

Grant Agreement Number: 257528

KHRESMOI

www.khresmoi.eu

**Report on the results of the initial user test of the
radiology search system**

Deliverable number	<i>D10.2</i>
Dissemination level	<i>Public</i>
Delivery date	<i>due 30.4.2013</i>
Status	<i>Final</i>
Authors	<i>Dimitrios Markonis, Markus Holzer, Frederic Baroz, Rafael Luis Ruiz De Castaneda, Georg Langs, Celia Boyer, Henning Mueller</i>



This project is supported by the European Commission under the Information and Communication Technologies (ICT) Theme of the 7th Framework Programme for Research and Technological Development.

Executive Summary

This deliverable describes the first round user-oriented evaluation of the KHRESMOI system prototype by radiologists. The design aspects, the setup details and the results are presented and analysed. An iterative approach was followed, using pilot user tests on the 2D image search prototype and interviews using mockups of the 3D image search system interface. A round of full user tests followed in the University hospitals of Geneva and the Medical University of Vienna.

In total, 17 persons participated in the user test. The outcomes of the pilot studies were used to refine the user study design and the system's development. The participants performed guided information retrieval tasks and answered questionnaires on the interaction with the system. Usability aspects, such as effectiveness, efficiency and user satisfaction were recorded and the results were analysed and are presented in this document.

The results indicate that the KHRESMOI system has tools that have the potential to improve the current visual information search success rate and were found to be novel and useful in practice. User satisfaction was positive by the majority of the participants and a large amount of feedback was given. Aspects where the system was less satisfactory to the users were identified and propositions were made for improvements and additions. The outcomes of this evaluation will be used to guide future system development.

Table of Contents

1	Introduction	5
2	Methods and Materials	6
2.1	Interviews and interface design	6
2.2	User study protocol	7
2.3	Tasks and data sets	9
2.3.1	2D image/article search	9
2.3.2	3D image search	9
2.4	Recording material	10
2.5	Survey forms	12
3	Results	12
3.1	2D Image/article search subsystem	14
3.1.1	Pilot user study	14
3.1.2	Full user tests	17
3.2	3D Image search subsystem	20
3.2.1	Interviews	21
3.2.2	Full user tests	22
4	Discussion	24
4.1	2D Image/article search subsystem	24
4.1.1	Pilot user study	24
4.1.2	Full user tests	25
4.2	3D Image search subsystem	26
4.2.1	Interviews	26
4.2.2	Full user tests	27
5	Conclusion	28
6	References	29
	Appendices	31
A	DEMOGRAPHIC SURVEY	31
B	USABILITY SURVEY	33
C	TASKS	37
D	INVESTIGATOR INSTRUCTIONS	40
E	CONSENT FORM	43

List of Figures

Fig.1	3D Mockups generated at MUW, and used during the interviews. (A) Query case and grid view on query results. (B) Query case and list view on the query results. (C) Query case, chosen case and given grid view on results. (D) In detail view on query case and one result case. (E) Same as <i>D</i> with Report displayed. (F) Same as <i>C</i> , but with visualized query ROI and ROIs in the result view that match the query.	7
Fig.2	The interface of the Morae Recorder software, which was installed to the participant's computer. A common study configuration file should be created for all the users and the recording starts by pressing red button.	11
Fig.3	The interface of the Morae Observer software, which was installed to the observer's computer. This tool allows the observer to take notes on the timeline of the recording and use different markers (e.g. found bug, participant comment, participant blocked etc.) The program in the screenshot is waiting for the recording to begin.	11
Fig.4	Screenshot of filling in a survey form using the Morae software.	12
Fig.5	Median values of measuring general user satisfaction about the system in Likert scale (1=strongly negative, 5=strongly positive).	13
Fig.6	Mode values for each participant over the global satisfaction questions in a Likert scale.	13
Fig.7	Mode frequency for each participant over the global satisfaction questions.	14
Fig.8	Pilot study: Median of measuring user satisfaction over specific system aspects in a Likert scale (1=strongly negative, 5=strongly positive).	15
Fig.9	Pilot study: Median values of measuring general user satisfaction about the system in Likert scale.	15
Fig.10	Pilot study: Mode values for each participant over the global satisfaction questions in a Likert scale.	16
Fig.11	Pilot study: Mode frequency for each participant over the global satisfaction questions.	17
Fig.12	Median of measuring user satisfaction over specific 2D image search prototype aspects in a Likert scale (1=strongly negative, 5=strongly positive).	18
Fig.13	The 3D prototype used for the user tests. The left pane is the query volume with an example ROI, the right pane the result pane with the thumbnails and the center pane the detail view with the report and overlaid matching ROIs.	22
Fig.14	Median of measuring user satisfaction over specific 3D image search prototype aspects in a Likert scale (1=strongly negative, 5=strongly positive).	23

Abbreviations

CBIR	Content-based image retrieval
UCD	User-centered design
SUS	System Usability Scale
QUIS	Questionnaire for User Interaction Satisfaction
ROI	Region of Interest
PACS	Picture archiving and communication system

1 Introduction

Part of this section has been accepted for publication in [14].

Images are produced in a quickly increasing variety and quantity and are essential in many aspects of medical diagnosis and treatment planning. Much of the knowledge stored in images is little exploited as access to the visual image information is rarely possible. Content-based image retrieval (CBIR) uses the visual content of a set of images as positive or negative examples to retrieve other images or cases. Over the past 15 years, CBIR has been considered promising for assisting information search in the medical field and several systems were developed [19]. However, most systems were rather technology-driven and very few applications have reached the end users or were integrated into the medical professionals workflow [1].

User-centered design (UCD) [25] has been used in industry [8, 11], but also in medical applications [5]. It is driven by the user requirements and feedback to improve the product's usability and the user experience. A few aspects of UCD have also been used for CBIR [6].

UCD in software development includes key elements in order to involve user feedback to the design and the development of the application. First, investigation and understanding of the user requirements [7, 24] needs to be achieved to identify the general design directions. User-centered evaluation is another important part of UCD, which needs to be performed in the early stages of the development [9] and is seen as an iterative process throughout the development cycle [11, 5]. The key elements are also described in the ISO standard for the Human-centered design for interactive systems (ISO 9241-210, 2010).

User-centered evaluation is often performed in the form of empirical usability tests, which include having a number of target users to interact with the system. Usability of the system is assessed with factors such as learnability, efficiency, effectiveness, memorability and satisfaction [9]. Various methods exist for conducting these tests, including thinking aloud, direct or recorded observation of the interaction, survey forms and log analysis. A survey on common usability testing techniques and tools is presented in [2]. A more detailed description of aspects to be taken into account when designing a usability test can be found in [12].

An important aspect when designing a usability test is the number of participants required. Early studies have discovered that a single person is not able to detect all usability problems but 3-4 are sufficient [22]. In [21] it is suggested that five users are enough, while studies have questioned this choice [23, 26]. The exact number of participants remains an open question, though in [20] it is explained that five participants are indeed enough for each iteration of an iterative user-centered evaluation.

In this report, the design choices, the setup and the results of the first round of the user-centered evaluation of the KHRESMOI search engine for the radiology use case are presented. The system combines text and CBIR search to retrieve images and articles from internal and external data sources. The general research questions that the evaluation tries to answer are:

- Does the KHRESMOI system improve current search for information in radiology (which is mainly patient-centered or using Google on the Internet)?
- Does it cover unmet information needs and to what extent?
- Which functionalities are more useful and which tools need to be improved, changed or added?

2 Methods and Materials

Part of this section has been accepted for publication in [14].

This section describes the methodology followed for designing, setting up and running the user tests. The datasets used for performing the tasks and the materials used for capturing the user interaction with the system are presented, as well as the acquired feedback. With regard to the user profiles and their information needs, described in the KHRESMOI deliverable report D9.1 [17], two parts of the KHRESMOI system were used: the 2D image/article search and the 3D image search subsystems.

The user oriented evaluation process followed an iterative approach and different preliminary steps were taken for each system. The 3D image search system was designed to mainly provide information search in the hospital image data during the clinical duty. Before running the user tests, interviews with radiologists were arranged in order to design an interface that would fit in the clinical workflow.

The 2D image/article search system was designed to cover information needs related to search into external sources. Therefore an interface inspired by the state-of-the-art medical image and literature search engines interfaces [15, 10] served as a basis for the system prototype. Pilot user tests were performed to evaluate the basic aspects of the interface and the system's functionalities. This also helped detecting shortcomings of the user tests design and refining the study protocol.

2.1 Interviews and interface design

The goal of the KHRESMOI retrieval system is to add content based retrieval functionality to the clinical radiologist workflow. We do not aim at replacing existing image management systems, but instead want to offer additional functionality. In the beginning we were interested in the workflow radiologists perform and during which they could imagine that KHRESMOI functionality would be helpful. A few draft layouts for the user interface as shown in Figure 1 were presented to the physicians that would allow the radiologist (1) to work with a query image (typically this is the image the radiologist is analyzing at the moment) and (2) to browse through retrieval results. Retrieval results are images that are identified by the retrieval engine, and share information such as pathological features, local appearance, and anatomical location with the query image.

The goal of the interviews was to get insights into the various ways radiologists can imagine integrating the tool into their workflow and what specific functionality would be necessary to do this. Optimally interviews should be a conversation, in which the needs of the radiologists and the know-how about what is possible should be discussed. The information extracted from the first interviews is discussed in Section 3.2.1.

Based on the output of the interviews the user interface for the first 3D image search prototype was designed. This prototype was applied for the user tests, where the users have to complete specific tasks. The tasks were designed in a way that makes it possible to extract relevant information about the usability, quality and improvement options for the prototype.

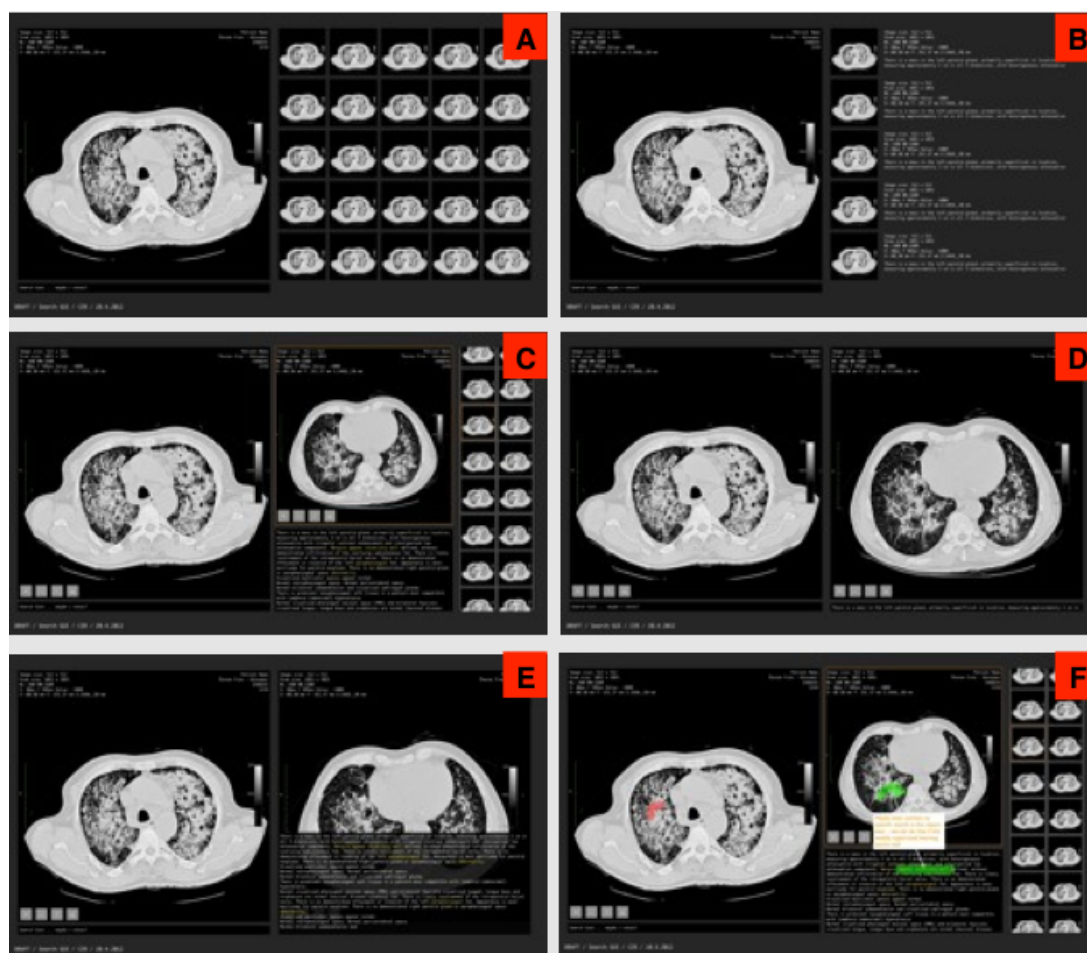


Figure 1: 3D Mockups generated at MUW, and used during the interviews. (A) Query case and grid view on query results. (B) Query case and list view on the query results. (C) Query case, chosen case and given grid view on results. (D) In detail view on query case and one result case. (E) Same as *D* with Report displayed. (F) Same as *C*, but with visualized query ROI and ROIs in the result view that match the query.

2.2 User study protocol

In order to investigate the research questions described in the introduction section, the following aspects were taken into account:

1. Success of information finding by radiologists using KHRESMOI.
2. Time to find relevant information using KHRESMOI.
3. User satisfaction of the KHRESMOI system performance.
4. Usability of the KHRESMOI system.
5. Missing useful functionalities in the current system.

In this user study, the methods of the above mentioned evaluation aspects needed to be decided. The final selection of methods, after being refined by the preliminary step of pilot user tests, is presented below:

- Participants were asked to perform information retrieval tasks for which at least one of the results is known. Therefore aspect no.1 could be evaluated.
- The time taken to fulfill each task was measured. For tasks whose time was fixed, the time taken to find the first relevant result was measured, instead. This method evaluated aspect no.2.
- Participants were asked to fill a questionnaire about their experience of using the system. This allowed to evaluate user satisfaction (aspect no.3) and detect usability problems found by the participants. Questions were included that requested feedback and propositions for system improvement (aspect no.4).
- Participants were observed and video recorded while using the system. Possible system flaws or usability problems that were not consciously detected by participants were identified through this technique (aspect no.4).

The user tests were conducted in the format of one-to-one sessions, one participant performing the tasks and one observer to facilitate the user test. The details of the session were also refined after the pilot tests by including and removing tasks, as well as modifying the time limitations. The final session outline is presented below:

1. Introduction to the KHRESMOI project, the existing search system and the user test goals (5 minutes).
2. Tutorial video on the system tools and functionalities (5 minutes).
3. Demographic survey (5 minutes).
4. Introductory task, simple use of the tools (5 minutes).
5. Guided user tests in clear scenarios (30-40 minutes).
6. Survey on the satisfaction with the tools and functionalities (10 minutes).
7. Free possibility to use the system (5+ minutes).
8. Survey on the satisfaction with the system, free discussion (10 minutes).

The introduction by the test facilitator intended to help the participant understand the concept of the system and motivate to do the test. Then, the video demonstration of the system introduced the tools offered by the application. The introductory task was introduced after the pilot user tests because the video tutorial alone did not contain enough information for the user to get familiar with the tools available. Throughout the session, the participant was being test facilitator by the observer to identify potential shortcomings of the system. The observer was

instructed to have a neutral attitude and was allowed to help only when the participant was blocked and could not proceed with a task.

The setup of the session included hardware and software preparation but also training sessions of the observer to get familiar with the recording tool and the study purpose. The hardware used in each session included two Windows computers one for the participant and one for the observer. The KHRESMOI client was downloaded to the participant's computer and the recording software was installed on both computers.

At the end of each session the file containing the recordings, the answers to the surveys and the observer's notes were acquired. The details of preparing, setting up and running a session were added into a document to ensure that the experiment can be reproduced under the same conditions. This document of instructions can be found in Appendix D.

2.3 Tasks and data sets

As mentioned in section 2.2, the user was requested to perform several information seeking tasks during the session. The design of the tasks took into account that they need to use most of the system tools and functionalities and cover the information needs of the target user group. They had to describe realistic scenarios that appear in clinical and academic workflows. Depending on the tasks and the subsystem used, different data sources were required. The tasks and data sets used are described in sections 2.3.1 and 2.3.2.

2.3.1 2D image/article search

For the evaluation of the KHRESMOI system in terms of radiology-related information search into external sources, the ImageCLEF2012¹ medical data set was used [18]. This data set contains more than 75,000 articles from PubMed open access journals and more than 300,000 images that are figures included in these articles. It represents a relatively realistic source for a medical literature search and especially for an initial test on the system's scalability and performance. More information about this data set can be found in KHRESMOI deliverable document D2.3 [13].

Two groups of information retrieval tasks were used: Three 2D image search tasks and two article search tasks. A subset of the ImageCLEF2012 medical image-based and case-based retrieval task topics was used respectively. The topics for the image-based task were selected after the log analysis of queries to a radiology image search engine [24], while case-based topics consisted of cases included in an educational database [18]. The full task descriptions can be found in Appendix C.

The guided scenarios of the user tests were based on these information retrieval tasks and included use of the various tool of the system, such as query-by-text, query-by-image-example, the personal library, the tray and others.

2.3.2 3D image search

The total anatomical dataset for the 3D image search prototype consisted of a total of 7936 MRI and CT Volumes with a total size of 470GB including 5817 radiological descriptions as reports.

¹<http://www.imageclef.org/>

A subset of 117 Lung CT volumes was used for the pathology dataset. The labeled pathologies for these volumes were PE (panlobular emphysema) and ZE (centrilobular emphysema).

The tasks were defined in order to answer distinct questions that allow quantifying the quality and features of the prototype.

1. How do the new features and functionalities added with the prototype compare to the standard tools of radiology. More exactly, how the addition of searching for similar images compares to only using tools like scrolling and changing of brightness and contrast.
2. This question is about the quality of the retrieval. Does it actually retrieve the desired pathologies and therefore help the radiologists in answering the question of finding similar diseases in the database. The radiologists can also use the radiological report of the found volumes.
3. Another question concerns the educational aspect of the prototype. The goal is to find out if it is possible for radiologists to find good examples for certain pathologies using the new features. This would allow them to create better cases and visual examples for demonstration purpose.
4. The quality of the anatomy retrieval is also an important part of the prototype. Therefore a question is designed where radiologists are asked to find similar anatomical regions based on some random examples.
5. Finally the participants should have time to freely use the system. This could give some new perspectives and additional information that could be used for further development of the prototype.

For the first user tests the goal was to decide whether the way the prototype is designed would be useful for their daily routine. This includes the design of the user-interface and the representation of the query image, the result list as well as the detail view. Therefore only a subset of the tasks was necessary in the beginning, which also reduced the time taken for each user-test. For this purpose the chosen tasks were (1), (2) and (5). The final tasks are described in detail in Appendix C.

2.4 Recording material

For observation and recording, the commercial software Morae mentioned in [2] was used. This software allows screen and face video recording of the participants (Figure 2) and also remote online observing on a different computer (Figure 3). Moreover it facilitates the conduction of the user tests by displaying the introduction and task description text as well as surveys on the participant's computer screen (Figure 4). All the surveys' answers, observer's notes and recordings are saved in a digital format which is compatible with commonly used statistical packages for result analysis and presentation.

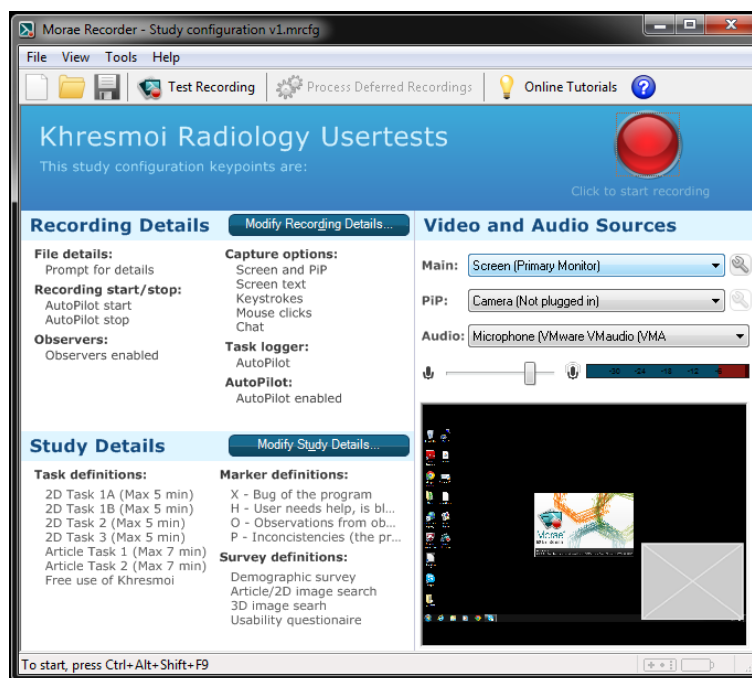


Figure 2: The interface of the Morae Recorder software, which was installed to the participant's computer. A common study configuration file should be created for all the users and the recording starts by pressing red button.

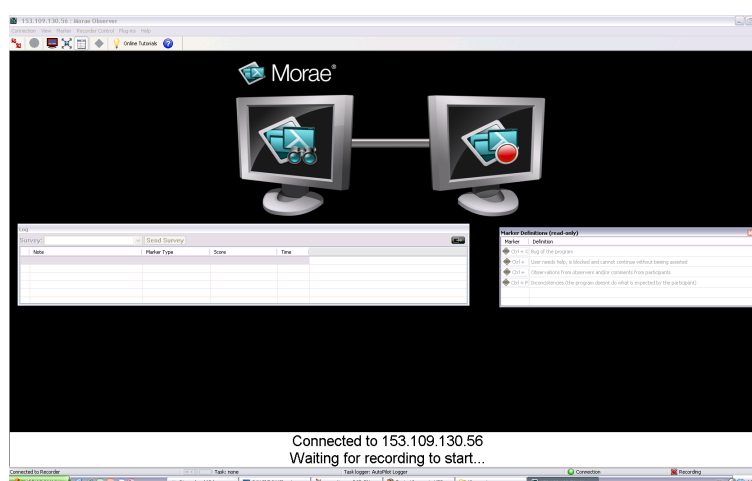


Figure 3: The interface of the Morae Observer software, which was installed to the observer's computer. This tool allows the observer to take notes on the timeline of the recording and use different markers (e.g. found bug, participant comment, participant blocked etc.) The program in the screenshot is waiting for the recording to begin.

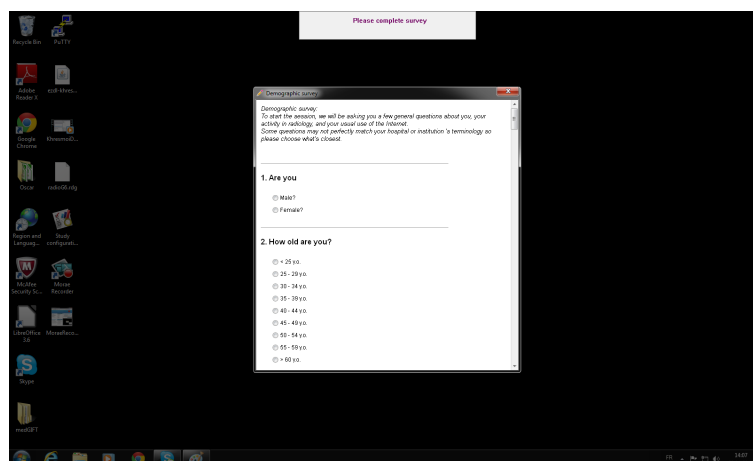


Figure 4: Screenshot of filling in a survey form using the Morae software.

2.5 Survey forms

Four survey forms were used in this study. The initial demographics survey form was used to get information on medical experience and computer use of the participants. Two survey forms were used to evaluate the subsystems' tools and functionalities usability and one to evaluate user satisfaction with the global system.

A combination of modified versions of the System Usability Scale (SUS) [3] and the Questionnaire for User Interaction Satisfaction (QUIS) [4] was used for the user satisfaction and usability survey forms. Open questions for providing comments on specific aspects of the system and suggestions for improvements were added. To get preliminary answers to the research goals, questions about the novelty, usefulness and intention of use of the tools were included. The final survey forms, after the refinement during the pilot user tests, can be found at the Appendices A and B.

3 Results

Part of this section has been accepted for publication in [14].

The user tests took place at the University hospitals of Geneva and the Medical university of Vienna. Twelve persons (3 females, 9 males) participated in the full user tests round. This number does not include the participants in the pilot user study and the interviews. They were all below 40 years old, with eight of them being below 30 and three below 35.

Five persons were interns, four were residents, one associate professor in radiology, one attending and one with no radiology background. Among the radiology specializations (participants could choose more than one field) the most common was thorax (3), radio oncology (3) and bone(2) while other chosen fields were echocardiography, neuroradiology, cardiac, pediatric, general and emergency radiology. All of the participants reported frequent computer use (more than once a day) and search for medical information (7 reported more than once a day, 3 once a day and 2 once a week). Due to technical difficulties, bandwidth problems and development schedule, not all of the participants were able to perform all the tasks.

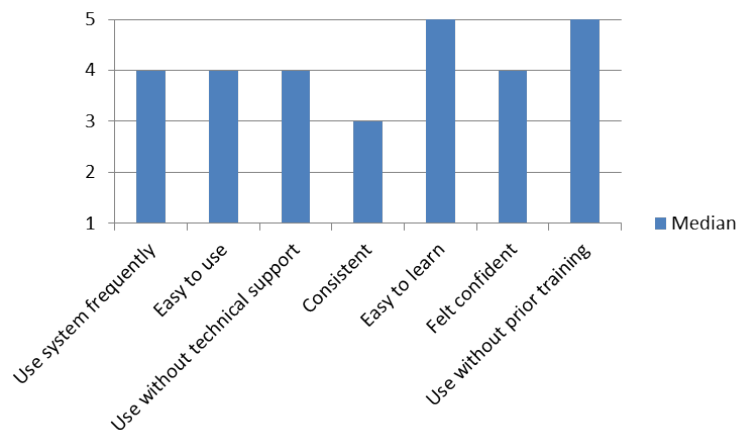


Figure 5: Median values of measuring general user satisfaction about the system in Likert scale (1=strongly negative, 5=strongly positive).

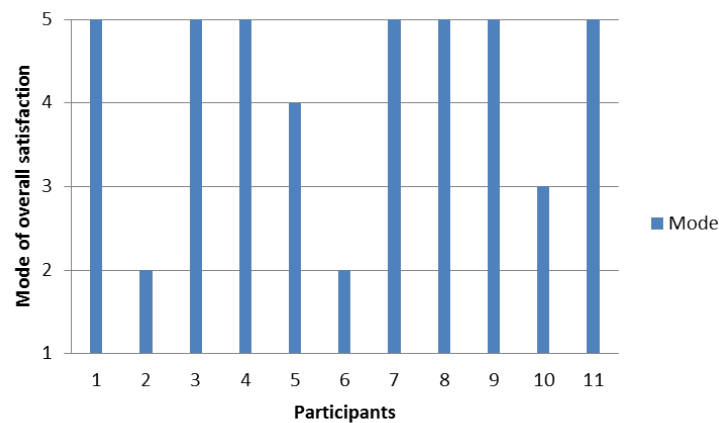


Figure 6: Mode values for each participant over the global satisfaction questions in a Likert scale.

User satisfaction results over key general aspects of the system are presented in Figure 5. The median for the question about intention to use the system frequently was 4. The same median was obtained for easiness to use, the ability to use the system without technical support and feeling confident when using the system. The median for easiness to learn and using the system without prior training was 5 for both and the general feeling about the system's consistency was 3. In order to assess the global satisfaction of each participant the mode over the general satisfaction questions was taken, measuring the most frequent grade given (Figure 6). Also, for measuring the consistency of this satisfaction, the frequency of mode was given (Figure 7).

In the following sections a detailed description of the results of each step of each subsystem prototype evaluation is given.

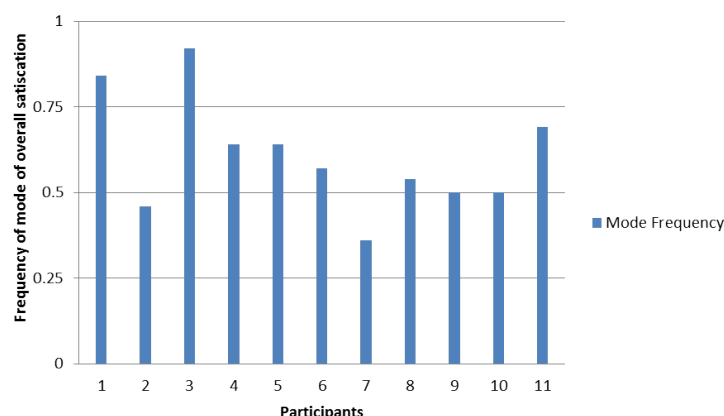


Figure 7: Mode frequency for each participant over the global satisfaction questions.

3.1 2D Image/article search subsystem

Two sets of user tests were run in the context of KHRESMOI first round of user-centered evaluation. The first set was a pilot user study that aimed at finding the most significant bugs and inconsistencies as well as the user study's protocol shortcomings. Then the full user tests were performed after refining the protocol using the results of the pilot study.

3.1.1 Pilot user study

Five persons (2 females, 3 males) participated in two sets of parallel sessions. All were below 30 years old, with two of them being below 25. Two participants had radiology background (one specializing in bones), one was a non-radiology intern and two were final year students in medicine. All participants declared frequent computer use. Three persons answered to search for medical info more than once per day, one once per day, and one answered once per week. The recruitment of participants was done via personal contacts and people who volunteered to take part in the study at the radiology department of the University hospitals of Geneva.

The mean time for retrieving the first relevant result during the 2D image search tasks was 158 seconds. This time included choosing image examples, investigating the results and judging a result as relevant. This time includes only the cases when a relevant result was found. For case-based retrieval tasks the respective mean time was 179 seconds. The mean number of results selected as relevant was 5 for the 2D image search tasks and 2.6 for the case-based search. One participant (one still studying medicine) did not select any relevant results for any of the tasks.

User satisfaction on the specific system aspects was measured on a Likert scale where 1 was strongly negative and 5 was strongly positive. Results are given in Figure 8. Questions about the user's use intention in academic, research and clinical work respectively obtained medians of 4. Finally a question regarding the practical usefulness of the novel features of the system obtained a median of 5 out of 6 due to a design error. This was excluded from the global user satisfaction evaluation. User satisfaction results over general aspects of the system are presented in Figure 9.

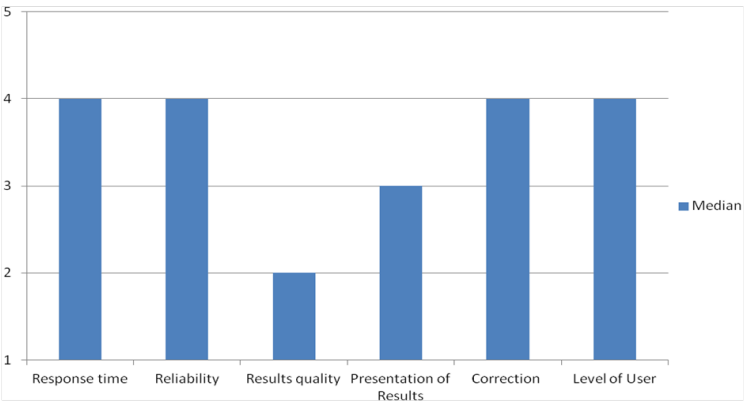


Figure 8: Pilot study: Median of measuring user satisfaction over specific system aspects in a Likert scale (1=strongly negative, 5=strongly positive).

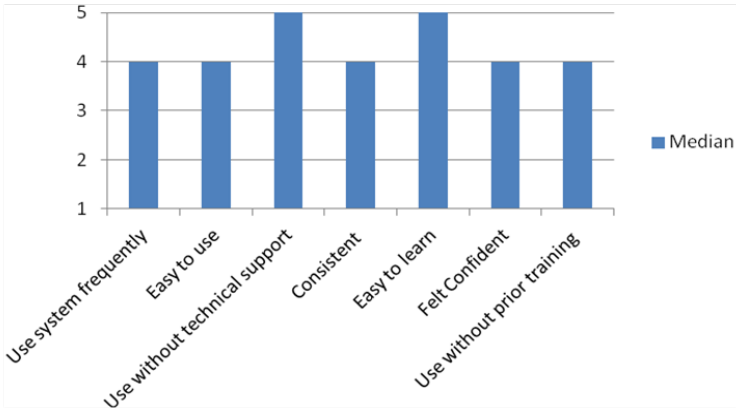


Figure 9: Pilot study: Median values of measuring general user satisfaction about the system in Likert scale.

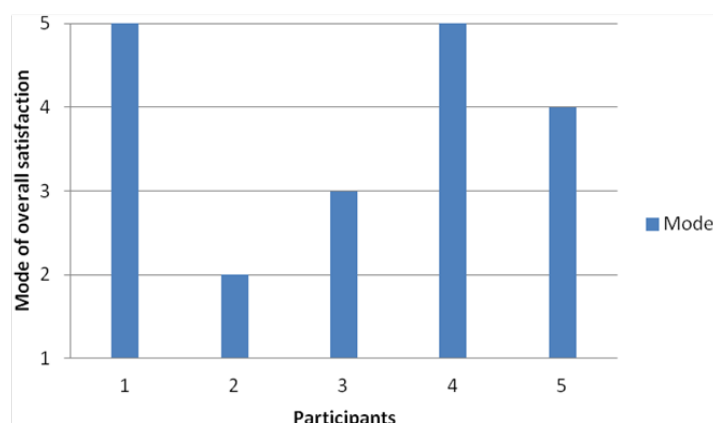


Figure 10: Pilot study: Mode values for each participant over the global satisfaction questions in a Likert scale.

In order to assess the global satisfaction of each participant the mode over the general satisfaction questions was taken, measuring the most frequent grade given (Figure 10). Also, for measuring the consistency of this satisfaction, the frequency of mode was given (Figure 11).

All open responses were classed into similar comments. Redundant comments were removed and all comments were transmitted to the development team. Frequent comments include:

- complaints about CBIR performance were frequent as often several irrelevant results were ranked highly;
- Zooming into images and basic manipulation such as level/window settings were considered important but are currently not possible in the interface;
- Displaying more information about the images in the result lists was also requested;
- Other propositions about functionalities such as backspace usage, radiology related functionalities (contrast adjusting etc.) were given;

Below, some of the comments are given in their raw form (translated from French):

- The search for associated articles is interesting at this stage, the search by images would also be useful if visual results were more relevant.
- As a student, search results have to be extremely relevant because we do not have the knowledge to exclude bad images on our own.
- It seems reliable more or less, I feel like it has difficulties distinguishing CT scan images from MRIs.
- More information on the description of images could be interesting to narrow down searches. A zoom in on an image in the 'details' section would be useful.

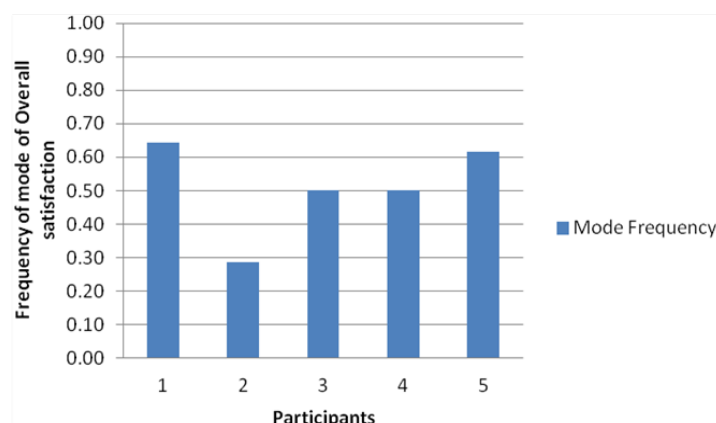


Figure 11: Pilot study: Mode frequency for each participant over the global satisfaction questions.

- There is no text below images in the list of result. It would be good to see the description when you hover the mouse over a result image. Difficult to get a good idea of the image at a glance when they are so small.
- The tool reacts very well to its use, no delay, no bug, tasks we are asked to do are rapidly performed.
- Takes 15 minutes to be comfortable.

3.1.2 Full user tests

Eleven out of the twelve persons that participated in the user study tested the 2D image and article search subsystem. However in two cases the participants did not performed all the tasks or answered all the questions due to technical difficulties. This resulted to 31 performed 2D image search tasks out of 33 (11 participants \times 3 tasks) and 19 article search tasks out of 22 (11 participants \times 2 tasks).

The mean success rate was 80.65% (25/31) for image search tasks and 78.95% (15/19) for article search tasks. Every task that the user found at least one relevant result was considered as successfull. The mean time for finding the first relevant result during the 2D image search tasks was 106 seconds. This time included choosing image examples, investigating the results and judging a result as relevant. It includes only the cases when a relevant result was found. For case-based retrieval tasks the respective mean time was 150 seconds. The mean number of results selected as relevant was 4 for the 2D image search tasks and 3.1 for the case-based search. This numbers include also the cases that no relevant result was found by the user. User satisfaction on key aspects of the 2D image search prototype and intention of use was measured on a Likert scale where 1 was strongly negative and 5 was strongly positive. Results are given in Figure 12. The median for system response time was 5 (mean 4.5). A median of 3 was reported for system reliability (mean 3.6). In terms of results quality and presentation the median was 3.5

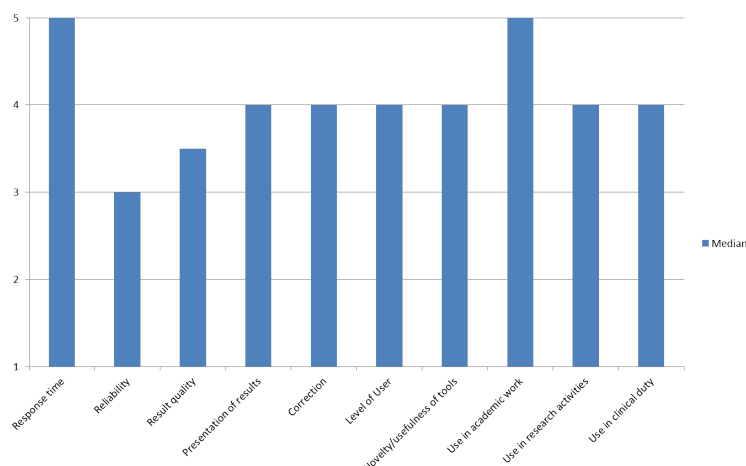


Figure 12: Median of measuring user satisfaction over specific 2D image search prototype aspects in a Likert scale (1=strongly negative, 5=strongly positive).

(mean 3.2) and 4 (mean 4.1) respectively, while ability to correct mistakes and system design to be used by all levels of users both obtained a median of 4 (means 4.1 and 4).

Question about the user's use intention in academic work obtained a median of 5 (mean 4.2), while the respective questions for research activity and clinical duty obtained medians of 4 (mean 3.9 for both). The question regarding the novelty and practical usefulness of the features of the system obtained a median of 4 (mean 4.5).

The same procedure with the pilot study was followed for the open responses (see Section 3.1.1). The comments were classified into Frontend and Backend-related. The most common comments can be summarized in the following points:

- Frontend

- Querying, such as advanced text querying and relevant/non-relevant marking of images; available options should be more explicit and easy to use;
- Basic and radiological-based image manipulation of the selected and query images should be available;
- Results presentation and views; images should be presented in grid as default and articles as lists.

- Backend

- Complains about CBIR returning many non relevant results; Non relevant marking didn't produce the desired results;
- Modality filtering requests;
- Propositions about Finding articles using images only or using example article ("Find similar articles");

Below, some of the comments are given in their raw form:

- Frontend – Querying

- When the mesh system is on, and is researching for auto completing, it is really hard to position the cursor in the middle of a term to refine it;
- Non-relevant marking not intuitive. It would be useful to be able to mark images as non-relevant directly from the results list;
- It is not very obvious when you are able to drag and drop the images to the query zone (multiple images, results list, personal library, details view etc.);
- Want to see the detail of an image from the query zone (images from the query zone drag n droppable also to tray);
- Re do "enter" when modifying query by for example drag and dropping images is not very comfortable;
- When searching with images in the query zone, it would be useful to have a contextual command (right click) to launch the search easily;
- Globally, the use of AND OR NOT, of text in bracket, and of non relevant marking of images should be more explicit (also what is default?);

- Frontend – Details View

- In the detail view, the participant would like to have the possibilities: scroll (CTs), scale, contrast, brightness;
- Didn't have time to read all the article: better summarization could maybe spare time to users;
- Abstract translation to German is useful;

- Frontend – Results View

- Thumbnails would be better a little bigger, so that we can already start a quick pre-analysis at this stage;
- Couldn't it be that we could resize thumbnail as we want them to be?
- When you select "Text" from the media type displaying options for results, we still find images in the first place. Couldn't that be changed, so that when you select one media type, this type appears first?
- Thumbnails presentation is not great. Everything is a bit mixed up and unclear, there is no organization, we don't know about modality, anatomy, or pathology;
- The button to switch mode in the results list (list or grid) is a bit confusing because it makes you think it is switch between images and articles;
- Articles: would be nice to have the whole title when hovering with the mouse;
- Multiple selection (shift or ctrl) would be nice in the result list (to drag n drop them all at once to the query zone);
- The button of grid mode in the result list is not reverted back after some clicks;
- Usually expect results as grid if image and as list if articles.

- Backend

- Marking graphs as non-relevant gives still a lot of graphs in result. This comment is the same with other types of images: e.g.: marking a given chest x-ray image as non-relevant does not even exclude the result itself from the resulting results list;
- It would be great to be able to specify the type of image that we want in result list (x-ray, MRI, CTs, graphs, statistics etc.);
- If i m using std x ray for visualizing a pulmonary parenchyma, then I probably want std x rays and CTs but no ultrasound or graphs etc.;
- Globally the system is more relevant with textual queries (with or without images) than with image queries only. This is the case for both article results and image results. Osteoporosis does not find many images, and only from source 2;
- It is impossible to return article with queries with only image-search. This is maybe not adapted and we should always be able to search for articles;
- The search over images only does not return anything that I could use. As soon as I add terms to my query (with or without images), I can find some more relevant results. Some results are even focused on the same anatomical area, but usually regards different pathologies;
- I would like to have something like search for similar articles, as it is now for images;
- It would be nice to be able to check "Only search articles" before the search, so that he does not have to do one more step by filtering results. Maybe it could spare a bit of execution time too, which is great.

- General

- Performing the is highly dependent on the user's skills to analyze de radiological image, and therefore depends a lot on experience;
- Consistency and integrity of the program, and execution times are good. Real problem is result relevancy;
- The learning curve of the system is quite rapid;
- During the 2D test, the high number of windows in the perspective is confusing;
- I can trust the image results because they come from scientific articles, in contrast to images I find from general purpose search engines;
- I understand the concept behind the system. I find the connection of images and articles useful and this is not currently available with the existing tools. It will be more useful if the results are better.

3.2 3D Image search subsystem

This section covers the specific results of the 3D image search prototype. First information about the design and integration of a 3D image search system was acquired with a series of interviews. Based on the results the 3D prototype was designed. After removing initial bugs

and refinement of the user test protocol the full usertest were performed. Note that the user tests focused on the evaluation of the workflow and user interaction. Only a preliminary version of the image search algorithms was deployed in the system.

3.2.1 Interviews

The interview results are grouped into (1) workflow, (2) 3D frontend and (3) organization. Workflow results present what is important for the radiologists in order to quickly and efficiently use the system. Furthermore, they describe what information needs to be available at each time and how it should be presented. This also includes whether or not to display the associated report. The 3D frontend examples are shown in Figure 1. In this paragraph the comments about which layout would work best, what is missing and what should be changed is summarized. Information concerning the organization and setup of the user tests is outlined in the last paragraph.

(1) Workflow: For the physicians it is important to quickly judge relevant cases. Therefore they need quick browsing through the thumbnail images. In order to reduce the number of thumbnail results to a smaller number of relevant cases, text search for the corresponding reports should be possible. Additional filters like gender, age or modality would also reduce the number to more relevant search results. Furthermore the reports could also be used for making differential diagnosis.

Optional improvements regarding the thumbnails were identified: enlargement of the thumbnails or having scrollable thumbnail to further facilitate relevance judgement of the results. Optional thumbnail size would also be a possibility, because this can depend on the modality and pathology that is currently worked on.

In the detail view more extra information would contribute to the decision making of the physicians, like: gender, age, slice id, modality, sequence information, windowing information (brightness/contrast), number of images, anatomical position, patient preparation, confirmation tests (lab tests, biopsy etc).

Keywords and ontology terms extracted from the reports could be shown in either the detail view or thumbnail view. A search history would be appreciated by the physicians where they can easily and quickly look up the latest volumes they looked at. Adjusting brightness and contrast, as well as zooming in and out of the image are very important for the daily routine of a radiologist. The method called *windowing* is widely used in different radiology systems and is an essential tool for any physician. The current settings of the query window should also be automatically transferred to the detail view in order to make the two volumes instantly comparable.

(2) 3D frontend: Figure 1 shows six drafts named from (A) to (F) for the first UI design. The desired thumbnail size tends to be as in (A) and (B). Reports are important especially during the detail view, as in (C), (E) and (F), although the report should not be overlaid with the image (E). In this way they can look at the volume and the report simultaneously. A short description of the report added to the thumbnails view is also appreciated (B). The presentation of the report to the user should be clearly structured for easy readability including highlighting of important parts (F).

Marking a region of interest (ROI) should be as simple as possible and the visualization of the result ROIs (F) could vastly reduce the time needed for the tasks. Indicating the relevance

of the matching ROIs (e.g. by brightness) could improve and speed up the interpretation of the results. The physician should be able to easily mark and browse one or more slices of the query volume.

If different layouts are used, fast switching between them is very important. During browsing the thumbnails additional information should be available, for example: age, gender, modality, sequence information, number of images, resolution. Primarily the axial view is preferred by the physicians, options to switch to one of the other two planes can be included.

(3) Organization: A handout for the physicians should be prepared, where the project and its goals are briefly described. Information concerning the user tests and the analysis of its result should be mentioned. The hardware and software used for the tests should also be defined, such that results from different locations are consistent and comparable.

3.2.2 Full user tests

The user interface used for the 3D image search prototype tests is shown in Figure 13.

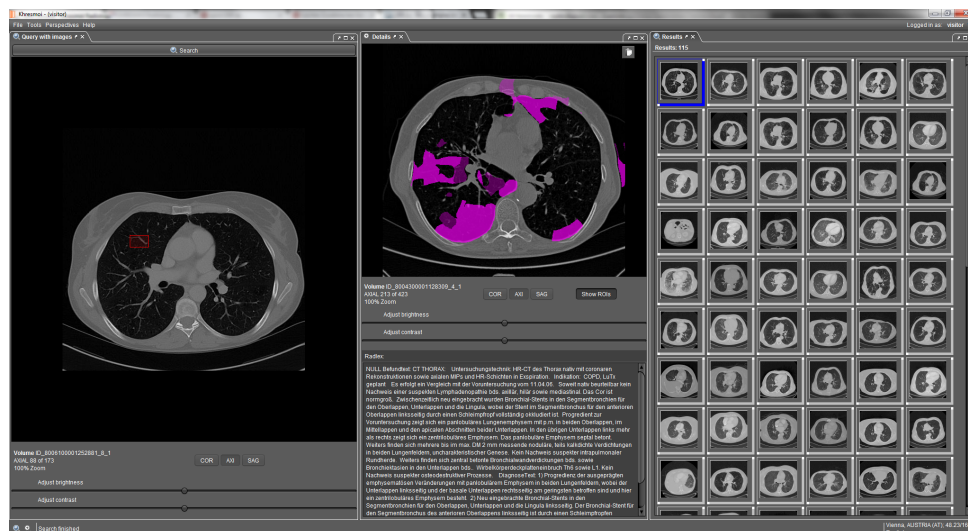


Figure 13: The 3D prototype used for the user tests. The left pane is the query volume with an example ROI, the right pane the result pane with the thumbnails and the center pane the detail view with the report and overlaid matching ROIs.

Six users participated in the user tests on the 3D image search prototype. User satisfaction on key aspects of the system and intention of use was measured on a Likert scale where 1 was strongly negative and 5 was strongly positive. Results are given in Figure 14. The median for system response time was 3 (mean 2.8). A median of 4 was reported for system reliability (mean 3.4). Results quality and presentation obtained medians of 2 (mean 2.4) and 4 (mean 3.4) respectively. The ability to correct mistakes and to be used by all levels of users obtained medians of 5 and 4 respectively (means 4.4 and 4).

Regarding the user's use intention in academic work obtained a median of 5 (mean 4.8), while the respective questions for research activity and clinical duty obtained medians of 4 (means 4.2 and 3.8 respectively). The question about the novelty and practical usefulness of the features of the system obtained a median of 5 (mean 4.6).

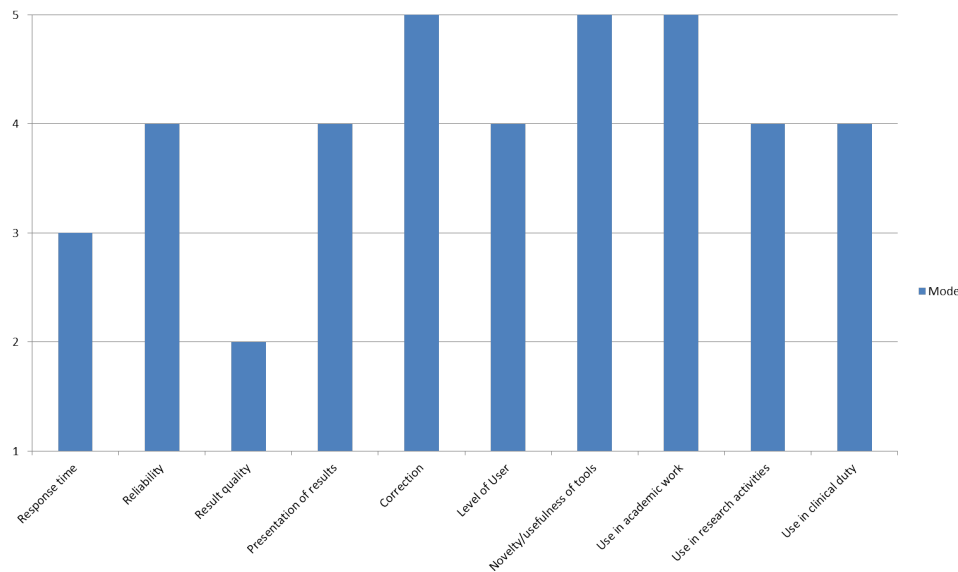


Figure 14: Median of measuring user satisfaction over specific 3D image search prototype aspects in a Likert scale (1=strongly negative, 5=strongly positive).

Similar to the interview results, the user test results are grouped into (1) workflow and (2) 3D frontend.

(1) Workflow: Changing brightness and contrast was working with the current implementation, although using *windowing* would have been preferred by most users. Also using the same values for the detail view automatically would be very useful, as the physicians always have to set it for every result image opened in the detail view. The physicians often asked the question, whether or not some of the results are from the same patient/study as the query image. Therefore an indication in the result list would be very useful. The shown thumbnail slices should also be the most similar slice corresponding to the query ROI and not the center slice by default. This helps to judge about the usefulness of the result volume. Many of the doctors intuitively tried to use a *drag and drop* feature from the results view to detail view. Implementing this would maybe result in a more user-friendly overall experience. Full text search for reports was requested by most of the test users. This would help to either reduce the number of relevant results or even make it possible to search for certain pathologies by text.

(2) 3D frontend: Scrolling through a high resolution volume with 500 slices and more takes a lot of time and is annoying to the user. To make this faster two options were suggested by the users. (a) Include a scroll bar for the slices on the side of the image or (b) while pressing the left mouse button scrolling should be possible with moving the mouse. The second type is how it is usually implemented in many radiology tools. The size of the thumbnails was too small. There were two suggestions concerning this problem. (a) Displaying larger thumbnails, replacing 4 thumbnails by 1 or (b) by allowing the users to adapt the thumbnail size to their needs. The latter would increase the flexibility for the tool, because thumbnail size could be application specific and therefore would need to be changed by the user. Because the loading of the detail view currently takes a lot of time (between 15 and 40 seconds), the physicians would like to have a more informative thumbnail view in order to quickly decide if the results are relevant for them or not. One suggestion was to make the thumbnails scrollable, meaning that for example

every 10th slice of the volume is available as a thumbnail and can be used for scrolling. Another was to transform the result view more to the style of the 2D image search prototype, including text from the report and other information like gender, age, modality and so on.

4 Discussion

Part of this section has been accepted for publication in [14].

A total number of 17 persons participated in the user tests of the KHRESMOI radiology prototype. The sample of users was relatively young and had a varying level of medical experience, with the participants being involved into various radiology specializations. Recruiting radiologists was a difficult task as radiologists are usually on a busy schedule, with a lot of clinical and academic. The number of participating users, even though it was less than what was aimed at, is mentioned as sufficient in the related literature, in regard to the current development stage, the specialized type of the system and the goals of the study. More participants would potentially be an overspend of human resources without resulting to significantly better insight.

Main tendencies on user satisfaction could be identified (Figure 5), which can be used to guide further technical development on improving certain aspects. In regard to the global aspects of the system, users found the system easy to use without the use of technical support. They felt confident using the system and are positive towards using it frequently. They seemed strongly positive about the system being easy to use without any prior training, despite the new tools offered (e.g. CBIR, relevance feedback and ROI marking). They were less satisfied with the consistency of the system, which is a logical outcome considering the current stage of the system's development. The main tendency of the users seems to be strongly positive with the majority of the users (6 persons) giving a mode grade of 5 with a frequency above 0.5 over the general satisfaction questions.

4.1 2D Image/article search subsystem

4.1.1 Pilot user study

One of the main outcomes of the pilot study was that a video tutorial alone was not enough and that a user required exploring himself the new functionalities before proceeding to complex information search tasks. This can limit the effectiveness of information finding during the early tasks and makes them less appropriate for performance comparison (text search vs. visual plus text search). For this purpose, a guided tutorial task after the video was included in the full tests, where it was asked from the user to perform very simple tasks using the tools (see Appendix C).

Some task descriptions and questions of the survey were not completely clear and this caused misunderstanding results retrieved by the participants. It was also observed that participants did not read the tasks in full detail and often performed slightly differing actions than the ones the task asked. This led to rewriting the task descriptions to be shorter and more clear. Moreover an oral description was given in the full tests, pointing out the important parts of each task.

Concerning the use of a commercial recording and observation software such as Morae, both advantages and drawbacks were found. All information that the participant needs for

performing the test can be found on his screen and no transition to paper is needed. This helps the user concentrate on the tasks and facilitates an uninterrupted flow of the tests. It provides results in a unified digital format that is easy to transfer to statistical packages, to analyze and present in a meaningful way. It allows for indirect observation (as the observer can remotely observe the user's screen and face), which removes some of the subject's stress of being observed and the incorrect feeling that he is being evaluated.

On the other hand, the use of such a tool increases the hardware (e.g. every session uses two computers with Internet connection instead of one) and software requirements, adds extra complexity to the setup of the tests and is prone to software crashes. Moreover, purchasing a commercial product depends on the available resources. It needs to be noted though, that all parts of the user test can be performed without the use of such software but require additional manual work. Overall, it is a helpful solution but it would be advised to also have a paper version of the user test material available as a backup plan if the software fails.

A general feeling expressed by a few participants was that they felt they were being evaluated instead of the system. This feeling can affect the subject's behavior, performance and answers, so this aspect was more explicitly clarified when the purpose of the study was explained in the beginning of each full tests session.

This pilot study was considered as partly internal because participants were chosen through personal contacts. For this reason, user satisfaction measurements were taken with skepticism, while feedback on improvements and proposed additions continue to be fully valid. Main satisfaction tendencies of the system could be observed.

Overall system satisfaction was high as it can be seen in Figures 8 9 and 10, with the majority of the participants having a mode above neutral and mode frequency above 0.5. However there was a clear drop in satisfaction about certain aspects, such as the results quality and presentation (with median 2 and 3 out of 5 respectively) and the development team concentrated on improving these aspects before the full user tests round.

4.1.2 Full user tests

The overall success in finding relevant images (80.65%) using the KHRESMOI 2D image search prototype indicates an improvement over the percent (75%) that was reported in [16] as self assessment of radiologists about their image finding success rate using current tools. Case-based retrieval, was shown to be a more challenging task (78.95% success rate) which was expected by the results reported in KHRESMOI deliverable document D2.3 [13].

The average time over the successful tasks for the participants' to select a relevant result was less than 3 minutes for both types of tasks (1 minute 46 seconds for image retrieval and 2 minutes 30 seconds for article retrieval). This is also below the average estimated time reported in [16] (between 5 and 10 minutes) and indicates an added value in terms of time efficiency when using the KHRESMOI system.

Regarding user satisfaction over the basic aspects of the 2D image search prototype (Figure 12), participants seemed strongly positive about the system response time. They had a positive opinion on the ability to correct mistakes and on the fact that the system can be used by all levels of users. The modification done to the results presentation and retrieval performance after the pilot study seem to have worked in a positive way, shifting the medians of satisfaction from 2 and 3 to 3.5 and 4.0 respectively. Participants were neutral over the system reliability,

which can be explained of the presence of bugs and inconsistencies in the prototype.

The users seemed to find the system novel and useful in practice giving a positive to strongly positive grade on this aspect (median 4 with a mean of 4.5). The activity that they gave a preference in intention of use was on academic work which goes along the design purposes.

A lot of feedback was given by the participants on the open questions, post-test discussions and spoken comments while performing the tests. Some confirmed the outputs of the pilot user study while many new comments and propositions were introduced.

On the graphical user interface aspect, the main comments were related to the image use, either requesting basic image manipulation features (which was also identified in the pilot tests but was not yet implemented for the full tests) or were about the image inconsistencies (e.g. drag and drop not being available on all views, detail views not being available for query images, non relevant marking being non-intuitive). Advanced text querying seemed to not be straightforward and several participants either used advanced queries or at least asked about the availability. These facts may indicate that a more comprehensive interface would be useful for radiologists.

Regarding the functionalities offered by the prototype, CBIR performance was again one of the common complains. Moreover, filtering out images of irrelevant modalities that appeared in the results was also a common request and relevance feedback results were questioned. Propositions about finding similar articles and finding articles using image examples were made by a couple of participants.

Overall, the system's concepts were appreciated, such as the connection of articles and images and the trustworthiness of the results. An improvement over the results quality would result in a system with even more practical use. Moreover, even though most of the tasks were successfully performed the quantity and quality of resources returned in several scenarios was considered insufficient.

4.2 3D Image search subsystem

In this chapter the results of the 3D image search prototype interviews and full user tests are interpreted and their implications on the development of the prototype are discussed.

4.2.1 Interviews

From the interviews already a lot of information was gathered. Due to time and work power constraints not all of them could be included in the first prototype version. Therefore the decision ones made to implement only one few. The final version is similar to (F) from Figure 1 and is shown in Figure 13. The user comments were prioritized and resulted in a final list of points that were included in the prototype:

- Marking one or more ROIs per slice and volume and options to remove the latest or all ROIs from the volume. Matching ROIs are overlaid in the detail view.
- Changing of brightness and contrast in the query and detail view.
- Options for changing between coronal, axial and sagittal pane.
- Corresponding report and extracted Radlex terms in the detail view.

- Information about zoom level and current slice ID in the query and detail view.
- A search history is added such that the user is possible to switch to previous detail views that were cached within the session.

4.2.2 Full user tests

Figure 14 presents the user satisfaction over basic aspects of the 3D image search prototype. System response time and results quality were the least satisfactory aspects while the participants were positive about the ability to correct mistakes, the results presentation and the system reliability.

The users seemed to strongly agree about the system's novelty and practical usefulness in practice. The activity that they gave a preference in intention of use was academic work. A reason behind this choice could be the slow response times, as in clinical duty time saving is critical.

Using the preliminary search engine backend, the most time consuming process is querying the database with a volume and a ROI. It takes about 40 seconds until the thumbnail results are shown in the result view and another 15 seconds for a selected detail view to load. The lack of quality was the retrieval of matching ROIs, which currently did not satisfy the users needs.

Since some features that were mentioned in the interviews could not be added for the first prototype, the full user tests also pointed to some of those points indicating their necessity for the physicians. The following list shows the most important features and changes that were requested by the physicians during the interviews and user tests that should be implemented in the future version of the 3D image search prototype:

- Increase thumbnail size by either making them 4 times larger, or by letting the user choose the thumbnail size.
- Show the most significant slice according to the query ROI as the result thumbnail instead of always the center slice.
- Include more case relevant information in the result view by adding indicators/text for gender, age, modality, study/series description and a report summary or important report keywords. Additional information like patient preparation and confirmation tests (laboratory tests, biopsies etc.) would be beneficial.
- Additionally indicate whether the results are from the same user/study as the query volume.
- Change of brightness/contrast using the *windowing* method.
- Automatically setting the brightness/contrast values for the detail view to the one from the query view.
- The overlaid ROIs of the detail view should represent the estimated precision of the algorithm (e.g. matchings with higher probability should have a higher color saturation)
- Implement a drag and drop feature for the result view.

- Include full text search for reports, highlighting important key words.
- Make scrolling through volumes with a large number of slices more user-friendly by either including a scroll bar or by implementing the hold-left-mouse-button scroll function.
- Improve the response time of loading a result in the detail view.
- Improve the quality of the retrieval engine.

5 Conclusion

The design, the setup and the results of the first round of user-centered evaluation of the KHRESMOI system by radiologists were presented and discussed. User tests were performed on the two main subsystems, the 2D image/article and 3D image search prototypes. An iterative approach was followed having a preliminary round of user tests and a set of interviews with radiologists using mockups. Key aspects on usability such as user satisfaction, efficiency, effectiveness and learnability were used.

The results show an overall positive opinion by the participants while aspects where the system lacks in quality were also identified. More importantly, concrete feedback and propositions by radiologists were given and provided useful insight and directions for further development.

The 2D image search prototype showed signs of being able to improve the current success rate of visual information finding by radiologists, in a time-efficient way. The results quality and the consistency in the interface and backend components are the aspects to which further development should be concentrated. Advanced query options need to be more explicit and the relevance feedback functionality improved both in the interface and the backend. CBIR poor performance, which was partly addressed after the pilot tests, needs to be further improved. Additional radiology-relevant resources need to be included to provide satisfactory retrieval results.

The 3D image search prototype was found to be very novel and useful in practice by the users. Participants were less satisfied with the response time. While the performance and quality for searching for volumes without a ROI was ok (15 seconds), refining the query using a ROI it took about 1 minute in total, until matching results could be viewed. As expected for the initial prototype backend, the lack of quality of the matching ROIs was pointed out by the users. Consequently, these need to be the main directions of improvement for the next steps of the system development. More image manipulation options were requested and results representation modifications were discussed. Linking between the two prototypes was proposed by users, something that is in line with the project plans for the next development steps.

The system concept and added value was understood by most of the participants. Comments on improving the current system tools were made and additional functionalities were proposed. The outcomes of this evaluation will be discussed with the technical development team and be prioritized for future development.

6 References

- [1] Alex M. Aisen, Lynn S. Broderick, Helen Winer-Muram, Carla E. Brodley, Avinash C. Kak, Christina Pavlopoulou, Jennifer Dy, Chi-Ren Shyu, and Alan Marchiori. Automated storage and retrieval of thin-section CT images to assist diagnosis: System description and preliminary assessment. *Radiology*, 228(1):265–270, July 2003.
- [2] C.J. Bastien. Usability testing: a review of some methodological and technical aspects of the method. *International Journal of Medical Informatics*, 79:18–23, 2010.
- [3] J. Brooke. A quick and dirty usability scale. *Usability evaluation in industry*, 189:194, 1996.
- [4] J.P. Chin, V.A. Diehl, and K.L. Norman. Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems, ACM*, pages 213–218, 1988.
- [5] A. De Vito Dabbs, B.A. Myers, K.R. Mc Curry, J. Dunbar-Jacob, R.P. Hawkins, and A. Begey. User-centered design and interactive health technologies for patients. *Computers, Informatics, Nursing*, 27, 2009.
- [6] J.C. Faga. Usability testing of a large, multidisciplinary library database: basic search and visual search. *Information technology and libraries*, 27:140–150, 2005.
- [7] Alba García Seco de Herrera, Dimitrios Markonis, Ivan Eggel, and Henning Müller. The medGIFT group in ImageCLEFmed 2012. In *Working Notes of CLEF 2012*, 2012.
- [8] M. Hertzum. A case study of laboratory, workshop, and field tests. In A. Kobsa and C. Stephanidis (Eds.), *User interfaces for all, Proceedings*, volume 228, pages 59–72, 1999.
- [9] A. Holzinger. Usability engineering methods for software developers. *Communications of the ACM*, 48:71–74, 2005.
- [10] Charles E. Kahn Jr. and Cheng Thao. Goldminer: A radiology image search engine. *American Journal of Roentgenology*, 188:1475–1478, 2008.
- [11] A. Kaikkonen, A. Kekalainen, M. Cankar, T. Kallio, and A. Kankainen. Usability testing of mobile applications: A comparison between laboratory and field testing. *Journal of Usability*, 1:4–17, 2005.
- [12] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3:1–224, 2009.
- [13] Georg Langs, Joachim Ofner, Andreas Burner, Rene Donner, Henning Müller, Adrien Depeursinge, Dimitrios Markonis, Celia Boyer, Alexandre Masselot, and Nolan Lawson. Report on results of the wp2 first evaluation phase. Deliverable d2.3 of the khresmoi project, Medical University of Vienna, 2012.

- [14] Dimitrios Markonis, Frederic Baroz, Rafael Luis Ruiz de Castaneda, Celia Boyer, and Henning Müller. User tests for assessing a medical image retrieval system: A pilot study. In *MEDINFO 2013, the 14th World Congress on Medical and Health Informatics*, 2013. accepted for publication.
- [15] Dimitrios Markonis, Adrien Depeursinge, Ivan Eggel, Antonio Foncubierta-Rodríguez, and Henning Müller. Accessing the medical literature with content-based visual retrieval and text retrieval techniques. In *Proceedings of the Radiological Society of North America (RSNA)*, November 2011.
- [16] Dimitrios Markonis, Markus Holzer, Sebastian Dung, Alejandro Vargas, Georg Langs, Sascha Kriewel, and Henning Müller. A survey on visual information search behavior and requirements of radiologists. *Methods of Information in Medicine*, 51(6):539–548, 2012.
- [17] Henning Müller. Report on image use behaviour and requirements. Deliverable d9.1 of the khresmoi project, University of Applied Sciences, Western Switzerland, 2011.
- [18] Henning Müller, Alba García Seco de Herrera, Jayashree Kalpathy-Cramer, Dina Demner Fushman, Sameer Antani, and Ivan Eggel. Overview of the imageclef 2012 medical image retrieval and classification tasks. In *Working Notes of CLEF 2012 (Cross Language Evaluation Forum)*, September 2012.
- [19] Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. A review of content-based image retrieval systems in medicine—clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1):1–23, 2004.
- [20] J. Nielsen. Alertbox. <http://www.useit.com/alertbox/number-of-test-users.html>, 2012. Online; accessed 09–December-2012.
- [21] J. Nielsen and J.K. Landauer. A mathematical model of the finding of usability problems. In *CHI '93 Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, pages 206–213, 1993.
- [22] J. Nielsen and R. Molich. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Empowering people*, pages 249–256, 1990.
- [23] J. Spool and W. Schroeder. Testing web sites, five users is nowhere near enough. In *CHI'01 extended abstracts on Human factors in computing systems, ACM*, pages 285–286, 2001.
- [24] Theodora Tsikrika, Henning Müller, and Charles E. Kahn Jr. Log analysis to understand medical professionals' image searching behaviour. In *Proceedings of the 24th European Medical Informatics Conference, MIE2012*, 2012.
- [25] K. Vredenburg, J.Y. Mao, P.W. Smith, and T. Carey. A survey of user-centered design practice. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves*, pages 471–478, 2002.
- [26] A. Woolrych and G. Cockton. Why and when five test users aren't enough. In *Proceedings of IHM-HCI 2001 conference*, pages 105–108, 2001.

Appendices

A DEMOGRAPHIC SURVEY

1. Are you ?

- ☐ Male
☐ Female

2. How old are you ?

- ☐ < 20 ☐ 20-30 ☐ 30-40
☐ 40-50 ☐ 50-60 ☐ >60

3. What is your native language ?

- ☐ English ☐ French ☐ German
☐ Spanish ☐ Czech ☐ Other

4. If you have chosen « other » please specify :

5. Your skills in English are ? <Scale>

- ☐ Basic, can comprehend simple issues ☐ ☐ ☐ Native language

6. What is the highest position you have had in a medical service other than radiology ?

7. What is the highest position you have had in a service of radiology ?

If non-applicable, enter N/A.

8. If you have a work experience in a radiology service, how long have you been working in radiology ?

- ☐ N/A ☐ 0-3 y ☐ 4-6 y ☐ 6-10 y ☐ >10 y

9. If you have a work experience in a radiology service, what field in radiology are you specialized in ?

- ☐ Bone ☐ Thorax
☐ Nuclear radiology ☐ Interventional radiology
☐ Radio-oncology ☐ Echography
☐ Emergency radiology ☐ Other

10. If you have checked « other » please specify :

11. Do you use a computer in your day-to-day life ?

☐ Never ☐ Once a month ☐ Once a week ☐ Once a day ☐ More than once a day

12. Do you use a computer for job or education related tasks ?

☐ Never ☐ Once a month ☐ Once a week ☐ Once a day ☐ More than once a day

13. Do you use Google search ? <Scale + free text>

☐ Never ☐ Once a month ☐ Once a week ☐ Once a day ☐ More than once a day

14. If you use other search engines, please specify below

15. Do you search the Internet for health related information ?

☐ Never ☐ Once a month ☐ Once a week ☐ Once a day ☐ More than once a day

16. If yes, please specify the websites you use below :

17. Do you use Google image search ? <Scale + free text>

☐ Never ☐ Once a month ☐ Once a week ☐ Once a day ☐ More than once a day

18. If you use other image search engines, please specify below :

19. Do you use Facebook ?

☐ Never ☐ Once a month ☐ Once a week ☐ Once a day ☐ More than once a day

20. If you use other social media network, please specify below :

B USABILITY SURVEY

USABILITY OF THE SOFTWARE

1. *I would like to use this system frequently.*

☐ Strongly disagree ☐ ☐ ☐ ☐ Strongly agree

2. *I found the system unnecessarily complex.*

☐ Strongly agree ☐ ☐ ☐ ☐ Strongly disagree

3. *The system was easy to use.*

☐ Strongly disagree ☐ ☐ ☐ ☐ Strongly agree

4. *I would need the support of a technical person to be able to use this system.*

☐ Strongly agree ☐ ☐ ☐ ☐ Strongly disagree

5. *The various functions in this system were well integrated, that is, the program works in a harmonious way which is logical to me.*

☐ Strongly agree ☐ ☐ ☐ ☐ Strongly disagree

6. *There was too much inconsistency in this system, that is, the program react in a way that I was not expecting and surprised me.*

☐ Strongly agree ☐ ☐ ☐ ☐ Strongly disagree

7. *I would imagine that most radiologists would learn to use this system very quickly.*

☐ Strongly agree ☐ ☐ ☐ ☐ Strongly disagree

8. *I found the system very awkward to use.*

☐ Strongly agree ☐ ☐ ☐ ☐ Strongly disagree

9. *I felt very confident on what I was doing, using the system.*

☐ Strongly agree ☐ ☐ ☐ ☐ Strongly disagree

10. *I needed to learn a lot of things before I could get going with this system. That is, the program requires a lot of training before an adequate use.*

☐ Strongly agree ☐ ☐ ☐ ☐ Strongly disagree

11. *Are there any tools that need to be improved/changed? If yes, how would you like them to be changed so that they will be more useful to your searches?*

12. *Are there any new functionalities/tools that would like this search system to have?*

*SCREEN PRESENTATION***13. Reading characters on the screen**

☐ difficult ☐ ☐ ☐ ☐ easy

14. Add free comments**15. Presentation of images (e.g. size, position, additional information provided)**

☐ poor ☐ ☐ ☐ ☐ excellent

16. Add free comments**17. Quality of translation**

☐ poor ☐ ☐ ☐ ☐ excellent

18. Add free comments**19. Performing task is straightforward**

☐ never ☐ ☐ ☐ ☐ always

20. Add free comments*SYSTEM CAPABILITIES (FOR 2D/ARTICLE SEARCH AND 3D SEARCH)***21. Does the system respond quickly to your requests? Are the results delivered quickly enough?**

☐ too slow ☐ ☐ ☐ ☐ fast enough

22. Add free comments**23. Do you find the system reliable? Does it react the way you expect it to?**

☐ unreliable ☐ ☐ ☐ ☐ reliable

24. Add free comments

25. Are the results satisfactory? Do they match the queries you formulated?

☐ Unreliable ☐ ☐ ☐ ☐ Reliable

26. Add free comments

27. Are the results well presented?

☐ Dislike how results are presented ☐ ☐ ☐ Like how results are presented

28. Add free comments

29. How easy is it to correct your mistakes, that is, undo, redo tasks ?

☐ Difficult ☐ ☐ ☐ Easy

30. Add free comments

31. I think the system is appropriately designed for all levels of user (e.g. containing both simple and more advanced features in tools, for beginners and advanced in radiology respectively).

☐ Strongly disagree ☐ ☐ ☐ Strongly agree

32. Add free comments

33. I think the system provides some tools and features that can be helpful in my work/research that are not available in the current tools I use.

☐ Strongly disagree ☐ ☐ ☐ Strongly agree

34. Add free comments

35. I would use the 2D image and article search for academic work (preparation of lectures etc.).

☐ Strongly disagree ☐ ☐ ☐ Strongly agree

36. Add free comments

37. I would use the 2D image and article search for research activities.

☐ Strongly disagree ☐ ☐ ☐ ☐ Strongly agree

38. Add free comments

39. I would use the 2D image and article search during clinical work.

☐ Strongly disagree ☐ ☐ ☐ ☐ Strongly agree

40. Add free comments

C TASKS

TUTORIAL TASK (FOR 2D/ARTICLE SEARCH AND 3D SEARCH)

During this session, you will be asked to fulfill tasks using both the 2D/article search and the 3D search tools. This task contains simple actions so that you get familiar with the functionalities of the two tools.

In the 2D/article search tool:

1. Enter the term "brain" in the query zone and start the search.
2. Select one image from the results view and find its details in the Details view.
3. Drag and drop this image to the query zone.
4. Add 1 positive and 1 negative example image to the query zone from the results view to refine your query.
5. Note: Negative examples can be given by right clicking an image in the query zone and choosing "mark as non relevant"
6. Drag and drop an image from the results view to the Tray.

In the 3D search tool:

1. Right click somewhere in the search zone, and select Load a volume. Use the following ID:
ID_8006000001239002_3_1
2. When the volume is loaded, try to scroll through it, play with contrast and brightness.
3. Select one or more regions of interest (ROIs) by selecting an area on a slice of the volume with the mouse.
4. Try to select another ROI on a different slice and to remove other ROIs that you might have selected.
5. When you are satisfied with your selection, hit the Search button.
6. Browse through results in the results view and select a relevant volume.
7. Play with the volume loaded in the details view, contrast, brightness.

2D IMAGE SEARCH TASKS

Task 1 (max 5 minutes):

1. Click on "Clear all content" from the File Menu.
2. Find 3 osteoporosis x-ray images.
3. Only use textual queries.
4. Place the images you find into the *Tray*.

Task 2 (max 5 minutes):

1. Click on "Clear all content" from the File Menu.
2. Find 3 osteoporosis x-ray images.
3. Use the example images under the tag "image_task_1" in your *Personal Library* to complement the query.
4. You may mark the results as relevant or non relevant to relaunch the search.
5. Place any relevant results into the *Tray*.

Task 3 (max 5 minutes):

1. Click on "Clear all content" from the File Menu.
2. Find 3 images that share the same diagnosis with the example images under the tag "image_task_2" in your *Personal Library*.
3. You may mark results as relevant or non-relevant to relaunch the search.
4. Place the image you find into the *Tray*.

ARTICLE SEARCH TASKS

Task 1 (max minutes):

1. Click on "Clear all content" from the File Menu.
2. You have the following:
3. *"A 43-year-old man with painless, gross hematuria. Abdominal CT scan revealed a large left renal mass with extension into the left renal pelvis and ureter."*
4. You can find the images associated with the case are placed in your *Personal Library* under the tag: "case_retrieval_task_1".
5. Find 3 relevant articles to the case above and place them into the *Tray*.

Task 2 (max 7 minutes):

1. Click on the "Clear all content" from the File Menu.
2. You have the following:
3. *"A 56-year-old woman with Hepatitis C, now with abdominal pain and jaundice. Abdominal MRI shows T1 and T2 hyperintense mass in the left lobe of the liver which is enhanced in the arterial phase."*
4. you can find the images associated with the case are placed in your *Personal Library* under the tag: "case retrieval_task_8".
5. Find 3 relevant articles to the case above and place them into the *Tray*.

3D SEARCH TASKS

Task 1 A (Max 7 min):

- Give a radiological description for the patient using only the given volume with simple image processing features (zoom, contrast, brightness).
- Data: ID_8006000001239002_3_1

Task 1 B (Max 10 min)

- Give a radiological description for the patient using one volume and the pathology image search including ROI
- Data: ID_8006000001239002_3_1

Task 2 (Max 10 min)

- Verify one or more of the given pathological descriptions using the image search feature, e.g. find similar pathologies
- Data: ID_8006100001247763_6_1
- Honey Combing, Emphysem, Milkglass (Milchglastrübung), ...

FREE USE OF KHRESMOI

In this task you may use the systems freely (tool for 3D search or 2D/articles search, or both).

In the 2D tool:

You may search for images or articles, using whichever of the functionalities available.

In the 3D tool:

- Find similar anatomical region using example Data or any of the search Results (enter ID Manually from the Details-View)

- Example Data:
- ID_RA10001171211450_2_1
- ID_8004900001184356_2_1
- ID_RA10001173454330_603_1
- ID_RA10001180048880_16_1
- ID_RA10001181474240_901_1

D INVESTIGATOR INSTRUCTIONS

Before a session the investigator should perform the following actions:

- 1) Obtain 2 Windows-based computers (one for the participant and one for the investigator) with internet connection (preferably not wifi).

For the participant's computer:

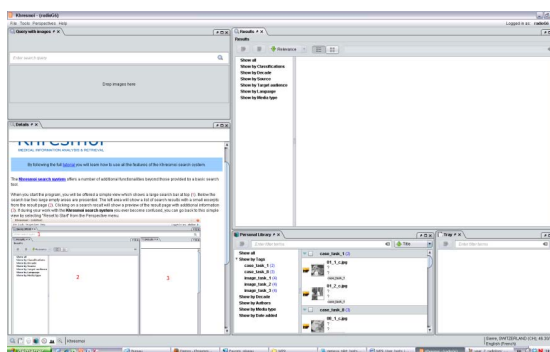
- a. Make sure that it has a conventional mouse, reasonably large, good resolution screen, a web camera and a microphone.
- b. Download the demo videos (2D/article and 3D) from here [link](#) save it on the desktop.
- c. Make sure that all Firewalls are disabled.

Morae:

- d. Install Morae Recorder.
- e. Download the Study Configuration File (the zip with the two files) from here [link](#) and save it to the desktop. **Important:** Do not make any changes to the file!
- f. Run Morae Recorder. Choose between the Study Configuration Files that you downloaded. If the participant ID number is even choose 2D-article-3D order. If the participant ID number is odd choose 3D-article-2D order. See point j) for details.
- g. Make sure that the web camera is adjusted to record the participants face and test the microphone.

ezDL 2D:

- h. Download ezdl client from here [link](#) and save it on the desktop.
- i. Start the ezDL client (Java 1.7 has to be installed for the ezDL client to run). Use the participant's account ID to login.
- j. The account IDs that were created for the user tests are in the form username/password: xxxxx/xxxxx
- k. Check here [link](#) for available accounts that have not been used yet. Remember to update the list with the accounts that you use.
- l. Deselect the TextSearch from the Source Selection Tool (Tools->Source Selection). Only MedSearch and ParaDISE should be selected as sources.
- m. Create a custom perspective for the user tests. Start with the Image search perspective



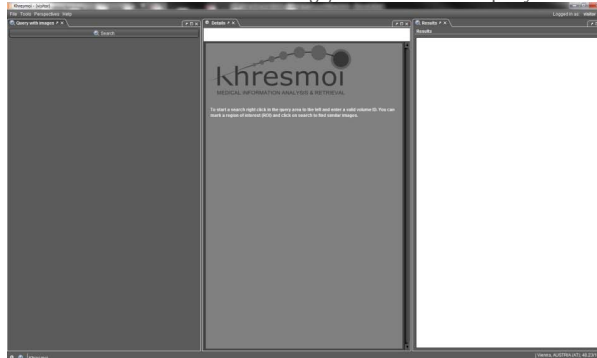
and add the Personal Library and Tray tools. The layout should look like this:

- n. Save the perspective

ezDL 3D:

- o. Start the ezdl client from here [link](#) by clicking on the ezDL Icon
- p. The account IDs that were created for the user tests are in the form username/password: xxxxx/xxxxx
- q. Make sure that the Image 3D source is selected (Tools -> source selection)

- r. Create and save a custom perspective that looks like this and save it (for better viewing the users are able to enlarge/minimize the query and detail views)



For the investigator's computer:

- a. Make sure that all Firewalls are disabled.

Morae:

- a. Install Morae Observer.
 - b. Run Morae Observer.
 - c. Connect to the participant's computer by using its ip address (you can find it the ip address of a computer by running cmd in the start menu of windows and then ipconfig – the ip address needed is the IPv4)
- 2) Download the consent form document from here [link](#) and print it.
 - 3) Make sure there is a way(e.g. a timer) to keep track of the duration of the tasks.

Note: It would be really useful for the investigator to have watched the demo video and played with the system before the session so he can better help the participant and understand misuses of the system. It would be also useful to have a colleague performing the tasks and play the role of the observer to get familiar with the note taking using the markers of the Morae Observer tool.

The 2D article demo video includes among others, demonstration of the following features:

- a. Features/Functionalities of Query by text to retrieve articles.
- b. Features/Functionalities in the Detail view of articles.
- c. Features/Functionalities of Query by text and or/image to retrieve images (Relevance Feedback).
- d. The Facet functionalities and the results presentation options in the Results view.
- e. Features/Functionalities in the Detail view of images.
- f. Personal library and Tray drag n drop use.

The 3D article demo video includes among others, demonstration of the following features:

- a. Features/Functionalities of the Query image view.
- b. Features/Functionalities of the Detail image view.
- c. Features/Functionalities of the Result view.

STEPS OF A SESSION

The steps of a session should be the following, in the following order:

- 1) The investigator should make an introduction to the Khresmoi project and the purpose of the user study. Key points should be
 - a. What are the Khresmoi project goal in radiology use case,
 - b. in which stage of the project we currently are (middle of the project – first prototype of the system, not finished)
 - c. why we are doing the user tests (evaluate/redesign the tool, not the radiologists)
- 2) At this point, the participant should be asked to sign the consent form.
- 3) The participant should be asked to watch the video demonstrating the system features. He is allowed to pause and rewatch if he wants, as well as ask questions to observers.
- 4) Once the video is finished and the user is ready to start, he should be asked to press the red button in Morae Recorder. The recording will then start.
- 5) The Introduction of the system will be shown in a prompt box and the completion of the demographics survey form should start.
- 6) Once the user has finished the completion of the demographics survey form he should be asked to maximize the Khresmoi system window (ezdl client) to start with performing the tasks.
- 7) During each task, the investigator should note all required information and observation notes using the Observer tool on his computer and call for the end of a task if the time is up before the participant completes the task. The observer should have a neutral attitude. He can talk to the participant (which will also encourage him to think out loud) but should not do operations for him. He may help him when he sees he is blocked and can't proceed. Observations can be of different types: free comments from the participant, free observations from the observers about the system or the user study itself, bugs and inconsistencies of the system. When an observation is noted into Morae, the observer should define its type in field next to where he entered his text. He can also define a score for each observation, but this step is not mandatory.
- 8) Once all the tasks are finished and questionnaires are filled in by the participant, the investigator could have an informal interview (off record) with the participant about the overall feeling of the use of the system, thoughts etc. The observer should take note of what the participant says during this phase of the test, since it is not recorded into Morae.
- 9) Two files need to be saved and be collected. The investigator should find the recording file on the participant's computer and save it as well as the observer log on the investigator's computer. The recording file should be saved as radioXy.rdg and the log file should be radioXy_observationLog.txt (e.g. radioV1.rdg, radioV1_observationLog.txt)

E CONSENT FORM

I hereby give my consent to participate in a scientific study conducted as part of the EU-Project KHRESMOI by the Health on the net Foundation. I have been informed about the content, the purpose and the extent of this study. I have been given sufficient time to reflect on my participation in the study. All my questions have been answered. I understand that my participation is voluntary and can be terminated at any time, without any specified reason. I understand that participation in this study does not influence or affect my health status. As there are no health-related risks involved, I agree with the absence of an insurance coverage.

The experiment will take about 1 hour and 30 minutes. It consists of two questionnaires and a number of recorded computer related tasks related to search some specific topics with a search engine prototype. The principal investigators, as well as all involved project members, pledge to use the collected data only in anonymized form and with the utmost discretion. They are required to treat the data and observations confidentially. No personally identifiable data is passed or sold to third parties.

I give permission for the involved project members of KHRESMOI to have access to my anonymized data collected during the study. I agree with my data being used for approved research to help advance insights on medical information retrieval.

Name organization and researcher

Date and signature

Date and signature

Participant ID – provided by the researcher

Place

This form should be signed in two copies. One is retained by a participant, another by the investigator.

In case of questions, please, contact the principal investigators at

[email1](#) and [email2](#) // 0041 22 372 62 50