NTNU – Trondheim
Norwegian University of
Science and Technology

Think Aloud Methods with Eye Tracking in Usability Testing

A comparison study with different task types

Terje Røsand

Master of Science in Informatics

Submission date:	December 2012
Supervisor:	Dag Svanæs, IDI
Co-supervisor:	Øystein Nytrø, IDI

Norwegian University of Science and Technology
Department of Computer and Information Science

Preface

This endeavor is the result of my Master of Science study in informatics at the Department of Computer and Information Science at the Norwegian University of Science and Technology - NTNU. The project was done in conjunction with the Evicare project.

I would like to thank my main supervisor Dag Svanæs for his guidance throughout the project, and my co-supervisor Øystein Nytrø for providing an interesting case.

I would also like to thank Zheng Wang, whom the experiment was designed in cooperation with. She wanted to find out how reading clinical guidelines affect the decisions made by clinicians. She collected her data by asking the participants to write down their “clinical questions” after task 2 and task 3, and by a questionnaire regarding the clinical guideline.

Trondheim, December 20th 2012,

Terje Røsand

Abstract

The concurrent think aloud (CTA) method is perhaps the single most valuable usability engineering method. The method has certain issues that may be avoided by using another method: Retrospective think aloud (RTA). RTA can be conducted un-cued, or cued by different stimuli. Several studies have been done to compare the different methods, but none of these have examined the methods under different circumstances. In this study three different methods, CTA, RTA and RTAg (RTA cued by video and gaze data) are compared when applied on three different task types: Easy interaction task, reading task and complex problem solving task. The comparison was done in terms of four measurements: Task performance, usability problems identified, amount of verbalization and the participants' subjective assessments. It was found that with the two retrospective methods there were elicited more usability problems by means of verbalization as compared to CTA. It was also found that during reading tasks and complex problem solving tasks the participants found it easier to verbalize using RTAg than using RTA. Some observations were done that invite to further research.

Sammendrag

Det å tenke høyt under brukbarhetstesting (concurrent think aloud – CTA) er sannsynligvis den mest verdifulle enkeltstående metoden som brukes til å evaluere brukbarhet. Metoden har noen svakheter som kan unngås ved å bruke en variant av metoden: Retrospective think aloud (RTA). Ved bruk av RTA løser brukeren først oppgaven, og verbaliserer deretter I retrospekt. RTA kan brukes uten noen hjelpemidler, eller det kan brukes forskjellige stimuli for å hjelpe brukeren med å verbalisere. Det mest vanlige er å bruke en video der brukeren selv løser oppgaven. Det har blitt gjort flere studier der ulike tenk høyt-metoder har blitt sammenlignet, men ingen studier der disse metodene sammenlignes under foskjellige omstendigheter. Denne studien sammenlignes CTA, RTA og RTAg (RTAg = RTA som bruker video og øyebevegelser som stimuli), når de brukes på oppgaver som har forkjellige karakteristika: en enkel interaksjonsoppgave, en leseoppgave og en kompleks broblemløsingsoppgave. Det ble målt gjennomføringstid og gjennomføringsrate for hver oppgave, antall identifiserte brukbarhetsproblemer, antall ord verbalisert, og deltakernes subjektive bedømmelse av hvordan de opplevde å bruke de ulike metodene. Det ble slått fast at de to retrospektive metodene førte til identifikasjon av flere brukbarhetsproblemer gjennom verbalisering I forhold til CTA. Det ble også konkludert med at testdeltakerne foretrekker RTAg når de løser en leseoppgave eller kompleks broblemløsingsoppgave. Det ble gjort endel observasjoner som inviterer til videre forskning.

Contents

Figures

1 Introduction

1.1 Background

Think aloud (TA) has been used in usability testing for about two decades, and is perhaps the most widely used usability evaluation method. The method was first utilized within the usability field by Clayton Lewis [1], and is based on the work on Protocol Analysis by Ericsson and Simon [2]. The concept is very simple: The users are asked to think out loud while they are executing a task, with the purpose of gaining insight in the participants' thoughts on what they are doing.

Several questions have been raised about the validity of the method. Russo et al. [3] are concerned with the reactivity of the method, while Boren and Ramey [4] point out that most literature on think aloud within usability does not conform to the model created by Ericsson and Simon, and that there is substantial variation in how think aloud is carried out by practitioners. Still it is widely used in research as well as in practice. Nielsen wrote in his book in 1993 [5]: "Thinking aloud may be the single most valuable usability engineering method", and still stand by this assessment in 2012 [6].

Think aloud is reactive if the act of verbalizing is affecting the performance of the primary task, and several studies show that this is the case[3], [7], [8]. In cases where it is desirable to measure task performance, a variation of the method may be utilized: retrospective think aloud (RTA). With this method the user first executes the task, and then verbalizes in retrospect. RTA is usually conducted by playing back a video of the participants executing the tasks. This may help the participants remembering what they did during task execution, and can contribute to reduce fabrication, which are the most prominent problems with RTA [3]. To emphasize that video playback is used, researchers often use terms such as stimulated RTA or cued RTA[9], [10].

In recent years eye trackers have gained popularity as a useful tool in usability testing. Non-intrusive eye trackers are accurate, available from several vendors, and while still rather expensive, a lot cheaper than they used to be. By using an eye tracker in a usability test, the moderator and the observers can see what the test participants are looking at during the test, and the software will record the gaze data into the produced video from the tests. The eye tracking software offers a few different ways of aggregating and visualizing the gaze data, and a substantial amount of different metrics that may be useful.

When using eye trackers in usability tests it is believed that the reactivity problems with concurrent think aloud (CTA) are more prominent. Not only does it affect the task performance in terms of the time the participants use for each task, but it will result in invalid gaze data due to random eye movements while performing think aloud [11], [12]. This may indicate that RTA is a better method when using eye trackers if you want to measure task working time and/or need reliable gaze data. With eye trackers the concept of cued RTA can be taken a step further by presenting the participants' gaze data during RTA. This may amplify the effects of helping the participants to remember and reducing fabrication. Usability practitioners at Bunnyfoot[13]invented a methodology they called PEEP (Post-Experience Eye-tracked Protocols), where they use RTA cued with video and gaze data. The PEEP method is presented by Ball et al. [14], and further examined by Eger et al. [15], where they found that retrospective reports may be less reactive and more informative than other verbalization techniques. Another study done by Olsen et al. [12] support these findings to some extent. In this study they compare four different versions of cued retrospective think aloud: un-cued RTA, video cued RTA, gaze plot cued RTA and gaze video cued RTA. Their findings show that using any cue is significantly better than using no cue at all. Further they suggest that the gaze plot or gaze video cue stimulates participants to identify more usability problems.

Later Bunnyfoot ceased to use the gaze data as cue to RTA [16] due to the test participants' tendency to focus on the eye tracking technology itself, thus turning attention away from what they were supposed to do: verbalizing. This is supported by a more recent study by Elling et al. [10]. They did a comparison of RTA with and without gaze data as cue, and in contrast to Eger et al. found no differences in terms of the number of usability problems, types of problems and the way the problems were detected. Neither did gaze replay result in more verbalization. They point at the confronting and distracting nature of showing the participants their own gaze data as one possible explanation of the result.

1.2 Motivation and research questions

Despite the contradictory results of the above mentioned studies, they have all contributed with knowledge about different aspects of using the different TA methods when conducting usability tests with eye tracking. However there is little knowledge about how the various TA methods may be utilized under different circumstances.

The main motivation for this study is observations done in earlier projects at the Norwegian EHR Research Centre (NSEP)[17]. In a study using eye tracking and retrospective think aloud with gaze playback, sometimes the gaze data were missing. This may be due to technical problems with the equipment, or that the participant was out of reach for the eye tracker, typically too close to the screen. In these cases the participants commented that they could not verbalize because they did not know what they were doing. This occurred while reading a rather big chunk of text. This led to the hypothesis that there are situations where one TA method is suitable and other situations where another TA method is suitable.

This study aims at providing advice on which TA method is better suited when applied on tasks with different characteristics. Three different TA methods are examined:

- **Concurrent think aloud (CTA):** participants verbalize their thoughts while completing the tasks.
- **Retrospective think aloud, video cued (RTA):** participants verbalize their thoughts after completing the tasks, cued by a video playback of their own task completion.
- **Retrospective think aloud, video and gaze cued (RTAg):** participants verbalize their thoughts after completing the tasks, cued by a video playback of their own task completion, where the gaze data is superimposed into the video.

In order to investigate the different task characteristics, three task types are designed:

1. **Easy interactive task:** This task type is not very cognitively demanding and it concerns interacting with a user interface.
2. **Reading task:** This task type is designed to force the participants into reading a lot of text, and hence contain little interactivity. It may also be argued that this task type is more cognitively demanding than the first task type.
3. **Complex problem solving task:** This task type is designed to be cognitively demanding and complex in the sense that the participants are solving two tasks simultaneously: 1. Solving a problem, 2: Interacting with the tool used to solve the task.

The different TA methods will be evaluated for their ability to:

1. **Produce valid results (reactivity):**
   This is assessed by measuring task completion time and task completion rate. The purpose of this assessment is to find out if verbalizing during task completion (CTA) affects the time used to complete the task and the ability to complete the task. This may be useful when designing a study where task performance is measured.
2. **Elicit usability problems:**
   This is assessed by counting usability problems, and identify whether the problems are elicited by observation, verbalization or a combination. This may be useful when choosing the think aloud method to use for a study, depending on the characteristics of the study tasks.
3. **Gain insight in the participants' thoughts:**
   This is assessed by counting words verbalized during the different think aloud sessions. The purpose of this assessment is to find out if seeing their own eye movements during retrospective think aloud may motivate the participants provide more verbal data. Another purpose is to find out whether task complexity affects the participants' ability to perform CTA. This may be useful when choosing the think aloud method for a study, depending on the characteristics of the study tasks.

The overall research question is:

Is there a significant difference in the appropriateness of the different think aloud methods applied on the different task types?

This can be decomposed into these research questions:

1. Validity: Is there a difference in task performance when using the different methods applied on the different task types in terms of task completion time and task completion rate?
2. Usability: Is there a difference in how the different methods applied on the different task types are able to help eliciting usability problems in terms of number of problems and the way the problems are elicited (observation, verbalization or a combination)?
3. Insight:
    a. Is there a difference between the two retrospective methods when applied on the reading task in terms of the amount of verbalization and the participants' subjective assessment?
    b. Is there a difference between the complex problem solving task and the easy interactive task in terms of amount of verbalization during CTA?

2 Usability and eye tracking

2.1 Usability

In popular speech the term user-friendliness is often used about websites, computer programs, or things like tools, gaming consoles, ticket machines etc. And it is often referred to as if user-friendliness is a property of the product. Within computer science, and other disciplines like product design and psychology, the similar term usability is used. Usability has some assumptions about who is using the product, in what context it is used, and which goals the user has. Usability is defined in the ISO 9241-11 standard[18] as:

Usability: Extent to which a product can be used by specified users to achieve specified goals with *effectiveness*, *efficiency* and *satisfaction* in a specified context of use.

Effectiveness: Accuracy and completeness with which users achieve specified goals.

Efficiency: Resources expended in relation to the accuracy and completeness with which users achieve goals.

Satisfaction: Freedom from discomfort, and positive attitudes towards the use of the product.

This can be illustrated with an example. Let say that a doctor is expecting a patient, and it is a patient that is previously unknown for the doctor. The doctor wants to spend a few minutes to get familiar with the patient case. The doctor then goes to her computer and finds the actual patient in the electronic health record (EHR), and reads the patients record.

In this example the specified user is the doctor, the context of use is at the doctor's office expecting a patient in a few minutes, and the doctor's goal is to find information about the patient. Effectiveness relates to whether the doctor is able to use the electronic health record to find information about the patient. Efficiency relates to how fast she can find the information. Satisfaction relates to whether she finds it pleasant or unpleasant to use the EHR system for finding the information she wants.

The EHR system is probably a useful tool for this user with this goal and in this context. For an archaeologist situated in the wilderness trying to find the best price on helicopter transport with a mobile phone, the EHR system would probably be totally useless. This is why these assumptions are important, and that usability is not a property of the product per se.

2.2 Usability and user-centered design

The user centered design model (UCD) is an iterative life cycle model for interactive systems. It is described in the standard ISO 13407: Human Centered Design Process for Interactive Systems[25]. The UCD model describes key principles for user-centered design, and design activities that take place throughout the life cycle. The most important principles for UCD include:

- **Active user involvement**: Users can be directly involved as part of the design team (participatory design), or indirectly through observations and interviews, and other activities such as design workshops, focus groups and usability tests.
- **Balanced allocation of resources between system and user**: Ensure that the project resources are balanced between programming and user research.
- **Iterative design**: The design process takes place in several rounds, with an increasingly refinement of the product until the requirements are met.
- **Multi-disciplinary design teams**: Build teams with complementary skills.

The UCD model includes an entry condition, four iterative steps and an exit condition:

I. **Identify need for UCD**: This is the entry condition where it is taken a decision whether to use a UCD strategy for the development project.

II. **Specify context of use**: Identifying the users of the product, what they will use it for and under what conditions they will use it. Typical design activities for this step are field observations and user interviews.

III. **Specify requirements**: Identify user and organizational goals. Typical design activities for this step include design workshops and focus groups.

IV. **Produce design solutions**: This step involves building the product, starting with rough sketches, via prototypes of different fidelity, until the finished product.

V. **Evaluate design**: In this step the product is evaluated against the specified requirements. Usability testing is a typical design activity for this step.

VI. **System satisfies specified requirements**: This is the exit condition. When the specified requirements are met, the product is finished.

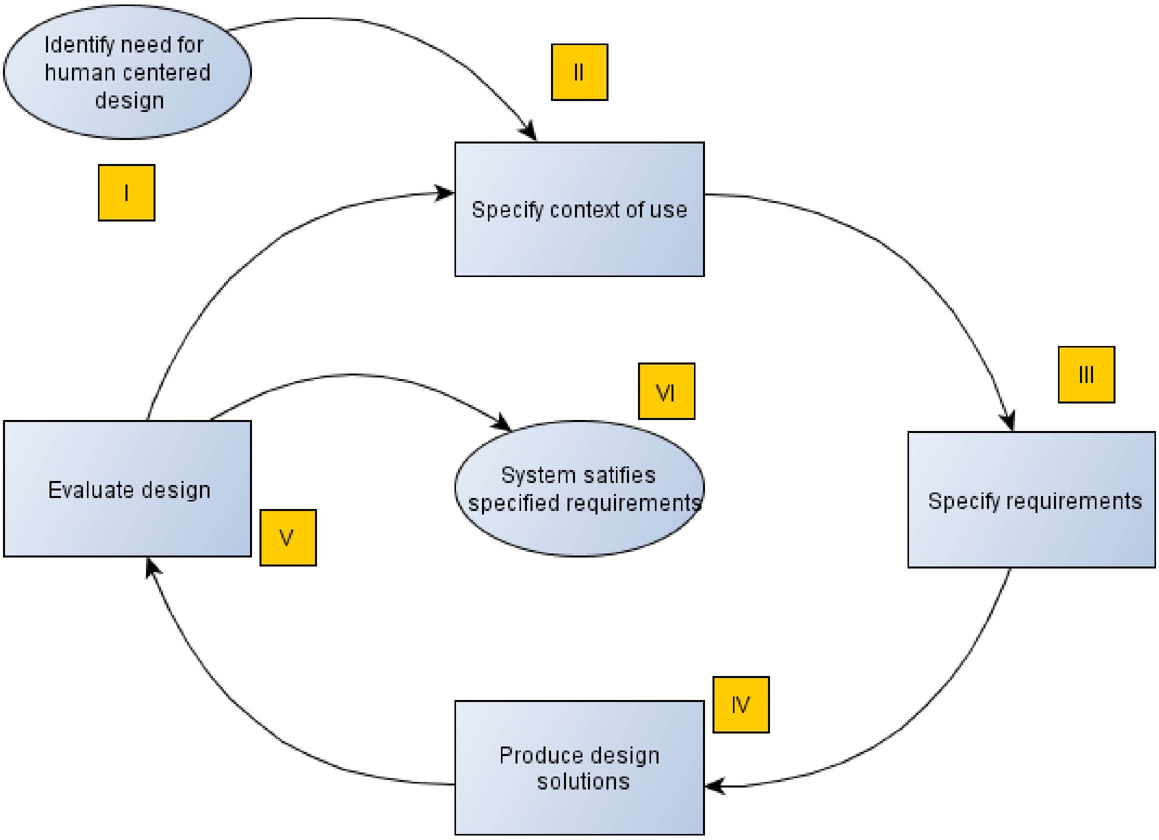


Figure 1 The user-centered design process

Usability and UCD are closely interrelated. The UCD model ensures that the user, and hence usability, is in focus throughout the development process, and usability testing is one important part of the UCD process.

2.3 Usability evaluation

In order to evaluate usability a number of methods may be used, but the most fundamental method is usability testing[19], often referred to as user testing. In a usability test representative users are solving representative tasks using some product. The purpose of the test is to find out how well the product is designed for interaction. The weaknesses of the products ability to offer easy interaction are then reported back to the developers of the product, often along with suggestions for improvement. The test is repeated with several users, typically around 8, to provide a pattern of how the product is used. In research larger samples of test users are often used to get more reliable results. During the tests the users are asked to think aloud while solving the tasks, with the purpose of getting insight in the user's thoughts during task completion. The tests are guided by a facilitator who is sitting next to the test participants during task completion, and sometimes there are additional observers present.

The most important result of a usability test is a set of usability problems. A usability problem occurs when the user has a problem with solving the task. The usability problems are often rated by severity and frequency. If a problem has low severity and low frequency may not be fixed at all. If a problem has high severity and/or high frequency it is of upmost importance to fix the problem. Another outcome of a usability test that may be useful is task performance. Task performance can be decomposed into task completion time, which reflects the efficiency part of the definition above, and task completion rate which is a measure of how many tasks a user completed in reasonable time during a usability test, and reflects the effectiveness part of the definition. In order to elicit additional information a questionnaire like SUS[20] may be used. At the end of the test it is common to conduct an interview with the participant to elaborate on certain issues and collecting loose ends.

Usability testing can roughly be divided into formative, summative and comparative testing. A formative usability test is conducted with the intention of informing further development of a product, as a part of an iterative development strategy. Summative usability testing sums up the usability of a finished product, while comparative usability testing is used for comparing the usability of two or more products.

Common Industry Format (CIF)[21] is a format for reporting findings from usability tests. It is a format agreed upon by major software suppliers and purchasers coordinated by the National Institute of Standards and Technology (NIST)[22]. The format is consistent with the definition of usability, and provides a standardized interface between usability professionals and stakeholders in an organization such as purchasing managers. CIF is mostly used for reporting results from summative or comparative usability tests.

2.4 The usability laboratory at NSEP

The usability tests are usually recorded, and often situated in a usability lab. A usability lab typically consists of two rooms, whereas the test room is where the test participant is solving the tasks, and the observation room is where usability practitioners or researchers are observing. A usability lab provides a controlled environment, equipped to produce high quality data material. Usability studies may also be done in the field, but as opposed to a lab, the field is not a controlled environment, and it may be difficult to produce video and sound recordings of good quality. In addition there may be legal issues such as getting permission to do video recordings in a hospital, which is omitted by using a lab. On the other hand the lab offers less realism.

The usability tests in this study were conducted at the usability lab at The Norwegian EHR Research Centre (NSEP)[23]. This lab is a joint venture between NSEP and the Department of Computer and Information Science at NTNU, and is especially equipped for doing medical or health oriented projects. This is a rather large lab, with 3 rooms in the test area. All of them equipped with video cameras and microphones which can be controlled from the combined control/observation room. The lab is designed in this manner in order to provide an environment for testing scenarios involving several test participants using several devices simultaneously. The lab is also well suited for arranging design workshops and focus groups.

In Figure 2 shows the test area from the NSEP Usability Lab. A camera and a microphone can be spotted in the ceiling. The lab is rigged for a usability test using the eye tracker, where the test participant is working on the computer on the left side, the facilitator sits to the right and can follow the participant's eye movement on the facilitator's monitor. The eye tracker can be spotted underneath the monitor on the test computer. There is a cable that leads into the observation room, where observers can see the test, including the participant's eye movements on a 52" screen.

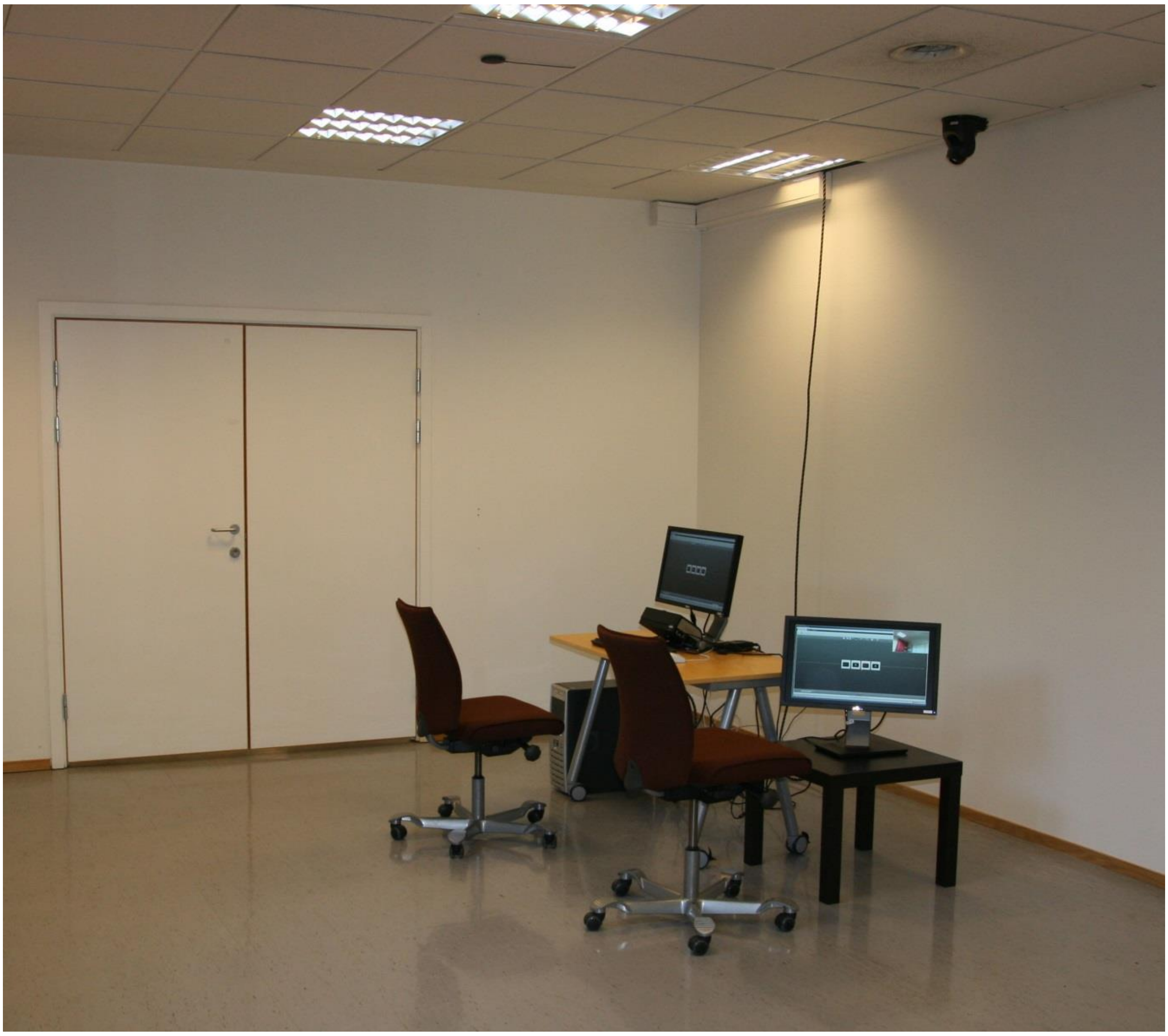


Figure 2 From the test area at the NSEP Usability Laboratory

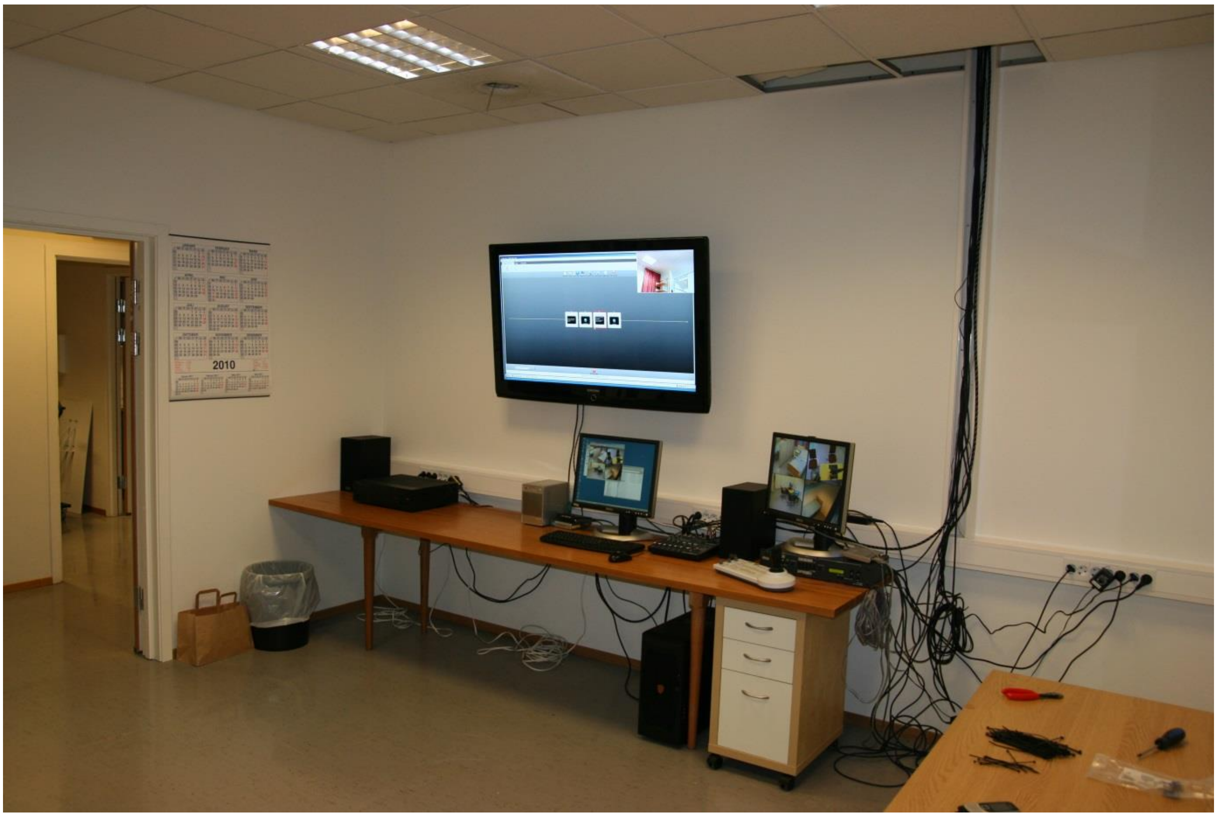


Figure 3 From the observation room of the NSEP Usability Laboratory

2.5 Eye tracking

An eye tracker is a device that measure where on the screen a test participant is looking. Since the study of eye movements started in the second half of the nineteenth century, a lot of different equipment and techniques have been used to track eye movements. This includes electrodes mounted around the eye, various kinds of contact lenses, and analog film. Most eye tracking systems today use video. The point-of-gaze is measured by the "corneal-reflection/pupil-center" method. When the eye is illuminated, usually by infrared or near infrared light because it is more convenient for the user than visible light, the pupil becomes bright, and there is also a small bright spot that is called the corneal reflection, or glint. When the pupil and the corneal reflection are located by the camera it is possible to calculate the point-of-gaze. Since there are two reference points measured, this technique allows for a certain degree of head movement.

Saccades are the rapid eye movements used when scanning a web site or reading text. The eye does not slide smoothly over the text but rather makes short stops with very rapid movements between stops. The stops are called fixations and typically have duration from 200-250 milliseconds or more, while the saccades have duration of 20-40 milliseconds. This is essentially what the eye tracker is measuring.

Modern eye trackers come in a variety of shapes and models. The most common models used for usability testing are either integrated with a computer monitor, or a stand-alone device that is positioned between the keyboard and the monitor on a standard desktop computer. The stand-alone models can be used with mobile devices when used with special brackets. There are also models that are built into glasses, and smaller models that can be used with laptops.

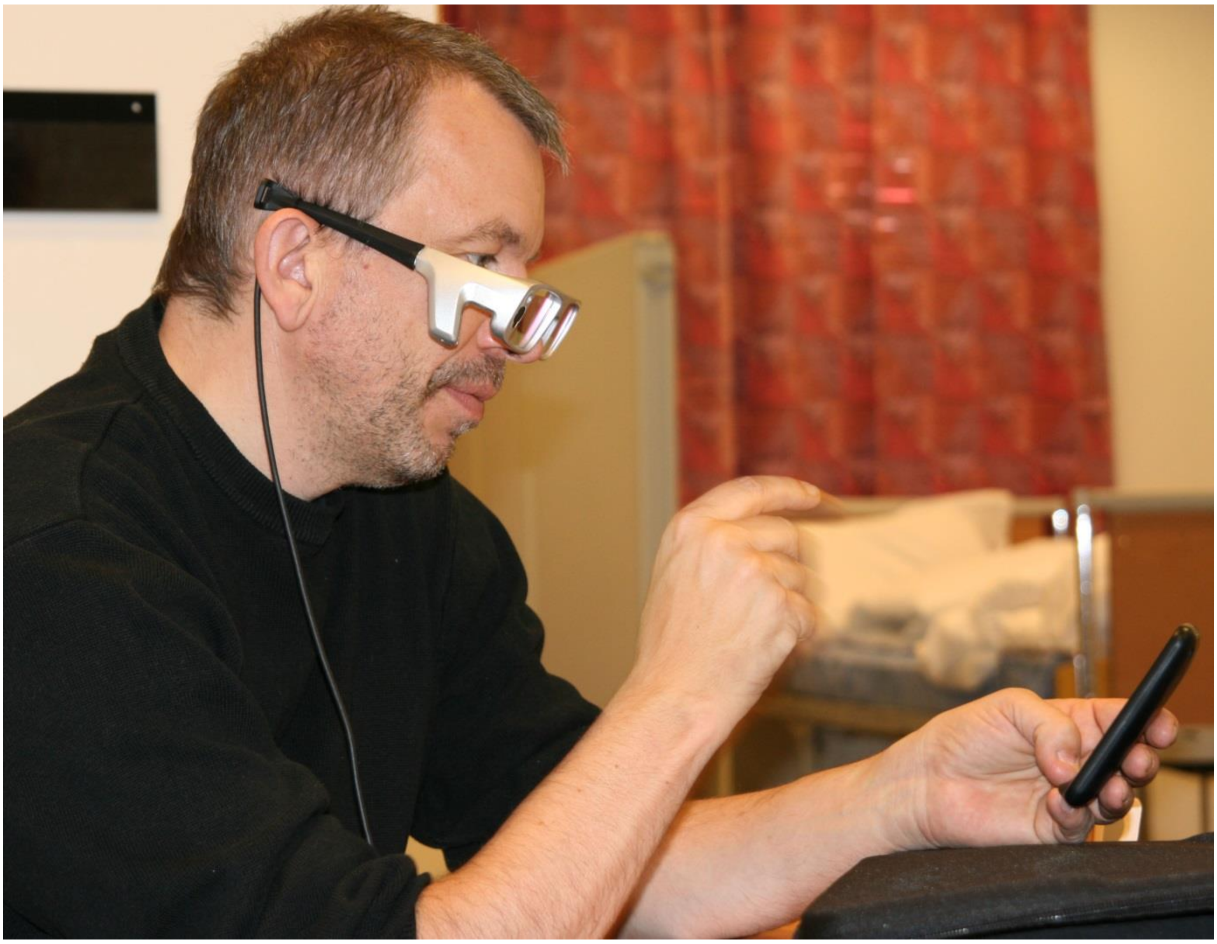


Figure 4 Tobii Glasses - eye tracker built into glasses

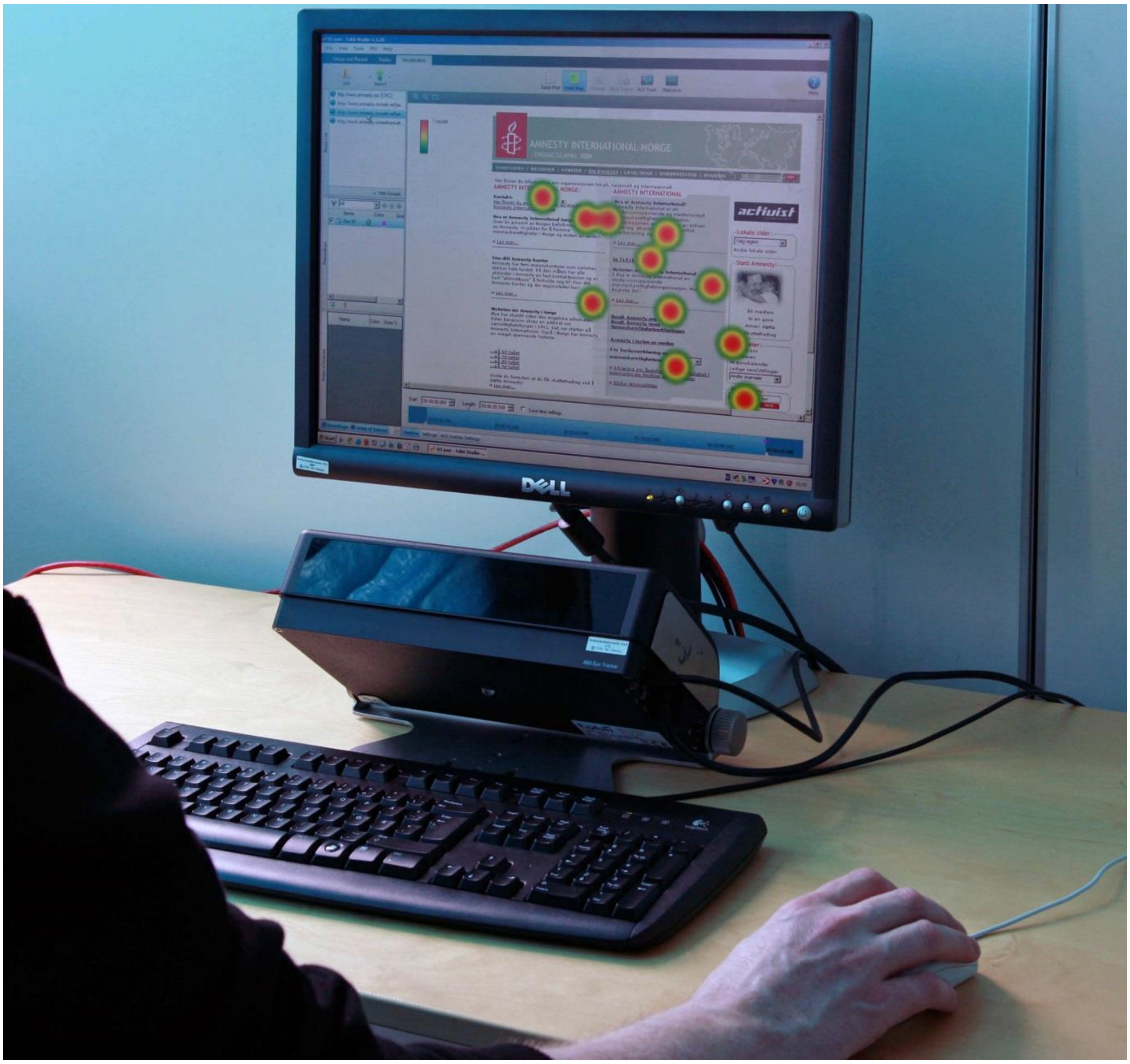
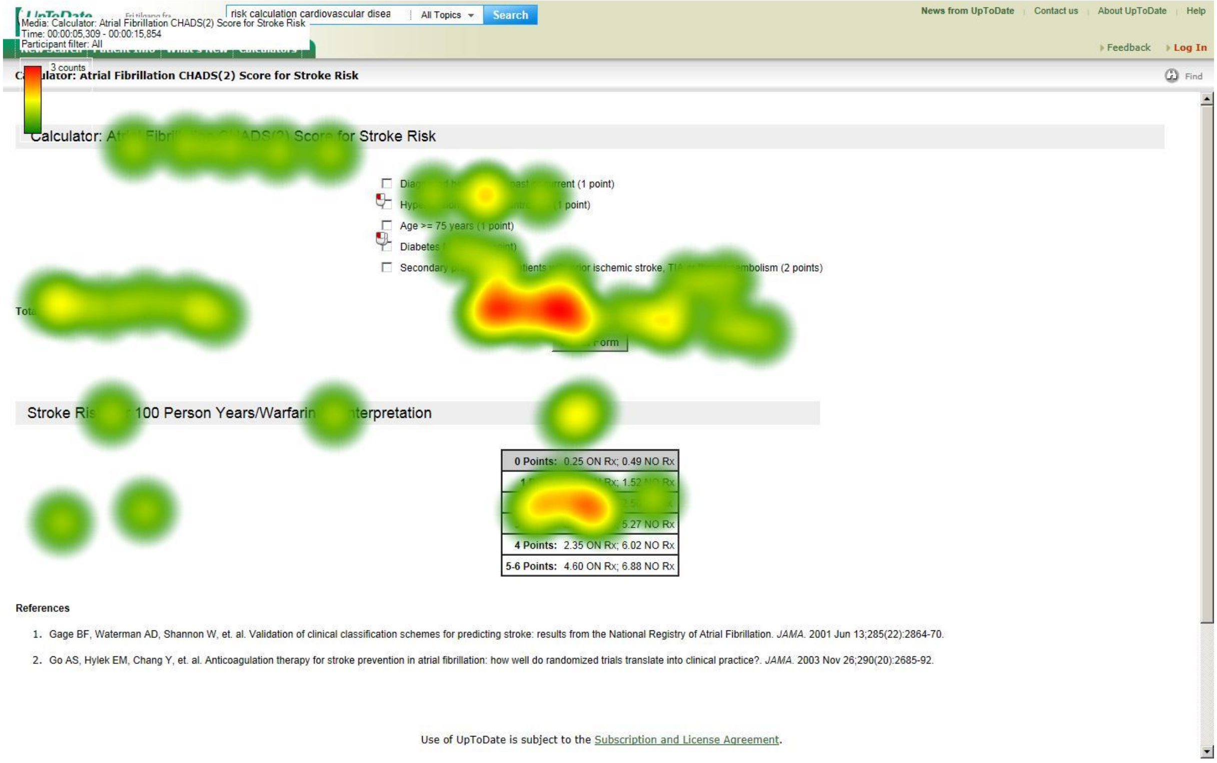


Figure 5 Stand-alone eye tracker - Tobii X60

At the NSEP usability lab there is a Tobii X60 stand-alone eye tracker as shown in Figure 5. It came bundled with Tobii Studio software that is used for conducting and analyzing usability tests. Tobii Studio offers a broad range of functionality:

- Managing projects, tests and users.
- Conducting usability tests, using a variety of stimuli.
- Replay and export of video.
- Visualization of gaze data using a number of techniques, the most common are heat maps and gaze plots.
- Remote viewing for observers.
- Retrospective recording.
- Statistics.


Media: Calculator: Atrial Fibrillation CHADS(2) Score for Stroke Risk
Time: 00:00:05,309 - 00:00:15,854
Participant filter: All
risk calculation cardiovascular disea
All Topics
Search
News from UpToDate | Contact us | About UpToDate | Help
Feedback
Log In
3 counts
Calculator: Atrial Fibrillation CHADS(2) Score for Stroke Risk
Find
Form

Calculator: At Fibri DS(2) Score for Stroke Risk

□ Diag rent (1 point)
Hyp 1 point)
□ Age >= 75 years (1 point)
Diabetes int)
□ Secondary tients ior ischemic stroke, TIA embolism (2 points)

Tota

Stroke Ris 100 Person Years/Warfarin nterpretation

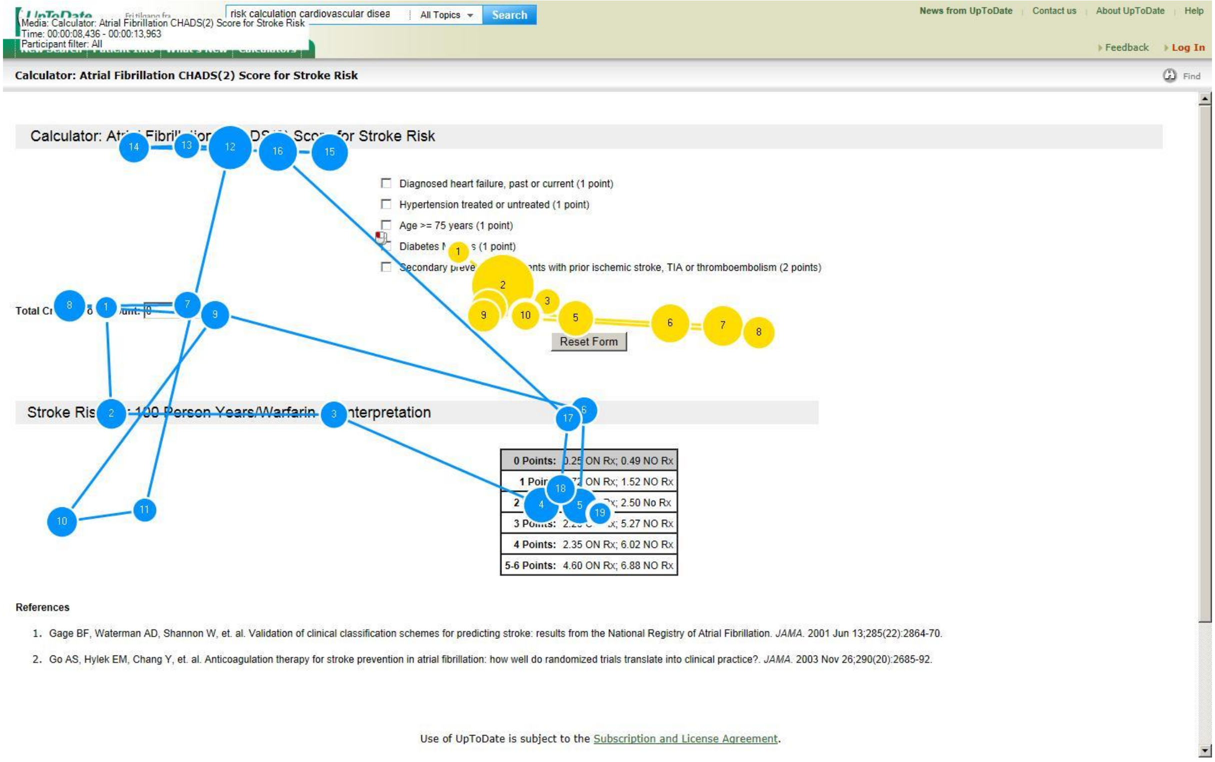
0 Points:	0.25 ON Rx; 0.49 NO Rx
1	Rx; 1.52 NO Rx
	5.27 NO Rx
4 Points:	2.35 ON Rx; 6.02 NO Rx
5-6 Points:	4.60 ON Rx; 6.88 NO Rx

References

1. Gage BF, Waterman AD, Shannon W, et. al. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. *JAMA.* 2001 Jun 13;285(22):2864-70.
2. Go AS, Hylek EM, Chang Y, et. al. Anticoagulation therapy for stroke prevention in atrial fibrillation: how well do randomized trials translate into clinical practice?. *JAMA.* 2003 Nov 26;290(20):2685-92.

Use of UpToDate is subject to the Subscription and License Agreement.

Figure 6 Heat map with data from three users


Media: Calculator: Atrial Fibrillation CHADS(2) Score for Stroke Risk
Time: 00:00:08,436 - 00:00:13,963
Participant filter: All
risk calculation cardiovascular disea
All Topics
Search
News from UpToDate | Contact us | About UpToDate | Help
Feedback
Log In
Calculator: Atrial Fibrillation CHADS(2) Score for Stroke Risk
Find
Reset Form
14
13
12
16
15
1
2
3
9
10
5
6
7
8
8
1
7
9
2
3
17
6
18
4
5
19
10
11

Calculator: At Fibri ior DS Sco for Stroke Risk

□ Diagnosed heart failure, past or current (1 point)
□ Hypertension treated or untreated (1 point)
□ Age >= 75 years (1 point)
Diabetes s (1 point)
□ Secondary preve nts with prior ischemic stroke, TIA or thromboembolism (2 points)

Total Cr unt: 0

Stroke Ris 100 Person Years/Warfarin nterpretation

0 Points:	0.25 ON Rx; 0.49 NO Rx
1 Poir	72 ON Rx; 1.52 NO Rx
2	Rx; 2.50 No Rx
3 Points:	2. Rx; 5.27 NO Rx
4 Points:	2.35 ON Rx; 6.02 NO Rx
5-6 Points:	4.60 ON Rx; 6.88 NO Rx

References

1. Gage BF, Waterman AD, Shannon W, et. al. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. *JAMA.* 2001 Jun 13;285(22):2864-70.
2. Go AS, Hylek EM, Chang Y, et. al. Anticoagulation therapy for stroke prevention in atrial fibrillation: how well do randomized trials translate into clinical practice?. *JAMA.* 2003 Nov 26;290(20):2685-92.

Use of UpToDate is subject to the Subscription and License Agreement.

Figure 7 Gaze plot with data from two users

The heat map in Figure 6 is generated by aggregation of the data from 3 users. The red areas contain most fixations, then yellow, and at last green. The heat map can when used properly give a good indication if there are important elements in the user interface that the users have overlooked, and opposite if there are less important elements that draw too much attention. This is a powerful tool for communicating usability problems to the customers or developers.

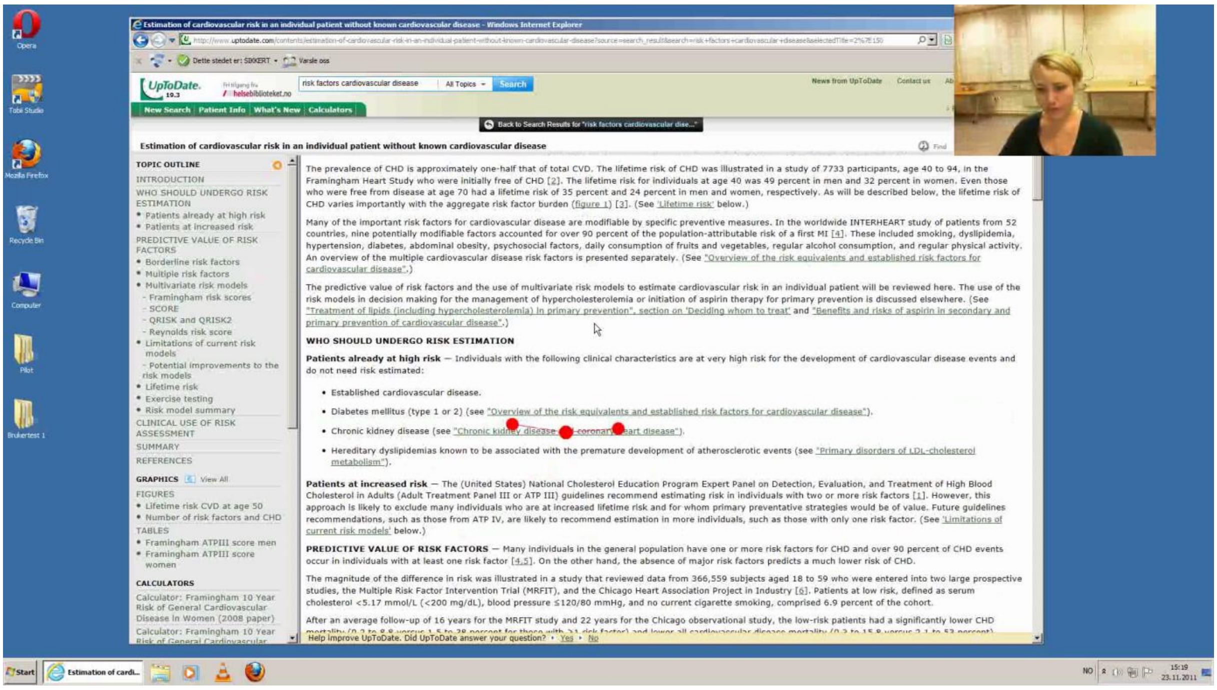
The gaze plot in Figure 7 is generated by aggregation of the data from two users. In this plot the users' gaze are drawn in different colors. Each fixation is drawn as a circle with the saccades as lines between the circles. The fixations are numbered, and the size of the circle indicates the duration of the fixations.

Tobii Studio has built-in functionality for recording retrospective sessions. After the task is completed there is a button that starts recording the retrospective session. The video produced during task execution is the cue for verbalization. This video can be paused, rewinded and forwarded as pleased. This allows for elaborating on certain issues, and the retrospective session can be conducted more like an interview.

2.6 Motivation for research question

The NSEP usability lab has been involved in a lot of studies regarding usability related to health services. In one study involving the Evicare project[24] there were conducted six usability tests where the test object was a web based clinical guideline from UpToDate. That study was a prequel to the current study. The test subjects were medical students, and the objective was to find out how the participants were able to find the information they needed in the guideline. The second objective was to inform the design of the current study.

All participants were verbalizing in retrospect using the RTAg method, and at several occasions the gaze data was lost. This resulted in several comments on how they were not able to verbalize because they did not know what was going on in the video. The tasks implied a substantial amount of test reading, and these observations led to the motivation of the current study.



Estimation of cardiovascular risk in an individual patient without known cardiovascular disease - Windows Internet Explorer

http://www.uptodate.com/contents/estimation-of-cardiovascular-risk-in-an-individual-patient-without-known-cardiovascular-disease?source=search_result&search=risk+factors+cardiovascular+disease&selectedTitle=2%7E150

Estimation of cardiovascular risk in an individual patient without known cardiovascular disease

TOPIC OUTLINE

INTRODUCTION

WHO SHOULD UNDERGO RISK ESTIMATION
- Patients already at high risk
- Patients at increased risk

PREDICTIVE VALUE OF RISK FACTORS
- Borderline risk factors
- Multiple risk factors
- Multivariate risk models
  - Framingham risk scores
  - SCORE
  - QRISK and QRISK2
  - Reynolds risk score
- Limitations of current risk models
  - Potential improvements to the risk models
- Lifetime risk
- Exercise testing
- Risk model summary

CLINICAL USE OF RISK ASSESSMENT

SUMMARY

REFERENCES

GRAPHICS View All

FIGURES
- Lifetime risk CVD at age 50
- Number of risk factors and CHD

TABLES
- Framingham ATPIII score men
- Framingham ATPIII score women

CALCULATORS

Calculator: Framingham 10 Year Risk of General Cardiovascular Disease in Women (2008 paper)

Calculator: Framingham 10 Year Risk of General Cardiovascular

The prevalence of CHD is approximately one-half that of total CVD. The lifetime risk of CHD was illustrated in a study of 7733 participants, age 40 to 94, in the Framingham Heart Study who were initially free of CHD [2]. The lifetime risk for individuals at age 40 was 49 percent in men and 32 percent in women. Even those who were free from disease at age 70 had a lifetime risk of 35 percent and 24 percent in men and women, respectively. As will be described below, the lifetime risk of CHD varies importantly with the aggregate risk factor burden (figure 1) [3]. (See 'Lifetime risk' below.)

Many of the important risk factors for cardiovascular disease are modifiable by specific preventive measures. In the worldwide INTERHEART study of patients from 52 countries, nine potentially modifiable factors accounted for over 90 percent of the population-attributable risk of a first MI [4]. These included smoking, dyslipidemia, hypertension, diabetes, abdominal obesity, psychosocial factors, daily consumption of fruits and vegetables, regular alcohol consumption, and regular physical activity. An overview of the multiple cardiovascular disease risk factors is presented separately. (See "Overview of the risk equivalents and established risk factors for cardiovascular disease".)

The predictive value of risk factors and the use of multivariate risk models to estimate cardiovascular risk in an individual patient will be reviewed here. The use of the risk models in decision making for the management of hypercholesterolemia or initiation of aspirin therapy for primary prevention is discussed elsewhere. (See "Treatment of lipids (including hypercholesterolemia) in primary prevention", section on 'Deciding whom to treat' and "Benefits and risks of aspirin in secondary and primary prevention of cardiovascular disease".)

WHO SHOULD UNDERGO RISK ESTIMATION

Patients already at high risk — Individuals with the following clinical characteristics are at very high risk for the development of cardiovascular disease events and do not need risk estimated:

- Established cardiovascular disease.
- Diabetes mellitus (type 1 or 2) (see "Overview of the risk equivalents and established risk factors for cardiovascular disease").
- Chronic kidney disease (see "Chronic kidney disease and coronary heart disease").
- Hereditary dyslipidemias known to be associated with the premature development of atherosclerotic events (see "Primary disorders of LDL-cholesterol metabolism").

Patients at increased risk — The (United States) National Cholesterol Education Program Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III or ATP III) guidelines recommend estimating risk in individuals with two or more risk factors [1]. However, this approach is likely to exclude many individuals who are at increased lifetime risk and for whom primary preventative strategies would be of value. Future guidelines recommendations, such as those from ATP IV, are likely to recommend estimation in more individuals, such as those with only one risk factor. (See 'Limitations of current risk models' below.)

PREDICTIVE VALUE OF RISK FACTORS — Many individuals in the general population have one or more risk factors for CHD and over 90 percent of CHD events occur in individuals with at least one risk factor [4,5]. On the other hand, the absence of major risk factors predicts a much lower risk of CHD.

The magnitude of the difference in risk was illustrated in a study that reviewed data from 366,559 subjects aged 18 to 59 who were entered into two large prospective studies, the Multiple Risk Factor Intervention Trial (MRFIT), and the Chicago Heart Association Project in Industry [6]. Patients at low risk, defined as serum cholesterol <5.17 mmol/L (<200 mg/dL), blood pressure ≤120/80 mmHg, and no current cigarette smoking, comprised 6.9 percent of the cohort.

After an average follow-up of 16 years for the MRFIT study and 22 years for the Chicago observational study, the low-risk patients had a significantly lower CHD mortality (0.2 to 8.8 versus 1.5 to 28 percent for those with ≥1 risk factor) and lower all cardiovascular disease mortality (0.3 to 15.8 versus 2.1 to 53 percent).

Help improve UpToDate. Did UpToDate answer your question? • Yes • No

Figure 8 From a previous project at NSEP. The picture is printed with permission from the participant.

3 Method

In order to provide answers to the research questions an experiment was designed and conducted. When designing the experiment the following variables were derived from the research questions:

Independent variables:

- Think aloud method
- Task type

Dependent variables:

- Task completion time and rate
- Number of usability problems detected
- Amount of verbalization
- Participants' subjective assessment

Table 1 What is measured in the different combinations of think aloud method and task type

	CTA	RTA	RTAg
Task 1	• Task completion time and rate • Usability problems • Verbalization • Subjective assessment	• Task completion time and rate • Usability problems • Verbalization • Subjective assessment	• Task completion time and rate • Usability problems • Verbalization • Subjective assessment
Task 2	• Task completion time • Verbalization • Subjective assessment	• Task completion time • Verbalization • Subjective assessment	• Task completion time • Verbalization • Subjective assessment
Task 3	• Task completion time • Verbalization • Subjective assessment	• Task completion time • Verbalization • Subjective assessment	• Task completion time • Verbalization • Subjective assessment

Other variables that may influence the results include variations in the participant group and the way the experiment was conducted. These variables were held as constant as possible.

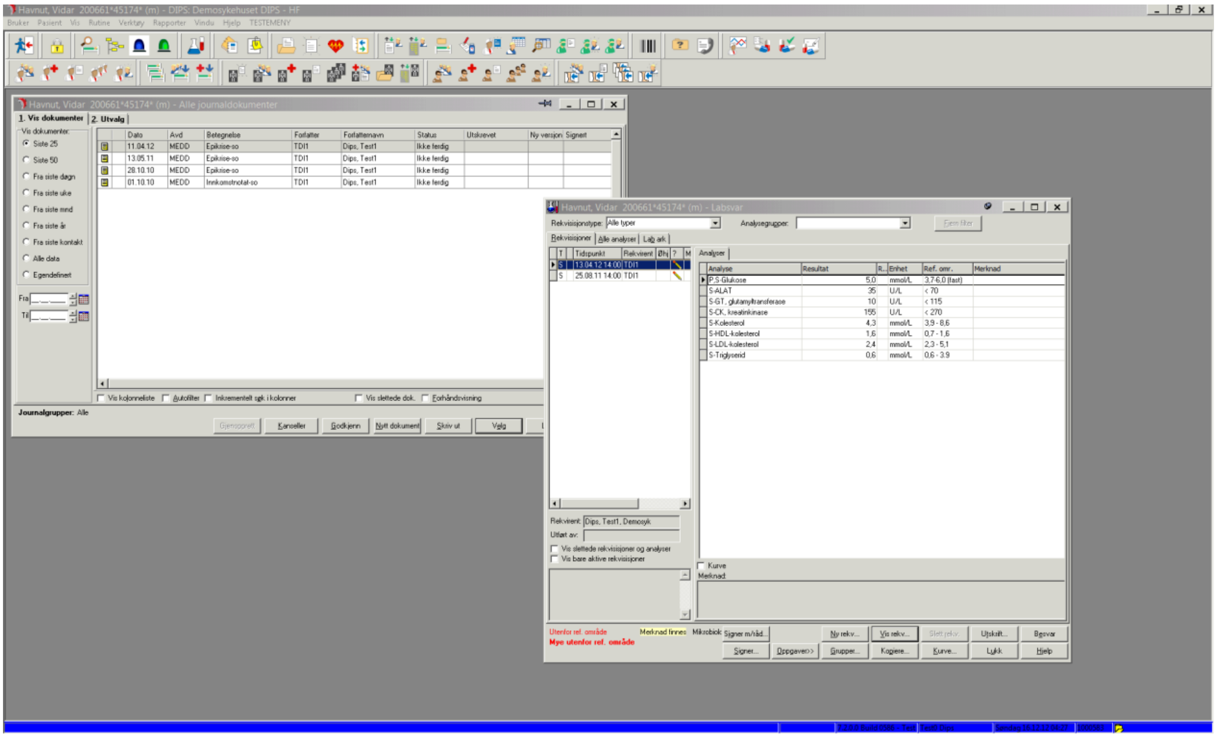
The experiment was designed as a usability test with three tasks, one task of each type. All tests were recorded on video, including the retrospective sessions. All the dependent variables could be extracted from the video material except the participants' subjective assessment, which was collected through a questionnaire.

3.1 Experiment

The setting for the experiment was that the participants were acting as a doctor at a neurological outpatient clinic, and the patient was about to come in for an appointment.

3.1.1 Test object

The object of the test was a major Electronic Health Record (EHR) system. The system is the most widely used hospital EHR system in Norway. Figure 9 shows the EHR system, with the active patient's collection of record documents in the left window, and the lab results in the right window.



Havnut, Vidar 200661*45174* (m) - DIPS: Demosykehuset DIPS - HF

Havnut, Vidar 200661*45174* (m) - Alle journaldokumenter

Vis dokumenter: ● Siste 25 ○ Siste 50 ○ Fra siste døgn ○ Fra siste uke ○ Fra siste mnd ○ Fra siste år ○ Fra siste kontakt ○ Alle data ○ Egendefinert

	Dato	Avd	Betegnelse	Forfatter	Forfatternavn	Status	Utskrevet	Ny versjon	Signert
	11.04.12	MEDD	Epikrise-so	TDI1	Dips, Test1	Ikke ferdig			
	13.05.11	MEDD	Epikrise-so	TDI1	Dips, Test1	Ikke ferdig			
	28.10.10	MEDD	Epikrise-so	TDI1	Dips, Test1	Ikke ferdig			
	01.10.10	MEDD	Innkomstnotat-so	TDI1	Dips, Test1	Ikke ferdig			

Journalgrupper: Alle

Havnut, Vidar 200661*45174* (m) - Labsvar

Rekvisisjonstype: Alle typer

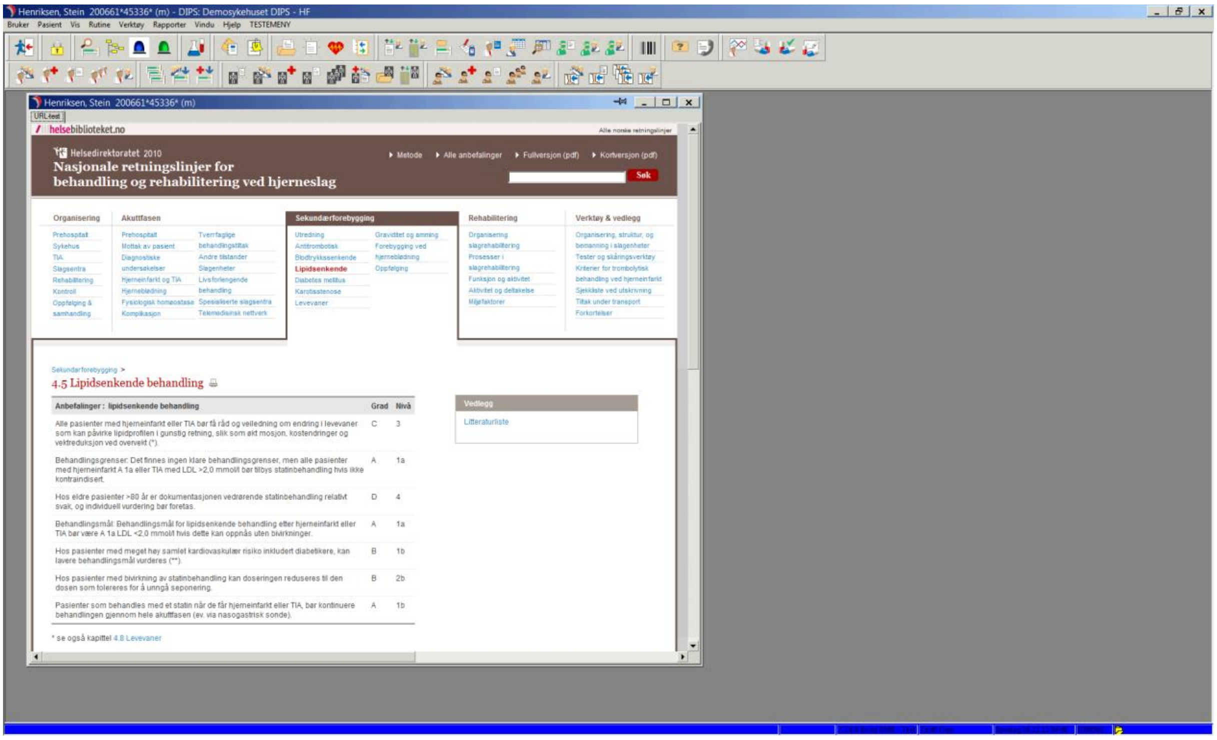
T	Tidspunkt	Rekvirent	Øhj	?	M
S	13.04.12 14:00	TDI1			
S	25.08.11 14:00	TDI1			

Analyse	Resultat	R..	Enhet	Ref. omr.	Merknad
P,S-Glukose	5,0		mmol/L	3,7-6,0 (fast)	
S-ALAT	35		U/L	< 70	
S-GT, glutamyltransferase	10		U/L	< 115	
S-CK, kreatinkinase	155		U/L	< 270	
S-Kolesterol	4,3		mmol/L	3,9 - 8,6	
S-HDL-kolesterol	1,6		mmol/L	0,7 - 1,6	
S-LDL-kolesterol	2,4		mmol/L	2,3 - 5,1	
S-Triglyserid	0,6		mmol/L	0,6 - 3.9	

Rekvirent: Dips, Test1, Demosyk

Figure 9 The EHR system

For this study we used a custom made prototype that is able to give the clinicians advice based on the content of the patients' health record. This prototype is executed from the menu system in the EHR system, through a menu item called "External patient information". When executed a browser window appears within the EHR systems' work space, showing a national clinical guideline for cerebral infarction (see Figure 10). The experiment was designed in such a way that half the participants were presented the whole guideline[26], while the other half was presented "Lipid-lowering treatment"[27] which is a sub chapter of the "Secondary prevention"-section of the guideline.



Helsedirektoratet 2010
Nasjonale retningslinjer for behandling og rehabilitering ved hjerneslag

4.5 Lipidsenkende behandling

Anbefalinger : lipidsenkende behandling	Grad	Nivå
Alle pasienter med hjerneinfarkt eller TIA bør få råd og veiledning om endring i levevaner som kan påvirke lipidprofilen i gunstig retning, slik som økt mosjon, kostendringer og vektreduksjon ved overvekt (*).	C	3
Behandlingsgrenser: Det finnes ingen klare behandlingsgrenser, men alle pasienter med hjerneinfarkt A 1a eller TIA med LDL >2,0 mmol/l bør tilbys statinbehandling hvis ikke kontraindisert.	A	1a
Hos eldre pasienter >80 år er dokumentasjonen vedrørende statinbehandling relativt svak, og individuell vurdering bør foretas.	D	4
Behandlingsmål: Behandlingsmål for lipidsenkende behandling etter hjerneinfarkt eller TIA bør være A 1a LDL <2,0 mmol/l hvis dette kan oppnås uten bivirkninger.	A	1a
Hos pasienter med meget høy samlet kardiovaskulær risiko inkludert diabetikere, kan lavere behandlingsmål vurderes (**).	B	1b
Hos pasienter med bivirkning av statinbehandling kan doseringen reduseres til den dosen som tolereres for å unngå seponering.	B	2b
Pasienter som behandles med et statin når de får hjerneinfarkt eller TIA, bør kontinuere behandlingen gjennom hele akuttfasen (ev. via nasogastrisk sonde).	A	1b

* se også kapittel 4.8 Levevaner

Figure 10 The EHR system showing the clinical guideline

Two identical patients were created in the EHR, but with different name and social security number. Each patient is designated a *patientid* number which was used to redirect to the right part of the guideline. A php-script was written to do this redirect (see appendix). These patients were used for task 2 and 3, and were based on real, anonymous data. For task 1 a third patient was created. This patient did not have real data. There was added one health record document, and a couple of lab results. For this patient the “external patient info” presented the national guideline for diabetes[28].

3.1.2 Participants

A total of 18 medical students from NTNU in Trondheim participated in the study. They were recruited from fourth, fifth and sixth year classes. At this level the students have enough clinical experience to be in the target group of the test object. As can be seen in Table 2 Participant characteristics, it is a fairly homogenous group, though it could be better balanced for gender.

Table 2 Participant characteristics

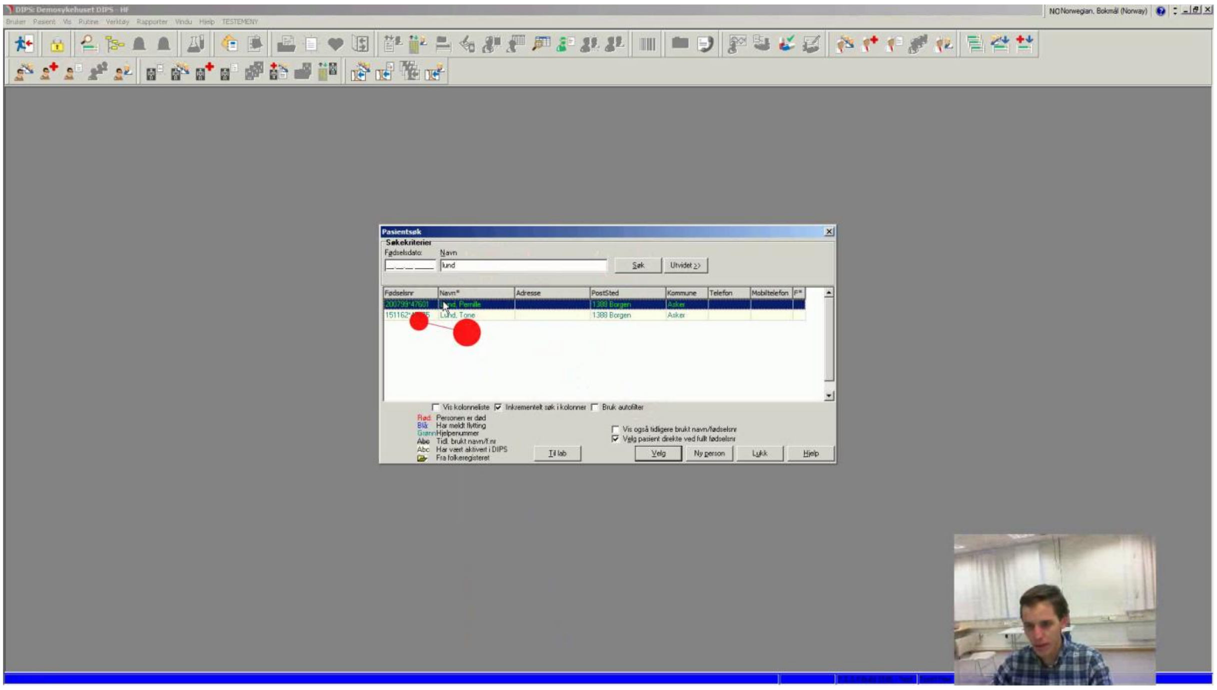
Gender		Age			Current year of study		
Female	Male	22-24	25-27	>27	4th	5th	6th
6	12	6	7	5	8	4	6

3.1.3 Tasks

The tasks are designed in a rather artificial fashion to ensure that they have the desired characteristics.

Task 1 (interactivity): Find patient #1 and locate the patient's record content and lab results, then go to "External patient info".

This was designed as an easy interactive task. This task is a typical usability test task, and is the benchmark that the other tasks are compared against. It is a highly interactive task that is not too cognitively demanding for the participants. It implies little reading and no problem solving beyond interacting with the user interface of the EHR system. Figure 11 shows a participant using the patient search engine.


DIPS: Demosykehuset DIPS - HF
Bruker Pasient Vis Rutine Verktøy Rapporter Vindu Hjelp TESTEMENY
NO Norwegian, Bokmål (Norway)
Pasientsøk

Søkekriterier

Fødselsdato: __.__.__ ____

Navn: lund

Fødselsnr	Navn*	Adresse	PostSted	Kommune	Telefon	Mobiltelefon	F*
200799*47601	L[illegible]nd, Pernille		1388 Borgen	Asker			
151162*[illegible]5	Lund, Tone		1388 Borgen	Asker			

□ Vis kolonneliste ☑ Inkrementelt søk i kolonner □ Bruk autofilter

Rød	Personen er død
Blå	Har meldt flytting
Grønn	Hjelpenummer
Abc	Tidl. brukt navn/f.nr
Abc	Har vært aktivert i DIPS
	Fra folkeregisteret

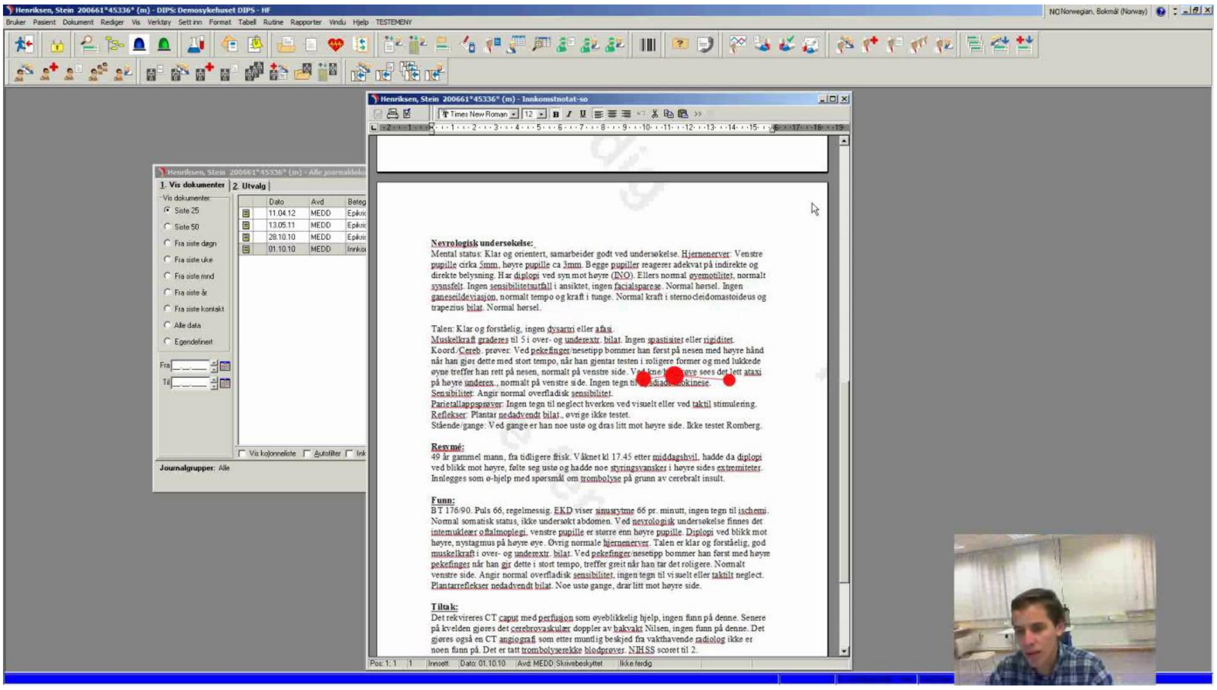
□ Vis også tidligere brukt navn/fødselsnr

☑ Velg pasient direkte ved fullt fødselsnr

Figure 11 The completion of task 1. The picture is printed with permission from the participant

Task 2 (reading): Find patient #2 and read the patients' record in order to get to know the case history.

This was designed to be a reading task. This task was designed with the intention of having a task that implies little interactivity, and hence produces a video that is quite static. The purpose of this task is to examine how the participants are able to verbalize in retrospect with a static video as a cue, with and without gaze playback. Figure 12 shows a participant reading the patient's record.



Nevrologisk undersøkelse:
Mental status: Klar og orientert, samarbeider godt ved undersøkelse. Hjernenerver: Venstre pupille cirka 5mm, høyre pupille ca 3mm. Begge pupiller reagerer adekvat på indirekte og direkte belysning. Har diplopi ved syn mot høyre (INO). Ellers normal øyemotilitet, normalt synsfelt. Ingen sensibilitetsutfall i ansiktet, ingen facialsparese. Normal hørsel. Ingen ganeseildeviasjon, normalt tempo og kraft i tunge. Normal kraft i sternocleidomastoideus og trapezius bilat. Normal hørsel.

Talen: Klar og forståelig, ingen dysartri eller afasi.
Muskelkraft graderes til 5 i over- og underextr. bilat. Ingen spastisitet eller rigiditet.
Koord./Cereb. prøver: Ved pekefinger/nesetipp bommer han først på nesen med høyre hånd når han gjør dette med stort tempo, når han gjentar testen i roligere former og med lukkede øyne treffer han rett på nesen, normalt på venstre side. Ved kne/hæl-prøve sees det lett ataxi på høyre underex., normalt på venstre side. Ingen tegn til dysdiadokokinese.
Sensibilitet: Angir normal overfladisk sensibilitet.
Parietallappsprøver: Ingen tegn til neglect hverken ved visuelt eller ved taktil stimulering.
Reflekser: Plantar nedadvendt bilat., øvrige ikke testet.
Stående/gange: Ved gange er han noe ustø og dras litt mot høyre side. Ikke testet Romberg.

Resymé:
49 år gammel mann, fra tidligere frisk. Våknet kl 17.45 etter middagshvil, hadde da diplopi ved blikk mot høyre, følte seg ustø og hadde noe styringsvansker i høyre sides extremiteter. Innlegges som ø-hjelp med spørsmål om trombolyse på grunn av cerebralt insult.

Funn:
BT 176/90. Puls 66, regelmessig. EKD viser sinusrytme 66 pr. minutt, ingen tegn til ischemi. Normal somatisk status, ikke undersøkt abdomen. Ved nevrologisk undersøkelse finnes det internukleær oftalmoplegi, venstre pupille er større enn høyre pupille. Diplopi ved blikk mot høyre, nystagmus på høyre øye. Øvrig normale hjernenerver. Talen er klar og forståelig, god muskelkraft i over- og underextr. bilat. Ved pekefinger/nesetipp bommer han først med høyre pekefinger når han gir dette i stort tempo, treffer greit når han tar det roligere. Normalt venstre side. Angir normal overfladisk sensibilitet, ingen tegn til visuelt eller taktilt neglect. Plantarreflekser nedadvendt bilat. Noe ustø gange, drar litt mot høyre side.

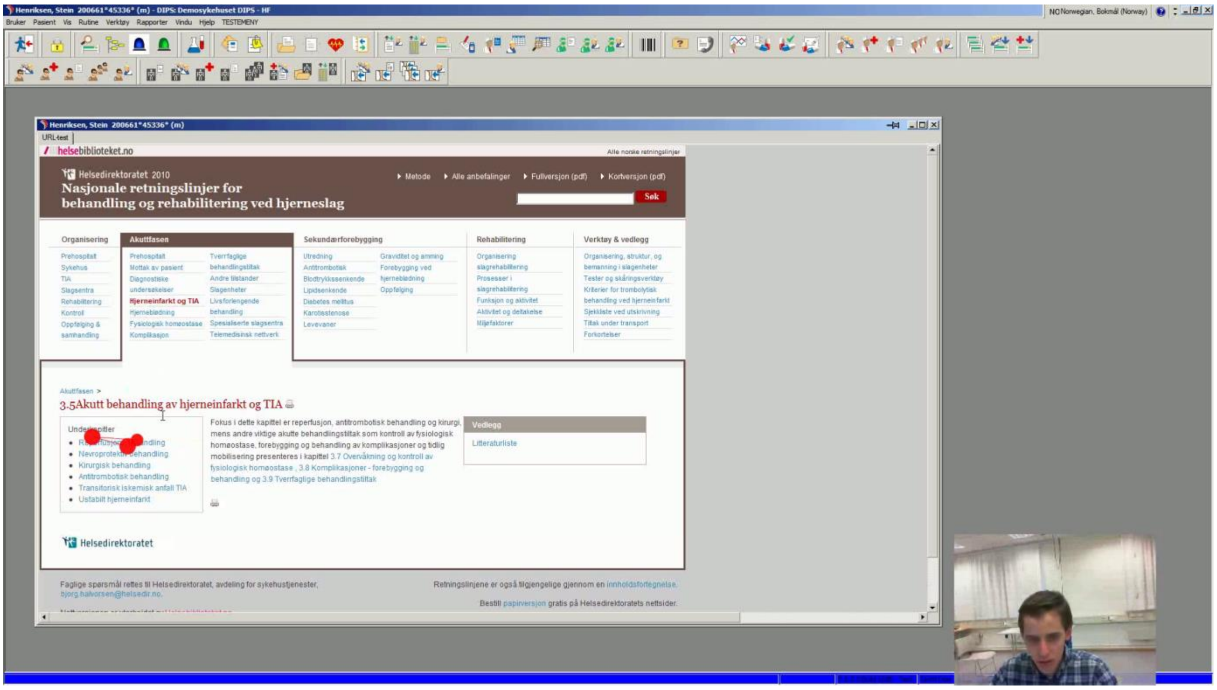
Tiltak:
Det rekvireres CT caput med perfusjon som øyeblikkelig hjelp, ingen funn på denne. Senere på kvelden gjøres det cerebrovaskulær doppler av bakvakt Nilsen, ingen funn på denne. Det gjøres også en CT angiografi som etter muntlig beskjed fra vakthavende radiolog ikke er noen funn på. Det er tatt trombolyserekke blodprøver. NIHSS scoret til 2.

Pos: 1: 1 | 1 | Innsett | Dato: 01.10.10 | Avd: MEDD Skrivebeskyttet | Ikke ferdig

Figure 12 The completion of task 2. The picture is printed with permission from the participant

Task 3 (problem solving): Open the "external patient info" and get oriented in the guideline. Then write a note where you sketch how you would treat this patient.

This was designed to be a complex problem solving task. It is complex in the sense that the participants are solving a medical problem, and at the same time interact with the tools they use for solving the medical problem, which are the clinical guideline and the EHR system. The purpose of this task is to examine how complexity affects the participants' ability to verbalize during CTA. Figure 13 shows the participant browsing and reading the clinical guideline.



Helsedirektoratet 2010
Nasjonale retningslinjer for behandling og rehabilitering ved hjerneslag

Akuttfasen >

3.5Akutt behandling av hjerneinfarkt og TIA

Underkapitler
- Reperfusjonsbehandling
- Nevroprotektiv behandling
- Kirurgisk behandling
- Antitrombotisk behandling
- Transitorisk iskemisk anfall TIA
- Ustabilt hjerneinfarkt

Fokus i dette kapittel er reperfusjon, antitrombotisk behandling og kirurgi, mens andre viktige akutte behandlingstiltak som kontroll av fysiologisk homøostase, forebygging og behandling av komplikasjoner og tidlig mobilisering presenteres i kapittel 3.7 Overvåkning og kontroll av fysiologisk homøostase , 3.8 Komplikasjoner - forebygging og behandling og 3.9 Tverrfaglige behandlingstiltak

Faglige spørsmål rettes til Helsedirektoratet, avdeling for sykehustjenester, bjorg.halvorsen@helsedir.no.

Retningslinjene er også tilgjengelige gjennom en innholdsfortegnelse.

Bestill papirversjon gratis på Helsedirektoratets nettsider.

Figure 13 The completion of task 3. The picture is printed with permission from the participant

3.1.4 Questionnaire

The participants were asked one question after each task depending on the TA method used. The questionnaire was designed as a five point likert scale:

CTA:

To what extent does thinking aloud influence the way you solved the task. The scale went from “unproblematic” to “disturbing”.

RTA:

To what extent do you find it easy to remember what you were thinking while solving the task? The scale went from “easy” to “hard”

RTAg:

To what extent do you find it useful to see your own eye movements while verbalizing? The scale went from “very useful” to “irrelevant”.

3.1.5 Apparatus

The test was conducted at NSEP (Norwegian EHR Research Centre) usability lab. The test is run on a PC with Windows 7 and a Tobii X60 eye tracker. Data acquisition and is done with Tobii Studio 2.3.2.0 software.

3.1.6 Procedure

With three methods and three tasks where the tasks are fixed and the methods are randomized give six permutations as shown in Table 3. It was decided to recruit for three full rounds, which is why there was 18 participants in the study.

Table 3 Randomization scheme

Participant #	Task 1	Task 2	Task 3
1	CTA	RTA	RTAg
2	CTA	RTAg	RTA
3	RTA	CTA	RTAg
4	RTA	RTAg	CTA
5	RTAg	CTA	RTA
6	RTAg	RTA	CTA
7	CTA	RTA	RTAg
8	CTA	RTAg	RTA
9	RTA	CTA	RTAg
10	RTA	RTAg	CTA
11	RTAg	CTA	RTA
12	RTAg	RTA	CTA
13	CTA	RTA	RTAg
14	CTA	RTAg	RTA
15	RTA	CTA	RTAg
16	RTA	RTAg	CTA
17	RTAg	CTA	RTA
18	RTAg	RTA	CTA

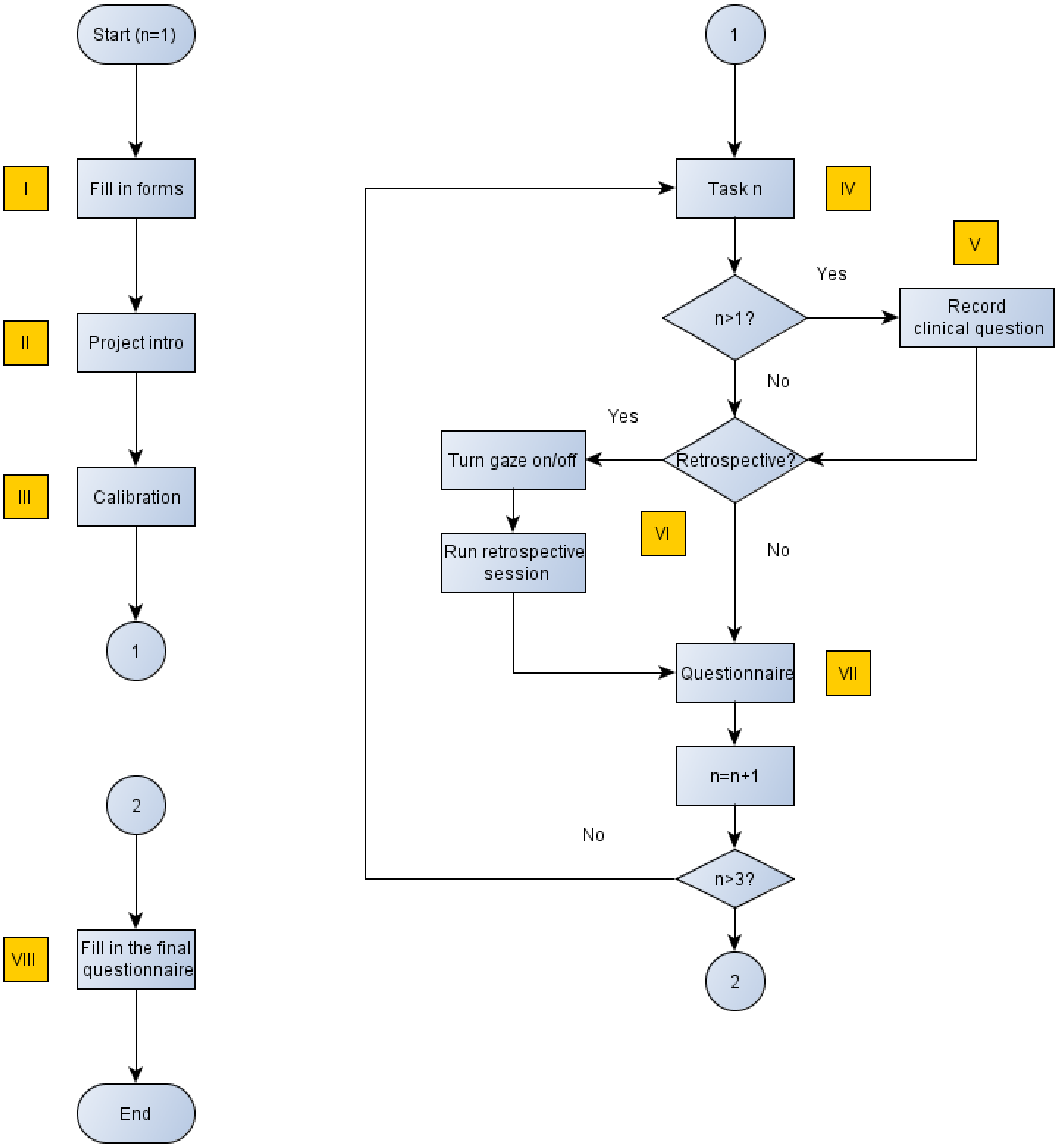


Figure 14 Experiment procedure

Explanations to the flow chart in Figure 14:

I. **Fill in forms:** All participants filled inn two forms when they arrived at the usability lab:

- Consent form (see appendix)
- Demographics and background form (see appendix)

II. **Project intro:** The project, the experiment and the setting is explained to the participant.

III. **Calibration:** The eye tracker has to be calibrated for each participant. This is a quick procedure that takes less than a minute.

IV. **Task n:** This is the actual task that the participant is performing. The task is written on a piece of paper. There are two sets of tasks that are given to every second participant. The difference on the sets is the name and social security number for the patient used in task 2 and 3 to ensure that the participant was given the right part of the clinical guideline on “external patient information”.

V. **Record clinical questions:** The participants are asked to write down their clinical questions. This is done after task 2 and task 3.

VI. **Retrospective:** The facilitator has to check the scheme which participant is using which method. If the method is RTA or RTAg the participant has to perform the retrospective session at this point. If the method is RTA the gaze playback must be turned off, and if the method is RTAg the gaze playback must be turned on.

VII. **Questionnaire:** The participant is asked one question depending on the method.

VIII. **Final questionnaire:** The participants are asked to fill in at questionnaire regarding the clinical guideline used in task 3.

3.1.7 Conducting retrospective sessions

In the retrospective sessions the video from task completion were played back to the participant as a cue for verbalization. In this study the retrospective sessions were conducted in a strict fashion. The video was started, and the participants were asked to verbalize what they were thinking while solving the task, and then stopped in the end.

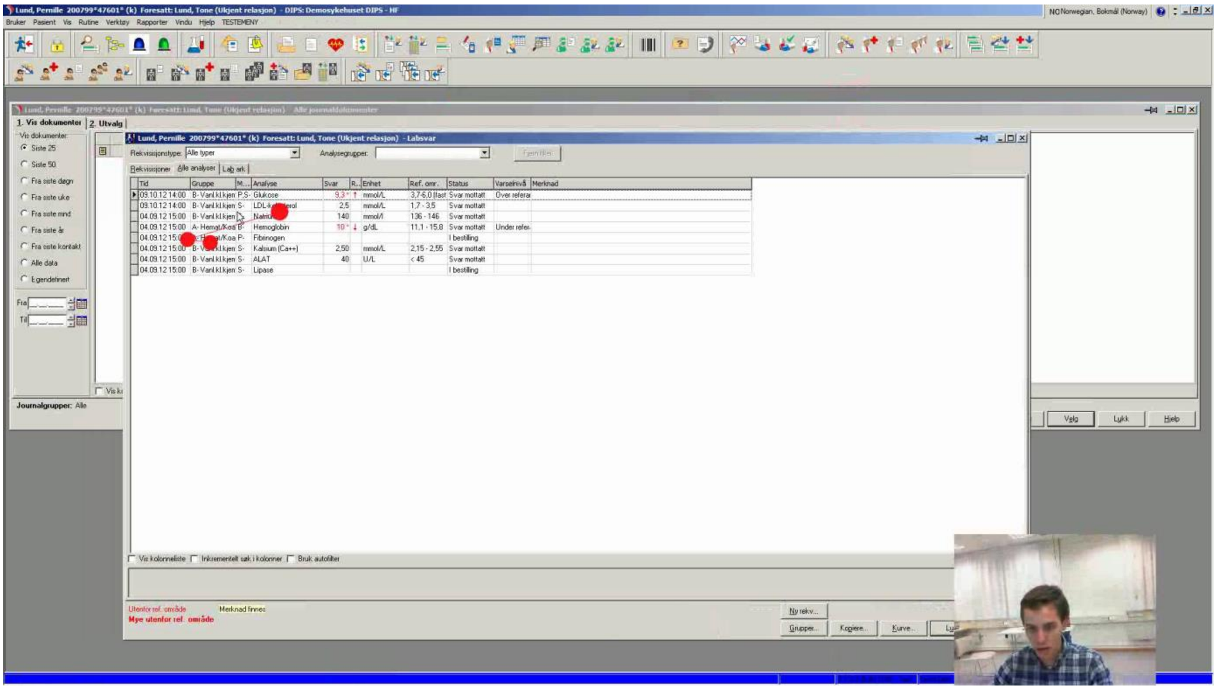
3.2 Analysis

Not every person can be eye tracked. According to Tobii 95% out of a random population can be tracked. Out of the 18 participants there was one that could not be tracked, so one additional participant were recruited. After the experiment was over it was discovered that one participants' data was missing, and one of the tasks from two of the other participants was also missing due to Tobii Studio crash. The following video material was analyzed:

- 17 tasks using CTA
- 16 tasks using RTA
- 16 tasks using RTAg
- 17 RTA retrospective sessions
- 17 RTAg retrospective sessions

3.2.1 Processing empirical data

The video exported from Tobii Studio (see Figure 15 Video exported from Tobii Studio) was imported into Transana version 2.51 for coding[29]. Transana is a qualitative analysis tool that is used for transcribing video data, but it may also be used for quantification. A transcript is added to the video where text and time stamps can be added.



Lund, Pernille 200799*47601* (k) Foresatt: Lund, Tone (Ukjent relasjon) - Labsvar

Rekvisisjonstype: Alle typer   Analysegrupper:

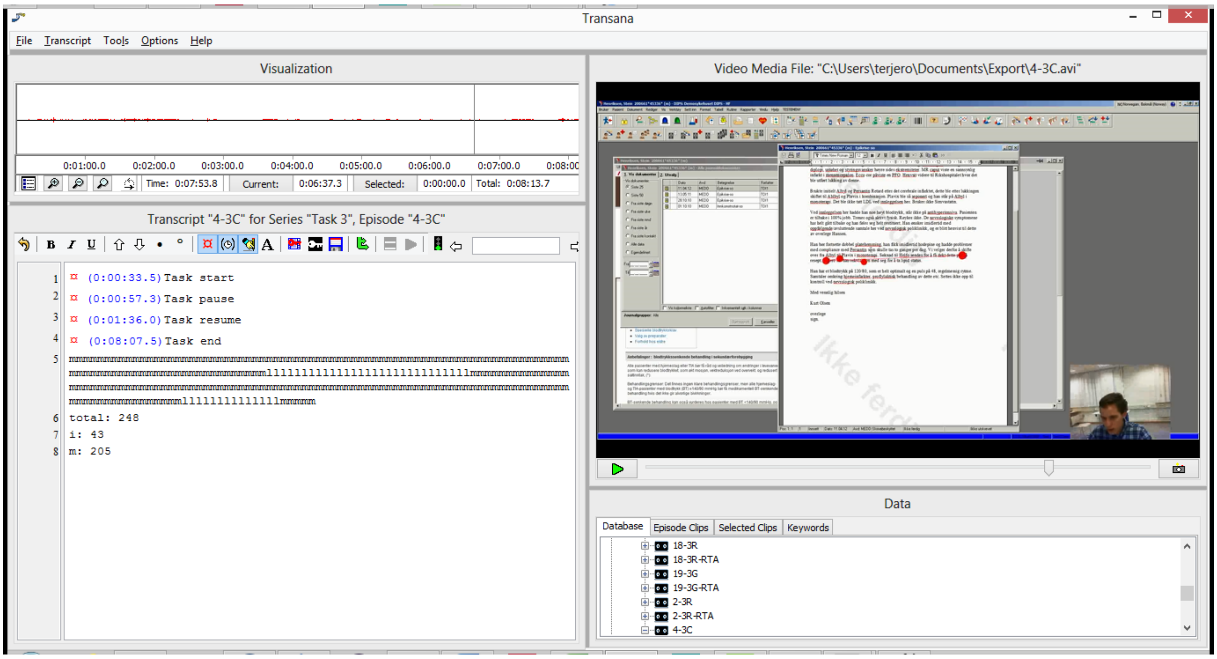
Tid	Gruppe	M...	Analyse	Svar	R...	Enhet	Ref. omr.	Status	Varselnivå	Merknad
09.10.12 14:00	B-Vanl.kl.kjen	P,S-	Glukose	9,3 *	↑	mmol/L	3,7-6,0 (fast	Svar mottatt	Over referan	
09.10.12 14:00	B-Vanl.kl.kjen	S-	LDL-k[illegible]erol	2,5		mmol/L	1,7 - 3,5	Svar mottatt		
04.09.12 15:00	B-Vanl.kl.kjen	[illegible]	Natri[illegible]	140		mmol/l	136 - 146	Svar mottatt		
04.09.12 15:00	A-Hemat/Koa	B-	Hemoglobin	10 *	↓	g/dL	11,1 - 15,8	Svar mottatt	Under refer.	
04.09.12 15:[illegible]	[illegible]t/Koa	P-	Fibrinogen					I bestilling		
04.09.12 15:00	B-V[illegible]kl.kjen	S-	Kalsium (Ca++)	2,50		mmol/L	2,15 - 2,55	Svar mottatt		
04.09.12 15:00	B-Vanl.kl.kjen	S-	ALAT	40		U/L	< 45	Svar mottatt		
04.09.12 15:00	B-Vanl.kl.kjen	S-	Lipase					I bestilling		

Utenfor ref. område   Merknad finnes
Mye utenfor ref. område

Figure 15 Video exported from Tobii Studio. The picture is printed with permission from the participant

The following codes were added to the video transcripts:

- Task start, task pause, task resume and task end.
- For task 2: Participant start reading, participant stop reading.
- One code for each word verbalized concerning interaction.
- One code for each word verbalized concerning the medical problem.
- Occurrence of usability problems, and whether the participant needed assistance.
- Recurring themes for qualitative analysis.
- Quotations.


Transana
File Transcript Tools Options Help
Visualization
0:01:00.0 0:02:00.0 0:03:00.0 0:04:00.0 0:05:00.0 0:06:00.0 0:07:00.0 0:08:00
Time: 0:07:53.8 Current: 0:06:37.3 Selected: 0:00:00.0 Total: 0:08:13.7
Transcript "4-3C" for Series "Task 3", Episode "4-3C"
Video Media File: "C:\Users\terjero\Documents\Export\4-3C.avi"
Data
Database Episode Clips Selected Clips Keywords
18-3R
18-3R-RTA
19-3G
19-3G-RTA
2-3R
2-3R-RTA
4-3C

```
¤ (0:00:33.5)Task start
¤ (0:00:57.3)Task pause
¤ (0:01:36.0)Task resume
¤ (0:08:07.5)Task end
mmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
mmmmmmmmmmmmmmmmmmmmmmmmmmmmlllllllllllllllllllllllllllllmmmmmmmmmmmmmm
mmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
mmmmmmmmmmmmmmmmllllllllllllllmmmmm
total: 248
i: 43
m: 205
```

Figure 16 Screen shot from Transana. The picture is printed with permission from the participant.

In Figure 16 the transcript can be seen on the lower left side. In this example there are four time stamps, one task start, one task pause, one task resume and one task end. Next is the verbalization word count. An “m” is used for words concerning the medical problem and “l” for words concerning interacting with the system. The “l” was used instead of “i” because it was more conveniently located on the keyboard. Last in the transcript are listed the total number of words, the number of words concerning interaction, and the number of words concerning the medical problem.

All data was then transferred into different Microsoft Excel spreadsheets along with the results from the questionnaire.

3.2.2 Statistical analyses

Most of the data was analyzed using the Minitab 16.2.1 statistics package, using one-way ANOVA (variance analysis) to find statistical significant differences between different groups of data. Tukey’s post-hoc test was used to identify which group of data was different when statistical significance was found. Minitab was also used to provide interval plots to illustrate how the different groups of data related to each other. All datasets were tested for normality, which is a premise for ANOVA to produce valid results. The Kolmogorov-Smirnov test was used for this purpose. If one or two outliers could make the dataset pass the normality test, these were removed from the analysis.

4 Results

4.1 Task completion time

In order to find out whether a method is able to produce valid results, task completion time was measured. RTA and RTAg becomes the baseline, as no think aloud is done with these methods. There should consequently be no difference between RTA and RTAg, and the datasets from these two methods are combined and designated RTAc. This applies to reactivity described by Russo et al[3], and the hypothesis is that the participants will use more time completing the task during CTA compared to the retrospective methods.

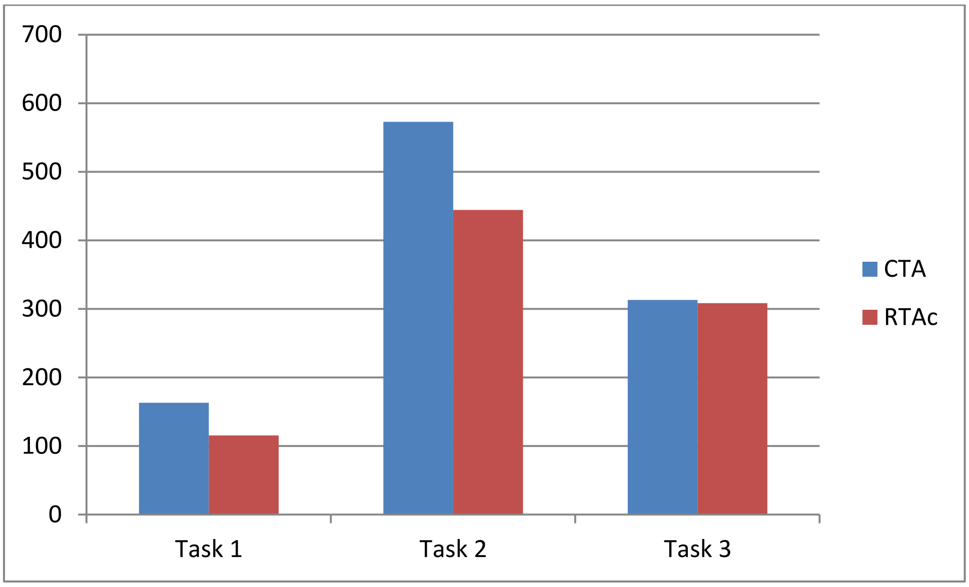


Figure 17 Average task completion times measured in seconds

Figure 17 shows the average task completion times for CTA versus the two retrospective methods. The task completion time is higher during CTA for all tasks. By removing two outliers, one in task 2 and one in task 3, the datasets for all tasks passed the normality test. When running one-way ANOVA no statistical significant difference was found.

4.2 Usability problems

A total of 5 distinct usability problems were identified in task 1. In task 2 there were some occurrences of some of the same problems as in task 1, but only 1 new problem. Task 3 concerned mostly navigating a clinical guideline, which was a totally different user interface. Therefore it was decided to concentrate on task 1.

Problems:

P1: The participant cannot see that the patient is already active.

P2: The patient search engine doesn't support given names.

P3: When clicking the birth date field, the cursor is placed where the participant clicks.

P4: Confusion between "patient" and "view" drop down menus.

P5: Too many buttons. This invites the participants into a lengthy search for the right button.

4.2.1 Usability problem P1

This was the most prominent usability problem in terms of number of instances, and this was the only problem where the participants needed assistance. This is a typical feedback problem, and occurred when the participants had found the patient, but did not notice that the patient was chosen. His is illustrated by Figure 18 and Figure 19. The EHR system can be configured to open certain windows by default when a patient is chosen.

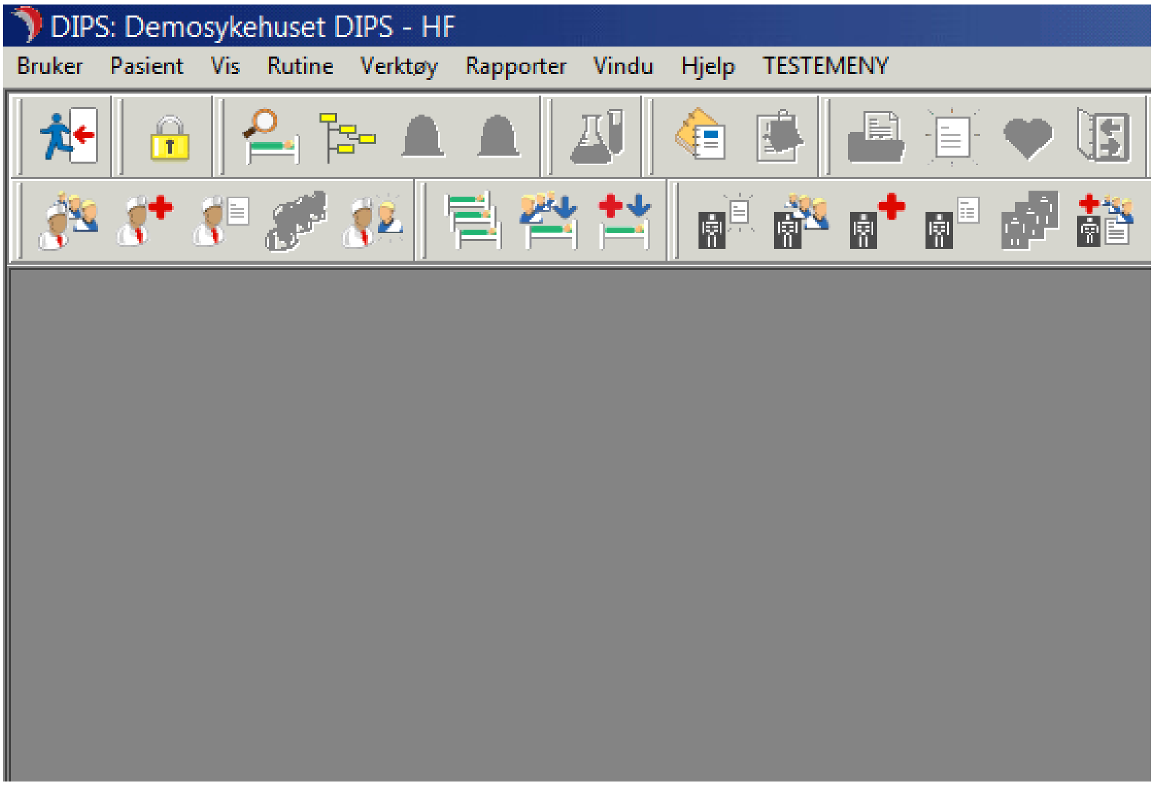
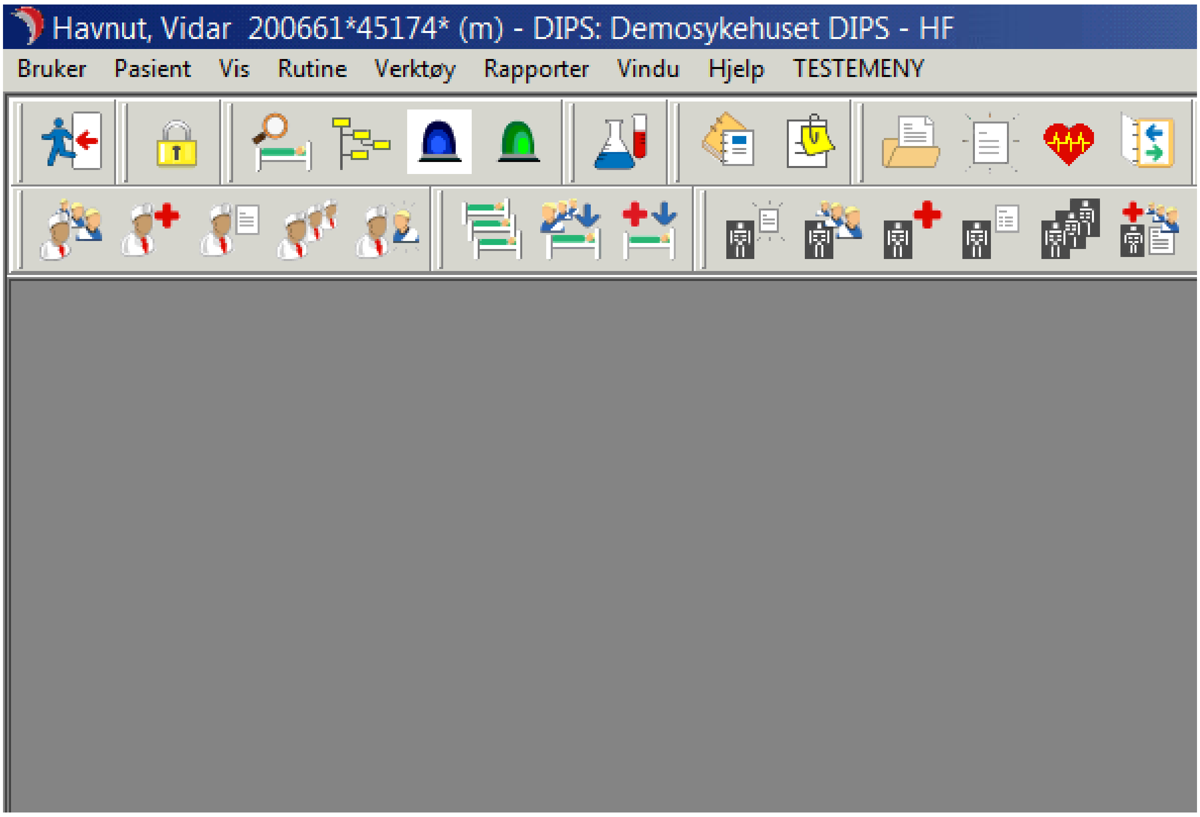


Figure 18 Section of the EHR system with no active patient. This relates to usability problem P1



Havnut, Vidar 200661*45174* (m) - DIPS: Demosykehuset DIPS - HF
Bruker
Pasient
Vis
Rutine
Verktøy
Rapporter
Vindu
Hjelp
TESTEMENY

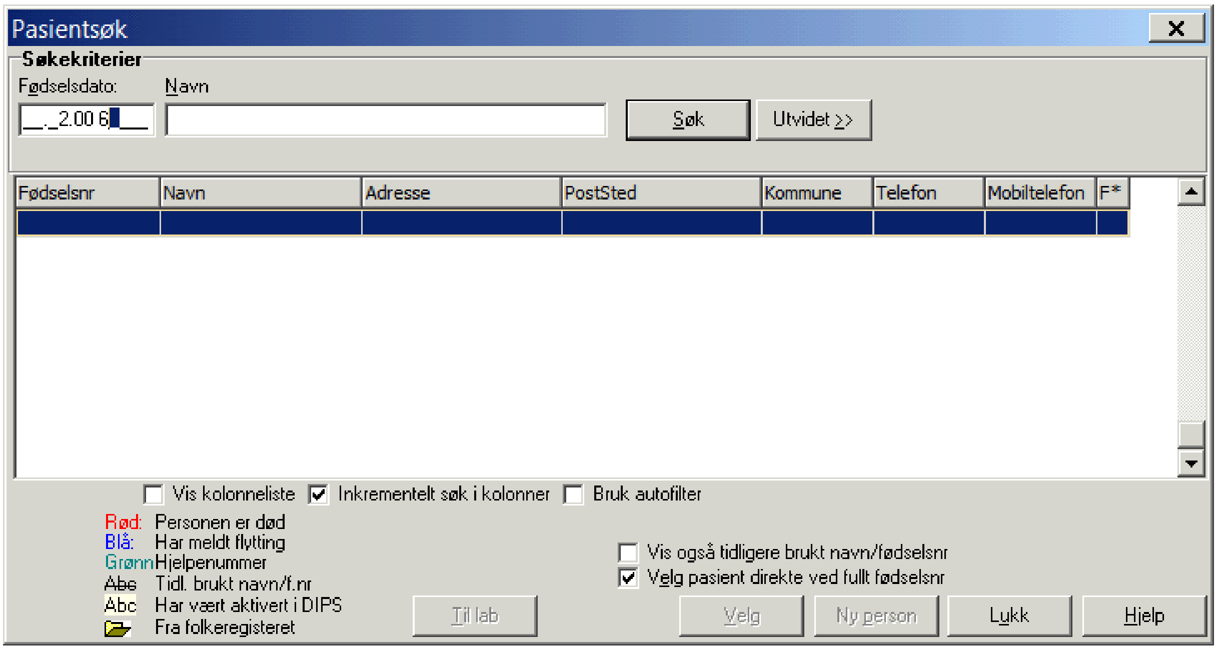
Figure 19 Section of the EHR system showing active patient. This relates to usability problem P1

4.2.2 Usability problem P2

The patient search engine does not support given names. This is not a very severe problem, but it would be expected from a modern information system that there is not such limitation.

4.2.3 Usability problem P3

The field for entering the patient's birth date and social security number is selected by default, so the user can start entering the birth date without further ado. Some of the participants did not realize that, and clicked on the field before entering the birth date. What happens is that the cursor is positioned where clicked. It is assumed that this is the desired functionality, but for some of the participants this was a problem. Figure 20show the birth date field at the upper left corner, and what happens if the user starts typing where clicked.



Fødselsdato: __.__2.00 6▌___

Navn:

☐ Vis kolonneliste ☑ Inkrementelt søk i kolonner ☐ Bruk autofilter

Rød: Personen er død
Blå: Har meldt flytting
Grønn Hjelpenummer
~~Abc~~ Tidl. brukt navn/f.nr
Abc Har vært aktivert i DIPS
🗁 Fra folkeregisteret

☐ Vis også tidligere brukt navn/fødselsnr
☑ Velg pasient direkte ved fullt fødselsnr

Figure 20 The patient search dialog box were involved in two of the usability problems: P2 and P3

4.2.4 Usability problem P4

Many of the participants expected to find the patient's record and lab results under the "patient" menu item. But this information was located under the menu item "view" which is the next menu item to the right. In Figure 21 patient = pasient and view = vis.

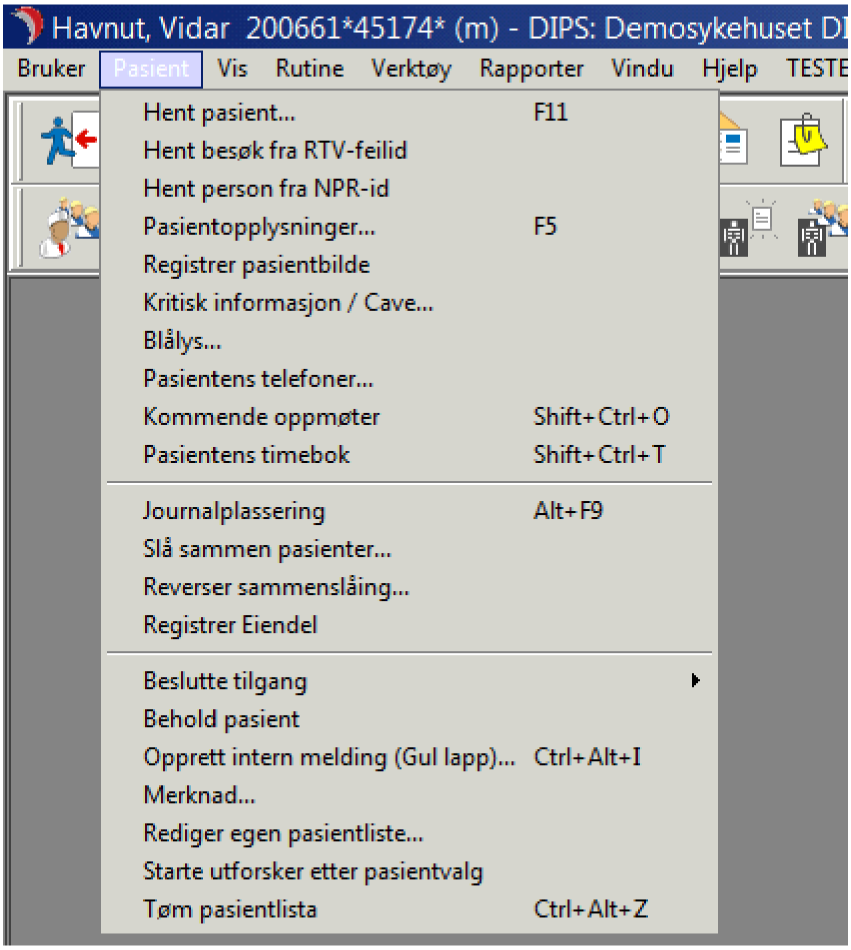

Havnut, Vidar 200661*45174* (m) - DIPS: Demosykehuset DI
Bruker Pasient Vis Rutine Verktøy Rapporter Vindu Hjelp TESTE
Hent pasient... F11
Hent besøk fra RTV-feilid
Hent person fra NPR-id
Pasientopplysninger... F5
Registrer pasientbilde
Kritisk informasjon / Cave...
Blålys...
Pasientens telefoner...
Kommende oppmøter Shift+Ctrl+O
Pasientens timebok Shift+Ctrl+T
Journalplassering Alt+F9
Slå sammen pasienter...
Reverser sammenslåing...
Registrer Eiendel
Beslutte tilgang
Behold pasient
Opprett intern melding (Gul lapp)... Ctrl+Alt+I
Merknad...
Rediger egen pasientliste...
Starte utforsker etter pasientvalg
Tøm pasientlista Ctrl+Alt+Z

Figure 21 The "patient" menu. This relates to usability problem P4

4.2.5 Usability problem P5

There was a lot of buttons or icons at the top of the EHR systems main window. Some participants spent a lot of time examining each and every icon in order to find the right one. In a real situation the system would have been configured with the most relevant icons visible depending on hospital department. In this project all icons were visible and represented a problem for some of the participants.

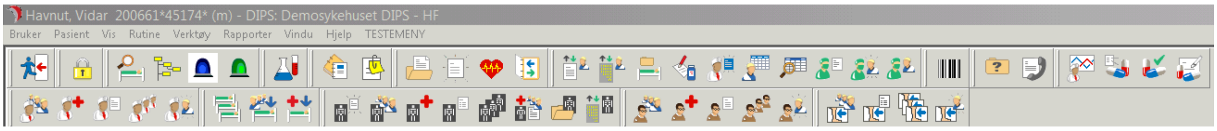

Havnut, Vidar 200661*45174* (m) - DIPS: Demosykehuset DIPS - HF
Bruker Pasient Vis Rutine Verktøy Rapporter Vindu Hjelp TESTEMENY

Figure 22 Icon jungle. This relates to usability problem P5

4.2.6 Instances of usability problems

Table 4 Instances of usability problems

	Method	P1	P2	P3	P4	P5
1	CTA	O, V				
2	CTA		O			
3	CTA	O, A	O		O	O
4	CTA				O	
5	CTA	O, A			O	O
6	CTA					
7	RTA		O, V			
8	RTA	O, V, A	O			O
9	RTA		O	O, V		
10	RTA	O, V, A	O			
11	RTAg	O, V	O		O	
12	RTAg	O, V			O	
13	RTAg					
14	RTAg				O, V	
15	RTAg			O, V		
16	RTAg	O, A		O		O, V

Table explanation: O= observed, V= verbalized, A= assistance given.

Out of a total of 28 instances of usability problems observed by the researcher, while only 10 of them were verbalized by the participants. One instance was verbalized during CTA, while the 9 others were almost equally distributed between the two retrospective methods. When saying that a usability problem is verbalized means that the usability problem could have been identified through verbalization alone.

The findings were analyzed using one-way ANOVA with the following results:

Observed: F (2, 13) = 0.10, p = 0,901

Verbalized: F (2, 13) = 8.13, p = 0,005

There is no difference between the methods in terms of usability problems identified, but there is a statistical significant difference in terms of usability problems verbalized. Tukey's post-hoc analysis reveals that there is significantly less usability problems verbalized during CTA compared to the two other methods, but no difference between RTA and RTAg.

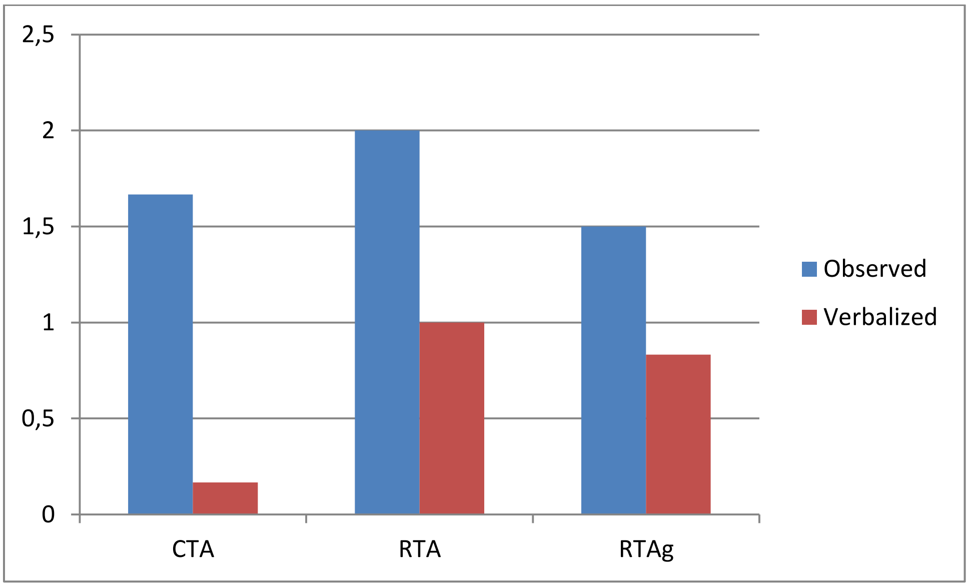


Figure 23 Mean number of usability problems identified by observation and by means of verbalization

Assistance was given to 5 users regarding P1. When participant had found the patient, the patients name and social security number appeared at the top of the window, with such small print that many did not see it. So they tried another search. If still not successful when initiating the third search, they were given assistance. Figure 18 and Figure 19 show a small section from the upper right corner of the EHR systems main window. In Figure 18 no patient is active. This is the state of the system when the test starts. In Figure 19 the patient "Vidar Havnut" is the active patient.

4.3 Verbalization word count

With the assumption that more words verbalized means more insight in the participants' thoughts, all the words verbalized were counted. The total word count was divided by the task completion time in order to calculate words verbalized per minute. Since there is variation in completion time between participants and between tasks, words per minute is a measure that can be used when comparing the different methods applied on the different tasks.

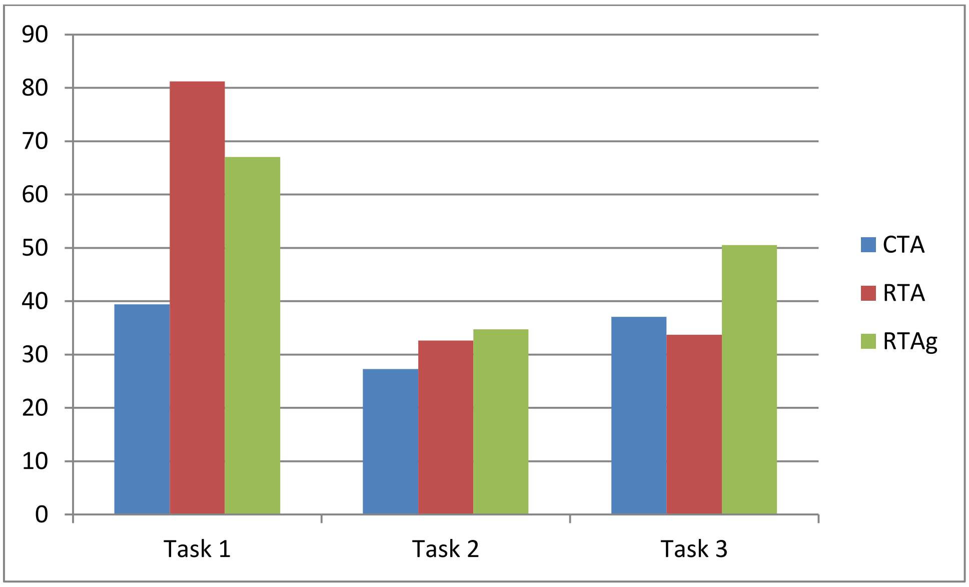


Figure 24 Average verbalization measured in words per minute

In Figure 24 Average verbalization measured in words per minute it seems like there is a tendency towards more verbalization with the retrospective methods compared to CTA in task 1, this is in agreement with the findings under usability problems. The datasets were tested for normality, and two outliers were removed from task 2. A one-way ANOVA indicates that there is a statistical significant difference in task 1: F (2, 13) = 4.77, p= 0.028. Tukey's post-hoc analysis reveals that the difference is between CTA and RTA, and that RTAg can be grouped with both CTA and RTA. The graph in Figure 24 suggests that RTAg is closer to RTA than CTA.

4.4 Participants' subjective assessment

The participants' subjective assessment of using the different methods was collected to give a richer data material. There was asked on question regarding the method used after each task. Since the methods were randomized, and all methods were not used on all tasks, the participants were not asked to compare the methods.

CTA:

To what extent does thinking aloud influence the way you solved the task? The scale goes from 0=unproblematic to 4=disturbing.

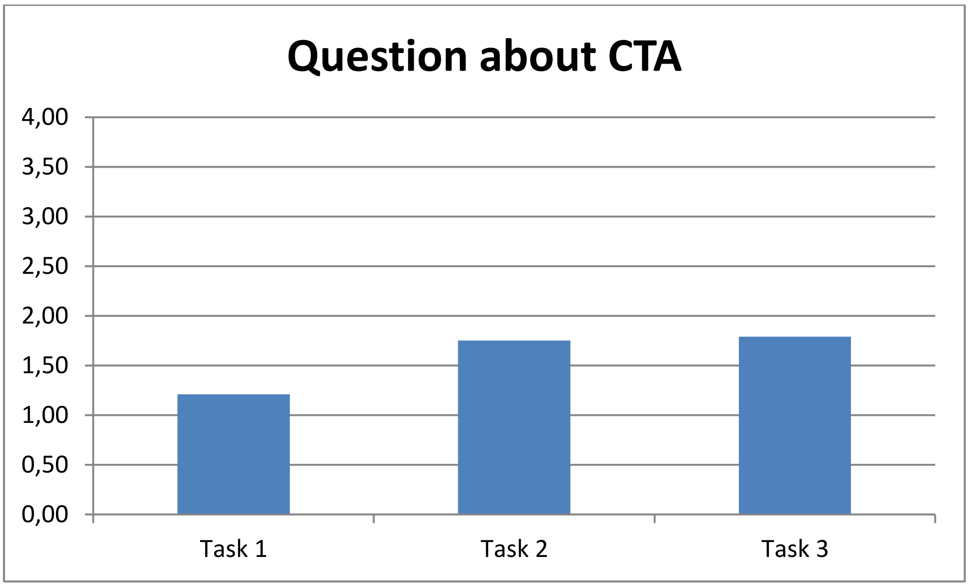


Figure 25 Response to question about CTA

As can be seen in Figure 25 the participants' subjective assessment that the act of thinking aloud during CTA does not affect their task performance to a large extent. There is a tendency towards that performing CTA is more disturbing during task 2 and 3, but the one-way ANOVA reveals no statistical significant difference.

RTA:

To what extent do you find it easy to remember what you were thinking while solving the task? The scale goes from 0=easy to 4=hard.

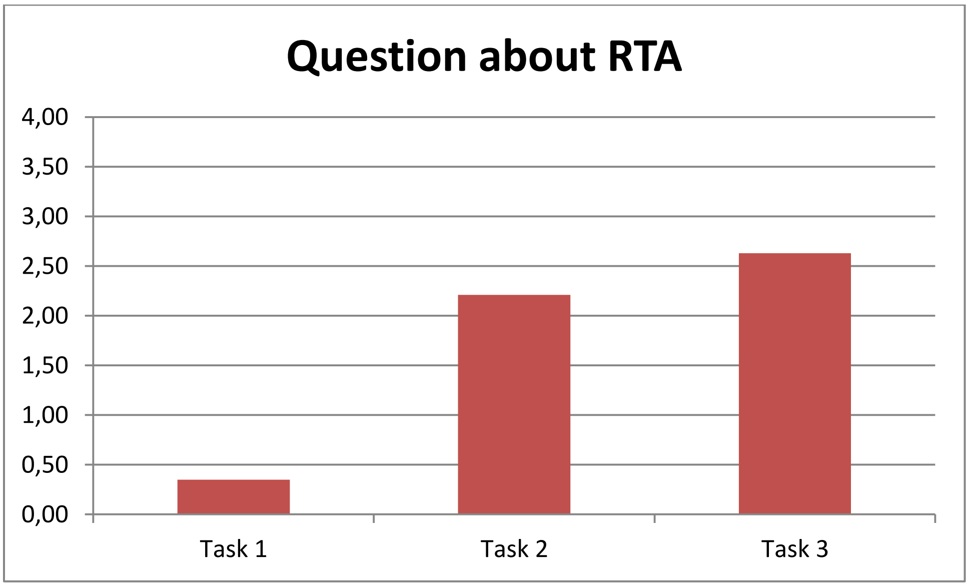


Figure 26 Response to question about RTA

The participants experienced less problems remembering what they were thinking when completing the task during task 1 than what they did for task 2 and 3. A one-way ANOVA resulted in statistical significant difference:

F (2, 14) = 16.24, p=0.000

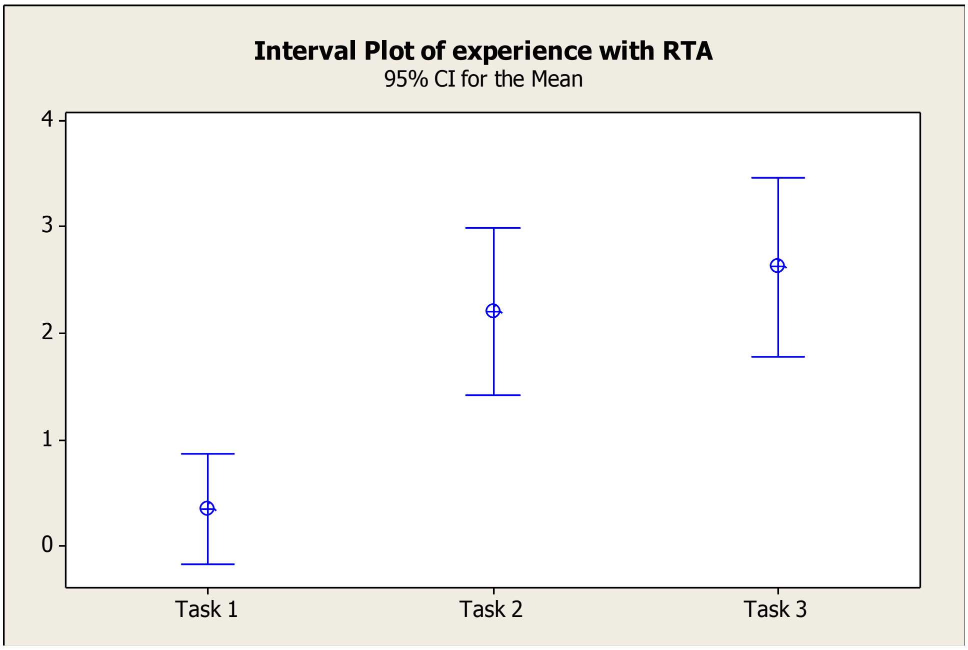


Figure 27 Interval plot for question about RTA

RTAg:

To what extent do you find it useful to see your own eye movements while verbalizing? The scale goes from 0=irrelevant to 4=very useful.

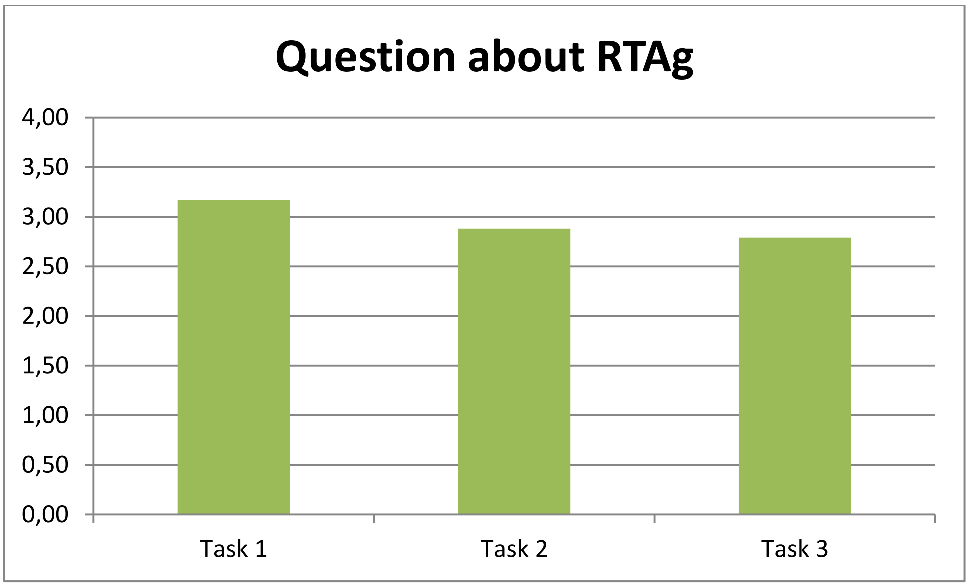


Figure 28 Response to question about RTAg

The participants show an overall positive attitude towards seeing their own eye movements for all methods.

4.5 Qualitative observations

Even though no qualitative analysis was done in this study some observations were done that may be important for explaining the other results, and recommending further research.

4.5.1 Observations during retrospective sessions

During the RTA retrospective sessions the participants were talking about what they did in more general terms, and often out of sync with was going on. In the RTAg retrospective sessions the verbalization was in sync and more specific about what they were thinking. There were a lot of comments during RTA that might indicate that verbalization were more difficult:

Selected comments from task 2:

- I was really just reading.
- I am reading for the most part.
- I am just sitting here and try to remember what I was reading.

Selected comments from task 3:

- I think I am reading.
- I do not know what I am doing now.
- Where I am reading now is difficult to say.

Comments like this was not found during task 1 for none of the TA methods, but there were occurrences of such comments with RTAg in task 2 and task 3, though not as many as with RTA.

During the RTAg retrospective sessions there were also some comments that relates to vanity, for example one participant said: “oh my god, I read so messy!” And there were some comments that may be rationalization, or it may be relevant information, for example one participant started to explain why the gaze went back and forth between two elements in the interface.

4.5.2 Observations during CTA

During CTA it was observed that the participants located the button or link they should click, but waited until finished verbalizing before actual clicking. It was found that 4 out of 6 participants had this behavior during task 1, for some of them this was quite prominent. The other two were just clicking and talking at the same time.

5 Discussion

5.1 Reactivity

A think aloud method is reactive if the act of verbalizing is influencing the task performance[3], and hence the method is not able to produce valid results. Task performance is:

- Task completion time: the time it takes for a participant to complete a task.
- Task completion rate: can be measured as how many tasks a participant completed, or how many participants completed a task.

In this study this related to research question 1, and is assessed by measuring task completion times and rates. The participants were also asked about their experience with using CTA on the different tasks, and they found it slightly more disturbing for task 2 and 3 compared to task 1.

5.1.1 Task completion time

Time completion times were measured, and the mean times were highest for CTA in all tasks, but there were no statistical significant differences. This is in agreement with the results of van den Haak et al.[8], Eger et al.[15] and Hurskykari et al. [30].

It seems to be difficult to measure task completion times. Nielsen says in his book: "It is a well-established phenomenon that there are huge individual differences between test users. It is not uncommon to find that the best user is 10 times as fast as the slowest user, and that the best 25% of the users are normally twice as fast as the slowest 25%"[5]. Borgman[31] refers to a number of studies of user performance where the ranges varies from 7:1 for text editing to 50:1 for programming tasks. 7:1 means that the fastest user is 7 times as fast as the slowest user.

If the study were designed in such a way that all participants completed one task of each task type using all TA methods the effect of the individual differences may have been reduced.

Based on this study we cannot conclude that the CTA is reactive in terms of task completion time.

5.1.2 Task completion rate

In this study the task completion rate were measured for task 1, and it was only one usability problem that resulted in unsuccessful task completion:

P1: The participant cannot see that the patient is already active.

There were 5 participants that received assistance with this problem. Two in the CTA condition, two in the RTA condition and one in the RTAg condition. These are very small numbers and any conclusions could not have been drawn on the basis of this anyway. Van den Haak et al.[8] and Eger et al.[15] did however find that CTA is reactive in terms of task completion rate. This indicates that task completion rate may be used as a measurement for the reactiveness of CTA.

5.2 Usability problems

There were no differences between the three TA methods in terms of identifying usability problems. All problems that were identified were observed and verbalized. There were a total of 27 instances of usability problems, an all of them were observed, but only 10 of the instances were verbalized. Out of the 10 verbalized instances of usability problems, 9 of them were verbalized during retrospective sessions, almost equally distributed between the two retrospective methods. This indicates that retrospective methods elicit more usability problems in terms of verbalization, which is in agreement with the results of van den Haak et al.[8] and Hyrskykari et al.[30]. Eger et al.[15] found that the RTAg method elicited more usability problems than CTA.

5.3 Insight

With the assumption that more verbalized words means more insight in the participants' thoughts, all words verbalized were counted for all tasks and TA methods. This relates to research question 3.

In RQ3a the intention was to find out if presenting the participants' gaze data during the retrospective session would elicit more verbal data while completing the reading task. It was found a tendency towards confirming that, but with no statistical significance. This tendency was more prominent in task 3, which was the complex problem solving task. When the participants were asked to give their subjective assessment on how they experienced remembering what they were thinking during the RTA retrospective session, there was a significant difference between task 1 and the two other tasks. They found it a lot easier to remember what they were thinking during task 1.

The hypothesis for RQ3a was that the video produced for using as cue for retrospective sessions during reading would be static in the sense that there was little feedback to the participant for verbalizing during RTA, and that using RTAg would make it easier to verbalize. It may seem like this also characterizes task 3. Another explanation may be that both task 2 and task 3 had longer duration than task 1.

As for task completion times, it is suspected that there are huge individual differences in verbalization as well. Some participants could report a problem with only 2-3 words, while others would use 10-15 words to describe the same phenomenon.

For a large population it may be true that more words verbalized result in more insight in the participants' thoughts, but doing a qualitative analysis would probably be more efficient. In order to gain insight the words must have meaning, so by reducing the verbalization to meaningful utterances through qualitative analysis, the result could perhaps be tested with statistical methods.

For RQ3b the hypothesis was that it would be harder for the participants to verbalize while solving a complex task, compared to an easy task. The results show that there was less verbalization during CTA for task 3 than for task one, but no statistically significant difference. The difference was more prominent between task 2 and task 1. This may be another indication of that task 2 and task 3 shares some characteristics. But this time it could not be due to task duration.

When the participants were asked to give their subjective assessment of the usefulness of seeing their own eye movements, the response were very positive for all tasks. In fact the mean score was highest for task 1. If all participants were solving all task types using all methods, this result may have been different.

Based on this study it can be concluded that the participants find it hard to remember what they were thinking during RTA for tasks that have the same characteristics as task 2 and 3. In order to draw any hard conclusions the characteristics of task 2 and 3 must be further examined, and the verbal data should be analyzed with qualitative methods.

5.4 Task types

5.4.1 Easy interactive task

The easy interactive task was designed to be the typical usability test task. The results from the participants' subjective assessment about the CTA method show a tendency towards that it is easier to perform CTA in this task compared to the two other tasks. This relates to RQ3b, and coincides with the results from the verbalization word count that show a slight tendency towards more verbalization in CTA for task 1 compared to the two other tasks. This may indicate that the CTA method is better suited for this task type than the two other task types.

The participants' subjective assessments about RTA leave no doubt. There was a statistical significant difference between task 1 and the other tasks. The participants reported that when performing RTA on task 1 it was a lot easier to remember what they were thinking compared to the two other tasks. The verbalization word count for this task was statistically significant higher during RTA than CTA. The word count was also higher for RTAg, but with no statistical significant difference. This may indicate that that if the suspected reactiveness of CTA is considered important to avoid, RTA is a good approach for this task type.

When asked to assess the RTAg method, the participants reported an overall positive response to seeing their own eye movements to help remembering what they were thinking during task execution. Surprisingly the highest score was for task 1. But the differences are small, and this is not considered a tendency. As with RTA, RTAg may be a good method to use if the suspected reactiveness of CTA is to be avoided.

5.4.2 Reading task

During CTA the task completion time was higher than with the retrospective methods. The difference was not statistically significant, but the difference was more prominent for this task compared to the two other tasks. In their subjective assessments the participants reported that performing CTA on this task influenced task performance more on this task than on task 1. This may indicate that using CTA may not be a good approach for reading tasks.

There is little difference between the methods in terms of verbalization in this task. The word count is slightly lower for RTA than for RTAg, which relates to RQ3a. In their subjective assessment the participants report it to be significantly more difficult to remember what they were thinking during the RTA retrospective session for this task than for task 1. This indicates that RTA is not the best method for this task type.

Since RTAg is better than CTA in terms of task completion time, and better than RTA in terms of word count and the participants' subjective assessments, this is probably the best method to use on reading tasks.

5.4.3 Complex problem solving task

As discussed in 5.3 there is suspected that there are similarities between task type 2 and task type 3. The analysis for the complex problem solving task will be the same as for the reading task in 5.4.2.

6 Conclusions

The purpose of this study was to examine whether some of the think aloud variants could be better suited for tasks with certain characteristics. Based on the issues discussed in the former chapter a recommendation matrix is developed (see Table 5 Method recommendation matrix).

Table 5 Method recommendation matrix

	CTA	RTA	RTAg
Easy interaction tasks	+	+	+
Reading tasks	-	-	+
Complex problem solving tasks	-	-	+

6.1.1 Why not just use CTA?

Even though this study could not conclude the CTA is reactive, some other studies show that CTA is reactive in terms of task completion rate. In order to avoid this effect, a retrospective method should be used. It is up to the practitioner or researcher to consider whether this effect is strong enough to alter the results of a study. Since there are no differences in task completion time, obviously using a retrospective method would be a lot more time consuming. On the other hand the retrospective methods produce more verbal data about usability problems.

6.1.2 RTA or RTAg?

As mentioned in the introduction, there are reasons not to play back gaze data during retrospective sessions. Stevens[16] experienced that the users were enamored by the eye tracking technology and the results it produced, and thus turned the attention away from the primary task.

When it comes to task characteristics it is recommended to use RTAg if the tasks contain little interactivity, and/or are lengthy.

6.1.3 Limitations of this project

6.1.3.1 Usability problems

In order to conclude on the reactiveness of CTA in terms of task completion rate, there should have been more data on usability problems, hence more tasks of type 1: Easy interactive tasks. Another limitation with regards to usability problems is that there was only one person that did all the analysis. With two or more researchers the list of usability problems might have been different.

6.1.3.2 Task characteristics

When the results were analyzed it emerged that there was no clear difference between task type 2 and task type 3. It seems like these two task types share some characteristics, but it could not be pinpointed exactly what this likeness implies.

6.1.3.3 Insight

The assumption that more words verbalized means more insight may be true for a large population. In this study the population may have been too small, or the hypothesis that the participants will verbalize more during the RTAg retrospective session on the reading task compared to RTA, is wrong.

6.2 Further research

As mentioned above it is suspected that there are huge individual differences in how the participants verbalize. It would have been interesting to develop a method for processing the verbal data into comparable datasets, where all unnecessary words are removed, and each sentence could be transformed into statements, perhaps in different categories.

It was observed that the participants had different verbalization style during RTA and RTAg retrospective sessions. In RTA they were talking about what they did in more general terms, and often out of sync with what was happening in the video. In RTAg there were elements of vanity and rationalization. A qualitative analysis of these differences could give deeper insight into what to expect from using one method or the other.

A deeper understanding of what characterizes different task would have been helpful in this study. The tasks that were designed had different levels of interaction and complexity. Further investigation of task characteristics along these axes, and perhaps other axes, could shed some light upon which characteristics are important when deciding which think aloud method to use.

A qualitative analysis of the find-wait-click phenomenon that was observed during CTA could perhaps provide contributions to the CTA reactiveness debate.

7 References

[1] C. Lewis and J. Rieman, *Task Centered User Interface Design : A practicle introduction TASK-CENTERED USER INTERFACE DESIGN A Practical Introduction TASK-CENTERED USER INTERFACE DESIGN A Practical Introduction Foreword :* 1993.

[2] K. A. Ericsson and H. A. Simon, *Protocol analysis: Verbal reports as data (Rev. ed.)*. Cambridge, Ma: MIt Press, 1993.

[3] J. E. Russo, E. J. Johnson, and D. L. Stephens, "The validity of verbal protocols.," *Memory & cognition*, vol. 17, no. 6, pp. 759–69, Nov. 1989.

[4] T. Boren and J. Ramey, "Thinking aloud: reconciling theory and practice," *IEEE Transactions on Professional Communication*, vol. 43, no. 3, pp. 261–278, 2000.

[5] J. Nielsen, *Usability Engineering*. Morgan Kaufmann Publishers Inc., 1993.

[6] J. Nielsen, "Thinking Aloud: The #1 Usability Tool (Jakob Nielsen's Alertbox)." [Online]. Available: http://www.useit.com/alertbox/thinking-aloud-tests.html. [Accessed: 11-Feb-2012].

[7] J. Nielsen, "The Usabilitv Engineering Life Cycle," *Computer*, 1992.

[8] M. Van Den Haak, M. De Jong, and P. J. Schellens, "Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue," *Behaviour & Information Technology*, vol. 22, no. SEPTEMBER–OCTOBER 2003, NO. 5, pp. 339–351, Sep. 2003.

[9] Z. Guan, S. Lee, E. Cuddihy, and J. Ramey, "The Validity of the Stimulated Retrospective Think-Aloud Method as Measured by Eye Tracking," *Methods*, pp. 1253–1262, 2006.

[10] S. Elling and L. Lentz, "Retrospective Think-Aloud Method : Using Eye Movements as an Extra Cue for Participants ' Verbalizations," in *CHI 2011*, 2011, pp. 1161–1170.

[11] B. Kim, Y. Dong, S. Kim, and K.-P. Lee, "Development of integrated analysis system and tool of perception, recognition, and behavior for web usability test: with emphasis on eye-tracking, mouse-tracking, and retrospective think aloud," pp. 113–121, Jul. 2007.

[12] A. Olsen, L. Smolentzov, and T. Strandvall, "Comparing different eye tracking cues when using theretrospective think aloud method in usability testing," in *BCS '10 Proceedings of the 24th BCS Interaction Specialist Group Conference*, 2010, pp. 45–53.

[13] "Eye tracking, usability testing, user centred design - from Bunnyfoot UK." [Online]. Available: http://www.bunnyfoot.com/. [Accessed: 12-Feb-2012].

[14] L. Ball, N. Eger, R. Stevens, and J. Dodd, "Applying the PEEP method in usability testing," *Interfaces 67*, 2006.

[15] N. Eger, L. J. Ball, R. Stevens, and J. Dodd, “Cueing Retrospective Verbal Reports in Usability Testing Through Eye-Movement Replay,” in *HCI 2007*, 2007, pp. 129–137.

[16] R. Stevens, “Why I’m Not in Love with Eye Tracking | Think Eye Tracking.” [Online]. Available: http://thinkeyetracking.com/2009/11/why-i’m-not-in-love-with-eye-tracking/. [Accessed: 12-Feb-2012].

[17] C. Fossen, “Norsk senter for elektronisk pasientjournal.”

[18] “INTERNATIONAL Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11 : Guidance on usability,” 1998.

[19] J. Nielsen, *Usability Engineering, Kindle edition*. Morgan Kaufmann Publishers Inc., 1993.

[20] J. Brooke, “SUS - A quick and dirty usability scale,” *Usability Evaluation in Industry*, 1996.

[21] N. Bevan, “Common Industry Format Usability Tests,” in *Proceedings of UPA’98*, 1999, pp. 0–5.

[22] “National Institute of Standards and Technology.” [Online]. Available: http://www.nist.gov/index.html. [Accessed: 18-Dec-2012].

[23] “Norsk senter for elektronisk pasientjournal.”

[24] “Evicare.” [Online]. Available: http://www.evicare.no/. [Accessed: 20-Dec-2012].

[25] “ISO 13407:1999 - Human-centred design processes for interactive systems.” [Online]. Available: http://www.iso.org/iso/catalogue_detail.htm?csnumber=21197. [Accessed: 18-Dec-2012].

[26] “Forord og innledning - Nasjonale retningslinjer for behandling og rehabilitering ved hjerneslag.” [Online]. Available: http://www.helsebiblioteket.no/Retningslinjer/Hjerneslag/Forord-og-innledning. [Accessed: 16-Dec-2012].

[27] “4.5 Lipidsenkende behandling - Nasjonale retningslinjer for behandling og rehabilitering ved hjerneslag.” [Online]. Available: http://www.helsebiblioteket.no/retningslinjer/hjerneslag/Sekundarforebygging/Lipidsenkende-behandling. [Accessed: 16-Dec-2012].

[28] “6.3 Diagnostiske grenser - Nasjonale faglige retningslinjer: Diabetes - Forebygging, diagnostikk og behandling.” [Online]. Available: http://www.helsebiblioteket.no/retningslinjer/diabetes/6.diagnostikk/6.3-diagnostiske-grenser. [Accessed: 16-Dec-2012].

[29] “Transana.” [Online]. Available: http://www.transana.org/. [Accessed: 06-Dec-2012].

[30] A. Hyrskykari, S. Ovaska, and T. M. Lehtinen, “Gaze Path Stimulation in Retrospective Think-Aloud,” *Journal of Eye Movement Research*, vol. 2, no. 4, pp. 1–18, 2008.

[31] C. L. Borgman, “ALL USERS OF INFORMATION RETRIEVAL SYSTEMS ARE NOT CREATED EQUAL : AN EXPLORATION INTO INDIVIDUAL,” *Information processing & Management*, vol. 25, no. 3, pp. 237–251, 1989.

Appendix A – Php script

```
<?php

        if ( strpos($_GET['patientId'], '1000272' ) !== FALSE ) {

                header ('Location:
http://www.helsebiblioteket.no/Retningslinjer/Hjerneslag/Forord-og-innledning');

                exit;

        }

        if ( strpos($_GET['patientId'], '1000270' ) !== FALSE ) {

                header ('Location:
http://www.helsebiblioteket.no/retningslinjer/hjerneslag/Sekundarforebygging/Lipidsenkende
-behandling');

                exit;

        }

        if ( strpos($_GET['patientId'], '1000218' ) !== FALSE ) {

                header ('Location:
http://www.helsebiblioteket.no/retningslinjer/diabetes/6.diagnostikk/6.3-diagnostiske-
grenser');

                exit;

        }

?>
```

Appendix B – Consent form

Samtykkeerklæring for studie om kliniske retningsliner

Jeg har mottatt skriftlig og muntlig informasjon, og fått anledning til å stille spørsmål. Jeg er klar over at det er frivillig å delta, og at jeg kan trekke meg når som helst uten å oppgi noen grunn. Jeg er klar over at det vil bli gjort lyd- og videoopptak av studien og at dette materialet vil behandles konfidensielt, og slettet etter at analyse er gjort. Jeg samtykker i å delta i studien.

Godtar du at stillbilder fra opptaket kan brukes i masteroppgaven?

Ja □   Nei □

Godtar du at videosekvenser fra opptaket kan brukes i presentasjonen av masteroppgaven?

Ja □   Nei □

Ditt navn:____

Sted/Dato:____

Signatur:____

Appendix C – Background and demographics form

Bakgrunnsopplysninger

Dette skjemaet brukes for å registrere bakgrunnsinformasjon. Opplysningene vil holdes anonymt. Takk for din deltakelse.

1. Kjønn

☐ Mann ☐ Kvinne

2. Alder

____

3. Hvilket årstrinn går du på?

☐ Fjerde ☐ Femte ☐ Sjette ☐ Annet

4. Fra hvilke(n) avdeling(er) i sykehus har du mest erfaring?

____

5. Hvor ofte bruker du kliniske retningslinjer i studiene eller praksis?

☐ Sjelden ☐ Av og til ☐ Ofte ☐ Aldri

6. Foretrekker du elektroniske eller papirbaserte retningslinjer?

☐ Papirbasert ☐ Elektronisk ☐ Likegyldig

7. Bruker du helsebiblioteket.no som kilde for kliniske retningslinjer?

☐ Sjelden ☐ Av og til ☐ Ofte ☐ Aldri

Appendix D – Questionnaire

På hvilken måte synes du at det å tenke høyt påvirket måten du løste oppgaven på?

|----------------------------|----------------------------|----------------------------|----------------------------|

Uproblematisk Forstyrrende

I hvilken grad føler du at du husket det du tenkte da du løste oppgaven?

|----------------------------|----------------------------|----------------------------|----------------------------|

Lett å huske Vanskelig å huske

I hvilken grad synes du det var nyttig å se dine egne øyebevegelser I forhold til å huske det du tenkte da du løste oppgaven?

|----------------------------|----------------------------|----------------------------|----------------------------|

Svært nyttig Helt irrelevant

Appendix E – Raw data

Task completion times

CTA			
Participant	Task	Time	Time [s]
1	1	00:02:09	129
2	1	00:02:02	122
7	1	00:04:35	275
8	1	00:01:12	72
13	1	00:04:10	250
14	1	00:02:12	132
5	2	00:19:50	1190
11	2	00:07:45	465
15	2	00:06:35	395
18	2	00:05:19	319
19	2	00:08:14	494
4	3	00:06:55	415
6	3	00:05:35	335
10	3	00:04:46	286
12	3	00:03:16	196
16	3	00:03:39	219
17	3	00:07:06	426

RTA			
Participant	Task	Time	Time[s]
4	1	00:01:05	65
10	1	00:00:50	50
16	1	00:02:27	147
19	1	00:02:06	126
1	2	00:09:00	540
6	2	00:12:10	730
7	2	00:09:43	583
12	2	00:04:15	255
13	2	00:05:39	339
17	2	00:07:50	470
2	3	00:05:32	332
5	3	00:08:04	484
8	3	00:02:07	127
11	3	00:12:40	760
14	3	00:06:13	373
18	3	00:02:46	166

RTAg			
Participant	Task	Time	Time [s]
5	1	00:03:04	184
6	1	00:02:53	173
11	1	00:01:08	68
12	1	00:01:01	61
17	1	00:01:01	61
18	1	00:04:18	258
2	2	00:07:26	446
4	2	00:04:29	269
8	2	00:06:00	360
10	2	00:09:05	545
14	2	00:04:16	256
16	2	00:08:57	537
7	3	00:05:19	319
13	3	00:03:04	184
15	3	00:03:13	193
19	3	00:04:37	277

Verbalization word count

CTA					
Participant	Task	Time	Time[s]	WC Total	Words/min
1	1	00:06:10	370	123	19,95
2	1	00:02:02	122	103	50,66
7	1	00:04:40	280	100	21,43
8	1	00:01:11	71	49	41,41
13	1	00:04:10	250	134	32,16
14	1	00:02:12	132	156	70,91
5	2	00:19:50	1190	979	49,36
11	2	00:07:45	465	174	22,45
15	2	00:06:35	395	74	11,24
18	2	00:05:19	319	215	40,44
19	2	00:08:14	494	106	12,87
4	3	00:06:55	415	248	35,86
6	3	00:05:35	335	77	13,79
10	3	00:04:46	286	391	82,03
12	3	00:03:16	196	116	35,51
16	3	00:03:39	219	141	38,63
17	3	00:07:00	420	116	16,57

RTA					
Participant	Task	Time	Time[s]	WC Total	Words/min
4	1	00:01:09	69	138	120,00
10	1	00:01:47	107	144	80,75
16	1	00:02:28	148	167	67,70
19	1	00:01:54	114	107	56,32
1	2	00:09:30	570	295	31,05
6	2	00:12:15	735	244	19,92
7	2	00:10:30	630	214	20,38
12	2	00:04:16	256	366	85,78
13	2	00:05:45	345	109	18,96
17	2	00:07:53	473	154	19,53
2	3	00:06:25	385	368	57,35
5	3	00:10:45	645	228	21,21
8	3	00:02:42	162	54	20,00
11	3	00:13:52	832	307	22,14
14	3	00:06:33	393	194	29,62
18	3	00:03:00	180	75	51,75

RTAg					
Participant	Task	Time	Time[s]	WC Total	Words/min
5	1	00:02:43	163	160	58,90
6	1	00:02:40	160	138	51,75
11	1	00:01:08	68	61	53,82
12	1	00:02:01	121	215	106,61
17	1	00:01:12	72	87	72,50
18	1	00:04:03	243	238	58,77
2	2	00:07:16	436	279	38,39
4	2	00:04:30	270	248	55,11
8	2	00:06:05	365	85	13,97
10	2	00:09:15	555	254	27,46
14	2	00:04:21	261	240	55,17
16	2	00:08:57	537	163	18,21
7	3	00:06:55	415	248	35,86
13	3	00:03:03	183	174	57,05
15	3	00:04:42	282	402	85,53
19	3	00:05:50	350	138	23,66

Participants' subjective assessment

Participant	Question - CTA	Task 1	Task 2	Task 3
1	How did thinking aloud interfere with your task performance?	0,75		
2	*0=unproblematic, 4=disturbing*	2,00		
4				1,00
5			0,75	
6				0,75
7		1,00		
8		0,00		
9			2,00	
10				2,00
11			3,00	
12				2,50
13		2,50		
14		1,00		
15			0,75	
16				1,00
17				3,50
18			0,00	
19			1,00	
	Mean	1,21	1,30	1,79
	Standard deviation	0,82	0,98	0,98

Question - RTA	Task 1	Task 2	Task 3
How did you experience remembering what you were thinking while solving the task?		3,00	
0 = easy to remember, 4 = hard to remember			3,00
	0,00		
			3,00
		1,00	
		3,00	
			1,00
	0,25		
	1,00		
			3,00
		2,25	
		2,00	
			2,75
	0,50		
		2,00	
			3,00
	0,00		
Mean	0,35	2,21	2,63
Standard deviation	0,37	0,68	0,73

Question RTAg	Task 1	Task 2	Task 3
To what extent did you find it useful to see your own eye movements while solving the task?			1,00
0 = totally irrelevant, 4 = very useful		2,00	
		4,00	
	3,25		
	3,25		
			4,00
		4,00	
			2,75
		3,00	
	3,00		
	3,00		
			3,00
		2,75	
			3,00
		1,50	
	2,50		
	4,00		
			3,00
Mean	3,17	2,88	2,79
Standard deviation	0,45	0,93	0,89