for individual survivors by use of the paper-based guideline.

Usability Analysis
For both the paper-based guideline evaluation and formative CDSS usability study, all participants’ verbal utterances were transcribed to verbal protocols. The verbal protocols were divided into segments, each segment representing a single comment from the participant. To detect what types of usability problems occurred in each user session, each protocol segment was linked to the video recordings (revealing the activities of healthcare providers on page-sections of the paper-based guideline or CDSS screens at certain time-stamps in the sessions). This analysis provided detailed insight in the origin of each usability problem encountered in the user-sessions of the paper-based guideline and prototype CDSS.

Each time a unique usability problem was revealed it was added to the list of usability problems. When a specific usability problem had already been revealed and listed (because the session with another participant already had revealed the same usability problem), the frequency of that particular usability problem was increased. Two lists of usability problems were constructed; one concerning usability problems revealed with the paper-based guideline and one with the prototype CDSS.

**Classification of Usability Problems (CUP)**

Each usability problem was mapped to the Classification of Usability Problems (CUP) scheme by two independent researchers (EK, RR), developed by Hvannberg and Law [16]. The classification of usability problems according to CUP of the two researchers was compared and differences in classification were resolved through discussion. The CUP scheme focuses on usability problems concerning the structure and information contents of an interactive tool. Because of this focus the CUP scheme was considered to be applicable to classify usability problems revealed both with the paper-based guideline and prototype CDSS.

The CUP scheme consists of attributes for describing a usability problem in detail. For the comparison of usability problems revealed with the paper-based guideline and the usability problems with the prototype CDSS, the CUP attributes Severity and Failure Qualifier were considered relevant in the context of this study. Cohen's kappa coefficient was calculated to assess inter-rater reliability for CUP classification of Failure Quality for detected usability problems. Table 1 provides a description for the Severity rating and Failure qualifier as defined by the CUP.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severity</td>
<td>A problem is</td>
</tr>
<tr>
<td></td>
<td>• Severe: prevents the participant from completing a task</td>
</tr>
<tr>
<td></td>
<td>• Moderate: hinders the completion of the task but workarounds</td>
</tr>
<tr>
<td>Failure Qualifier</td>
<td>Describes how the participant experienced a usability problem. Contains the following categories: Missing, Incongruent Mental Model, Irrelevant, Wrong, Better Way and Overlooked.</td>
</tr>
<tr>
<td>-------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
|                   | • Incongruent: something in the tool is unclear, because it does not match with the test participant’s mental model or previous experience  
|                   | • Missing: participant fails to find something in the tool that was expected to be present  
|                   | • Better way: participant suggests that something could have been done differently in the tool  
|                   | • Wrong: participant can notice something went wrong with the tool, e.g. an apparent programming bug or typing error  
|                   | • Irrelevant: tool contains information that is not needed  
|                   | • Overlooked: participant is given a task but overlooks an entity, i.e. the participant fails to see existing entity or fails to realize to interact with it |

The User-Centered Design of the prototype CDSS.

Figure 1 provides an overview of the four stages within the UCD process followed in development of the prototype CDSS (1) assessment of user needs and specification of the context of use, (2) specification of CDSS requirements and user needs, (3) design of the CDSS prototype and (4) evaluation of the CDSS prototype.
The in-depth analysis of the protocol segments and video recordings of the user sessions with the paper-based guideline provided insight into the usability problems participants encountered and their unmet information needs in defining the screening procedures for the fictitious patient scenarios. Based on this analysis, the functional and non-functional requirements for the CDSS prototype Graphical User Interface (GUI) and information contents were specified and used as input for the design of a working prototype CDSS. Based on the usability analysis of the paper-based guideline, the organizational structure according to clinical domains was completely abandoned in the design of the prototype CDSS. Instead, the GUI of the prototype CDSS was designed as a tab-based structure with 6 tabs, to support ease of use in navigating and retrieving guideline information. A knowledge base was developed containing all the clinical rules that were extracted from the paper-based guideline. The clinical rules were extracted from the guideline and formalised to a computer-interpretable format by using the Logical Elements Rule Method (LERM) [17]. A decision engine applies the clinical rules by using individual patient data whenever a patient record is retrieved. The prototype CDSS contains a home screen explaining its functionality. From the home screen menu, users can search for a specific patient by a unique identification number, a patient’s date of birth or surname. After the CDSS retrieves the patient record, the tab-based structure containing 6 tabs appears. The first tab shows patient identification data and information about the childhood cancer. The second tab contains information on the cancer treatment the patient received. The third tab...
contains a list describing the screening procedures to be performed for the patient. The fourth tab contains points of interests for the anamnesis and physical examination specific to that patient. The fifth tab shows the specific advices that the healthcare practitioner needs to give the patient. The sixth and final tab contains the possible late effects that a patient could develop as result of the earlier cancer treatment. The CDSS offers users the option to add, edit and delete content on the tabs for screening procedures, points of interests for anamnesis and physical examination and advices to accommodate for patient-specific situations. For instance, the tab Screening Procedures lists all the screening procedures that need to be performed for a patient. Healthcare practitioners are allowed to adapt this list when a patient’s condition would not allow the performance of a certain screening procedure. Figure 2 shows a screenshot with the Screening Procedures tab opened. To support trust in the CDSS, healthcare practitioners can also look up the evidence (decision rules) available for each screening procedure by clicking on information buttons (i-buttons) displayed next to each screening procedure.

![Figure 2. Screenshot of prototype system, screening procedures tab](image)

### Outcome measures

Effectiveness was measured in terms of completeness with which participants were able to retrieve relevant screening procedures for the patient scenarios by using the paper-based guideline and CDSS prototype. After each session with the paper-based guideline and CDSS prototype, participants were asked to write down all patient-specific screening procedures for each scenario. For each patient scenario, participants’ lists of screening procedures were compared to the list of all relevant screening procedures as defined by the guideline for that patient scenario (verified by an expert physician).
each participant, the number of correctly identified screening procedures per patient scenario was counted, both while using the paper-based guideline or CDSS prototype. Per patient scenario and per participant, the percentage of correctly identified screening procedures with the paper-based guideline and with the prototype CDSS was calculated.

Efficiency was measured in terms of the time (minutes and seconds) participants needed to define the overall follow-up screening procedure for each patient scenario by using the paper-based guideline or prototype CDSS. The Wilcoxon signed rank test was used to test for statistical significance of differences in effectiveness and efficiency between the paper-based guideline and prototype CDSS.

**Results**

Thirteen healthcare practitioners of six follow-up clinics responsible for preparing patient follow-up visits participated in this study: three male and three female expert physicians, and seven female expert physician assistants. The participants had a mean age of 44.9 years, a mean of 4.8 years experience in long-term follow-up of cancer survivors. On a scale of 1-5, the mean self-reported computer experience was 4.4. Participants’ demographic data are given in Table 2.

**Table 2. Characteristics of healthcare practitioners participating in the paper-based guideline evaluation and formative usability study of the prototype CDSS.**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>13</td>
<td>44.9</td>
<td>7.53</td>
<td>27 - 56</td>
</tr>
<tr>
<td>Years of experience</td>
<td>13</td>
<td>*</td>
<td>4.67</td>
<td>0.5 – 15.0</td>
</tr>
<tr>
<td>Self-reported computer experience (scale 1-5)</td>
<td>13</td>
<td>4.4</td>
<td>0.51</td>
<td>4 - 5</td>
</tr>
</tbody>
</table>

* regarding years of experience, mean is not measured.

**Usability problems with the paper-based guideline and prototype CDSS**

Table 3 shows the number of unique usability problems per failure qualifier and severity that participants experienced with the paper-based guideline and prototype CDSS. Inter-rater reliability of the classification of the paper-based guideline Failure Quality and CDSS were determined by $\kappa = 0.394$ (P .000) and by $\kappa = 0.349$ (P .000) respectively, both considered as fair agreement [18]. In using the paper-based guideline, participants encountered 31 usability problems, 4 severe, 15 moderate and 12 minor. With the prototype CDSS participants encountered 22 usability problems, 5 severe, 10 moderate and 7 minor.
Table 3. Number of unique usability problems encountered with paper-based guideline versus CDSS prototype, per CUP failure qualifier and severity (minor, moder = moderate, severe)

<table>
<thead>
<tr>
<th>Failure Qualifier</th>
<th>Severity</th>
<th>paper-based guidelines</th>
<th>prototype CDSS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Minor</td>
<td>moder</td>
<td>severe</td>
</tr>
<tr>
<td>Incongruent</td>
<td>1</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Better way</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Wrong</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Irrelevant</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Overlooked</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>12</strong></td>
<td><strong>15</strong></td>
<td><strong>3</strong></td>
</tr>
</tbody>
</table>

We classified 17 usability problems (1 minor, 13 moderate and 3 severe) with the paper-based guideline as Incongruent, indicating that information presented in the guideline was unclear because it did not match with participants’ mental model or previous experience within the LATER clinic. The majority of these problems related to the content structured according to clinical domains. Participants’ initial search patterns through the paper-based guideline revealed that they tried to define individual screening procedures on the basis of the cancer treatment a patient had received. Participants would give up this strategy when they realized that the inclusion criteria for certain screening procedures were structured according to organ system and not according to treatment. The remaining 13 usability problems were classified as Missing (1 moderate), i.e. participants failed to find something that was expected to be present; Better way (4 minor and 1 moderate), i.e. participant suggested that something could have been done differently; and Irrelevant (7 minor), i.e. the tool contains information that is not needed by users.

Nine usability problems (1 minor, 5 moderate and 3 severe) with the prototype CDSS were classified as Incongruent. Compared to the incongruent problems detected in the paper-based guideline, most of the Incongruent usability problems with the prototype CDSS concerned the terminology used for labelling screening procedures. Participants were sometimes unfamiliar with the terminology used to label screening procedures, because of differing naming within their institute/clinic. Four usability
problems (1 minor, 2 moderate and 1 severe) were classified as Missing. We also encountered problems that we classified as Wrong (1 severe), i.e. participant can notice that something went wrong, and Overlooked (2 minor problems), i.e. participant fails to see an existing entity or fails to realize to interact with it. We found one Missing usability problem when one participant expressed concerns about the CDSS prototype’s validity and indicated to miss a link in the CDSS by which he could see scanned pages of the paper-based guideline to verify the CDSS advices. A similar problem was classified as Overlooked, which related to an information-button with which participant’s could look up evidence on why a screening procedure should be performed. One participant was looking for this information but failed to notice the information-button. The Wrong usability problem was classified as severe because it resulted in lower completeness of retrieved screening procedures. A participant accidentally deleted a screening procedure from the list and the prototype CDSS provided no option to retrieve the screening procedure. The participant could not recall which screening procedure had been deleted, which resulted in a lower completeness. The remaining 6 usability problems with the prototype CDSS were classified as Better way (3 minor and 3 moderate).

Effectiveness and efficiency of use of the paper-based guideline and the prototype CDSS

Effectiveness: Table 4 shows the mean percentage of completeness of relevant screening procedures retrieved while using the paper-based guideline and prototype CDSS per scenario. When using the paper-based guideline, participants on average retrieved 36% more relevant screening procedures for the complex scenario than for the simple scenario. With the prototype CDSS participants on average retrieved 5% more relevant screening procedures for the complex than for the simple scenario. For both the simple and complex scenarios, the Wilcoxon signed rank tests showed a statistically significant improvement in completeness of retrieved screening procedures by use of the CDSS prototype compared to use of the paper-based guideline. On average, participants showed an improvement of 58% in retrieving the relevant screening procedures for the simple scenario by use of the prototype CDSS compared to use of the paper-based guideline. For the complex scenario, participants showed an improvement of 17% by use of the prototype CDSS compared to use of the paper-based guideline.
Table 4. Percentage total completeness of retrieving screening guidelines and time on task per condition

<table>
<thead>
<tr>
<th>Condition</th>
<th>N</th>
<th>Scenario</th>
<th>Completeness (%)</th>
<th>Time (min:sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper-based guideline</td>
<td>13</td>
<td>Simple</td>
<td>42%</td>
<td>7:21</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>Complex</td>
<td>78%</td>
<td>12:59</td>
</tr>
<tr>
<td>Prototype CDSS</td>
<td>13</td>
<td>Simple</td>
<td>100%*</td>
<td>8:19</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>Complex</td>
<td>95%*</td>
<td>9:09</td>
</tr>
</tbody>
</table>

* p < .01

**Efficiency:** Table 4 also shows the mean time it took participants to retrieve screening procedures with the paper-based guideline and the prototype CDSS per scenario. When using the paper-based guideline, participants on average needed 5:38 minutes more to retrieve the screening procedures for the complex scenario than for the simple scenario. With the prototype CDSS, participants on average needed 0:50 minutes more to retrieve the screening procedures for the complex scenario than for the simple scenario. It took participants on average 0:58 minutes more to retrieve screening procedures for the simple scenario with the prototype CDSS; however resulting in 58% improvement in scenario completeness compared to their use of the paper-based guideline. For the complex scenarios, participants on average needed 3:50 minutes less to retrieve the relevant screening procedures, with 17% higher completeness when they used the prototype CDSS than when they used the paper-based guideline. For both the simple and complex scenarios, the wilcoxon signed rank tests showed no statistically significant difference between the times it took participants to retrieve the screening guidelines with the CDSS prototype compared to the paper-based guideline.

**Discussion**

**Main findings**

Paper-based formats of evidence based clinical guidelines can limit their usability since the knowledge contained in the guideline may not be easily accessible to healthcare practitioners [19]. Because of this, extracting information from a paper-based guideline and determining relevance for a specific patient can require additional effort from a healthcare practitioner. In this study we performed a think-aloud analysis in order to gain insight into current usability problems when using a paper-based
From Expert based guideline to User Centered Decision Support

guideline for follow-up screening of childhood cancer survivors. The results of this analysis provided input into a user-centred prototype CDSS to support healthcare providers in defining the overall follow-up screening procedures for childhood cancer survivors.

We analyzed types of usability problems healthcare practitioners encountered with use of the expert-driven paper-based guideline and user-centered prototype CDSS by CUP Classification. With the paper-based guideline we found 17 usability problems that we classified as Incongruent. With the prototype CDSS the number of Incongruent usability problems was reduced to 9, indicating a better fit with participants’ information processing strategies. We also found less Irrelevant usability problems with the prototype CDSS (0 usability problems) compared to the paper-based guideline (7 usability problems), indicating that the prototype CDSS did not contain unnecessary information. Although the amount of usability problems was reduced in the categories of Incongruent and Irrelevant, we did find new problems with the prototype CDSS in other categories. There were more usability problems with the prototype CDSS compared to the paper-based guideline in the categories of Missing (4 vs. 1), Better way (6 vs. 5), Wrong (1 vs. 0) and Overlooked (2 vs. 0).

The degree to which the healthcare practitioners’ effectiveness and efficiency in retrieving screening procedures was impacted was also assessed. With the paper-based guideline, healthcare practitioners on average retrieved 36% more relevant screening procedures for the simple scenario compared to the complex scenario. The verbal protocols revealed that in performing the simple scenario participants often relied on their clinical knowledge to look up screening procedures specific to the treatment the patient had received. Study participants often skipped the screening procedures that need to be performed for all childhood cancer survivors. Participants were more thorough in their search for information in the paper-based guideline for the complex scenario, resulting in a higher completeness. This is in line with the results of Patel et al. [20], who found that reliance on experience in solving clinical problems contributed to the skipping of steps from the clinical guidelines. With the prototype CDSS, the completeness with which study participants retrieved screening procedures improved with 58% for the simple scenario and 17% for the complex scenario.

Implementations of CDSS have shown to improve healthcare practitioners’ performance particularly concerning preventive care [5;21]. The study of Verhoeven [3] showed that a user-centered design process of a website yielded a more efficient and effective means of communication of infection control guidelines; healthcare practitioners’ performance in completing tasks as recommended by the guideline improved while it took them less time to complete the tasks. Their results showed that a user-centered design process of a website supporting healthcare providers in evidence-based practice maximized the effectiveness and efficiency of healthcare practice. We likewise used a user-centered design approach for development of a CDSS to support healthcare practitioners in retrieving follow-up screening procedures for childhood cancer survivors. Retrieving of screening procedures by healthcare practitioners significantly increased in terms of effectiveness (correctly identified screening
procedures) when using the CDSS prototype compared to using the paper-based guideline. However, a CDSS design is more likely to lead to an improvement in efficiency and effectiveness measurements compared to the manual process employed in a paper-based guideline, provided that its design optimally supports healthcare practitioners [22]. Suboptimal design may lead to new and severe usability problems, inhibiting the potential improvement in healthcare efficiency and effectiveness [22]. The user-centered design process in this study provided input to a prototype CDSS design in which healthcare practitioners could indeed review patient-specific screening procedures, points of interest for anamnesis, physical examination and advices, and possible late effects more effectively by use of the tab-based structure. The comparison of usability problems revealed in both the paper-based guideline and the user-centered prototype CDSS indicated that the CDSS tab-based structure better mapped to the mental model of participants in effectively retrieving screening procedures, but also introduced new usability problems. With regard to the UCD process of the CDSS; these new usability problems will be tackled in a redesign of the CDSS.

Minimizing the time spent on consulting clinical guidelines is crucial when attempting to improve the uptake of clinical guidelines in everyday practice [19]. The results of Verhoeven [3] showed a decrease of approximately 4 minutes for executing tasks with a website that communicated infection control guidelines instead of using the paper-based guidelines. In our study, the average time it took healthcare practitioners to retrieve screening guidelines for the simple and complex scenarios did not significantly improve with the prototype CDSS, although they needed on average 3.50 minutes less to complete the complex scenario.

A trusted knowledge base and trust in the peers developing a CDSS are considered important factors for an effective implementation of a CDSS [23]. When using the paper-based guideline, participants didn’t express any concerns regarding the validity of the information content. However, some participants indicated that they were not sure about the validity of the guideline information provided by the CDSS prototype. They expressed a need for the paper-based guideline to validate the information provided by the CDSS prototype. The clinical rules contained in the knowledge base were constructed according to the LERM procedure [17] and verified by a group of experts on late effects of childhood cancer treatment. This approach guarantees the validity of the CDSS knowledge base. It has been found that links to referential material is highly valued by users of decision support system [24]. The prototype CDSS did provide referential material to the evidence behind specific screening procedures. The material was provided through small information-icons listed next to the screening procedure. However, the information-buttons for disclosing this evidence were often overlooked by the healthcare practitioners. Moreover, the CDSS prototype did not provide evidence based referential information for the advices, anamnesis, physical examination and possible late effects. This information will be available in a redesign of the CDSS prototype system.
Limitations

The healthcare practitioners that participated in this study were recruited through an email that was send to the coordinators of the follow-up clinics. Participation in the study was voluntary, potentially leading to bias of the results. However, only a limited number of healthcare practitioners are specialized in late effects of treatment for childhood cancer (approximately 4 in each follow-up clinic). The number of participating healthcare practitioners is therefore representative of the entire population.

In classifying the usability problems, not all detected problems could be adequately mapped to CUP. One specific problem was actually more related to the unavailability of resources in a specific follow-up clinic; Ordering of a specific blood test specified by the guideline was not possible in one specific center. During the usability session of the paper-based guideline, one participant did mention the screening procedure but choose not to document the procedure while stating that it could not be ordered anyway. This resulted in a lower effectiveness in task completion respectively lack of adherence to the guideline. In this study we mapped this problem by incongruent, though we underline the fact that CUP needs to be augmented to additionally specify and include workflow related problems.

Also, our reported observed agreement in the classification of Failure quality for the usability problems was considered fair agreement. The CUP classifies usability problems on a high aggregation level, as such the reliability of the inter-rater agreement can be negatively affected [23]. However, all disagreements in usability problem classification were resolved after discussion which contributed to detailed analysis of the final CUP classification. The study participants did not receive any training or explanation prior to the usability analysis of the prototype CDSS. Training was not provided because we were specifically interested in observing first time user experiences regarding the ease of learning to use the system with the user-centered prototype design of the CDSS. We were specifically interested whether the CDSS design would intuitively and as such efficiently and effectively support end-users in determining follow-up procedures. However, in relation to assess the ease of learning, the formative usability results of the CDSS prototype are potentially influenced by participants’ exploratory behaviour of the CDSS interface and its functionalities. This resulted in more time to retrieve patient-specific screening procedures with the prototype CDSS than participants might have needed when they would not have explored the CDSS. Research has shown that until software is learned, additional time is typically required to complete daily tasks [25]. The additional time that participants needed to learn to use the CDSS prototype could also explain the slight decrease in efficiency for the simple scenario compared to the situation in which they used the paper-based guideline. Such a learning effect might have had a negative impact on the time participants needed to complete the complex scenario by use of the CDSS. In addition, this could potentially have influenced the type of usability problems that were found. Usability problems that were found during the
prototype CDSS analysis are problems with first time use. After pro-longed use of the CDSS in practice these usability problems might be overcome by end-users.

**Implications**

The prototype CDSS increased the effectiveness with which healthcare practitioners retrieved patient specific screening guidelines. Whether the system will also lead to increased adherence of healthcare practitioners to the LATER guideline in clinical practice needs to be further investigated. However, this study shows that it is important to incorporate healthcare participants’ information processing strategies when designing clinical guidelines in order to increase effectiveness of a guideline containing a high amount of clinical decision rules.

Evidence from clinical studies keeps accumulating and clinical guidelines often need to be revised [26]. The knowledge bases of CDSSs need to adapt to dynamics in clinical guideline revisions in order to sustain high system validity [27]. Participants in our study stated that they had concerns with the validity of information provided by the prototype CDSS, even though the clinical rules contained in the knowledge base had been validated by experts in late effects of childhood cancer treatments. Further research into contributing factors to distrust in guideline validity in CDSS development is likewise necessary.

**Conclusion**

This study provides sight on the use and application of usability evaluation of current practices in paper-based guidelines to provide input in a UCD of prototype CDSS. Our research demonstrated that usability problems experienced by healthcare practitioners when using a paper-based clinical guideline could be overcome by implementing the guideline in a user-centered CDSS design. Other usability problems were found in the CDSS prototype and should be taken care off in a redesign. Even though the CDSS was only a prototype healthcare practitioners were more effective with the prototype than with the original paper-based guideline in defining the screening procedures for individual patient scenarios. Healthcare providers did not become more efficient in retrieving these procedures by use of the CDSS, but their exploration behaviour and time needed to learn to operate the CDSS might explain these results.

**Acknowledgments**

The authors thank the healthcare practitioners of the LATER follow-up clinics for their participation in the paper-based guideline evaluation and formative usability study of the prototype CDSS.
REFERENCES


Chapter 9

Discussion - Summary of findings and Conclusion

Introduction

Health Information Systems (HIS) are increasingly being introduced in healthcare, among them patient information websites, systems supporting the ordering of laboratory tests and medication and clinical decision support systems for support of evidence based clinical practices. Even though their manifold benefits in increasing safety and quality of care are widely recognized, their integration into healthcare continues to be hampered by factors associated with their design and implementation process [1,2]. Disruption of clinical work practices, system immaturity, poor system usability, system inefficiency, and non-effectiveness in attaining quality benefits, all appear to impede successful implementation of HIS in healthcare [3,4].

User Centered Design (UCD) of HIS is advocated as the approach to tackle these problems by ensuring development of usable HIS [5]. UCD is at the core of the activities of the field of Human Factors engineering – mandated to design products and processes for human use [6]. Within UCD Human Factors (HF) engineering and usability evaluation is an essential step to achieve efficiency, effectiveness, satisfaction and safety of HIS in relation to their context of use. These evaluations may incorporate usability inspection and testing methods in different phases of HIS development. Attention to UCD and HF engineering and usability evaluation of HIS is growing vastly [7-9]. New research objectives and initiatives are evolving on how to integrate and apply HF/usability methods in UCD approaches of HIS [10, 11]. The main objective is to provide insight in methods ensuring the design and development of usable HIS, which might overcome the barriers to their successful implementation.

This thesis aimed to contribute to this goal, by conglomerating case-studies on (re)design and usability evaluations of differing HIS and analyzing and discussing the results in the context of UCD approach of these systems. Specific attention is on advancing the knowledge on the application and performance
of HF/usability methods in UCD of HIS and on assessing and comparing the methods value in identifying types of usability problems concerning various HIS and contexts-of-use. The final aim of this thesis is to provide key lessons learned on applying HF/usability methods in UCD of HIS.

In chapter 1, five research questions were formulated concerning HF/usability methods application in UCD of HIS. In this chapter the results from the case-studies of the previous chapters are synthesized and discussed to provide answers to these research questions. Subsequently the strengths and weaknesses of the research are assessed and the meaning of the findings is discussed. Finally, this thesis concludes with challenges and implications for new research on HF/usability methods in UCD of HIS.

**Summary of findings**

**Question 1 - How can HF methods contribute to investigating the human, social and organizational issues that influence HIS implementation?**

In chapter 2 three complimentary HF methods; interviews, ethnographic observation including self reported problem lists, and document analysis were used in a longitudinal study of a CPOE-Laboratory system. Triangulation of these methods allowed assessment and disclosure of human, social, organizational, and workflow related factors impeding its implementation. Semi-structured interviews with stakeholders involved in the design and implementation process of the CPOE system were held. These interviews were based on 36 CPOE implementation themes’ heuristically drawn from CPOE literature. The analysis of interview transcriptions combined with document analysis provided sight on eight themes that contributed to the CPOE system implementation failure. The central problem was a misunderstanding of the clinical work flow of laboratory ordering by system designers. Mismatches of system designers’ model of the ordering workflow and context-of-use of the CPOE system resulted in an immature system with low usability and lacking required functionalities. These aspects had a major impact on the vast delays in the system implementation process and its final failure in practice. The ethnographical study performed during the CPOE-L pilot implementation aimed at analyzing the extent of CPOE-L system integration into order workflow and adoption among its users. This analysis revealed that end-users indeed missed essential system functionalities and experienced usability problems in their daily use of the system. Self reported problem lists by end-users showed problems concerning user-system interaction and the system’s instability. In addition an
increase of 15 min in patients’ waiting time at the laboratories was reported, resulting from suboptimal user-system interactions.

Gaining sight on the intricate workflow processes in a healthcare setting is a fairly complex and elaborate task. In this study the application and triangulation of HF methods provided sight on factors surrounding the laboratory ordering workflow and how they interrelated and impacted the CPOE implementation process and its final failure. The experiences with the laboratory ordering system show important issues for CPOE systems to be successfully introduced and to be taken into account in future CPOE implementations. Understanding and consideration of (clinical) workflow aspects by project managers and the involved clinical organization is of extreme importance from the very start of a CPOE implementation process. In UCD approach of HIS, key lessons are to focus on an understanding of work flow and specification of the context of system use, on specifying the user requirements concerning a system’s functionalities and usability in detail. These activities are to be performed at predefined phases in the UCD and should continuously inform system design to build systems that will fit workflows. Further, a broad variety of implementation preconditions should be fulfilled, including sufficient time and human resources, commitment, good leadership and organizational readiness. In managing a system implementation process, focus should not be on (technical) sub-problems surrounding the implementation, but rather on the overarching relation between the development of the system and the clinical workflow the system is to be implemented in.

**Question 2 -** How do usability methods compare regarding their potential to identify usability problems concerning (prototype) HIS and what are ensuing merits of their application to UCD of HIS?

In **Chapter 3** the use of Cognitive Walkthrough (CW), a usability inspection method, and Think Aloud (TA) a usability testing method, are explored in detecting usability problems of a different nature in a prototype CPOE Laboratory system (OM/lab). A total of 33 usability problems were detected, of which 25 problems were revealed both by the CW, which all were validated in the TA sessions. Usability problems revealed with the CW concerned interface design flaws related to misallocation of buttons on the screen, incomprehensibility of button labels and feedback containing no relevant information to the user about the cause of errors made and consequences of a user's action. The TA additionally provided sight on user-system interaction problems concerning mismatches of the system model and system terminology with clinical work practices and user terminology. For example, the grouping of laboratory tests within the OM.lab system did not correspond with the way clinicians tended to group these tests in clinical work practice. The clinical domain specific usability problems were the main cause for omissions and errors in orders and cancelled orders in the usability evaluation study and in the pilot implementation of OM/Lab (described in **chapter 2**).
CW was also applied in combination with TA in chapter 5 in usability evaluation of the CPOE medication system. 30 out 42 usability problems detected by the CW in this study were validated by the TA user sessions. The TA however detected 15 more usability issues with in general a higher severity compared to CW. The TA specifically provided sight on usability issues that concerned users misunderstanding of system recommendations, unclear system terminology, and incorrect user-system interaction patterns leading to wrong medication selection. The CW focussed more on visibility of buttons and screen layout. In chapters 3 and 5 the TA user test sessions added to the results of the CW by revealing usability problems of a more domain specific and clinical context nature. The value of the TA method was in revealing usability problems in the system related to users’ domain knowledge and clinical working practices, which were not revealed in the CW evaluation of these systems. Usability problems revealed by the CW, which were all validated with the TA user sessions, were of a more general nature compared to the additional problems revealed with the TA user sessions. The CW method focuses on the cognitive activities of first-time users, especially on their goals and knowledge when performing a specific task. A CW simulates the cognitive behaviour of the user by responding to questions related to the user's cognitive model. TA user testing does not simulate the cognitive behaviour of end users but instead allows direct access to the cognitive processes and behaviour of end users while they interact with a system. In contrast to a CW, TA can reveal domain specific system design flaws because it provides sight on the effects of users’ domain knowledge and expertise on their behaviour and decision making while they interact with a system.

In Chapter 4 Retrospective Think Aloud (RTA) and Concurrent think aloud (CTA) were compared on effectiveness in a formative usability evaluation case-study of an Intensive Care registry Query Tool. Overall 43 unique usability problems were detected in the query tool before its redesign. CTA performed significantly better than RTA in detecting usability problems. CTA usability problem detection effectiveness was 0.80 vs. 0.62 (p<0.05) respectively. In addition CTA outperformed the RTA in detection of usability problems of a moderate (0.85 vs. 0.7) and severe nature (0.71 vs. 0.57). Both methods differed in their detection rate of different types of usability problems. CTA detected unique usability problems concerning graphics/symbols, user-navigation issues, system error messages and the organization of information on the Query Tool’s screens. RTA detected unique issues concerning mismatches of system language and terminology with users’ language and terminology. Qualitative analysis of the verbal protocols showed that RTA verbal protocols contained significantly more explanatory comments regarding the causes of usability problems and user comments concerning additional system requirements. This case-study shows that CTA and RTA differ in their detection scope of types of usability problems revealed in a HIS. Furthermore, the CTA and RTA differ in the extent to which participant think aloud verbalizations complement the observable usability problems. During CTA the verbal protocols reflect users’ cognitive processes of task performance thereby leading to a higher effectiveness in detecting usability issues related to the direct user task.
performance. Compared to CTA, the RTA seemed to evoke a more rational and reflective report on task information processing behaviour as the computer supported tasks had already been completed.

CTA method is the primary method to be advised in evaluating a design (solution) or prototype HIS. The CTA was not only more thorough in detecting usability issues, it application was less resource intensive compared to RTA. Application of RTA almost doubled the analysis time to assess usability problems. However, in upgrading a system, or in defining new functionalities of a system, the RTA method might prove more beneficial to system redesign efforts. RTA provides not only more explanatory user comments, but also provided insight into new or potential system functionalities that exceeded the first objective of improving the usability of the Query Tool. New user-requirements for specific users groups could be determined. More research is necessary to assess the applicability of the RTA for these objectives in system redesign efforts.

In chapter 6, Mock-ups, focus groups and Heuristic evaluation (HE) with double experts and user testing with Think Aloud (TA) were applied in a concise UCD approach for a patient information website for childhood cancer survivors. First, extensive information needs analysis was performed, by means of a survey completed by 145 potential end-users. Understanding of the context of use of the patient website was the main focus of the information needs analysis. The resulting analysis provided input into the content and format of the website to be developed. This content was edited and peer-reviewed by a team of medical professionals. To produce a first design, mock-ups for the website were built based on the results of the information needs analysis. These mock-ups were subsequently discussed in focus groups in which users and clinicians could offer their opinion on the future contents and design of the website. In UCD focus groups are often employed near the end of the development process, after system usability testing or during a system redesign iteration. In addition, the mock-ups of the website provided input to the ‘producing a prototype’ phase of the UCD. Focus group participants were able to pinpoint to possible (usability) problems with both the navigational structure of the site and search structure for late effects in formation before the design of the prototype website started. This study showed that involvement of a focus group with all the stakeholders early in the development process resulted in a more detailed specification of requirements, and in a close match of user requirements with final website content, functionality and navigational structure. The involvement of users and the information gathered during these two initial phases was valuable for the development of a usable patient information website.

Heuristics (HE) inspection method and Think Aloud (TA) user-testing method were applied in the evaluation of the prototype website. The HE was performed by three double experts (knowledgeable both in usability engineering and childhood cancer survivorship), by use of the Nielsen Heuristics. The HE provided sight on 40 occurrences of usability problems of which 29 issues represented unique problem types, of which the majority had a low severity. The TA detected eleven unique usability problem types. All major usability issues discovered in the heuristic evaluation were likewise revealed
by the think aloud user test sessions and vice versa. These findings are in line with previous research on HE, were HE is known to find a higher number of usability issues with both low- and high severity, at much lower costs than the TA[12]. TA user testing of systems is nevertheless considered imperative for uncovering usability problems directly perceived during interaction of users with a system. The result from our case-study is not in line with those of previous studies. The TA in our study did not add to the results of the HE besides revealing two minor usability issues concerning website contents. This difference in study outcomes might be due to the inclusion of two double experts in our study, presumably enabling them to take on the role of a patient in testing our prototype website. When time and resources are limited for usability evaluation of a HIS and double experts are not available, the involvement of a ‘work domain expert’ in HE of a system’s usability might be a cost-effective means to reveal both general and domain specific usability problems in a system’s design. In circumstances that evaluators are not familiar with the system domain, a work-domain expert could assist the evaluator in tackling clinical context specific problems. These work domain experts in these so-called participatory heuristic evaluations may help usability evaluators in considering how the system contributes in reaching certain goals in their particular area of skills. The concise UCD methodology applied, in combination with the extensive information needs analysis, resulted in a website that accommodated end users’ information needs and which they found easy to use.

**Question 3 - How should usability problem descriptions be classified so that they can be effectively reported on?**

A known model (based on phases of human-system interaction), the User-Action Framework (UAF), was augmented with a severity rating and classification for expressing the potential impact of usability problems on final task outcomes. This augmented model was applied in classification of usability problems of a CPOE system for medication ordering (chapter 5). All 57 usability problems detected in the CPOE system could be classified according to the phases of the UAF; planning (establishing goals, tasks, and/or intentions), translation (translating intentions into plans for physical actions), physical action (making physical input actions) and assessment (perceiving, understanding and evaluating system response to the physical action). This study showed that usability problems in the CPOE system concerning the assessment phase of user-system- interaction (users missing or misunderstanding system responses following their actions) need more attention in the redesign process than other usability problems because of their severity and potential contribution to medication ordering errors.

In chapter 7, UAF classification was applied in the analysis of usability problems revealed in the NICE registry query Tool before and after its redesign. UAF classified usability problems were
Discussion and Conclusions

assessed in relation to intensivists’ task efficacy and efficiency to determine their impact on intensivists cognitive workload pre and post system redesign. Twenty-three of the 36 usability problems in the pre-test concerned the ‘planning/translation’ phase (64%); these problems were responsible for the high cognitive workload of system users, exemplified by low values in task efficacy and efficiency. After redesign of the tool, 30 new usability issues were detected in the post usability evaluation of the redesigned tool. The majority of these usability issues yet concerned the ‘assessment’ phase (63%) of user-system interaction. These usability problems did not or only minimally affect intensivists’ cognitive task workload, illustrated by high values of task efficiency and effectiveness (98% of the users were able to accurately perform the tasks with a reduction of 30.36 min in overall time on task). The usability problems related to the assessment phase did not increase users’ cognitive workload in performing the query tasks by use of the redesigned tool. Usability problems classified in the ‘Assessment’ phase of user-system interaction were mostly related to users’ preferences concerning interface layout, such as graph colour usage and the display of system information and feedback on the screen. Usability problems exercising a high cognitive burden on system use were thus successfully reduced in just one redesign iteration of the Query Tool. Future studies that apply the TA method in user testing of systems in the UCD phase of ‘producing design solutions’, should focus redesign efforts on those aspects of the system that affect the cognitive task load of users. This is especially important when high cognitive task workload of complex HI system tasks is seen as a major barrier for system use.

The classification of usability problems according to phases of human-system interaction in chapter 7 focussed on their impact on healthcare providers’ work load instead of their impact on patient safety as with the CPOE system studied in chapter 5. The CPOE system produced user problems related to the assessment phase of human-system interaction that could have a major effect on its safe use. Usability problems in the assessment phase concerned users overlooking or misunderstanding system feedback on CPOE specific medication alerts which could impact the completeness and accuracy of orders and ultimately lead to inaccurate and wrong medication orders. These usability problems in the ‘assessment’ phase in the CPOE system had a more negative impact on task performance compared to the usability problems in the assessment phase of the Query Tool studied in chapter 7. Since the Query tool is focused on providing intensivists benchmark information on the quality of care delivered, its usage will not directly impact patient safety. The application of the augmented UAF for classification of the tool’s usability problems in relation to their impact on patient safety would therefore have been of no value. The objective of the usability evaluation of the Query tool was thus focused on its context of use: efficient and effective Query development by its users. In fact, UAF classification of the Query tool’s usability problems showed that the problems revealed in the ‘planning’ phase prevented users from using it effectively and efficiently. These usability problems concerned users not knowing what action to perform next to achieve a certain task in the system. The tool appeared to lack visual cues to guide them to the correct action. The ensuing exploration and error
Chapter 9

recovery behaviour of users in planning the next action contributed to the high cognitive task workload. The analysis and classification of problems by UAF provided input to the redesign of the tool, which showed to support users better in understanding how to accomplish the creation of a clinical query. The study succeeded in designing a highly intuitive user interface for the Query tool, with minimal cognitive user task effort. Overall, application of the UAF in classification of usability problems concerning differing types of HIS may contribute to accurate and effective problem reporting and supports system (re)designers in understanding usability problems, their underlying causes and prioritizing problems to tackle in a system redesign.

The studies indicated that the augmented UAF is able to improve classification, analysis and prioritizing of usability issues with potential threats to task outcomes and patient safety. Since the UAF concentrates on the phases of human-system interaction in which usability problems occur, it likewise provides a grounded starting-point to assess system usability problems related to users’ cognitive and or physical efforts in performance of tasks according to phases of human-system interaction and profound insights into underlying causes of usability problems. These insights subsequently can support system designers in making reliable informed decisions on (re)design of a system. Especially in healthcare contexts, informed decisions on (re)design of a system are helpful in tackling usability problems exercising safety-critical effects on health practices.

The augmented UAF classification applied in these studies seems to overcome limitations of other classification models such as bottom-up, severity and heuristic classification of system usability problems. The UAF classifies problems in mutually exclusive problem categories. Use of UAF minimizes bias of evaluators’ subjective classification of problems by providing a layered classification model that takes into account the severity and impact of a problem on task outcomes. Application of this framework however is labour-intensive and its use in usability evaluation studies with limited time and resources available, needs further consideration. However, consistent use of the augmented AUF framework in publications on usability studies may support the analysis of trends and patterns across problem sets of HIS usability evaluation case studies. This will support the building of HF/usability knowledge bases in the healthcare domain.

**Question 4 - How does Think Aloud usability testing contribute to redesign of HIS in terms of usability problem types detected, efficiency and effectiveness of users’ task performance before and after redesign?**

In chapter 7 TA usability testing was performed to guide the redesign of the web-based Data Query Tool of a National ICU Quality Registry. The TA usability test of the original tool gave sight on 36 usability problems, whereas 35 usability problems were revealed after redesign of the tool. Assessing the usability of a system before and after its redesign in terms of number of problems detected poses a bias to the study results. Usability evaluation is an iterative process; subsequent changes to a user
interface design might reveal other problems that again need user testing [12]. The accuracy of the comparison of the usability problems pre and post redesign was enhanced by classifying them by UAF. Of the usability problems detected in the pre-test 34 (94%) were resolved by redesigning the Query Tool, 5 (14%) were considered overlapping with usability problems found in the post-test and 30 new usability problems were revealed in the post TA Test. The input of the TA in guidance of redesign of the query tool led to a clear reduction in the tool’s complexity and hence in the cognitive task workload of users working with the redesigned tool. Optimization of the task accuracy was obtained from 50% of tasks completed in the pre-test to 96% of similar tasks completed in the post test. Users’ efficiency, in terms of overall time on task, was highly improved with a reduction of 30.36 min after redesign of the Tool. This indicates that the new usability problems detected in the post TA test minimally affected their cognitive task workload in use of the new tool.

In chapter 8 a UCD approach was applied in the development of a Clinical Decision Support System (CDSS) for the guideline based retrieval of childhood cancer survivors follow-up screening procedures. First, TA was applied to reveal usability problems and clinicians effectiveness and efficiency in defining follow up screening procedures with the paper based guideline. Simple as well as more complex patient scenarios were applied in the user testing. Main usability problems of the paper-based guideline resulted from: unclear patient inclusion criteria for specific follow-up procedures, inconsistencies between the screening procedures, and unclear descriptions of the frequency with which specific procedures needed to be performed. Also, the general organization of the paper-based guideline severely hampered clinicians in effectively defining a follow-up procedure for a specific patient scenario. The content structured according to clinical domains proved to be inconvenient and did not match participants’ information processing strategies in defining screening procedures for the individual patient scenarios. The resulting problems in the paper-based guideline informed the design of the CDSS prototype. The prototype design solution was subsequently evaluated by means of TA user testing to assess the type of usability problems experienced by clinicians in the CDSS. Their effectiveness and efficiency in retrieving screening procedures for the patient scenarios by use of the CDSS prototype in comparison to usage of the paper-based guideline was also assessed. CDSS’ usability problems mainly concerned: unclear content terminology, unclear button functionality and system inflexibility related to users not being able to retrieve a guideline once they had deleted it. Analysis of participants’ effectiveness when using the CDSS instead of the paper-based guideline showed that they on average significantly improved in completeness of retrieval of screening procedures. Time measures on efficiency in retrieving screening procedures by the CDSS instead of the paper-based guideline proved not significantly different. This study illustrated a succinct UCD approach in which TA user testing provided sight on clinician information needs and information processing patterns in defining follow-up procedures for childhood cancer by use of the paper-based
guideline. The prototype CDSS supported users in effectively defining the follow-up screening procedures for cancer survivors. Further testing of the prototype system in practice is needed to gain sight on its actual usage and support of evidence-based practice.

**Question 5 – What are the key lessons learned of application of HF/usability methods concerning the UCD phases of HIS?**

Key issues to consider in conduct of each UCD phase:

1. **Understanding and specifying context of use**
   - Deficiencies in understanding clinical workflow processes and the context of system use leads to inadequate definition of system requirements and may result in system immaturity, under functionality, low system usability, and vast delays in system introduction, all of which may contribute to system failure. (chapter 2)
   - Triangulation of HF methods, (e.g. interviews, ethnographic observation and document analysis) provides sight on human, social, and organizational factors and their interrelations impacting the outcome of a system implementation process. These insights offer system implementers the possibility to adjust their implementation strategy and reach successful system implementation. (chapter 2)
   - Application of the Think Aloud method in this UCD phase provides early input to a HIS design that is usable and complies with user needs and requirements. (chapter 8)

2. **To specify the user requirements**
   - Involvement of focus groups in this UCD phase may result in a more detailed specification of user and system requirements, resulting in a closer match between HIS content, functionality and navigational structure meeting end-users expectations and needs. (chapter 6)
   - Application of the Think Aloud method can provide sight on user cognitive task behaviour in a certain clinical context as input to defining system requirements and its subsequent design. (chapter 8)

3. **To produce design solutions.**
   - Application of the augmented UAF classification framework provides a consistent means to assess the potential impact of usability problems of HIS on safety critical incidences (Chapter 5) and cognitive task workload (Chapter 7) for prioritizing these problems in HIS redesign efforts.
   - Mock-ups can be adequately used in this UCD phase in revealing possible (usability) problems with the simulated interface structure of HIS before design of a working prototype. (Chapter 6)
   - The CW provides sight on usability problems likewise experienced by users in TA testing and is a useful method for usability evaluation of HIS. (Chapters 3 and 5)
The TA user test sessions added to the results of a CW by revealing usability problems of a more domain specific and clinical context nature that may severely impact users’ performance of computer-supported tasks. (Chapter 3)

Assessment of the impact of usability problems on system users’ task performance may help in prioritizing problems to be solved in iterative system (re)design. (Chapter 3)

Concurrent (CTA) and Retrospective (RTA) Think Aloud methods differ in their detection focus of different types of system usability issues and may be applied in different phases of UCD of HIS. (Chapter 4)

− CTA is the primary method to be used in evaluating a design (solution) or prototype system. The CTA not only is more thorough in detecting usability issues, its application is less resource intensive compared to RTA. (Chapter 4)
− RTA provides more explanatory user comments on usability problems experienced by end-users. In addition, RTA provides insight into new or potential system functionalities, or new user-requirements for specific users groups. More research is required to assess the applicability of the RTA for these objectives in system redesign efforts. (Chapter 4)

The results of a HE of a system design may be improved by involving usability experts familiar with the clinical domain of system implementation. (Chapter 6)

If a UCD of a system is accurately performed, with fundamental input from end-users from the early phases on, and double experts for HE of a system’s usability are available, the performance of a TA may not be cost-effective (Chapter 6).

4. To evaluate design against requirements;

− The concise UCD methodology applied in development of a patient information website resulted in a website with high usability. (Chapter 6)
− TA user tests of a tool can provide useful input to the final tool design in terms of a more easy-to-use tool requiring less cognitive task workload of its users. (Chapters 7 & 8)

Strengths and Weaknesses of this thesis

The case-studies in this thesis have been conducted on differing types of HIS and provide a broad overview of the application of HF/Usability methods in HIS UCD processes. The case-study evaluations were conducted on a relatively small scale, which is typical for many software development efforts at large academic and healthcare centers [13]. Though the studies are all case-studies, the research advanced the knowledge on the value of HF methods to UCD approaches of HIS. By applying the lessons learned in new case-studies and researching new HF and usability methods, key lessons may be expanded, and contribute to enhancing the success of future HIS.
With regard to the HF/usability methods applied in this thesis, the performance of the Heuristic Evaluation, Cognitive Walkthrough, and Think Aloud methods all require evaluators with expertise in Human Factors and usability engineering. The validity of results of usability inspection and testing depends on the proper implementation of these methods in each UCD phase. To evaluate the potential of Think Aloud user testing in revealing user problems in different types of HIS and contexts, careful consideration should be given to how the method was applied and to what level of detail verbal protocols were analyzed. Ericsson and Simon put forth a complicated process for analyzing verbal protocols of TA studies that would be too time-consuming for most usability studies of HIS. In the usability evaluation studies of this thesis, time-pressures to deliver results was experienced. The usability study of the CPOE-L system did not contribute to the system redesign effort due to time restrictions related to the progress the implementation team needed to make with the system implementation. The study on the Intensive Care registry Query Tool was not hampered by time restrictions or pressures of the system design team. The time needed for the usability evaluation and the redesign of the system was taken in consideration by the software development team and other stakeholders. Also, the stakeholders of the system acknowledged the importance of an intensive user needs analysis with input from a usability test of the pre-existing ‘Query Tool’ for guidance of its redesign. 'Overall, a strength of this research was that the HF/usability methods were consistently applied, facilitating the comparison of HF/usability methods outcomes and assessment of their value in UCD of HIS.

Conclusion and Directions for further research

The studies in this thesis all aim to advance the knowledge on HF/ usability method application and their merits in UCD of HIS. Key lessons learned are presented, on which new research objectives may be based and further insights on HF/ usability method application in development of HIS may be generated. Attention to HF/ usability in UCD of HIS is vastly growing due to several recent initiatives of the National Institute of Standards and Technology (NIST), the Agency for Healthcare Research and Quality (AHRQ) and the Healthcare Information and Management System Society (HIMSS) [14]. In 2011, NIST released the (NISTIR 7769) Human Factors Guidance to Prevent Healthcare Disparities with the Adoption of Electronic Health Records (EHRs). Their goal is to identify best practices, guidance and standards in software usability and accessibility to ensure HIS support meaningful use by users that reflect the make-up of the healthcare. The AHRQ has provided support to HIT research and dissemination, including in the areas of HF and usability of EHRs [15], HF of clinical decision support [16;17], usability and workflow integration of HIT [18], and HF in the design of consumer health IT [19]. HIMSS usability task force focuses on promoting and initiating usability of HIS in health organizations [20]. Ultimately, the primary contribution to HF/ usability knowledge
related to HIS comes from scientific reports. It is therefore important that publications of HF/usability studies are of a certain degree of quality. A systematic review on usability studies in Health Informatics showed that publications in this domain lack in quality and structure. Information in these studies concerning system application domains (e.g. the type of system evaluated), case study objectives, stages of development, research methods applied, and study outcomes extremely differentiate [21]. In addition, these studies appear to be inconsistent and incomprehensive in their reporting. Because of this, comparison of study outcomes and method application, and overall generalization of study findings is hampered. To build an evidence base of sound HIT design and usability principles, reporting of these kinds of academic studies should be of a certain degree of quality: complete, homogeneous and unambiguous.

In continuation of advancing the knowledge base of HF/usability methods, an important aim is to build on the evidence base of HF/usability studies concerning HIT. Without a framework providing guidance of the reporting of HIT design and usability evaluation studies, the building of a proper evidence base of usability and design principles of HIT that lead to safe use in practice would remain hampered. The author of this thesis, part of the Usability working Group HITLab, managed by prof. dr. Monique Jaspers, started a collaborative effort with Evalab, a User Centred Usability working group from Lille, managed by Dr. Marie-Catherine Beuscart Zéphir, to develop a framework of good practice of reporting on HIT design, development and usability studies, in collaboration with the IMIA and EFMI WGs on Human (Organizational) Factors for Health Informatics. The framework aims to provide a set of principles to follow for comprehensive and unambiguous reporting of HIT design and usability evaluation studies with the objective to reduce variation and improve on the publication reporting quality of these studies.

To develop the framework, issues to be addressed in publications of HIT design and usability evaluation studies were identified by analysing all ISO and ANSI/HFES Standards concerning usability and human factor engineering and/or evaluation within the entire human centred software development life cycle. On this basis, a Delphi study aimed at establishing an international consensus regarding guidelines for good scientific reports of HF/usability studies of HIT is initiated.

The ultimate goal of the framework is to raise the quality of scientific research and publications in this domain, enhance the evidence base of HF and usability studies in health informatics and to learn from HIT design principles that improve or compromise its safe and efficient use in clinical practice.

To conclude, the HIS road to success remains under construction.
REFERENCES


16. Karsh BT. Clinical practice improvement and redesign: How change in workflow can be supported by clinical decision support. Rockville, Maryland: Agency for Healthcare Research and Quality; 2009. Report No.: AHRQ publication No 09-0054-EF.


**Samenvatting - Dutch Summary**

**Inleiding**

Tientallen jaren van onderzoek naar de introductie van Zorg Informatie Systemen (Health Information Systems, HIS) heeft ons talrijke lessen geleerd en factoren onthuld die samenhangen met waarom deze systemen falen of succesvol gebruikt worden in de klinische praktijk. Toch blijft het succes van deze systemen voor de zorg een voortdurend onderwerp van discussie. Hoewel nieuwe publicaties licht werpen op de complexiteit van het ontwerp en de invoering van zorg informatie systemen, hebben de geleerde lessen niet geleid tot betere en meer succesvolle implementaties. Schattingen uit 2010 tonen een alarmerend percentage dat 40% van deze systemen falen in de praktijk en daaruit worden teruggetrokken. Desondanks wordt de invoering van HIS beschouwd als de manier om de gezondheidszorg te voorzien van veelbelovende tools voor het verbeteren van de efficiëntie, effectiviteit en veiligheid van (geleverde) zorgprocessen. Verschillende studies hebben inderdaad positieve effecten van HIS op de gezondheidzorg aangetoond, maar er worden ook case-studies gepubliceerd die de link leggen tussen HIS introductie en de onbedoelde gevolgen en schadelijke effecten voor de gezondheidzorg. Deze onbedoelde gevolgen omvatten onvoorziene wijzigingen in werk en communicatiepatronen, veranderingen in de rollen en verantwoordelijkheden in de organisatiestructuur en technologie geïnduceerde fouten als gevolg van slechte bruikbaarheid (usability) van HIS door zorgprofessionals. Usability in deze context verwijst naar de mate waarin een systeem of systeem-functionaliteit kan worden gebruikt door bepaalde gebruikers om bepaalde doelen met effectiviteit, efficiëntie en tevredenheid te bereiken in een bepaalde context van het gebruik.

In plaats van het verbeteren van de kwaliteit van de zorg, kunnen gebrekkig ontwikkelde HIS dus slechts marginaal ondersteuning bieden aan hun gebruikers, wat leidt tot extra werk en mogelijke belemmeringen van (klinische) gezondheidszorg en het zelfs in gevaar brengen van de veiligheid van de patiënt. Als gevolg daarvan is 'Usability' uitgegroeid tot een belangrijke factor die van invloed kan zijn op de implementatie van HIS en de acceptatie daarvan door beroepsbeoefenaren in de gezondheidszorg. De ontwikkeling van HIS elke eenvoudig te gebruiken zijn en waarbij de systeem functionaliteiten effectieve, efficiënte en veilige (klinische) taakuitvoering ondersteunen, is een van de belangrijkste uitdagingen van nu. Het begrijpen en optimaliseren van de manier waarop mensen omgaan met technologie is het onderwerp van het onderzoeksgebied van Human Factors (HF) en usability engineering. Studies over informatiesysteem ontwerp en implementatie in andere veiligheidskritische industrieën, zoals de luchtvaart, hebben aangetoond dat door het toepassen van methoden uit de HF / usability onderzoeksgebied in gebruikers gericht ontwerp (User-Centered Design (UCD)), aanzienlijke voordelen en vooruitgang kan worden behaald in het ontwerp van effectieve,
efficiënte en veilige HIS. UCD is een gestructureerde methode voor het ontwikkelen van systemen, die zich richt op toekomstige gebruikers van een systeem. De focus van UCD is om een systeem te ontwikkelen dat gebruikers behoeften voldoet, terwijl rekening wordt gehouden met hun beperkingen en voorkeuren. In een UCD benadering van een HIS wordt HF/usability engineering toegepast tijdens de ontwikkeling van een systeem.

In dit proefschrift worden verschillende onderzoeken beschreven die zicht bieden op de belangrijkste lessen die zijn geleerd bij de toepassing van HF/usability methoden in UCD aanpak van zorg informatie systemen, HIS. De studies hebben betrekking op verschillende categorieën van HIS, waaronder order management systemen voor lab aanvragen en medicatie, een klinische query tool voor analyse en benchmarking van Intensive Care registergegevens, een patiënt georiënteerde webgeoriënteerde HIS en een klinisch beslissing ondersteunend systeem. Door de toepassing van verschillende HF/usability methoden, in verschillende fasen van UCD van deze informatiesystemen, wordt een breed perspectief van de toepassing en de waarde van HF/bruikbaarheid methoden in UCD van HIS gegeven. Het uiteindelijke doel van dit proefschrift is om de kennis over de ontwikkeling, evaluatie en implementatie van HIS te verrijken en toekomstig succesvolle introductie van HIS te bevorderen.

**Hoofdstukken**

Hoewel order-management-systemen vaak onderwerp van studie zijn, blijkt het ontwikkelen van deze systemen die artsen optimaal ondersteunen in het aanvragen van testen nog steeds een uitdaging. In **Hoofdstuk 2** worden de resultaten besproken van een evaluatie van een prototype order-management-systeem voor het elektronisch aanvragen van laboratoriumtests. Er zijn twee complementaire cognitieve evaluatiemethoden toegepast, namelijk de Cognitive Walkthrough (CW) en de Think-Aloud of hardop-denken-methode. Hiermee is een analyse gemaakt van de usability problemen in het prototype systeem en van het potentieel effect van deze problemen op de kwaliteit van de aanvragen. Deze studie richt zich voornamelijk op het effect dat deze usability problemen hadden op de kwaliteit van het aanvragen in termen van efficiëntie en fouten in de aanvragen. Het toepassen van de Cognitive Walkthrough-methode leidde tot een raamwerk van potentiële usability problemen, dat vervolgens werd toegepast in de analyse van de hardop-denken-protocollen van 7 gebruikers in interactie met het systeem. In totaal deden zich 33 usability problemen voor, die inderdaad in meer of mindere mate tot ondoelmatig aanvraaggedrag en onvolledige aanvragen leidden. De belangrijkste problemen hadden betrekking op misvattingen van de gebruiker over de acties die deze in het systeem moest uitvoeren bij het aanvragen van laboratoriumtests. De belangrijkste conclusie is dat het merendeel van de problemen in het gebruik van het prototype terug te voeren zijn op het feit dat het prototype onvoldoende aansluit bij de huidige manier van werken.
In Hoofdstuk 3 worden de menselijke, sociale en organisatorische vraagstukken rond de invoering van het order management systeem voor laboratorium aanvragen, dat in hoofdstuk 2 is geëvalueerd, gedurende het implementatieproces geanalyseerd en worden de onderling effecten op de uiteindelijke uitkomst van de implementatie onderzocht. Dit onderzoek richt zich in het bijzonder op de geleerde lessen binnen dit proces en doet aanbevelingen hoe dergelijke uitdagingen van menselijke, sociale en organisatorische aard rondom een order-management-systeem implementatie ondersteund kunnen worden. Voor de basis van het onderzoek zijn implementatie thema’s die bekend zijn van de order-management-systeem (Computerized Physician Order Entry, CPOE) literatuur, geïdentificeerd. De resulterende set van thema’s werd toegepast als referentiemodel voor 20 semigestructureerde interviews tijdens het implementatieproces van het systeem met 11 personen die betrokken zijn bij het project, en het werd toegepast bij de analyse van alle gerelateerd project documentatie. Additionele gegevens werden verzameld door het uitdoen van vragenlijsten onder gebruikers, gebruikers gespreksrondes en via een etnografische studie uitgevoerd aan de betrokken klinische en laboratorium afdelingen. Bij het analyseren van interview transcripten, project documentatie, gegevens van de gebruiker vragenlijsten en de discussie rondes is een 'grounded theory' benadering toegepast om knelpunten of problemen in het implementatie proces te identificeren. De centrale probleemgebieden met betrekking tot de implementatie van het systeem en hun onderlinge relaties zijn afgebeeld door middel van een conceptueel interpretatief model. Het gebrek aan inzicht in de klinische workflow, oftewel de stroom van klinische activiteiten, werd geïdentificeerd als een belangrijk thema dat onder invloed van organisatorische, menselijke en sociale problematiek, het gehele implementatieproces op een negatieve manier heeft beïnvloed. Sterke vertragingen in de introductie van het systeem op de werkvloer, onuitgewerkt systeem ontwerp en onderfunctionaliteit kan allemaal rechtstreeks worden toegeschreven aan een oppervlakkig begrip van de klinische workflow. Dientengevolge werd de integratie van het systeem in klinische en laboratorium workflows geremd door zowel eindgebruikers als afdelingsmanagers en was de uiteindelijke terugtrekking van het systeem onvermijdelijk. Deze case studie toont aan dat menselijke, sociale en organisatorische aspecten die relevant zijn voor de uitvoering CPOE systemen cumulatief leidde tot een mislukking in de implementatie van de order management systeem voor lab aanvragen. De ervaringen en overwegingen die in dit artikel beschreven worden zijn van belang voor de succesvolle introductie van order-management-systemen (CPOE) in de zorg en dienen te worden gebruikt als basis in toekomstige CPOE implementaties. Het begrip en de afweging van (klinische) workflow aspecten door project managers en de betrokken klinische organisatie is van zeer groot belang vanaf het begin van een CPOE implementatieproces.

In Hoofdstuk 4 ligt de focus op het vergelijken van de prestaties van twee verschillende usability evaluatie methoden, de Concurrent (CTA), of te wel tijdens taak uitvoering Hard op Denken, en

De usability evaluatie van de Data Query tool leverde een totaal van 43 unieke usability problemen op. CTA was significant beter dan RTA in het opsporen van usability problemen. De effectiviteit van CTA in detectie van usability problemen was 0,80 versus 0,62 (p <0,05) respectievelijk met een gemiddeld verschil van 42% minder tijd die deelnemers nodig hadden in vergelijking met RTA. Daarnaast was CTA grondiger in het opsporen van usability problemen van een matige (0,85 versus 0,7) en ernstige aard (0,71 versus 0,57). Met betrekking tot probleem-type detectie, werd door CTA unieke usability problemen in tekens/symbolen, navigatie problemen, foutmeldingen en de organisatie van de informatie op de schermen van de Query Tool's gedetecteerd. RTA detecteerde unieke problemen met betrekking tot systeem match met gebruikers' taal en toegepast terminologie. Vanuit de kwalitatieve analyse van de CTA en RTA verbale protocollen bleek dat RTA verbale protocollen significant meer toelichten over de oorzaak van een usability probleem en daarnaast meer opmerkingen over aanvullende systeemvereisten bevatte. Concluderend, CTA is effectiever in usability probleem detecteren, maar niet beter dan RTA. RTA biedt bovendien zicht op unieke usability problemen en nieuwe behoeften van de gebruikers voor specifieke gebruikersgroepen. Op basis van de resultaten van dit onderzoek adviseren wij het gebruik van CTA in formatieve usability evaluatiestudies van zorg informatie systemen. Wij adviseren echter verder onderzoek naar het gebruik van RTA in usability studies gericht op gebruikers profiel op maat in (her) ontwerp.

In Hoofdstuk 5 wordt een bestaand classificatiesysteem uitgebreid ten behoeve van het classificeren van usability probleembeschrijvingen bij het evalueren van zorg informatiesystemen. Door middel van deze classificatie is het mogelijk te bepalen welke problemen met voorrang dienen te worden opgelost gedurende systeem ontwikkeling en herontwikkeling. Het schema is getest op usability problemen die tijdens de usability evaluatie van een order-management-systeem voor medicatie aanvragen, zijn gedetecteerd. De gedetecteerde usability problemen in het systeem werden geclassificeerd met behulp van een classificatie raamwerk ontwikkeld in het Usability Engineering Domein (het User Action
Framework, UAF). Om de prioriteit van het oplossen van bepaalde problemen te kunnen vaststellen, achteraf Hardop denken, werden de invloed van het probleem op de interactie van de gebruiker met het systeem en de invloed van het probleem op de taak uitkomst bepaald. OM de invloed van het probleem op de interactie van het systeem met de gebruiker te kunnen bepalen werd de ernst van de problemen bepaald op basis van de mate van voorkomen, de persistentie en de uitwerking op de gebruiker. Aangezien usability problemen van zorg informatie systemen indirect de gezondheid van patiënten kunnen beïnvloeden, werden de usability problemen ook gerangschikt op basis van hun mogelijke invloed op de uiteindelijke uitkomst van een door gebruikers uitgevoerde taak in het systeem. Deze mogelijke invloed op de uitkomst van aan taak werd bepaald voor elke fase van de gebruiker-systeem interactie. De mogelijke invloed werd geclassificeerd als: verkeerde geneesmiddelennaam, verkeerde dosis, verkeerde toedieningfrequentie, verkeerde duur en verkeerde toedieningroute. De beoordelaars konden alle usability problemen die tijdens de usability evaluatie van het order-management-systeem werden gedetecteerd met behulp van deze classificatie classificeren en een prioriteit geven voor herontwikkeling. De overeenkomst tussen de verschillende beoordelaars was hoog. Dit onderzoek wijst erop dat het voorgestelde classificatieschema de nauwkeurigheid en effectiviteit van het rapporteren van usability problemen kan verbeteren. Daarnaast kan het schema ontwikkelaars van zorg informatie systemen helpen gedetecteerde usability problemen in een ontwerp, de onderliggende oorzaken van deze problemen, en de urgentie van her ontwerp, beter te begrijpen. Dit werk is ook van waarde voor andere onderzoekers die de resultaten van hun usability evaluatie studies beter willen rapporteren.

De toepassing van user-centered design principes is een essentieel onderdeel van de moderne website ontwikkeling en evaluatie. Een bestaand Website Ontwikkel Model voor de Zorgconsument (WDMHC) is een uitgebreid en met succes geëvalueerd kader dat dergelijke ontwerpprincipes bevat. Echter, vanwege de uitgestrektheid van de toepassing in de praktijk is dit model niet direct toepasbaar in specifieke settings. In Hoofdstuk 6 passen we een beperkte subset van het WDMHC kader toe in een case-study over de ontwikkeling en evaluatie van een website gericht op kanker bij kinderen overlevenden (CCS). De doelstelling van deze studie is om te beoordelen of een beperkte subset van het WDMHC-kader voldoende is om een kwalitatief hoogwaardige website met weinig usability problemen, die gericht is op een specifieke patiëntenpopulatie, op te leveren. De website is ontwikkeld met behulp van een zes-staps benadering verdeeld in drie fasen, die allen werden afgeleid van de WDMHC: 1) informatiebehoefte analyse, mock-up maken en focusgroep discussie, 2) website prototype ontwikkeling, 3) Heuristische Evaluatie (HE) en Hard opdenk protocol analyse (Think Aloud). De HE werd uitgevoerd door drie dubbele deskundigen (kennis, zowel in usability engineering en kanker bij kinderen overleving), op basis van bestaande heuristieken ontwikkeld voor websites door een expert genaamd Nielsen. Voor de TA analyse werden acht eindgebruikers uitgenodigd om drie scenario’s te voltooien die de gehele functionaliteit van de website
Samenvatting - Dutch Summary

bevatten. Voor de informatie analyse van de behoeften, werd een enquête uitgedaan met een respons van 145 potentiële eindgebruikers. Dit leidde tot een gestructureerde en geprioriteerde Requirements document (document van systeem eisen). Dit document diende als input voor het ontwikkelen van een mock-up (voorbeeld webpagina) en als input voor de agenda voor de focusgroep discussie, tijdens welke verfijningen voor beide werden voorgesteld. Nadat het prototype website was gebouwd, werden de HE en TA gelijktijdig uitgevoerd. De HE resulteerde in 29 unieke usability problemen, de TA detecteerde 11 unieke problemen. Alle usability problemen met een grote mate van ernst voor systeem ontwikkeling, die ontdekt waren in de HE werden eveneens geopenbaard door TA gebruiker testsessies en vice versa. Vier additionele problemen werden gedetecteerd door de HE, maar deze waren vooral gericht op cosmetisch ontwerpfouten. De TA detecteerde twee extra problemen welke betrekking hadden op de inhoud van website. Concluderend; door de betrokkenheid van zowel eindgebruikers en ‘dubbele’ deskundigen gedurende het hele project, waren we in staat om een prototype website te ontwikkelen die nauw aansluit bij de verwachting van de eindgebruikers en in relatief weinig usability problemen resulteerde tijdens eindgebruiker test. Met de succesvolle toepassing van deze beperkte subset van de WDMHC bieden we ontwikkelaars een kort, duidelijk en gemakkelijk toepasbaar kader voor de ontwikkeling van gezondheidszorg websites met een hoge gebruiksvriendelijkheid, die gericht zijn op specifieke medische populaties.

Interactieve Zorg informatie systemen worden vaak beschouwd als cognitief complex door hun gebruikers, wat leidt tot hoge cognitieve belasting en toegenomen werkdruk. In Hoofdstuk 7 wordt geanalyseerd of het Hard op Denk (TA) protocol in usability testen waardevolle input biedt om effectief een zorginformatiesysteem te herontwerpen. Het systeem geanalyseerd in hoofdstuk 4, de Intensive Zorg Registratie Data Query Tool, is in dit onderzoek herontworpen met input van de TA om de cognitieve werkbelasting van intensivisten tijdens systeem interactie te verminderen. Usability evaluatie pre- en postherontwerp van het systeem toonde een belangrijke vermindering van de cognitieve taak werklast na herontwikkeling van de Query Tool. De gedetecteerde usability problemen werden pre en post geclassificeerd aan de hand van het User Action Framework, UAF, besproken in Hoofdstuk 5. Deze indeling maakte het mogelijk om de usability problemen in te delen in fases van cognitieve gebruikers interactie. Deze studie toont dat de usability problemen die gericht zijn op de planning en translatie van een gebruiker in het ondernemen van een actie voornamelijk de cognitieve taak werklast van een gebruiker beïnvloeden. Door deze problemen in het systeem ontwerp aan te pakken en te herontwerpen is de Query Tool herontwikkeld op zodanige wijze dat het efficiënt, effectief en een hoge mate van bruikbaarheid heeft voor de intensivisten. Dit onderzoek draagt bij aan het belang van het toepassen van TA in zorg informatie systeem herontwikkeling en toont dat deze inspanningen leiden tot een systeem waarbij de cognitieve complexiteit is verlaagd en de bruikbaarheid is verhoogd.
In **Hoofdstuk 8** wordt een vergelijkend onderzoek gedaan tussen een op gebruikers gericht prototype voor een Klinisch beslissing ondersteunend systeem (Clinical Decision Support System (CDSS)) in het bepalen van follow-up screening procedures bij kinderen die kanker hebben overleefd, met de door klinische experts ontwikkeld papieren richtlijn. Het onderzoek richt zich met name op de type usability problemen in beide informatie systemen, en de efficiëntie en effectiviteit van zorgverleners in hun gebruik. Het user-centered (gebruikers gericht) ontwikkels proces is toegepast bij het ontwerpen van een werkend prototype CDSS. Usability problemen in het ophalen en vinden van informatie in de papieren richtlijn werden geanalyseerd met behulp van het Hard op Denk protocol (TA) uitgevoerd in een usability evaluatie met 13 deelnemers, gerelateerde zorgprofessionals. Zowel eenvoudige als meer complexe taken werden in het onderzoek toegepast. Deze analyse leverde input in de user-centered design van het prototype. De bruikbaarheid van het prototype CDSS ontwerp werd vervolgens geëvalueerd door TA analyse waarin dezelfde dertien deelnemers deelnamen. Usability problemen van de papieren richtlijn en het prototype CDSS werden vergeleken door classificatie van de usability problemen aan de hand van het ‘Classificatie Usability Problemen - CUP’ classificatieschema. Efficiëntie en effectiviteit van de deelnemers in het vinden van informatie in de papieren richtlijn en het user-centered prototype werden eveneens vergeleken.

Usability problemen in zowel de papieren richtlijn, evenals het CDSS prototype werden vooral gekenmerkt door 'incongruent met het mentale model'; mismatch met het mentale model van eindgebruikers. Het CDSS ontwerp vermindere dit soort problemen van 17 tot 6 problemen met lage ernst voor het gebruik. De resultaten toonden dat de tijd om eenvoudige informatie zoek taken uit te voeren in het CDSS verhoogde met gemiddeld 0:58 seconde, maar een 58% verbetering in taak volledigheid opleverde ten opzichte van de op papier gebaseerde richtlijn. Taakuitvoering van complexe scenario's daalde met 3:50 minuten in het CDSS, met 17% hogere volledigheid vergeleken met de papieren richtlijn. Concluderend; deze studie toont een user centered design van een CDSS met toepassing van de TA methode. Uit de analyse van de usability problemen van de papieren richtlijn is gebleken dat de inhoudsorganisatiestructuur van het document niet matchte met de informatie zoek strategieën van de zorgverleners. Gebruikers van het CDSS ervaarden tevens mismatch problemen in de user-centered CDSS design. Deze problemen hebben echter geen remmende werking op effectieve en efficiëntie in het uitvoeren van taken in het systeem.

**Concluderend over alle hoofdstukken;** De studies in dit proefschrift bieden kennis over HF/ usability methoden, hun toepassing en hun meerwaarde in User Centered Design (UCD) van zorg informatie systemen (HIS). De belangrijkste lessen worden weergegeven, waarop nieuw onderzoek kan worden gebaseerd en nieuwe inzichten over de toepassing van HF/usability methoden in de ontwikkeling van HIS kunnen worden gegeneereerd.
Acknowledgements

A dear and respected friend sat in my backyard a couple of years ago drinking tea (since I was inhibited from buying a nice Dutch beer that afternoon). The advice he gave me then, about the role of a PhD research in once life, somehow eluded me at that moment. However, after almost 7 years part-time research, 3 houses, a marriage, a divorce, and a new marriage, 2 children and now with a third one growing insight of me, I must admit, he was absolutely right. It is not the goal of the PhD to be reached, it is the life and research experienced during that time of which the thesis steadily grows and develops its inner material. Dear Christian, thank you for your unwavering optimism.

During this period I have been granted the honour to meet and learn from a lot of people, who all in their own unique manner contributed to the works presented in this thesis. Below, I mention only a few for a special reason.

First of all, I wish to thank my professor and mentor Monique Jaspers for offering me the opportunity to learn from her experience and guide me through this PhD. I have never met someone with so much of life’s energy, giving full commitment to her work while being a mother raising two beautiful daughters. She has been an example in all facets of life for me. Monique, I especially thank you for your warmth, support and understanding when life was really tough and the special moments we shared.

I would also like to thank professor Piet Bakker for offering me the opportunity to start this PhD and his support given when needed. From the start on I have encountered numerous obstacles to tackle. Because of the dependence of my research projects on running implementation projects in the AMC, I had to start over twice. I would therefore also like to thank dr. Nicolette de Keizer for asking me to perform the usability analysis of NICE Online. Working on this Project really made me aware of how much pleasure it can be doing practical research aiming to support clinicians in their work by understanding their way of working and implementing this in a system’s design. Of course, this project could have never been as successful as it was without the NICE Chief Engineer Eric van der Zwan. Eric, in the last seven years we have become friends and partners at design. I thank you so much for always listening to my ideas and struggles and just for being there. I do believe that you are the spill at our department; numerous of past PhD projects have depended on your knowledge and work and current PhD projects still and will do. You are great!

Also, I have worked with a lot of young and admirable Medical Informatics’ students and researchers, of whom I especially want to mention: Richard Spithoven, Jens Lauber, Sebastiaan Knijnenburg, Maurice Langemeijer, Reza Khajouei, Ellen Kilsdonk, Bas de Veer, Mariette van Engen-Verheul, Wouter Gude, Rinke Riezebos and Keiko Driest. Working with you all has been a great experience, thank you.
I would also like to thank my parents, Mary and Hans, for always being there. For your loving and cherishing and your support. Mam, dad, I love you both so much.

A very special person in my life is my older sister Selma, who I dedicated this thesis to. Selma, as children we were very different and it must have been sometimes hard for you having this younger sister always walking in your footsteps ;) As a child you were already brilliant, and I much respect you for your intelligence.

In the last seven years we have become as close as sisters could be. I believe this started when your son Sven was born. I never knew how special a child could be, until he came into this world. Not soon after my daughter Zenna was born, a bit unexpected, but welcomed with love. The pregnancy and birth of your daughter Sterre, was the start of a new route in life for us both, I believe. With strength you went on, in your heart believing that everything would be alright and you would be able to take care of your children. Though I will not go into detail what you went through, I need to say that I could not have undergone what you experienced with such dignity, courage and persistence you showed. I love and admire you so much for that.

I also wish to mention my best -friend Brechje Bos. Brechje, you are like a sister to me. Throughout this study we have laughed, cried, joked and done some crazy things. Brechje, you have been my bright star, always shining to make things better.

Also, I wish to thank my friend and Parampilh Nicolen Schermer for we have gone through so much together, and I could always rely on your intellectual insights, inspiration and discussions supporting me during this study.

Janet, Joyce and Eva, you three have come into my life in the recent years, but I am very grateful that you did. Being a Mother and doing a PhD study, you have been there for me with the children and support me when you could throughout this final phase.

There are no words to describe how much I appreciate and love my husband Joost. Joost, we started out as colleagues and friends. You were there at the right moment in time, kind and understanding what my family went through, like no other could. After that, it didn’t take long before we knew we were actually made for each other, and life had brought our roads together.

My love, you are the roots of my life. It has been your confidence, support and dedication to our small family that gave me the final strength to finish this PhD study. My loving thanks for that. Our son, now growing steadily in my womb, is our new small treasure to our family. You make me so happy with all that you do. Thank you for enjoying life and all its extraordinary aspects of it with me.
Curriculum Vitae

Linda Wilma Petronella Peute was born on November 4th, 1981, in Haarlem in the Netherlands. After graduation from the Linnaeus College high school in the year 2000, she started studying Medical Informatics (MI) at the University of Amsterdam (UvA). In 2001 she was elected in the University Central Student Counsel and became the vice-president of the committee of Research and Education. For this, she took part in the University Committee of ICT in Student Education (ICTO), was a member of the Bachelor-Master University Guiding Committee for accreditation and conducted Quality Analyses of UvA Bachelor-Master programs to advise the University Board of Directors. In the study year '02-'03 she was elected as party leader for the MFAS (medical Student Union) Faculty Student Counsel. She fulfilled the role of vice-president during this period. She also became a member of the Bachelor Medical Informatics development Committee. Together with a MI student who was part of the Educational Institute of Medical Informatics (OWI), Milena Albers, she developed and presented a new model for the Bachelor Program Medical Informatics.

In the subsequent years she became advisor to the OWI and supported the Educational Officer in developing a concept document in which Academic skills of Medical Informatics students were defined. In '04 she started with her Graduate Research in an Honour Student Program Medical Informatics. Her project was entitled: Evaluation of a prototype Computerized Laboratory Physician Order Entry system: usability, socio-technical factors and user satisfaction. She presented her graduate research at the World Health (Medical) Informatics conference (MedInfo) 2004 in San Francisco.

During and after her Graduate research she was a Quality Manager for the Clinical Registry Office of the Department of Medical Informatics. She organized and conducted Internal Audits and supported the ISO-9001-2000 Certification on 26th of August 2004 by Lloyd’s Register Quality Assurance.

In September 2005 she started her PhD study as a combined work/research project on the usability and design of the AMC Care-desktop. In September of 2006 she gave birth to her Daughter Zenna after which she re-started and aimed to perform her PhD study part-time on Human Factor Methods in Health Information system design and evaluation. In addition, she provided lectures on Usability and Human Factor Engineering and super-vised MI students’ graduate research. In 2008 she gave birth to her Son Colin. In 2010 she was awarded for second best paper at the World Health (Medical) Informatics conference (MedInfo) 2010 in Cape town. In 2011 she became junior Assistant Professor at the Department of Medical Informatics. In November first of 2011 she married her husband, Joost Dusseljee. Now, seven months pregnant, she will defend her PhD thesis on the 10th of April 2013 in the Agnietenkapel in Amsterdam.
Portfolio

Name: PhD student: Linda Wilma Petronella Peute
Name PhD supervisor: Prof. dr. Monique W.M. Jaspers

<table>
<thead>
<tr>
<th>Courses/ Seminars</th>
<th>Year</th>
<th>ECTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMC World of Science</td>
<td>2005</td>
<td>0.7</td>
</tr>
<tr>
<td>International Master classes Medical informatics (IPHIE)</td>
<td>2005</td>
<td>0.5</td>
</tr>
<tr>
<td>IFMI Workshop Human Factors and Medical Informatics, Lille, France</td>
<td>2006</td>
<td>1</td>
</tr>
<tr>
<td>Ph.D. Student Course Human Factor Engineering, Aarhus, Denmark</td>
<td>2007</td>
<td>1</td>
</tr>
<tr>
<td>Systematic reviews instruction (Cochrane)</td>
<td>2007</td>
<td>0.5</td>
</tr>
<tr>
<td>BKO, Basic course Qualification Assistant Professor</td>
<td>2011</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>International conferences</th>
<th>Year</th>
<th>ECTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical Informatics Europe (MIE) 2005, Genève; full paper presentation; Useability evaluation of a laboratory order entry system: cognitive walkthrough and think aloud combined.</td>
<td>2005</td>
<td>1</td>
</tr>
<tr>
<td>IFMI Workshop Human Factors and Medical Informatics 2006, Lille; presentation; Observational Study of Clinicians’ Information Needs and Information Review Behavior in Daily Assessment of the Patient Care Plan.</td>
<td>2006</td>
<td>1</td>
</tr>
<tr>
<td>2nd Conference on Human Factors Engineering in Health Informatics, Aarhus, full paper; Hospital-Wide Evaluation of the Usability of an Electronic Medical Record System</td>
<td>2007</td>
<td>1</td>
</tr>
<tr>
<td>Medical Informatics Europe (MIE) 2008, Göteborg, full paper; Usability studies on interactive health information systems; where do we stand?</td>
<td>2007</td>
<td>1</td>
</tr>
<tr>
<td>World Congress on Medical and health Informatics (Medinfo) 2010, Cape town; Full paper; Cognitive evaluation of a physician data query tool for a</td>
<td>2010</td>
<td>1</td>
</tr>
</tbody>
</table>
Compared national ICU registry: *Comparing two think aloud variants and their application in redesign.*

5th international symposium on Human Factors Engineering in Health Informatics (Trondheim) full paper presentation; *Effectiveness of Participatory Heuristic Evaluation: Case study with a Guideline Based Health Information System*

Medical Informatics Europe (MIE) 2011, Oslo, Norway; Full paper presentation; *Reducing clinicians' cognitive workload by system redesign; a pre-post think aloud usability study*

<table>
<thead>
<tr>
<th>Teaching experience</th>
<th>Year</th>
<th>ECTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecturer/ Assistant professor (2011 - 2013) Human Factor Evaluation studies in Medical Informatics, Master program Medical Informatics, MAM01 and MAM05</td>
<td>2006-2013</td>
<td>4</td>
</tr>
<tr>
<td>International lectures on Human Factor Evaluation studies in Health Informatics at the Universities of Braunschweig and Heidelberg (International Partnership for Health Informatics Education IPHIE)</td>
<td>2005-2006</td>
<td>2</td>
</tr>
<tr>
<td>Assistant Coordinator Bachelor Medical Informatics, Course: Third year Bachelor internships,</td>
<td>2006</td>
<td>2</td>
</tr>
<tr>
<td>Master MI student Graduate committee member</td>
<td>2008-2013</td>
<td>0.5</td>
</tr>
<tr>
<td>Supervisor Internships</td>
<td>2007</td>
<td>3</td>
</tr>
<tr>
<td>Supervisor Bachelor and Master Thesis students</td>
<td>2007-2013</td>
<td>8</td>
</tr>
<tr>
<td>Assistant Coordinator Course: Databases and Network computing,</td>
<td>2010</td>
<td>2</td>
</tr>
</tbody>
</table>
List of Publications


15. Peute LW, Bakker PJM, Jaspers MWM. Hospital-Wide Evaluation of the Usability of an Electronic Medical Record System in: Bertelsen P, Elkin PL, Madsen I, Nøhr C, editors. HFE2007 Proceedings : 2nd Conference on Human Factors Engineering in Health Informatics : Aarhus University Hospital, Skejby, Denmark, 7-8 June 2007. Aalborg: Virtual Center for Health Informatics, Aalborg University; 2007, p. 30-31


