Grant Agreement Number: 257528

KHRESMOI

www.khresmoi.eu

Report on user tests with initial search system

Deliverable number	*D10.1*
Dissemination level	*Public*
Delivery date	*April 2013, updated September 2013*
Status	*Final*
Author(s)	*Frederic Baroz, Celia Boyer, Manfred Samia Chahlal,Gschwandtner, Lorraine Goeuriot, Jan Hajic, Allan Hanbury, Marlene Kritz, Jérémy Leixa, João Palotti, Natalia Pletneva, Rafael Ruiz de Castañeda, Alexander Sachs, Matthias Samwald, Priscille Schneller, Veronika Stefanov, Zdenka Uresova*

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.

Abstract

This deliverable reports on user tests conducted from October to March 2013 in the EU funded project Khresmoi (2010-2014). The goal of these evaluations was to collect the end-users' feedback on the prototypes developed during the first two years of the project. Khresmoi targets three main user groups, namely the general public, medical practitioners, and specifically radiologists. In this deliverable the results for the first two groups are presented. The evaluations were carefully designed and refined during pilot studies. 14 Austrian physicians performed medical information retrieval tasks using the prototype for medical practitioners. General public evaluation was conducted in two different steps: 29 medical students of the University of Geneva – representing the general public -conducted a different type of evaluation, blindly comparing the search results from two search engines in certain health situations; further on 28 users (patients and online health information seekers) performed health information retrieval tasks using the prototype designed for the general public (this evaluation was similar to physicians evaluations). The feedback obtained during the evaluation tests has been analysed to provide further directions to the prototypes' development. The main areas of improvement identified are increasing the index, improving the relevancy and clarifying filter functionalities. Follow-up evaluations will be conducted to continue collecting end-users' feedback.

The deliverable was first submitted in April 2013, this is an updated version with an additional chapter describing the results of "full user tests" conducted with the representatives of general public.

Table of Contents

List of Abbreviations

BMI	Body mass index
ezDL	Easy access to digital libraries
GP	General practitioner
HON	Health On the Net
SUS	System usability scale
WP	Work package

List of Figures

List of Tables

1 Executive Summary

This document describes the user tests conducted to evaluate the initial search system for the general public and medical professionals (reported in deliverable D8.3 (D8.3, 2012)). The tests collected feedback from the targeted user groups to provide recommendations for work package 8 and others to guide the further development of the search system.

For the physicians, a total of 14 test users recruited by the Society of Physicians in Vienna provided roughly hour-long sessions each, 11 of which completed all 4 tasks. A large amount of feedback, positive and negative comments, recommendations and errors were collected in spoken form, via questionnaire, and by annotation of the session recordings.

The physician user tests showed that the system is functional and can be used to solve the tasks, but with room for improvement. The prototype is not yet ready for a medical real-life situation, as only more than half of the tasks were solved correctly within the timeframe given. In the collected user feedback (Section 2.3.3) the participants "expressed their dissatisfaction with the data found so far, citing "… that they would not make a decision based on their findings…". The available data sources, the result list ranking and preview metadata will be crucial points for further improvements. The classification and filtering of results as well as the translation feature could enhance the user experience and win many users for Khresmoi if they were to be improved sufficiently.

The general public user tests reported in this deliverable consist of two parts: A blind comparison of search results between Khresmoi for Everyone and Google, in which the preferences of a total of 26 undergraduate students of the Faculty of Medicine of the University of Geneva were evaluated, as well as the results of the pilot and "full user" tests conducted in the period from October 2012 to July 2013.

The blind comparison showed that the search results provided by Google were preferred due to wider coverage of resources. Contradictory to the stated position of the study participants, quality and trustworthiness do not appear to be criteria for selecting health web pages amongst the medical students.

The pilot tests for the full user tests with the general public led to the tasks and questionnaires being significantly simplified. Also, the decision was taken to exclude the ezDL prototype from the general public evaluations at this stage, as it would be too difficult and overwhelming for end-users. Full user tests were further conducted with general public in Prague, Geneva and Paris. While overall feedback was positive, the most crucial points for prototype improvement remain the relevancy, increasing the index, improving the filters and improving translation and localization.

2 Introduction

Khresmoi is a 4-year-long EU-funded large-scale project aiming at developing an information access and retrieval system for biomedical information, targeting the general public, physicians, and specifically radiologists. During the first two years of the project the work was focused on identifying user requirements and developing the system (both writing from scratch and integrating already existing components by the partners). During the end of the second year the evaluation of technical components was performed (Niraj Aswani, 2012) (Georg Langs J. O., 2012) (Thomas Beckers, 2012) (Pavel Pecina, 2012) (Konstantin Pentchev, 2012). The third year has begun with designing and performing user-centred evaluations, aiming at collecting feedback from the end-users. To be a valuable product, Khresmoi needs to satisfy users' needs; that is why this intermediate evaluation is so important, while there is still a year and a half of the project ahead to fix the identified problems and improve the overall prototype. Feedback received from the test users aims at providing further directions for Khresmoi development.

This deliverable describes design and results of the user-centred evaluations performed with physicians and the general public. The tests of the prototype for radiologists is described in a different document. This introduction describes the common aspects and differences of the evaluation tests performed, section 3 and 4 describe the physician and the general public evaluations respectively, and section 5 summarises the conclusions and future steps. A large amount of detailed content is attached in the appendix.

2.1 Timeline of user tests development

The work on the development of the evaluations started in May-June 2012, a few months before the prototypes were completed. The meeting to agree on and formalize the main steps was held in June 2012. From July to September the strategy for both physician and general public evaluations was defined.

However, after conducting the first pilots in October-November, some aspects of the protocol and tasks were reconsidered, especially for the general public. These pilot tests also provided first insights into what should be changed in the prototypes being tested.

After the project meeting in November 2012, the decision to conduct the evaluation for the general public in two stages was taken: the first stage – a blind comparison of search results retrieved from Google and Khresmoi by the end-users – and the second stage – the so-called "full user tests" – aimed on full testing of the upgraded prototype by the end-users. The main reason behind this decision was that a new Khresmoi asset was identified, namely the access to trustworthy and curated online health content. The hypothesis was that access to such content would be valued more than any of the specific tools developed. To evaluate this asset, or more precisely the end-users' attitude and satisfaction with it, would be not possible using "full user tests," as participants would have been informed about the benefits of Khresmoi well in advance. To overcome this bias, it was decided that the blind comparison would serve as an objective measure to see whether the end-users are likely to distinguish "curated" search results, and whether they prefer them.

From February to early March 2013 the physician user tests were conducted smoothly but later as originally planned, after the pilot tests in October and November 2012 had discovered a number of necessary changes and updates to the prototype software, data content, and the test protocol.

As for the general public evaluations, December 2012 and January 2013 were largely dedicated to protocol development for the blind comparison evaluation tests, as well as the platform allowing comparison of the two search engines' results. During February 2013 the blind comparison evaluation

tests were conducted. "Full user" tests significantly refined during the period of pilot tests were conducted in May-July 2013.

During the project meeting in February 2013 as a follow-up to the first pilot studies, it was decided that Khresmoi's current desktop interface (based on ezDL, (ezDL 2012)) is too complex for the general public, hence it will be mostly used and marketed towards physicians and researchers (and renamed "Khresmoi for health experts"), while the Khresmoi based on the HON Search prototype will be targeted to the general public (D8.3, 2012)(renamed "Khresmoi for Everyone").

2.2 Common approach

The evaluation strategy for the user-centered evaluations was defined through the following steps:

1. collecting user requirements (as described in deliverable D 8.2 (D8.2, 2012)
2. defining types of user and possible scenarios (as described in D8.2)
3. identifying data requirements and completing the list of data sources crawled by Khresmoi
4. further elaboration of types of users depending on initial knowledge about a subject and other factors
5. defining tasks for the evaluation tests

A common approach to structuring and running the sessions was agreed on:

1. Introductory part
    - Welcome
    - Signing the informed consent
    - Filling in demographic and background questionnaire
2. Testing
    - Performing predefined tasks, with all interactions are recorded by software (see section 2.4)
3. Feedback
    - Informal feedback given during the session
    - Answering a standard usability scale questionnaire (with some additional questions specifically for Khresmoi)
    - Some concluding remarks on what should be changed

The target groups of participants for each use case were defined:

- **Physicians**: Working physicians of various age groups, specialists and general practitioners, physicians in training and medical professors, all recruited by the Society of Physicians in Vienna. Participants received an Amazon voucher worth 50 Euros.
- **General public**: special focus on diabetes and cancer patients due to agreement with patients organizations, however not exclusively, both males and females of all ages and different health and web experience and knowledge speaking French and Czech as a mother tongue; the main inclusion criteria: at least occasional search for online health information.

2.3 Differences between physician and general public evaluations

The main difference between the evaluations conducted with the members of the general public and physicians is the additional evaluation stage of blind comparison for the general public. Also, as mentioned above, physician evaluations have been run with the ezDL prototype, which is a desktop application, while general public evaluations have been and are mostly run with Khresmoi for everyone, a web application used in a browser.

2.4 Testing Software

It was decided to use a recording software for all full user tests, in order to capture as much information as possible and to be able to store and revisit the test sessions later. The Morae usability testing software (Morae, 2013) from TechSmith was selected and tested during the initial tests and pilot tests, where it did not cause any problems. The software combines capturing and recording features with test management such as displaying task descriptions and reminders and prompting users with questionnaires at predefined steps.

The recording component of the software runs on the computer used by the test user and can record the screen, various interaction events such as mouse clicks or window dialogues, and optionally also the user's face and voice via webcam and microphone. The sequence of tasks and questions as predefined by the experimenters can be advanced manually or on autopilot. Additionally, comments and observations by the experimenters, so-called markers, can be integrated into the recording, either from a second computer during the recording, or at any time later.

This software has been used for physician evaluations, as well as for pilot evaluations for the general public.

All physician user tests were recorded. The setup was with one or two computers, with full recording of the screen and the user webcam and audio, from the beginning of the demographics questionnaire until the final answer to the feedback form after the tasks. Thus, only the conversation at the beginning, when the test setup and the Khresmoi software were introduced and the consent form was signed, are missing from the recordings.

The recordings proved valuable for deciphering the comments made by the users in the feedback forms. Due to the language situation, with German-speaking participants expressing their feedback in written English, many comments in the questionnaires are very brief and/or ambiguous. But since many participants spoke quite a lot during the tests, either prompted by the experimenter, or spontaneously when encountering a noteworthy situation, they provided an additional much richer level of feedback in spoken form. To make use of this source of information, all recordings were reviewed in detail at least one more time during the test evaluation, to improve the probability that no comments were lost or misinterpreted.

Figure 1 shows the interface of the analysis component of the software. The video is stopped at 21:23, and task 3 is selected, which can be seen as the brown bar in the timeline underneath the video. The timeline also displays the audio volume, making it easier to jump to the next discussion. The diamond shapes underneath the timeline show the position of comment markers. Markers and events can be searched by full text or types.

Figure 2 and Figure 3 show how tasks and questionnaires are defined and scheduled. For example, the questions about task 4 come after the end of task 4, and the overall survey "How did you like Khresmoi" is to be displayed at the end of the recording.

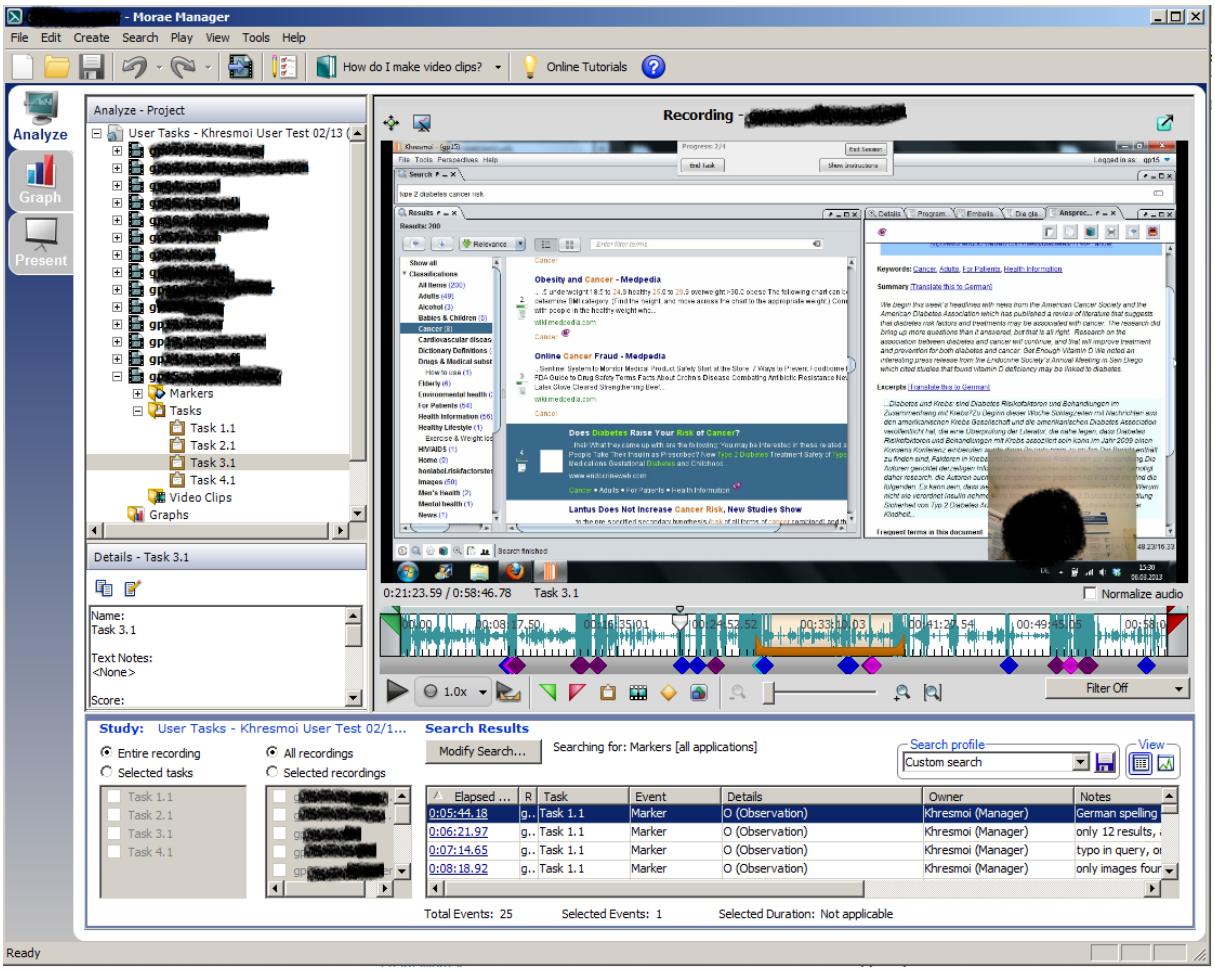


Figure 1 Recording software: Analysing the recording

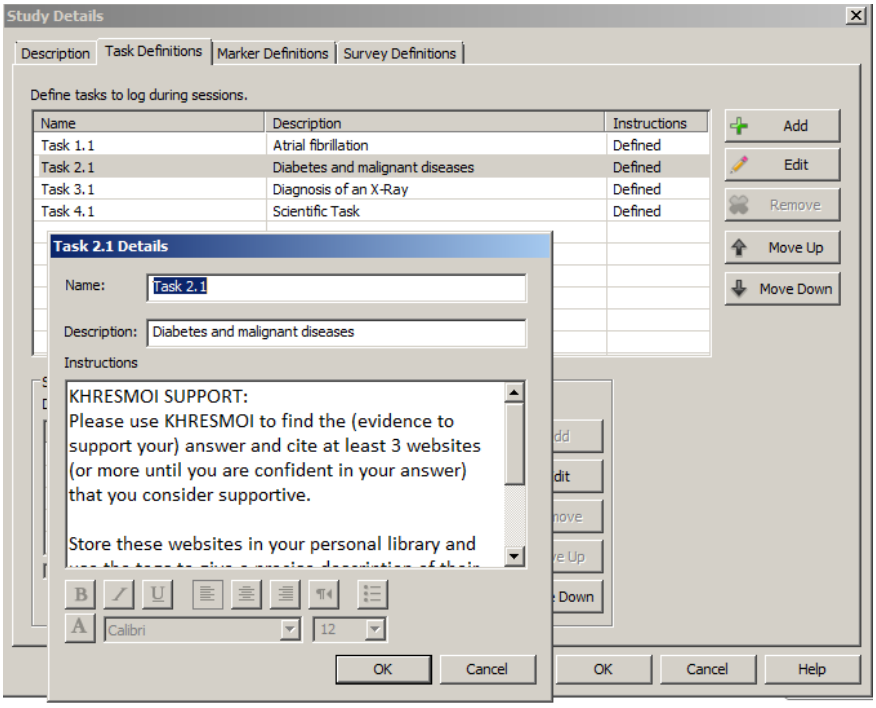


Figure 2 Defining tasks and instructions

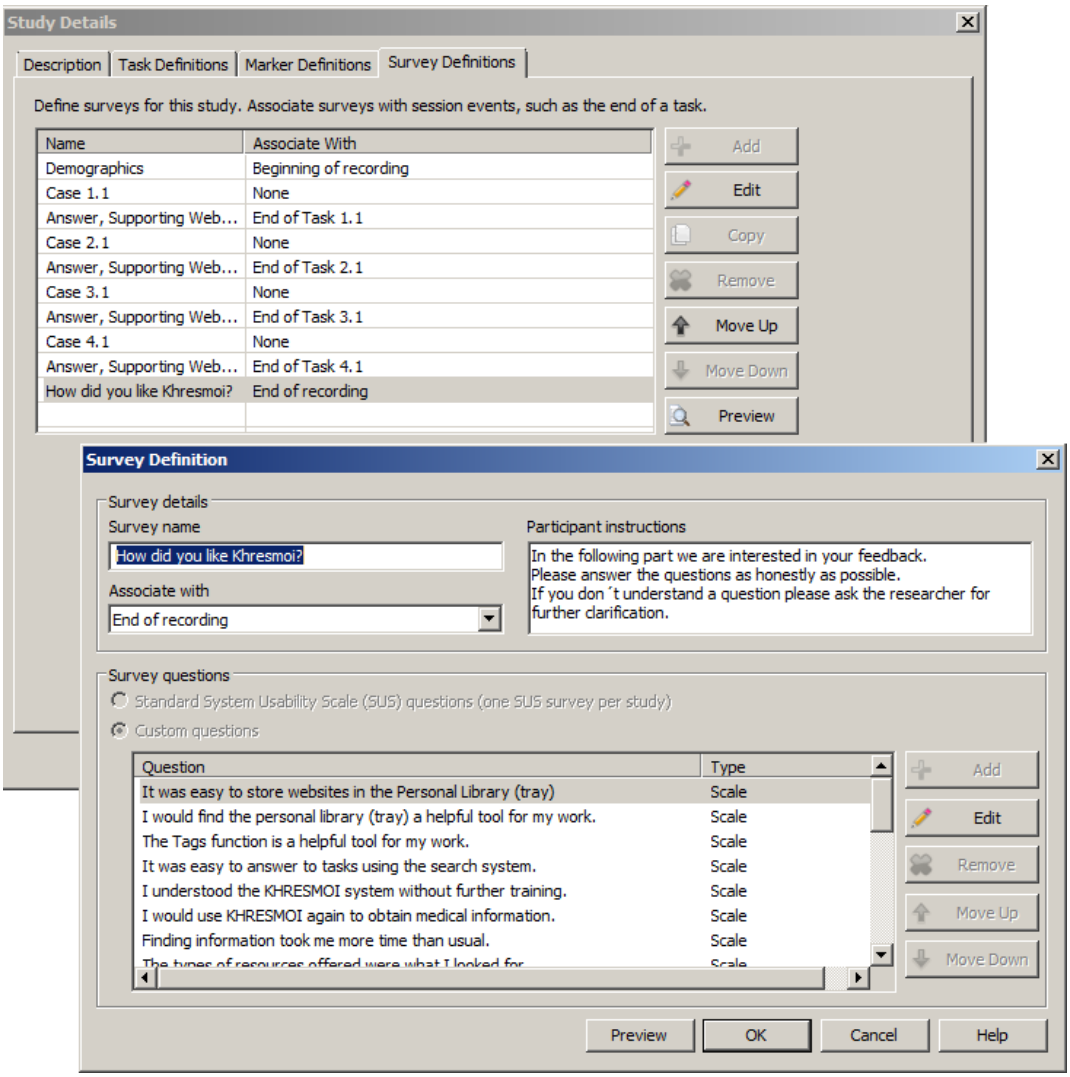


Figure 3 Definition and scheduling of Survey questionnaires

3 Physician evaluation

3.1 Hypothesis and research questions

As described in the previous section, the research questions and aspects to be evaluated were derived from the user requirements, scenarios and data requirements described in deliverable D8.2 (D8.2, 2012). The evaluation goals and overall experiment design remained the same, and over the course of the second half of 2012, user tasks and questionnaires were, on the basis of pilot study feedback, iteratively adapted.

3.1.1 Research questions

The research questions that the evaluation tries to answer can be categorized in two types, those focused on the usability of the search system, and those on the user satisfaction.

3.1.1.1 Usability of search system

- **Image:** Does image search assist users in obtaining relevant information?
- **Language:** Do users gain any benefit from translation support tools? (Query translation, document translation?) Which ones do they continue to use over further tasks?
- **Tools:** Which of the ezDL (ezDL, 2012) tools yield the greatest benefit for medical search?
- **Interface:** What will be more important for users: quality of results or quality of tools?
- **Training:** Is the user able to learn to use the search engine without any training?
- **Overall:** What aspects of the system need to be improved?

3.1.1.2 User satisfaction

- **Layout:** Is the user satisfied with layout and navigation of the search engine?
- **Resources:** Does the Khresmoi system allow users to find sufficient documents to meet their information needs within the given time?
- **Initial satisfaction at the starting-point of search:** Does the search engine provide sufficient help with the formulation of the initial query? Is the user satisfied with the first search result retrieved?
- **Final satisfaction at end-point of search:** Is the user able to find the correct answer? Is the user satisfied with the answer retrieved?

3.1.2 Hypotheses

Following from the research questions, the hypotheses can be grouped into the same two areas, those focused on the usability of the search system, and those on the user satisfaction.

3.1.2.1 Usability of search system:

- A comprehensive interface is of more utility than a simple interface with no additional tools OR at least some of the available tools are useful
- More professional users want more comprehensive tools

3.1.2.2 User satisfaction:

- The start point of the search requirement is determined by level of expertise on a given medical topic. There is a relationship between level of expertise (i.e. initial knowledge) and initial query input and resource expectation. Experts will use more specific terminology than novices and are expected to seek primary resources, from the start. Novices will use general terminology and start their search at a tertiary level.
- The end point of the search requirement is determined by level of academic qualification. Higher levels of academic qualification and expertise is expected to be associated with an increased access to scientific resources and a larger number of websites presented to support an answer.

3.1.3 Aspects to evaluate

- **Quality of results:** Is the user able to solve the task using the system? This aspect can be measured by the number of tasks solved, or the time required to solve the tasks.
- **Usability of Interface:** How much training is required for the user to understand the interface? How satisfied is the user with the interface? This can be measured by user feedback and observation.
- **Functionality:** Which functionalities (query suggestion, spelling correction, tray, personal library, query history, categorization etc.) are liked/disliked? This can be measured by observation (what is used/not used), user feedback and SUS questionnaire (SUS, 1996)

3.2 Methodology

Prior to the user evaluations, 5 comprehensive pilot studies were carried out with voluntary physicians to improve methodology, improve reliability of user tasks, test methodology and the screen capturing software. The pilot studies themselves were based on experience gained from internal tests conducted with attendees at earlier Khresmoi project meetings.

3.2.1 Overall procedure

The following procedure was applied to evaluate the research questions and test hypothesis described earlier:

- Participants were asked to perform information retrieval tasks for which at least one of the results was known.
- Time taken to fulfil each task was measured.
- Participants were asked to fill out a questionnaire about their experience of using the system. This allowed us to evaluate user satisfaction and detect potential usability problems and provided us with feedback and suggestions towards a system improvement.
- Participants were observed while using the system and the screen and interactions were recorded by a screen recording software. Through this technique, possible system flaws or usability problems that were not consciously detected by participants were addressed. It also allowed us to record signs of user disappointment and satisfaction.
- User interactions with the system were stored in a system log. This provided a more detailed report on how users interact with the system.
- Users were encouraged to propose new, useful, possibly missing, functionality. This gave room for further user-driven ideas/tools that could be integrated to improve the system.

3.2.2 Tasks

The initial version of the tasks was based on the evaluation session protocol created in cooperation with WP 7. The user tasks were written based on the findings of the surveys (D8.1.2, 2011) and the requirements defined previously (D8.2, 2012) (e.g. information needs, different groups of physicians). The idea was to determine the initial knowledge of the users and integrate user scenarios. Regarding the datasets to be used, the tasks assumed the availability of the complete set of HONcode-certified websites, as well as a set of additional websites ranked as the ”high priority” within the Khresmoi physician website list (D8.2, 2012), (D8.3, 2012).

During the pilot user tests, the limitations of such an approach became obvious. As described in Section 3.3.1, the scenarios and tasks were consequently simplified and adapted to the available resources and functionality. The number of tasks was reduced to four.

In the final user tests, all users were asked to solve the same four tasks. Each task had a context story attached to it and asked the user to employ one or more of the Khresmoi tools during the tasks, as listed in Table 1. For example, in task 4 the story is that the user is collaborating with a colleague and should therefore try to use the export function to share the findings.

Name	Description	Instructions	Tools to be used
Task 1	Atrial fibrillation	cite at least 3 websites	personal library
Task 2	Diabetes and malignant diseases	cite at least 3 websites	personal library, tags, translation from English to German.
Task 3	Diagnosis of an X-Ray	cite at least 1 website.	tray function to store 5 websites temporarily
Task 4	Scientific Task	cite relevant websites	export function

Table 1 Task definitions

Users were presented with the task instructions before beginning the tasks, and always had access to a short description in a pull-down (by the Morae software (Morae, 2013)) if they needed to re-read details during the task. (Figure 2.1 and Figure 2.2). The full text of these instructions is available in the appendix 5.1.1 of this document.

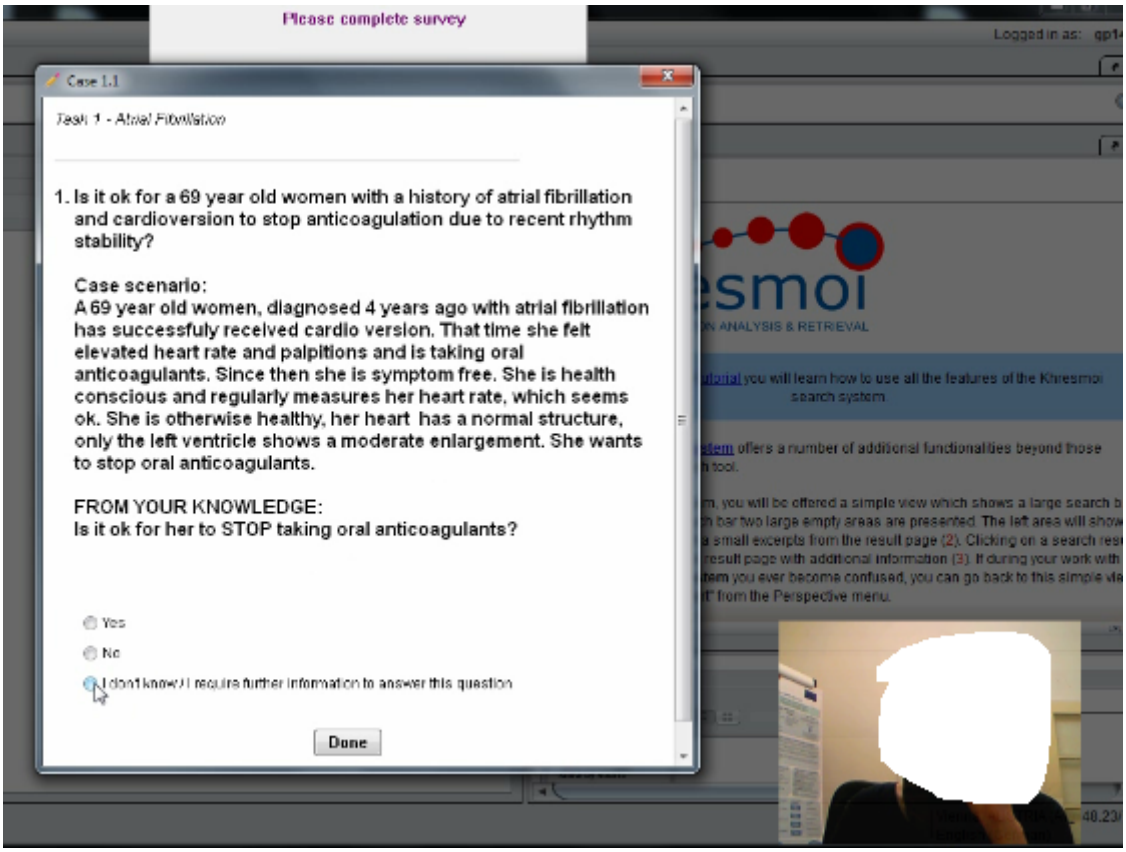


Figure 4 Task description window at the beginning of Task 1

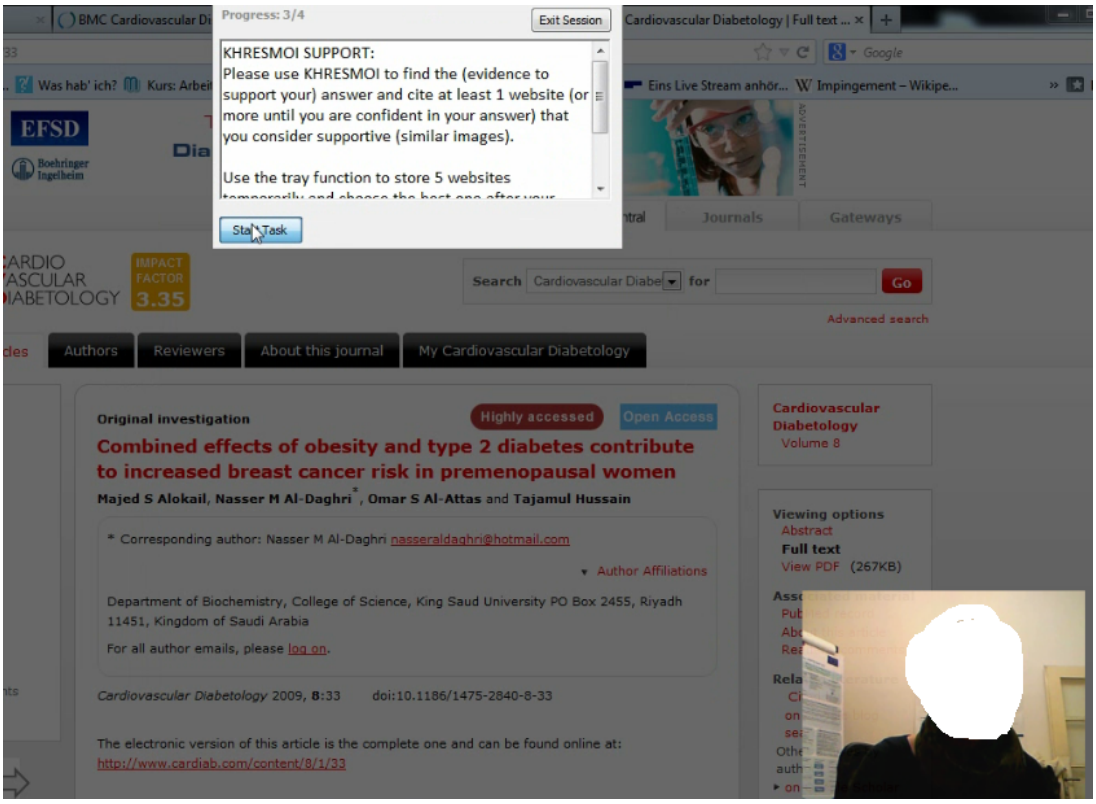


Figure 5 Opened pull-down with task description details reminder during Task 3

3.2.3 Questionnaires

Similar to the tasks, the questionnaires were continuously improved on the basis of user feedback provided in the pilot tests. The wording was clarified, and the number of questions was optimized to a compromise between the researchers' and the participants' needs.

The user tests started with a demographics questionnaire. (See Appendix 7.1.1.1). For each task users were then asked three feedback questions - what they liked, what they disliked, and what functionality or aspect they had missed (see Appendix 7.1.2.2). This allowed instant feedback in relation to solving a concrete task. After completing the last task, users were asked to complete a longer questionnaire to provide an overall feedback on the system. This was a combination of the SUS (System usability scale, (SUS, 1996)) questionnaire and additional Khresmoi specific questions (see Appendix 7.1.3)

Questionnaire	how often	when	Number of questions
Demographics	once	before the first task	12
Task feedback	for each task	after the task	3 (freetext)
Overall feedback	once	after the last task	21 (one freetext)

Table 2 Overall structure of user evaluation

3.2.4 Session Outline

Each user test consisted of three parts:

In the first part the user was introduced to the test. After a brief general information about Khresmoi and the goals of the user tests a consent form was signed. All information collected was used exclusively for the purpose of the study and was kept confidential. The user was then introduced to the interface of the search engine for physicians and led through a short case of a possible query with basic information on how to use the interface. After, that, the recording with Morae Software was started. The software then lead the participant through the whole user test, while the tasks were started manually by the examiner.

Also during the first part, the participant had to fill in the demographics questionnaire. This part took 10 minutes.

In the second part the participant was confronted with four different tasks (treatment, diagnosis, image diagnosis, scientific), representing real-life scenarios. During the 10 minutes attributed to each task first the initial knowledge on the given topic (without any external help) was evaluated. After that the Khresmoi search engine was used to find the answer or support existing knowledge. In every task different prototype modalities were integrated which the participant has to fulfil. After each task, feedback was provided either in autopilot questionnaire and/or verbally to the examiner present in the same room.

In the third part the participant was asked to complete a questionnaire evaluating how satisfied he or she was with the search system (SUS questionnaire with some additional questions specific to Khresmoi). This overall feedback represented the end of the test and had to be given in 10 minutes. With the end of the autopilot function the recording was terminated and stored for future analysis.

All in all the user test required a total of 60 minutes to complete.

3.2.5 Setup

All sessions took place in the offices of the Society of Physicians in Vienna. The examiner who provided the introduction remained in the same room as the participant during the test to be available for questions and comments. However, he remained in the background and pretended to work on something else, to avoid distracting the participant by direct observation. During the recording with the Morae software, a second observer was connected to the participant's computer from another room, watching the test via video and audio live streaming and adding annotations to the recording. Participants were aware of this setup.

3.3 Report on the evaluation tests conducted

3.3.1 Pilot User Tests

During October and November 2012 five pilot tests were conducted. They were performed in order to improve methodology, confirm/improve reliability of user tasks and test the screen capturing software.

The first pilot test already revealed that the user tests will be quite hard to perform within the allowed timeframe of 60 minutes with 4 tasks and direct evaluation, as well as letting the participants perform 4 tasks on their own. In consequence the user test setup was not ready to be performed by many other participants. Four more pilots had to follow in order to check the feasibility of solving the tasks within the timeframe, as well as testing the capturing software with its possibilities, and to optimize the time management.

In response to the second and third pilot tests, time management was adapted and confirmed: 10 minutes for the introduction and demographics, 40 minutes for four tasks (10 minutes each) and 10 minutes for the overall feedback at the end of test. This time management worked properly for the rest of pilot tests, so that the examiners were able to perform a user test in 60 minutes (+/- 5 minutes).

Concerning the tasks the pilot tests showed different results in their solvability: Task 1 (atrial fibrillation) and 2 (diabetes) were solvable with the search engine. Resources which could give a correct answer appeared in the list. Depending on the query the ranking appeared not useful, so that the users had difficulties in finding valuable resources. These tasks were not changed.

Task 3 (x-ray) and 4 (scientific) showed that there were no valuable resources inside the system to solve them.

Therefore, the content was slightly changed for task 3 to make sure that relevant data existed in the datasets used for the user tests. Concerning task 4 the resource for the correct answer didn't appear in the system. This task had to be changed completely with another scientific topic, where several scientific resources could be found inside the system.

Adapting two tasks in a way that they could be solved with resources which are already in the search system is not the same as a situation with an unexpected question during medical real life situation.

With the modified questions several possibilities in solving the tasks were open and the participants could also focus on the prototype itself with several ideas of improvement.

During all pilot tests the use of the recording software was tested and optimized. This part worked perfectly and was considered as excellent for the upcoming user tests.

In addition to that, several issues about the improvement of the prototype were reported with screenshots, direct feedback to the examiner or logs of unexpected processes (see Appendix 7.1.4 for complete pilot protocols). This long list of improvements was ranked. "Blockers" were problems that could not let the evaluation go forward, "Critical" issues hindered the evaluation seriously or

compromised results and "Minor" issues that would improve the system, but it was not urgent to be fixed before the user tests.

Consequently, the "Blocker" and "Critical" issues were identified and worked on in December 2012 and January 2013, so that the user tests could start with an acceptable performance in February 2013.

3.3.2 User Tests

From February 7th to March 6th 2013 14 users participated in the tests.

3.3.2.1 Demographics

The following charts summarize the demographics of the user test participants.

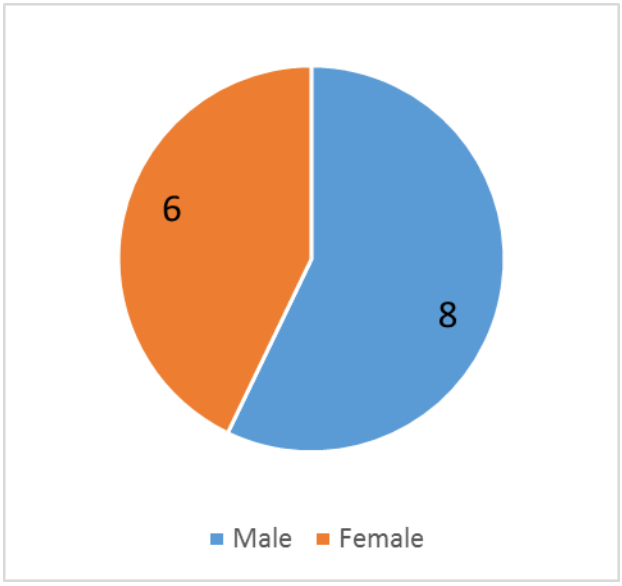


Figure 6 Number of participants by gender

0

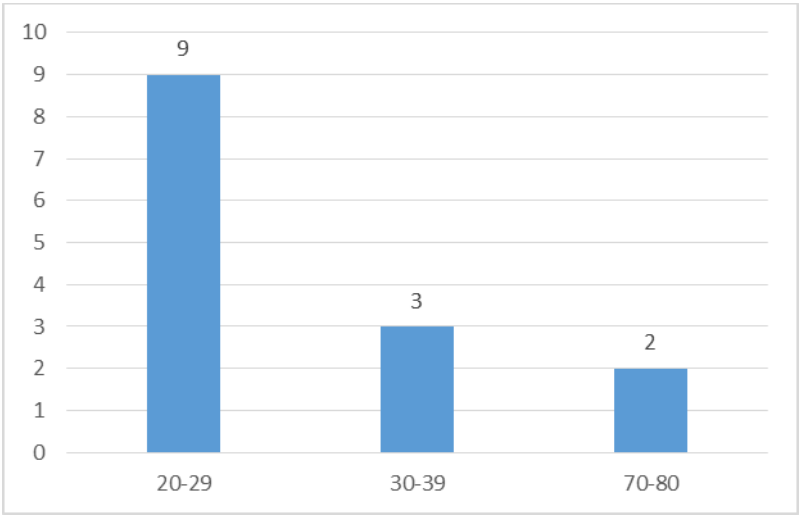


Figure 7 Number of participants by age group

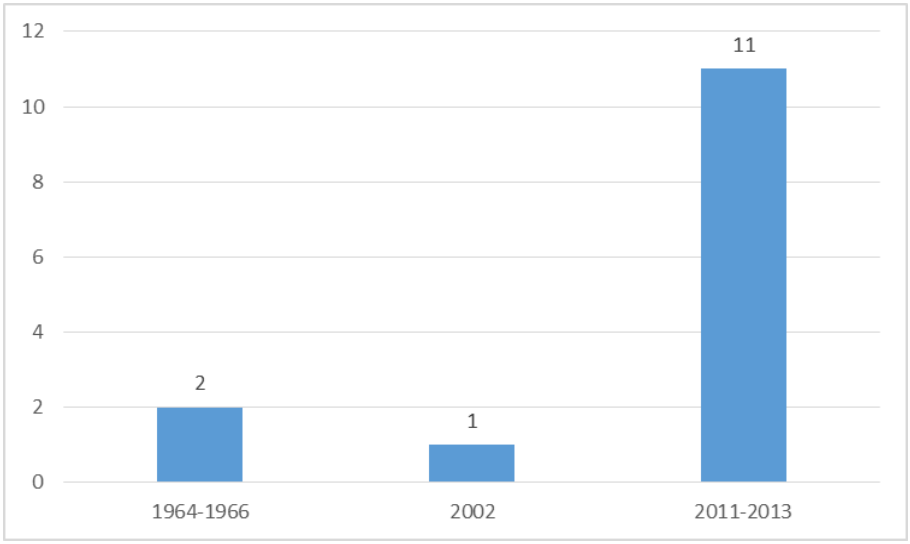


Figure 8 Number of participants by year range when they completed their medical degree

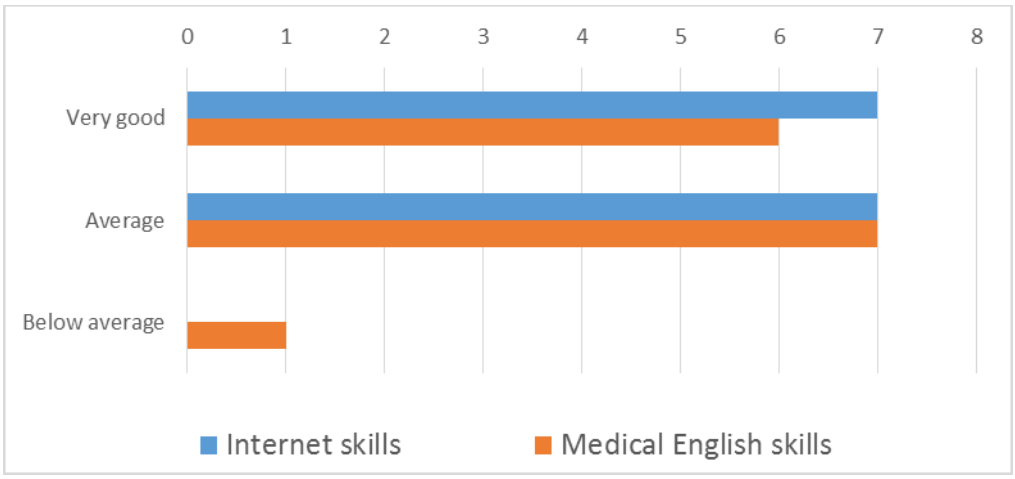


Figure 9 Number of participants by Internet and Medical English skills

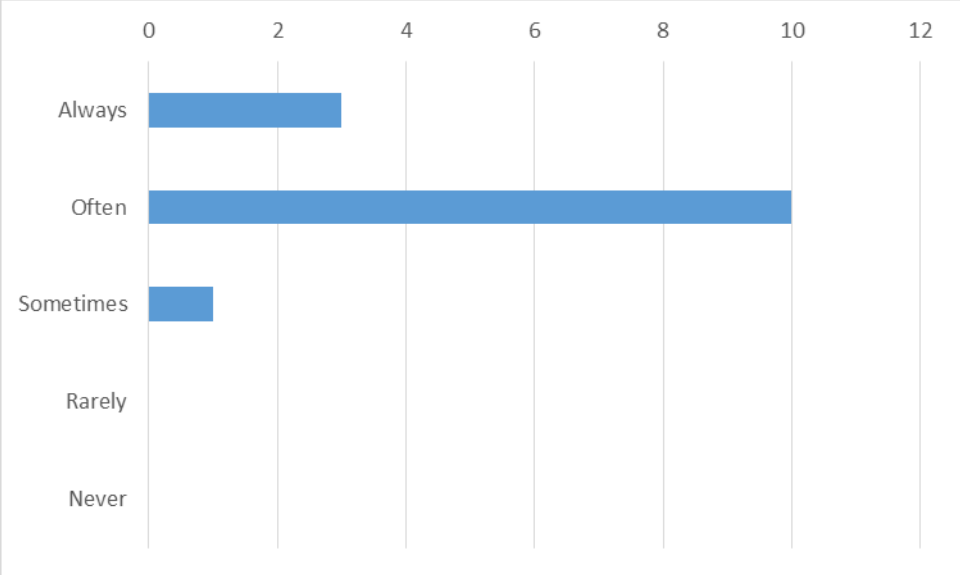


Figure 10 Number of Participants by use of the Internet to search for medical information

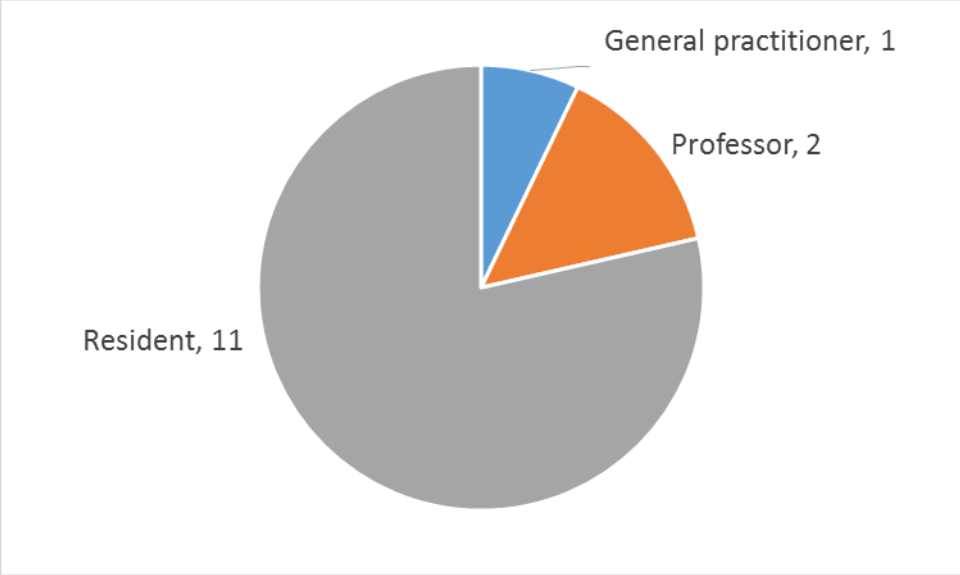


Figure 11 Number of participants by occupation

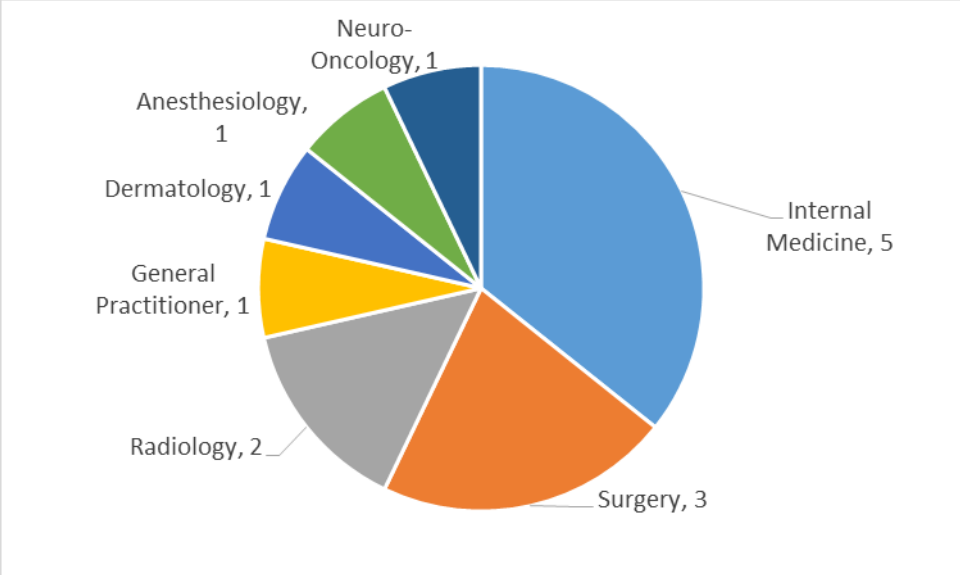


Figure 12 Number of participants by medical specialization

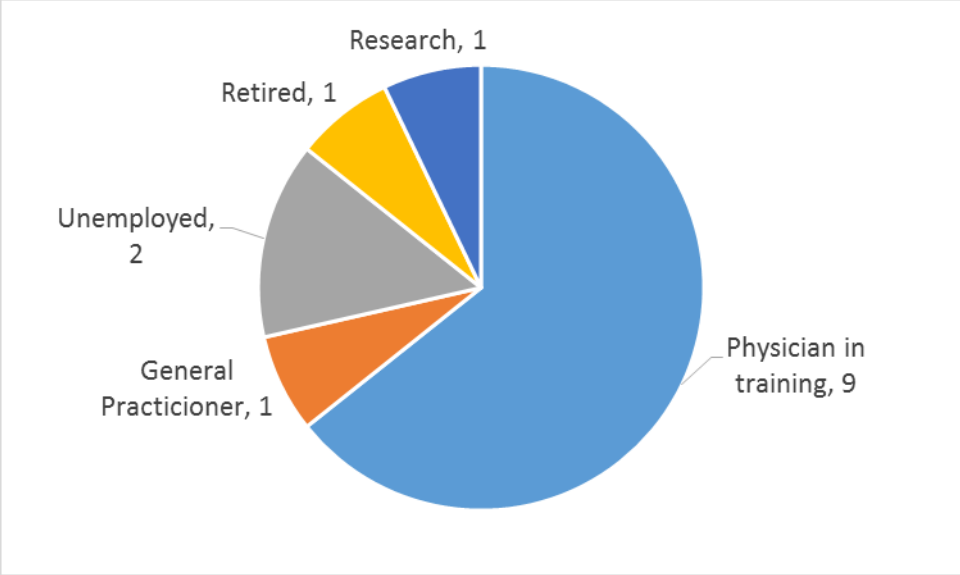


Figure 13 Number of participants by current occupation

All users had German as their mother tongue (9 Austrians and 5 Germans). The user interface of the Khresmoi prototype, the task descriptions and the questionnaires were all in English, as well as the sources in the search engine. Some popup windows from the Morae software had the question text in English and the answer buttons in German (Ja, Nein). The test users' user accounts were set to German, to enable the translation feature into German. The introduction given verbally by the examiner at the beginning of the tests was in German. All users spoke German during the test, asking questions and providing feedback to the examiner. The written feedback in the questionnaires was given in English by some test users and in German by others. Some users had difficulties expressing their thoughts in written English and used dictionary websites to find the words they were looking for while filling out the feedback forms. Overall, the test users seemed to be familiar with mixed language situations, where they read in English and discuss in German.

3.3.2.2 Tasks solved

Each user was asked to perform the same four tasks (see section 3.2.2), 12 users completed all four tasks, two users ran out of time and stopped after three, and one user had to end the test after the first task due to technical problems with the server. Table 3 below lists the time spent by the users on the tasks and the answers given before and after the task:

UserID	Task 1	Task 2	Task 3	Task 4	Total incl. questionnaires
physA	14:17 correct/correct	10:32 idk/nf	7:52 idk/correct	5:07 idk/correct	1:06:13
physB	26:40 wrong/wrong	11:45 idk/nf	12:06 correct/correct	8:39 idk/correct	1:24:34
physC	16:01 wrong/nf	10:05 idk/correct	10:35 correct/correct	---	1:06:17
physD	13:51 idk/wrong	13:00 correct/nf	12:16 idk/wrong	7:26 idk/~correct	1:05:26
physE	12:23 correct/nf	13:56 idk/correct	13:03 correct/correct	7:21 idk/~correct	1:14:18
physF	16:17 correct/nf	11:04 correct/nf	13:21 correct/wrong	7:27 idk/correct	1:17:39
physG	17:02 idk/---	---	---	---	0:19:49
physH	9:36 correct/correct	12:51 idk/wrong	7:08 correct/correct	7:16 idk/correct	1:02:34
physI	11:51 idk/nf	9:52 wrong/wrong	7:07 correct/correct	5:30 correct/correct	0:57:17
physJ	13:21 correct/nf	10:11 wrong/correct	11:37 correct/correct	---	0:59:05
physK	14:43 wrong/nf	8:44 idk/correct	9:03 correct/correct	6:40 idk/correct	0:58:18
physL	14:56 wrong/nf	13:56 correct/nf	8:13 correct/correct	6:31 idk/correct	1:04:26
physM	3:59 idk/correct	7:57 idk/correct	9:22 idk/correct	3:57 idk/correct	0:49:13
physN	9:26 idk/nf	6:49 correct/correct	11:30 idk/wrong	8:29 idk/nf	0:58:46

Table 3 Correct answer, wrong answer, I don't know (idk) or not found (nf)

Out of 51 total cases, the 17 (~33.3%) cases where using Khresmoi resulted in improved answers given by the test persons (wrong -> correct, idk -> correct) are highlighted in green. 5 (~9.8%) cases where using Khresmoi resulted in worse answers (correct -> wrong, idk -> wrong) are highlighted in red.

It should be noted that even though the tasks were defined with a 10-minute time limit in mind, this limit was not strictly enforced. If users liked to talk more during a task, the decision was usually that their feedback was more valuable than the time spent. Users were told that they were approaching the time limit, and then asked if they were finished, but nobody was forced to stop.

The following charts illustrate the relation between the answers given by the users before and after searching with Khresmoi, separated by tasks. The results greatly improved from task 1 to task 4.

Additionally, for each task the "favourite" most commonly used websites where users searched for answers are listed.

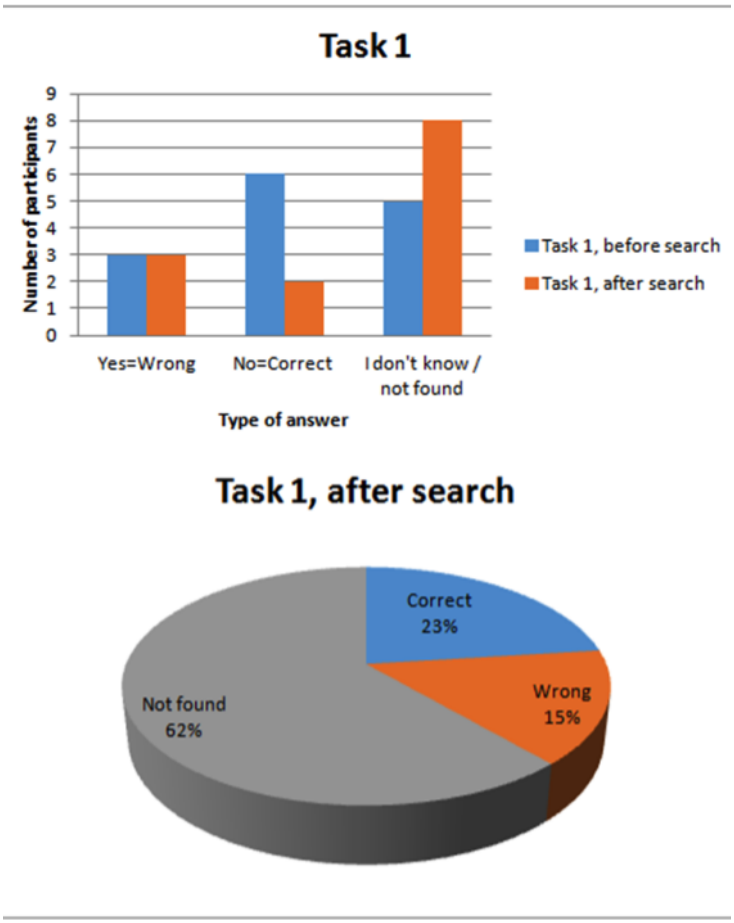


Figure 14 Initial knowledge and success in finding the correct answer for task 1

Most popular websites during search for Task 1:

0 http://www.tripanswers.org/ (7x)

1 http://fast.hevs.ch/images (3x)

2 www.uptodate.com (3x)

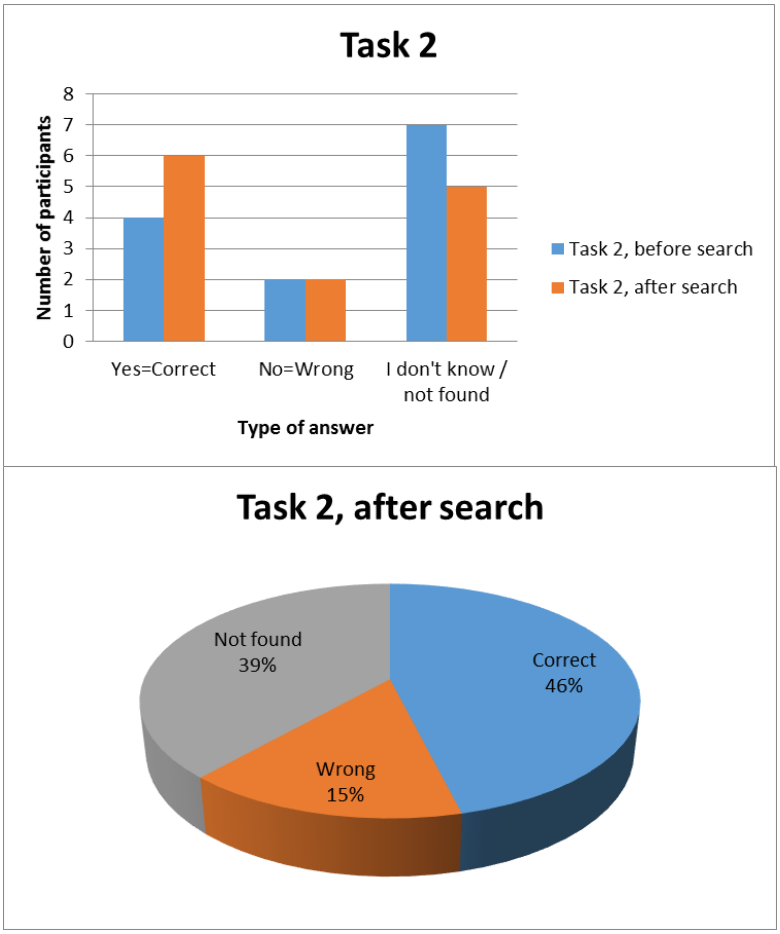


Figure 15 Initial knowledge and success in finding the correct answer for task 2

Most popular websites during search for Task 2:

0 http://www.endocrineweb.com/news/diabetes/1746-does-diabetes-raise-your-risk-cancer (3x)

1 http://www.cardiab.com/content/9/1/53 (2x)

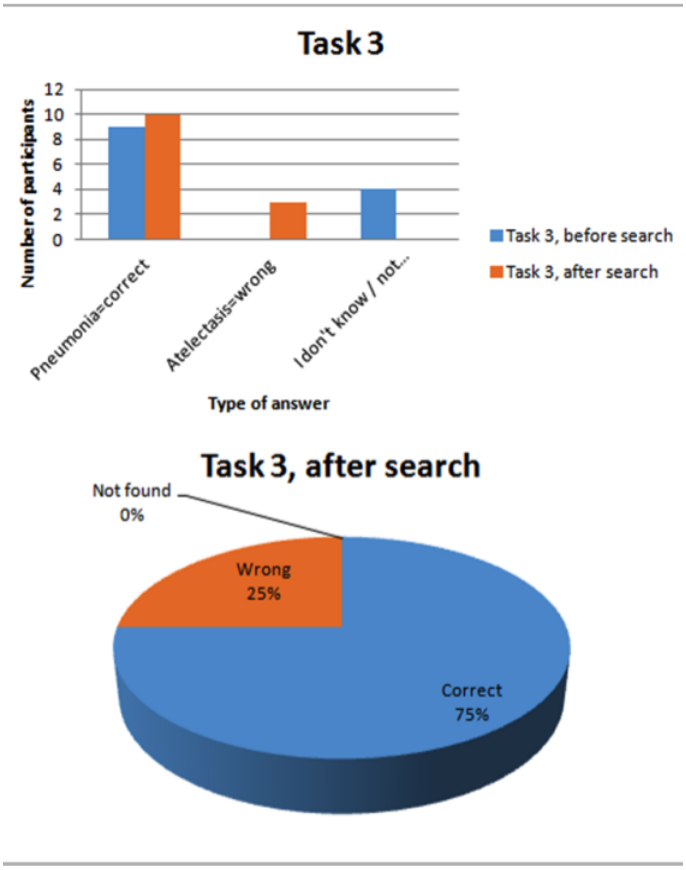


Figure 16 Initial knowledge and success in finding the correct answer for task 3

Most popular websites during search for Task 3:

0 http://fast.hevs.ch/images/ (9x)

1 http://www.ncbi.nlm.nih.gov/pmc/ (2x)

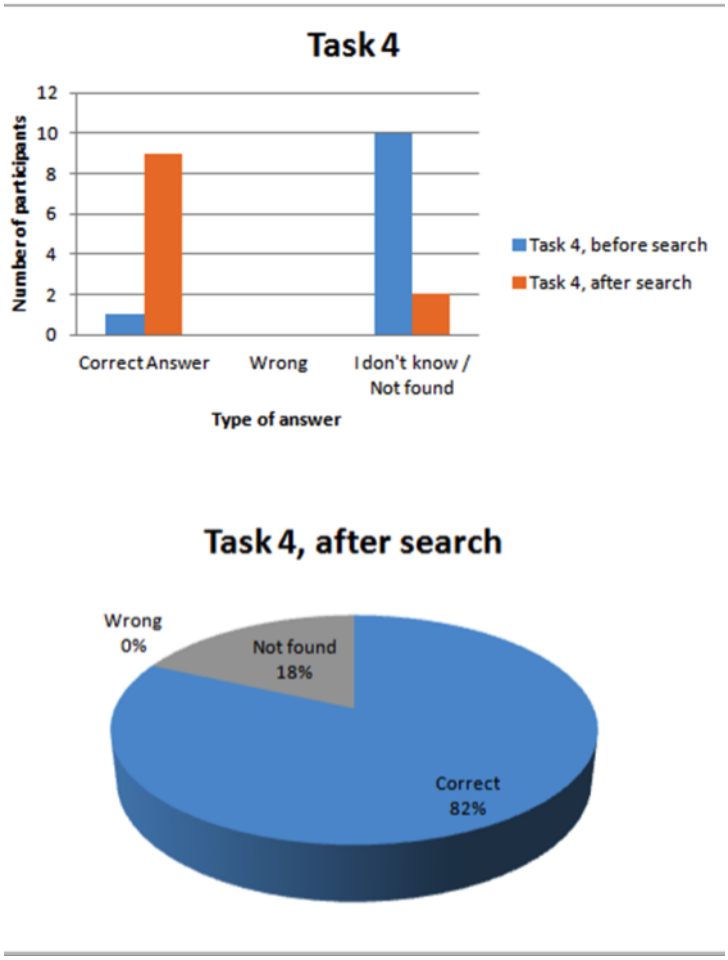


Figure 17 Initial knowledge and success in finding the correct answer for task 4

Most popular websites during search for Task 4:

0 http://ccforum.com/content/10/4/R112 (10x)

1 http://bestbets.org/bets/bet.php?id=1711 (3x)

2 http://fast.hevs.ch/images/other/imageclef2012/cc5668-7.jpg (2x)

3.3.2.3 Log files

The log files collected by ezDL on the server side during the user tests were subjected to elementary log file analysis methods. Due to the small number of users and queries, the overall amount of data was not large enough for many more advanced analyses. This first experience with ezDL and Khresmoi log files highlighted several issues and will help with the decisions on what to collect and analyse during future user tests. For example, with the logs from these tests, the lines belonging to different tasks had to be selected manually, because the task assignment and advancement is done by the Morae software and therefore remains unknown to ezDL. As more and more users will test the software during the coming months, this step needs to be automated.

3.3.2.3.1 Clicks

ezDL provides extensive logs of queries, ranks, clicks, and click-through. Figure 18 below shows the percentage of clicks in each result rank. It means, for example, that around 18% of the users clicked in the first result for the task 4 (in red), while more that 30% of the users clicked in results ranked 11 or more. The participants actually had a better experience, since most of them employed some filter or classification function on the result lists, thereby hiding many of the not relevant documents from the list. This compression of the ranking caused by the filters cannot easily be reconstructed from the log files, and will also need to be recorded differently to be analysed in the future.

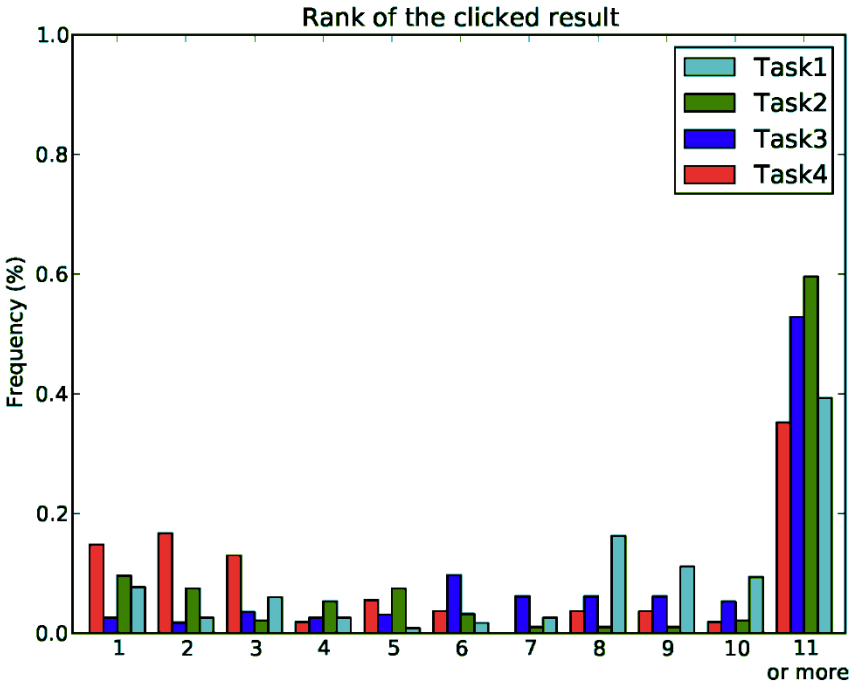


Figure 18 Percentage of clicks in each result rank

3.3.2.3.2 Spelling and entity suggestions

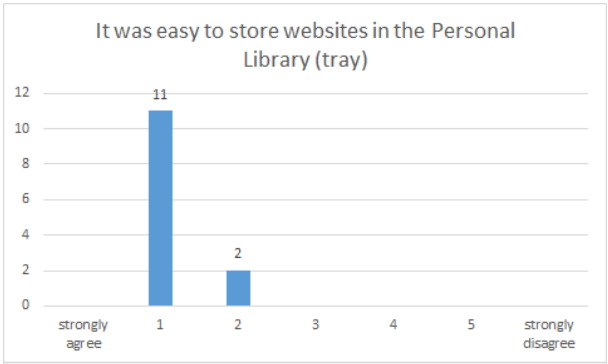
The prototype offered spelling corrections and semantic entities for queries to the users while they were typing. As can be seen from the user feedback, this feature was not very popular. From the ezDL log files it can be seen that a large part of the spelling mistakes made by the users are actually language mistakes: They use the German spelling of words that are very similar in German and English. The recordings of the tests show that the users often did not recognize these spelling mistakes even when prompted. Overall, the log files show many confusing replacements, suggesting a further need for improvement of this feature. Some example replacements are shown in the table below.

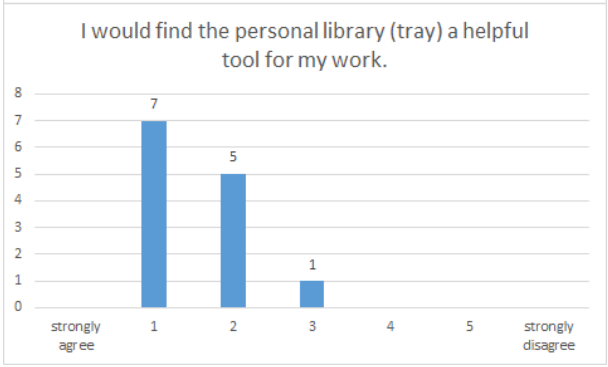
Query term as typed	Replacement suggestion accepted by user
assosiated	associated
malignent	maligne
anticoagulation	antikoagulation
fibrillation	Permanente atriale Fibrillation <[http://linkedlifedata.com/resource/umls/id/C2586056: Permanente atriale Fibrillation]>
x-rays	sprays
anticoagulation	anticoagulation<[language: DE]>
prophylax	prophylaxis
Diabetes	diabetes mellitus<[http://linkedlifedata.com/resource/umls/id/: Biologic Function, http://linkedlifedata.com/resource/umls/id/: Natural Phenomenon or Process, http://linkedlifedata.com/resource/umls/id/C0011849: Diabetes mellitus, http://linkedlifedata.com/resource/umls/id/: Phenomenon or Process, http://linkedlifedata.com/resource/umls/id/: Event, http://linkedlifedata.com/resource/umls/id/: Pathologic Function, http://linkedlifedata.com/resource/umls/id/: Disease or Syndrome]>
antikoagulation	anticoagulation<[language: DE]>
illnesses	illness

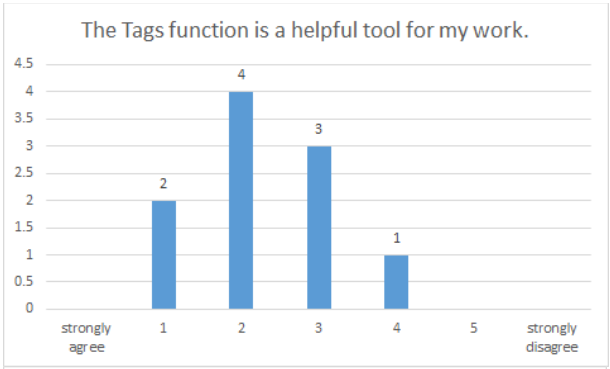
Table 4 Example query replacements accepted by users

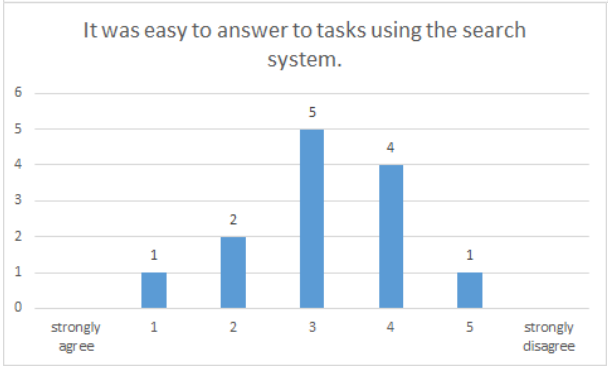
3.3.2.4 Overall final feedback questionnaire results

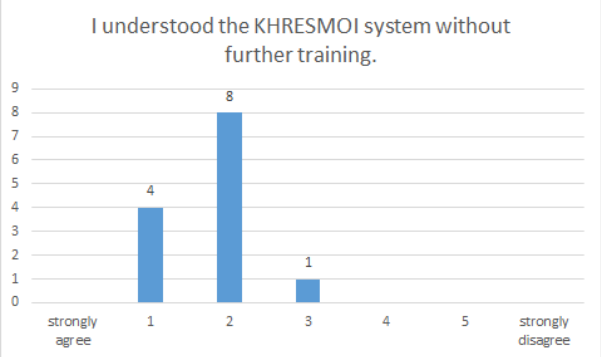
13 users answered the questionnaire for overall feedback that was presented to them at the end of the test, after completion of the last task. Not all test users answered all questions; for example not all had used the tagging feature that allowed them to add tags to elements stored in the Personal Library, so they did not answer whether they found this feature helpful. Additional, verbal feedback provided by users is summarized in section 3.3.3. The answers to the overall feedback questionnaire are summarized in the following charts.

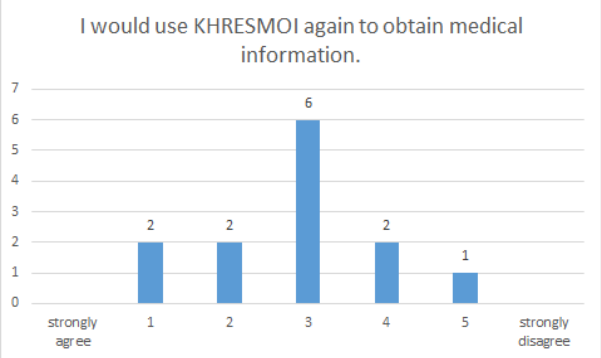

It was easy to store websites in the Personal Library (tray)
12
10
8
6
4
2
0
11
2
strongly agree
1
2
3
4
5
strongly disagree

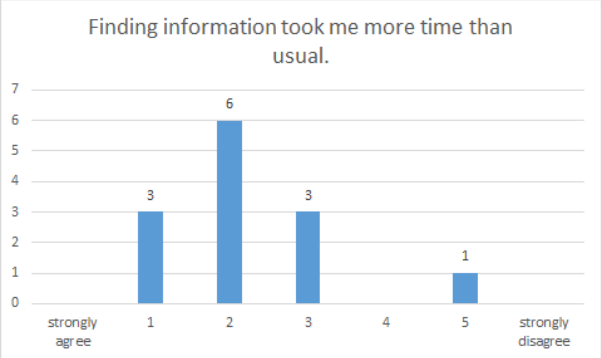

I would find the personal library (tray) a helpful tool for my work.
8
7
6
5
4
3
2
1
0
7
5
1
strongly agree
1
2
3
4
5
strongly disagree

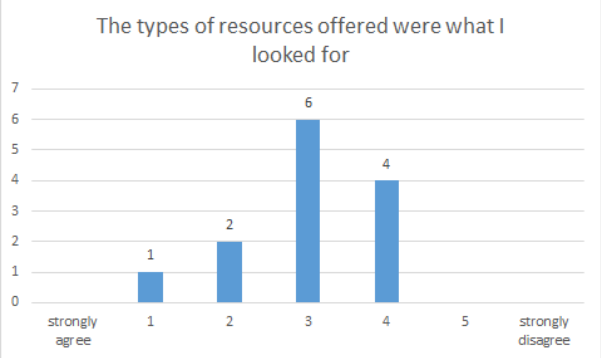

The Tags function is a helpful tool for my work.
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
2
4
3
1
strongly agree
1
2
3
4
5
strongly disagree

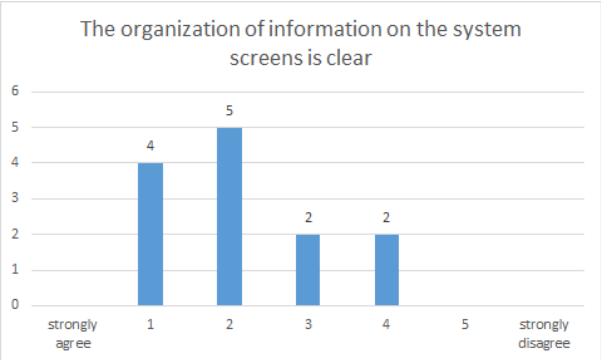

It was easy to answer to tasks using the search system.
6
5
4
3
2
1
0
1
2
5
4
1
strongly agree
1
2
3
4
5
strongly disagree

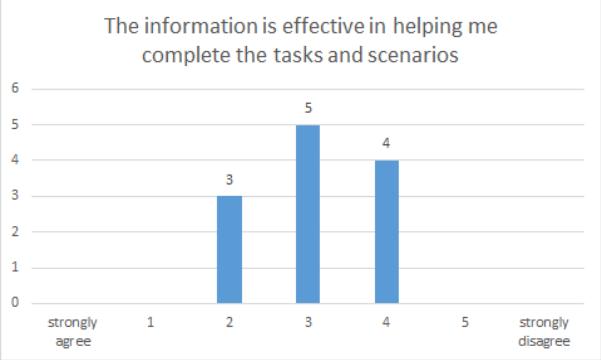

I understood the KHRESMOI system without further training.
9
8
7
6
5
4
3
2
1
0
4
8
1
strongly agree
1
2
3
4
5
strongly disagree

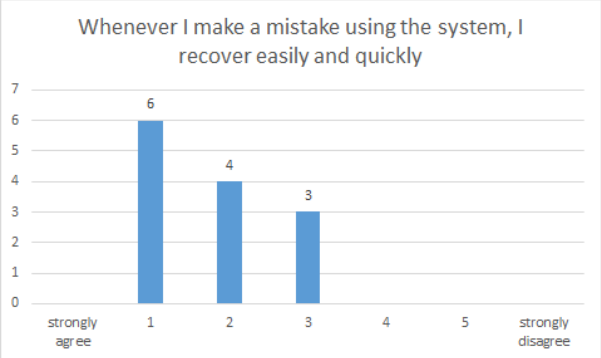

I would use KHRESMOI again to obtain medical information.
7
6
5
4
3
2
1
0
2
2
6
2
1
strongly agree
1
2
3
4
5
strongly disagree

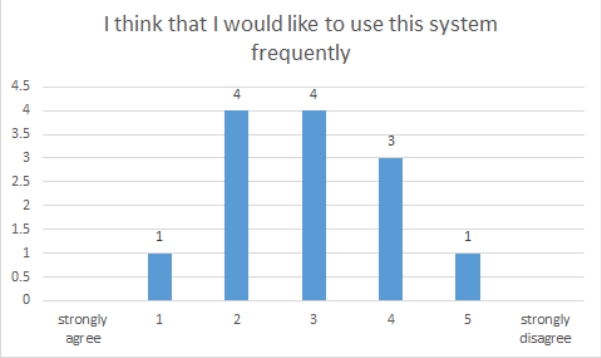

Finding information took me more time than usual.
7
6
5
4
3
2
1
0
3
6
3
1
strongly agree
1
2
3
4
5
strongly disagree

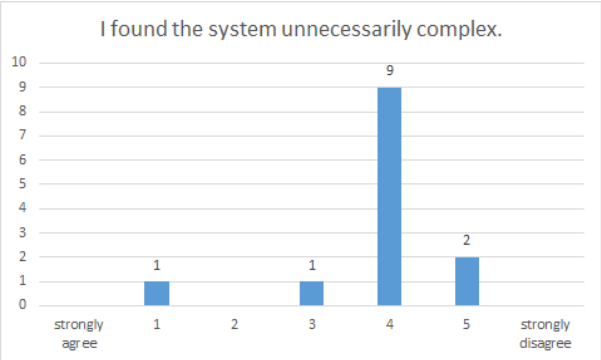

The types of resources offered were what I looked for
7
6
5
4
3
2
1
0
1
2
6
4
strongly agree
1
2
3
4
5
strongly disagree

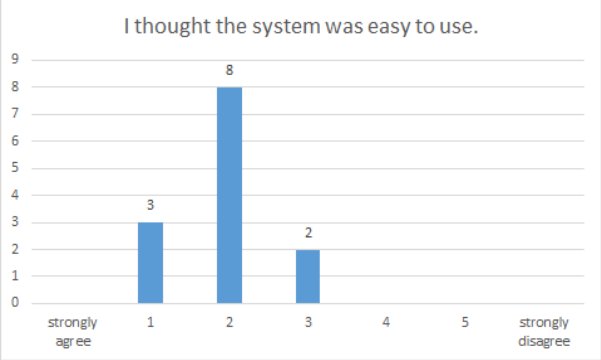

The organization of information on the system screens is clear
6
5
4
3
2
1
0
4
5
2
2
strongly agree
1
2
3
4
5
strongly disagree

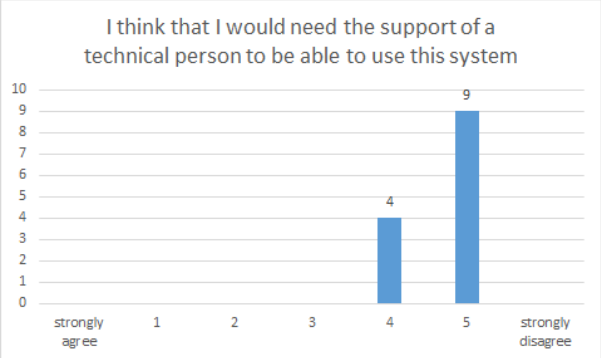

The information is effective in helping me complete the tasks and scenarios
6
5
4
3
2
1
0
3
5
4
strongly agree
1
2
3
4
5
strongly disagree

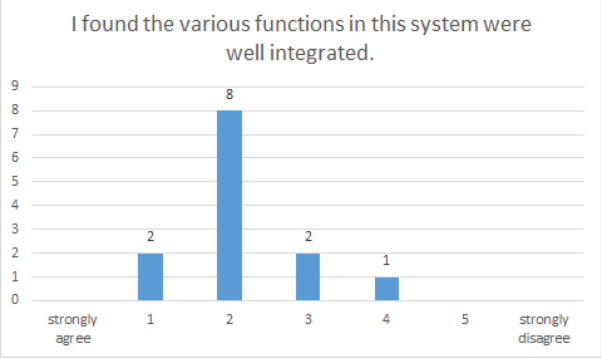

Whenever I make a mistake using the system, I recover easily and quickly
7
6
5
4
3
2
1
0
6
4
3
strongly agree
1
2
3
4
5
strongly disagree


I think that I would like to use this system frequently
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
1
4
4
3
1
strongly agree
1
2
3
4
5
strongly disagree


I found the system unnecessarily complex.
10
9
8
7
6
5
4
3
2
1
0
1
1
9
2
strongly agree
1
2
3
4
5
strongly disagree


I thought the system was easy to use.
9
8
7
6
5
4
3
2
1
0
3
8
2
strongly agree
1
2
3
4
5
strongly disagree


I think that I would need the support of a technical person to be able to use this system
10
9
8
7
6
5
4
3
2
1
0
4
9
strongly agree
1
2
3
4
5
strongly disagree


I found the various functions in this system were well integrated.
9
8
7
6
5
4
3
2
1
0
2
8
2
1
strongly agree
1
2
3
4
5
strongly disagree

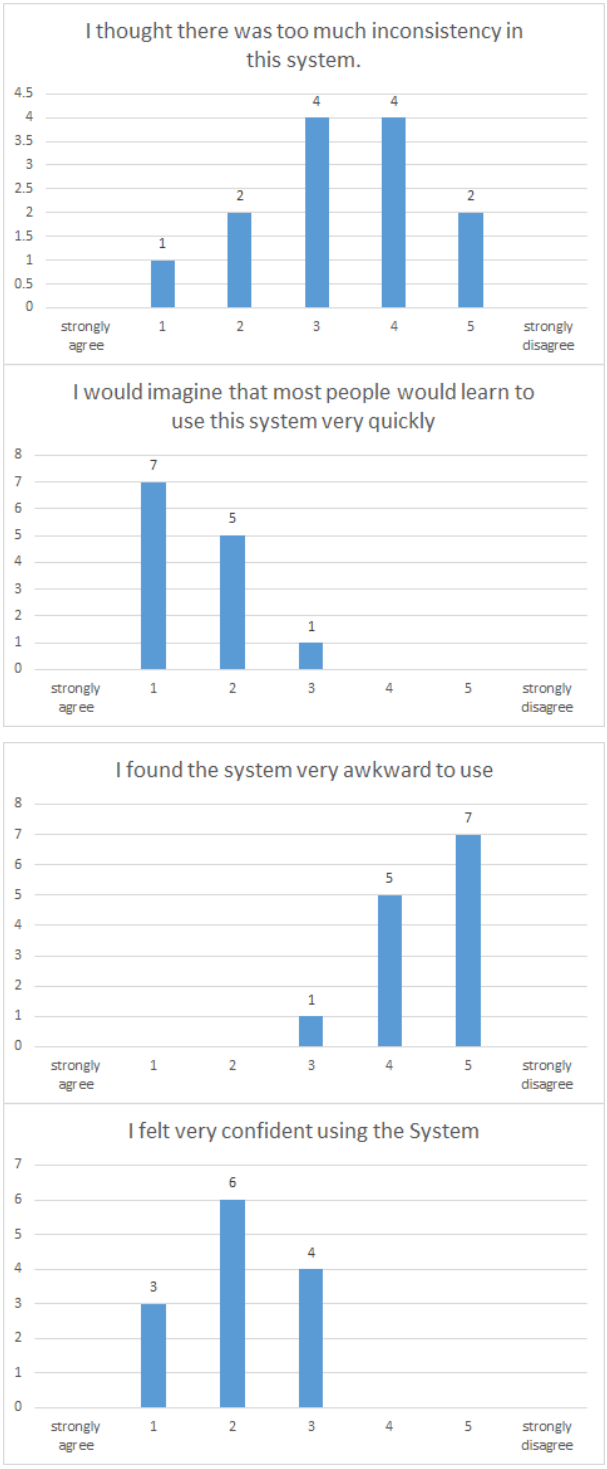


Figure 19 Number of participants by answers given in the overall feedback questionnaire (20 charts)

Participants agreed or strongly agreed with the following statements with a positive sentiment:

Question	Median	Mean
It was easy to store websites in the Personal Library (tray)	1	1.15
I would find the personal library (tray) a helpful tool for my work.	1	1.54

I would imagine that most people would learn to use this system very quickly	1	1.54
I understood the KHRESMOI system without further training.	2	1.77
Whenever I make a mistake using the system, I recover easily and quickly	2	1.77
I thought the system was easy to use.	2	1.92
I felt very confident using the System	2	2.08
The organization of information on the system screens is clear	2	2.15
I found the various functions in this system were well integrated.	2	2.15
The Tags function is a helpful tool for my work.	2	2.30

Figure 20 Positive statements with which the participants agreed (2) or strongly agreed (1)

They agreed with the following statement with a negative sentiment:

Question	Median	Mean
Finding information took me more time than usual.	2	2.23

Figure 21 negative statements with which the participants agreed

Participants were more or less indifferent to the following statements with both positive and negative sentiments:

Question	Median	Mean
I would use KHRESMOI again to obtain medical information.	3	2.85
I think that I would like to use this system frequently	3	2.92
The types of resources offered were what I looked for	3	3.00
The information is effective in helping me complete the tasks and scenarios	3	3.08
It was easy to answer to tasks using the search system.	3	3.15
I thought there was too much inconsistency in this system.	3	3.31

Figure 22 Statements to which the participants were largely indifferent

Participants disagreed or strongly disagreed with the following statements with a negative sentiment:

Question	Median	Mean
I found the system unnecessarily complex.	4	3.85
I found the system very awkward to use	5	4.46
I think that I would need the support of a technical person to be able to use this system	5	4.69

Figure 23 Negative statements with which the participants disagreed (4) or strongly disagreed (5)

3.3.3 Collected user feedback

This section lists the feedback given by the test users during the tests. It is a summary of the written feedback at the end of each individual task as well as the spoken comments during the whole test. Project members took notes during the tests using the Morae software, and afterwards all recordings of the tests were processed and carefully combed for comments of the users.

These comments were not evaluated statistically, as most numbers were too small and the risk of adding artificial weight to single statements was to be avoided.

3.3.3.1 Overall Satisfaction

The overall user satisfaction was very much linked to the perceived success in answering the task questions. When asked what they liked/disliked about Khresmoi, many times users simply responded "that I found the answer" or "that I didn't find an answer" or "that I found a supporting source for the answer I would have given".

- One user found the tool useful for a more in-depth search, but not for quick searches.
- Another user would only use it if it became much faster for mouse and keyboard inputs.
- At the end of a task, many users expressed their dissatisfaction with the data found so far, citing that they would need more scientific backing, and that they would not make a decision based on their findings, but then chose an answer for the sake of the test case.
- For task 1, one user claimed that the question was unanswerable without further information about the patient's current status (like a current ECG), then proceeded with the search, didn't find satisfying results and kept the original opinion.
- Several users mentioned that they normally use PubMed, are used to PubMed, would have looked on PubMed first, or that they would now look for a PubMed review on the subject. One user suggested that it would be a good idea of Khresmoi to work similarly to PubMed. "It would have been easier to just use PubMed", "Compared with PubMed, this system is rather unspecific and in terms of scientific research not equally satisfying".
- Users looking for only journal articles perceived Khresmoi as an unnecessary detour, having one more step via a website.

3.3.3.2 User interface and screen layout

- Two users thought that a higher screen resolution would have been useful to reduce the need for scrolling horizontally and vertically. One also believes that this makes mobile use not very likely.
- Another user would have liked an easier way to temporarily hide unused features from the interface to free up space for the task currently at hand.
- Several users mentioned multiple times that they found the user interface clear and well structured. Nevertheless, most users needed some help with the views of the interface.
- One user didn't find the query text box and then didn't know how to start the query.
- Many users needed help on how to open the personal library.
- Several users mentioned that the icons were too small or hard to find for them.
- One user did not like that the order of the icons (tray left or right of the library) was different in different situations.
- Two users inadvertently switched to the image search perspective by clicking on "add to image search" and didn't know how to switch back.
- Many users wrote in the feedback that they had problems deleting the old query before entering a new one. The query text field does not respond to clicks as expected (which would be two clicks to select a word, and three for the whole query), so these users proceeded to use the keyboard to delete every letter separately. There is an icon on the right end of the query text field for resetting the query, but these users apparently didn't expect such functionality and did not ask.

- The highlighted search keywords were mentioned as something that supports getting a quick overview.
- One user mentioned usually needing more time to get to know a program, and currently feeling that too much information is displayed at once.

3.3.3.3 Retrieval quality

- There were requests for both more and fewer results to be retrieved/displayed.
- Several users mentioned not seeing the relation between the query and a result document, with many documents on different topics that just happen to contain the query terms somewhere.
- There was an idiosyncrasy of the prototype discovered only after the first user tests: The three sources queried (displayed as "Medsearch", "ParaDISE" and "Textsearch") treated the queries differently. Textsearch by default combined the terms with Boolean AND, whereas ParaDISE (the image search) and Medsearch (medical articles) combined them with OR. This led to numerous occasions where due to one or more spelling mistakes in the query terms, only images were retrieved. How many results were found in each of the sources is displayed at query time and can later be seen by restricting the results by source, but this was far from obvious to the users. The resulting situation was that the users did not recognize the spelling mistakes, since a whole list of results was returned anyway, and spent considerable time analysing these results, becoming more and more puzzled about why these were displayed for their query and why they were all images. Due to the frequency of spelling mistakes and the Boolean OR, the connection to the query could be very distant, with only one of several query terms matching. After this issue was identified and discussed, it was decided to keep the settings but to watch for this situation and proactively point out any spelling mistakes in the query.

3.3.3.4 Data Export

- Two users requested the possibility to export selected library entries to Endnote.
- Another user requested exporting directly to PDF attached to an e-mail.
- The same user would like the HTML export to include just the link, not the whole page.

3.3.3.5 Multilingual features

- One user mentioned that he liked the translation help feature.
- Another user - who had selected "below average" for "understanding of medical English" - severely missed the possibility to formulate queries in German, as well as a translation option for all items found. The text in images cannot be translated, yet some searches returned many text-heavy images of charts or graphs. Also, for websites, the summary snippet can be translated, but the link to the whole page leads to the original page in English. This user mentioned typically searching the web in German only, with permanent settings to ignore English websites. This user thought that this need is very likely shared by a large proportion of Austrian general practitioners
- The lack of a "did you mean...?" feature was also mentioned in context with the need to formulate queries in English.
- One user mentioned not trusting the translation.

3.3.3.6 Spelling correction

- The spelling correction mainly was used when users accidentally used the German spelling of a word that is very similar in English.
- In one case, the spelling suggestion was wrong (it suggested the non-existent English word "treatmentschme").

3.3.3.7 Personal library

- The personal library was mentioned as a helpful feature by one user.
- Another user would have liked a sorting function within the library.

3.3.3.8 Query support features

- One user specifically requested a better possibility for multi-word/phrase search (after problems with restricting the search to type 2 diabetes in task 2).
- Several users requested a kind of helping function that would support them with the choice of queries, and also with the search process as a whole, as they struggled with finding the best balance between more general and more specific queries. The multi-level search possibilities with a full-text search first, followed by various filtering options (search within results, classification by topic, source, type, etc.) left many users feeling unsure about their search strategy. For task 4, one user mentioned that the two text query fields finally made sense for this task.
- For task 3 (x-ray image), one user would have liked the possibility to use the image directly as a query.
- Two users would have liked an option to search in the titles of the documents only.

3.3.3.9 Data Sources

The data sources accessible through the Khresmoi prototype was a major topic which most users commented on/asked about during many tasks.

- Some users liked the mix of high level overview websites and others with more detailed information.
- Another user liked the variety of content that makes it possible to view a subject from different perspectives (doctor, patient, etc.), and imagined using it to find web pages or forums to suggest to patients, but thought that for informing oneself on medical topics, other systems might be better.
- Another liked the lack of advertisements, as well as the good percentage of scientific material.
- Several users stumbled over external pages where access was denied. Users specifically requested access to UpToDate.
- Users would have liked more links to medical societies and clinical guidelines.
- One user specifically mentioned not liking the uncertainty about the quality of the sources and would have liked to know that they are of textbook quality. Another user asked for data on the reliability of the sources.
- Overall, many users had doubts about the quantity of the data and repeatedly asked the facilitator whether certain sources were contained in the prototype and supposed to be findable. This was most likely caused by the prototype cutting off the result list at 200 results which - especially if the list was then further filtered - left the users with a handful of

documents for topics where they knew that many more exist. This led some to believe that the search did not find more than 200 documents to begin with, even when told otherwise.

- Users liked that there were many different images for a topic, as well as finding ultrasound images.
- Another user did not find pictures helpful to solve complex problems.
- Several users would have liked descriptions or tags for the radiological images.
- Several users did not like finding a website which links to a scientific article and would have preferred the direct link.
- The full-text search occasionally returned documents on different topics where the search term was only contained in a link.
- The three-backend sources queried by the prototype were displayed at query time as "Medsearch", "Textsearch" and "ParaDISE", with no further explanation, which confused some users.
- One user was really happy to find out that "Medsearch" contains only scientific articles and from then on only used this source, ignoring the others.
- One user believed that there were only US sources and no European ones available because only results about "warfarin" were found, which he claimed is used in the US but not in Europe. The user did not succeed in using the name of the drug usually prescribed in Austria (ThromboASS).

3.3.3.10 Classification and Filtering of Results

Opinions were very much divided on the classification/filtering options next to the result list.

- Several users achieved very fast, good search results by combining more general text queries with a click on the right classification subgroup. The classifications thus were mentioned by several users as a feature they liked and as a highlight of their search experience.
- Other users were not so lucky, and thus there are also comments that the filters do not work, should be reduced, removed, or offered as advanced option only by user request.
- Several users mentioned that the choice of subgroups was unclear to them, e.g., why "Radiology" was a subgroup during the x-ray task. One user requested a topic "Therapy", another user would have used "Guidelines", and another would have expected different information under "Disease Biology". For task 3 (diagnosis of an x-ray image) one user would have searched for the answer under a "Diagnosis" subgroup.
- Most users liked the possibility to restrict the results to images only.
- In sync with the overall desire for trustworthy, scientific sources, many used or tried to use the classification feature to focus on scientific articles and studies only, but did not always know how to do so.
- One user would have liked a negative filter option ("these results but without those that contain the word X").
- One user collapsed the classification and did not know how to reopen it.
- One user would have liked a "back" or "undo" function for the filter, or even for all actions.
- One user mentioned not looking at the filter options at all because of not needing them.

3.3.3.11 Speed

- There are several comments on how fast the search was and how quickly the user was able to find an answer.

- Other users found the system “annoyingly slow”, or mentioned a fast search with a slow user interface. Specifically, one user found the cursor input too slow, and another the scrolling.
- One user mentioned that with only 10 minutes for the task, Wikipedia would normally be the tool of choice, as it provides good summaries.

3.3.3.12 Result presentation and Preview

- The prototype displays some metadata about the items on the result list, but several users would have liked more.
- They requested a clear indication of whether the result page has a “scientific background”, the type of source in general (e.g., a forum), as well as a differentiation in scientific articles vs. textbooks.
- There were numerous requests for the publication year to be displayed very prominently.
- The descriptions/summaries given for the items on the result list would have been preferred to be more substantial, to avoid having to open many documents to check, even more so for images results.
- For studies, tables should be included in the preview.
- In one case, a user found that the summary was misleading and that the whole page was about something else.
- A user requested PubMed reviews dealing with the search topic to be displayed on top.

3.3.3.13 Usability

- Several users mentioned that if a classification filter is applied to the result list, it should not be possible to get the removed items just by scrolling up or down (a click on the classification subgroup scrolls the list to the section with those items, but it does not remove the others).
- One user found it annoying that if a preview is opened, it is no longer possible to use the down button to navigate the result list.
- One user would like the filtering to occur in a separate window.
- Several users were surprised/confused by having to switch to the browser to look at websites and then back again to continue the search.
- Some users expected a tabbed window interface in the Khresmoi search application for opening several previews in tabs.
- One user found that the mouseover-event in the result list (popup of the title) was too slow and got in the way while scrolling. Several users wondered why the title was cut off in the popup.
- In the preview window, one user accidentally clicked on the keywords instead of the link to the article and did not know how to return. The user restarted the query to retrace the steps instead.

3.3.3.14 Images

- One user would have liked that clicking on an image gets the full size image.
- Another user would have preferred to find higher resolution images.
- One user mentioned that for the x-ray image task it was helpful to find many images and see the thumbnails in the result list.
- Another user liked the possibility to compare x-rays.

3.4 Discussion

There is a sample bias with an over-representation of young, Austrian physicians in training: Most (9/13) participants were below 29 years or older. Middle-aged physicians were under-represented. Some participants were particularly hard to acquire; there were no working medical professors, medical specialists and only one general practitioner included in this round of user tests. Most participants spoke English well and were adept at using the Internet. This might explain the perceived skilled use of Khresmoi. They found the platform well organized and 12/13 stated that they understood the search system without further training.

Spelling mistakes in query inputs were mainly linguistic language-confusion mistakes. Users mixed German and English spelling and jinxed the spelling correction/language detection.

From the final survey, the usability of the system seems to be high. Most cited that the system was easy to use, needs no further training, functions were well integrated, it was consistent and they felt confident using the system. But only 3/13 users reported that they found it easy to solve tasks using the system.

Only 4/13 users stated that they would use Khresmoi to retrieve medical information (most users remained neutral on this issue). The most negative comments revolve around resources and ranking, and most users (9/13) reported that it took them longer than usual to find answers. Many users were accustomed to information access solutions such as PubMed and UpToDate and found that Khresmoi did not offer an advantage to them. Only 3/13 users reported seeing the types of resources they were looking for and only 3/13 users reported that the provided information helped them to solve the tasks. Most users consulted secondary and tertiary resources (only 2 cases cited primary resources as sources of evidence).

There might be a difference between general practitioners and specialists, but the participants of this round of user tests do not allow us to make this distinction. Also, the comparatively young professional age of most participants might have produced a bias towards resources used at university. Only one user talked about finding webpages to recommend to her patients, whereas most other users found websites not scientific enough and sought to avoid them.

Finally, in 31% of cases using Khresmoi resulted in better answers to the tasks, but in 10% in worse.

3.5 Conclusions

In general we can conclude that the search system as it was made available for the physicians' user tests is functional and can be used to solve the tasks given to the participants.

Nevertheless, the results indicate that the prototype is not yet ready for a medical real-life situation. Only more than half of the tasks were solved correctly within the timeframe given (see Section 3.3.2.2), and more than half of the participants answered the – less modified – tasks 1 and 2 wrong or did not find the answer.

Important results in the overall feedback questionnaire (Section 3.3.2.4) like the "time spent to find the information" or "the types of resources offered where what I looked for" are negative. In the collected user feedback (Section 3.3.3) the participants "expressed their dissatisfaction with the data found so far, citing...that they would not make a decision based on their findings..". So there is room for improvement.

First and foremost the data sources included in the index, their topicality, relevance, timeliness and completeness need to be extended and improved. Local and most recent resources should be displayed first, and more secondary and tertiary sources are needed.

During the time the user tests were conducted on the production version of the system, promising advancements were made in the development versions. Combined with an improved ranking --

especially of the merged search results from heterogeneous sources -- and an augmented offer of metadata in the result list and previews, these changes could satisfy many requests and dispel many of the doubts voiced by the participants.

The classification and filtering of results hold a lot of leverage with regard to both user satisfaction and quality of results. Those users who had used these features well fared better than those that had not. The usability of the interface apparently was quite different for different users.

Translation help is a feature that could win many users for Khresmoi if it becomes more widely and deeply integrated. As it is now, users who depend on the translations are left alone with many untranslated sources such as text-heavy images as well as the full websites they select based on the translated snippets.

It will be crucial to make the differences between Khresmoi and other information access solutions such as Pubmed obvious to the target audience, and to explain how and in which cases these differences can be useful for them.

Overall, this round of user tests has been very successful in uncovering bugs and errors, many of which have already been fixed. Additionally, on the meta-level, improved internal processes for reporting problems, managing support requests, and scheduling of downtimes were defined during this time.

The methodology of the user tests themselves was tested, and plans for future tests were improved, such as which information needs to be logged, or how much time is needed to schedule test sessions with busy physicians.

Further rounds of tests will be crucial for the continued improvement of the “Khresmoi for health experts” search system. As the project approaches the end of year three, many new features become available in the prototype, and these should be tested on an on-going basis, before the large scale tests of the final product are conducted in 2014. This will require an even tighter knit communication between the technical WPs and the use cases.

These future tests should distinguish between use cases for different user groups, such as general practitioners vs. specialists, together with a widened range of tasks (which was not possible now due to the lack of data sources). A possible approach could be to combine the regular dissemination activities at medical conferences with focused user tests.

The next steps (which will be planned at a project meeting immediately after the deadline for this deliverable) will be to define a framework for small, on-going tests of new or improved versions.

4 General public evaluation tests

4.1 Background

Based on the layman survey results published in deliverable D8.1.1 (D8.1.1, 2011), end users want the search results to be more relevant, more trustworthy, more readable and more comprehensible and in sufficient amounts. Better user experience hence can be defined as overall system satisfaction (ease of use, speed) and convenience with useful functionalities and tools (spelling correction, query completion and translation, translation of results, etc.)

We have defined three general characteristics of the search process:

- system performance which can be measured both subjectively (how fast an answer was found) and objectively (recorded in logs)
- search success which can also be measured both subjectively (whether a user has been able to successfully find information and answer the question in his/her opinion) and objectively (whether user's answer is correct or not)
- user satisfaction can be measured only subjectively

Based on this, the evaluation goals were:

- Evaluate the search success, as well as the success in using the tools; which tool yields the greatest benefit for online health information research.
- Evaluate what are the possible benefits of Khresmoi over Google. What are the criteria for the users to judge the list of search results? Do users tend to prefer Khresmoi to Google results in terms of their quality and trustworthiness? What are users' relevancy criteria? Is the coverage of resources available in Khresmoi sufficient?
- Study user queries and overall search behaviour when asked to perform health-related tasks.

It is important to know variables such as age, gender, level of English, Internet experience and health experience (ideally health literacy, but it will become too complex and unfeasible given the limited time of the evaluation session).

4.2 Research questions

1. Do layman users get better results and have better user experience (in terms of more relevant search results, faster times and more comprehensible results) using Khresmoi compared to their previous experience of online health searches? What are their criteria for quality and relevance of search results? Are they satisfied with Khresmoi results compared with general search engine results?

2. Does the outcome of the search correlate with user profile (age, gender, mother tongue and level of English, Internet/web search experience and health knowledge and experience in a given topic)?

3. Which aspects of the system are already "good enough" and used by users? Which aspects of the system need to be changed? Which tools and functionalities are not "good enough"? What is missing?

4. Study users behaviour (and compare with physicians): how they change queries, how they use tools, which interface they prefer and collect other "difficult to structure" feedback on what is good and what is not good to report back to the developers team.

5. Study the use of translation services: (a) Is it useful to allow the user to ask questions in their native language / their language of choice? (b) is the quality of the query translation (and other support multilingual tools) sufficient for users to bother (for any reason - which)? (c) Is the possibility to get translation of foreign-language documents (= contents) a plus? (d) Is the quality of the translation of the foreign contents sufficient for the user to (d.1) open it and possibly ask for more translation, automatic or otherwise and (d.2) get the information s/he wanted directly from the Khresmoi translation? (e) Is the fact that the interface is in user's language preferred, or would users prefer an English interface regardless of whether they use their language for formulating queries/translating results? All five questions should be based on user's native language and English language abilities (in addition to all other demographic and experience factors).

4.3 Hypotheses

1) Usability of the search system: at least some of the available tools are helpful (for example, translation of English document for users with poor English).

2) User satisfaction depends on certain variables:

- Health/disease knowledgeable and experienced users are going to type medical vs. layman terms and to look for more detailed information in more complex resources, and also seek for other patients experience on web 2.0 platforms (user-generated content, e.g. forums and blogs)
- Internet-experienced users are going to look for more search functionalities
- Disease-knowledgeable and technically-savvy person would be interested in trying and using the advanced options integrating semantic search capability.

3) Users who are more knowledgeable in the health/medical domain appreciate higher quality of Khresmoi search results over a general search engine.

4) Having learnt about the "trustworthiness" and rigor selection of resources for Khresmoi, participants will turn to Khresmoi.

4.4 Aspects to evaluate

- Success with task completion (search results relevant to the question asked or not, has been able to find the information about given topic or not)
- Rapidity with task completion and system performance (how quickly the system responded and how long users had to wait for a task to be completed)
- Overall user satisfaction (like/ do not like, what requires change)

- Tools offered by Khresmoi: whether users use them and whether they are satisfied with their quality and presentation, namely, query suggestion and completion, definitions, classification/filters of results, images, and query and results translation.
- Overall preference to Khresmoi search results or Google, user's selection criteria of certain results from the list of results

4.5 Experiment Design

Initially, the evaluation tests were planned in October-December 2012. Having conducted the first internal pilots in October 2012, we have seen that the system requires quite a few adjustments before being exposed to and tested by the "real" patients. Hence, two stages of evaluation have been proposed and agreed on:

1) blind comparison of search results, and users' preference towards Google or Khresmoi search results, which was designed and conducted from December to February 2013. Further stages of this evaluation are foreseen during Autumn-Winter 2013-2014.

2) "full user tests" which will study all evaluation aspects.

The second one was designed beginning in June 2012 (meeting in Dublin, Ireland) and was gradually refined and changed based on the intermediate evaluation results. The pilot tests were taken with overall 11 participants. Initially 11 tasks were developed to cover various health topics with the idea of randomizing tasks among the participants. Due to Morae limitations, it has been decided to choose and include 5 tasks, which were further simplified after the first pilot tests. It was also noted that 5 tasks were time consuming and extensive for some of the participants, so eventually 3 simple tasks remained for the user evaluations. The "full user tests" were conducted by HON, CUNI and ELDA in May-July in Prague, Geneva and Paris.

4.6 Methods

The two following methods were used to quantify the measurements mentioned above:

1) blind comparison of search results allowed to compare the results of Khresmoi for Everyone and Google and let the end users decide which list they prefer and what are the results where they could find an answer. *Khresmoi for Everyone* contains only the 8,300 websites, which have been certified according to the HONcode standard created by the HON Foundation. It was conducted in French only. For the first stage the participants were recruited among the students of the Faculty of Medicine of the University of Geneva (Switzerland) via a newsletter emailed to the student groups by the library and also flyers (appendix 7.2) and posters distributed in the library. Follow-up evaluations will include a broader sample of participants to represent various types of online health information seekers. Findings of the tests conducted in February 2013 are presented in this deliverable.

2) Full user tests with the Morae recording software – similar to physician and radiologists evaluation. The intermediate and final outcomes of these evaluations are presented in the section 4.8 later in this deliverable.

4.7 Blind comparison of search results

4.7.1 Description of the Session

Each session lasted for 45-50 minutes and included:

0 Welcome and consent form (appendix 7.3) signing.
1 Step 1: short demographic and Internet usage questionnaire,
2 Step 2: presenting a health scenario and corresponding query,
3 Step 3: participants examining the two lists of the first ten results provided by each search engine under comparison and choosing the preferred results,
4 Step 4: repeating steps 2 and 3. Participants are asked to do overall three scenarios.
5 Step 5: once the session is over, collecting informal feedback about evaluation design
6 Thanking participants with refreshments

4.7.2 The platform for evaluation

The evaluation tool was developed internally as a PHP 5.3 and JavaScript web application, relying on a MySQL database and on the Prototype 1.7 framework for JavaScript portability. The web server used was one belonging to the Faculty of Medicine of the University of Geneva. It is hosted at the following URL: http://cmulinux4.unige.ch/kgc_hybrid/home.php Screenshots of the platform are available in the appendix.

The session begins with the participant entering his/her *username*, composed of the first letter of his first-name followed by the first letter of his last-name, and an alphanumeric *password*, MD5. The application identifies whether it is the first time that the user used the application or if he had already done one scenario in the past.

The first time that the participant performs the test, he is asked to answer to a demographic survey and an Internet-usage survey (See appendix). The participant could not move further if all questions were not answered.

Out of the eight scenarios available, one was randomly assigned to the participant. The second and the third scenarios were randomly assigned from the remaining ones after the completion of the previous one. None of the participants are ever assigned the same scenario.

The main task is to explore search results displayed in two different indistinguishable lists, the one provided by Khresmoi (www.khresmoi.honservices.org) and Google (www.google.ch) with the regional option to French in both interfaces. Each list of first ten results was provided using predefined queries. For each scenario, the first 10 results' title, URL and snippet of both search engines were transferred to the evaluation tool database. All graphic or style information were ignored. In case there were more than ten results on the first page (i.e. advertisements provided by Google as top results), only the first ten links were taken into consideration, hence advertisement links were included. This action took place between the 8th of February 2013 8:00 p.m. and 10th of February 2013 12:00 a.m.

In the evaluation tool, the two lists are identified as A and B, and are randomly presented to the user on the left or right side of the main tab. The user can therefore have all the results at a glance for exploration and then click on a result title. The website is shown in a popup window, allowing the user to navigate the webpage. Each time the user clicks a result,

the system saves the click location (URL of the clicked result) and the time (in the form of a timestamp).

After exploration of the search results, the participant was asked to select a total of five preferred results from both lists through an HTML form. The system records the time and location of each click made over the form. In this study we analysed and presented the three top results (hereafter defined as preference one, two and three in descending order of preference) as they were the most important and significant for the participants based on their informal feedback.

4.7.3 Health scenarios

In order to demonstrate a possible preference by the participants for the top results of *Khresmoi* or *Google* for online health information in general, the pre-defined health scenarios were designed to represent common online health information needs of end users. Health scenarios were identified based on the most popular health topics on the Internet according to Pew Internet & American Life Project (Fox, 2013). We intended to make scenarios real-life and of equivalent difficulty.

Based on telephone interviews, a nationwide US survey developed by Pew in 2012, identified the 12 most-searched online health topics (Fox, 2013). Although the study strictly focused on the US population, we could assume a close similarity in the most searched online health topics between USA and Western Europe where socio-economic characteristics, as well as the use of the Internet, are similar. No equivalent large-scale survey is yet available for Europe (Higgins, 2011).

The following list of most searched **health topics** was identified in the report:

- specific disease or medical problem
- certain medical treatment or procedure
- how to lose weight or how to control your weight
- health insurance
- food safety or recalls
- drug safety or recalls
- a drug you saw advertised
- medical test results
- caring for an aging relative or friends
- pregnancy and childbirth
- how to reduce your health care costs
- any other health issue

These topics also go in line with the findings of the large-scale surveys conducted by HON in 2010 and 2011 (Pletneva, 2011), (Pletneva, 2012).

From the list of 12 most searched health topics, we have decided to exclude "health insurance" and "how to reduce your health care costs" as these specifically address the US context and focus on healthcare systems, rather than health problems themselves and their treatments topics. We have also excluded "any other health issue" as it was too vague to define a scenario. The health topics "drug safety and recalls" and "a drug you saw advertised" were jointly considered to define a single scenario.

"Specific disease or medical problem" and "certain medical treatment of procedure" were the online health topics most searched by Americans in 2012. Based on the information provided by WebMD.com for 2010, a well-known American health website (www.webmd.com), the 2012 survey by PEW further identified the most sought health queries:

Specific disease or medical problem

- Shingles
- Gallbladder
- Gout
- Haemorrhoids
- Lupus
- Skin problems
- Allergies
- Heart disease
- Diabetes
- Sleep disorders

Certain medical treatment or procedure

- Pain relievers
- Antidepressants
- High blood pressure medication
- Corticosteroids
- Hysterectomy
- Diabetes medication
- ADHD medication
- Antibiotics
- Colonoscopy
- Cholesterol-lowering medication

We have randomly selected a medical problem from the two lists above to determine a health scenario for each of the two health topics: "Gout" for "Specific disease or medical problem" and "ADHD medication" for "certain medical treatment of procedure". As we do not have more detailed data on queries for other health topics identified by PEW, we have

identified one of possible queries for each health topic. Health forums, news and other types of information on the Internet have served as inspiration to define these hypothetical health scenarios. The queries that we defined were short and concise, including keywords (opposite to natural language question), trying to imitate the most common end-user search behaviour (research in progress on the analysis of HON logs, D8.2).

Eventually we have formulated eight queries and corresponding scenarios, one per each health topic of interest (Table 5 and more detailed in the appendix 7.6). This potentially ensures adequate representation of the "real-life" online health information needs of the general public. The scenarios were validated by a physician, a pharmacist and a biologist in the HON team. The scenarios are adapted to a general level of understanding.

Scenario	Health topic defined by Pew	Health query (full scenarios in the appendix)
1	specific disease or medical problem	*goutte régime (gout diet)*
2	certain medical treatment or procedure	*TDAH médicament stimulants effects (ADHD stimulant drug effects)*
3	how to lose weight or how to control your weight	*perdre poids rapidement (lose weight fast)*
4	food safety or recalls	*lait cru risques (raw milk risks)*
5	drug safety or recalls / a drug you saw advertised	*risques trop antidouleurs (too many painkillers risks)*
6	pregnancy and childbirth	*amniocentèse risques (amniocentesis risks)*
7	medical test results	*Sida aide Gèneve (AIDS help Geneva)*
8	caring for an aging relative or friends	*aide personne-agée maison (at-home elderly care)*

Table 5 Health topics and corresponding queries

4.7.4 Results

A total of 29 participants took part in the evaluations. Three participants were excluded because they did not match the inclusion criteria. Specifically, one of these three participants was a physician and not an undergraduate student of medicine anymore. The two other participants excluded were an experienced physician and a pharmacist of 48 and 38 years old respectively coming from Africa to take part in postgraduate medical studies at the Faculty of Medicine of the University of Geneva. The final study group included a total of 26 undergraduate students of the Faculty of Medicine of the University of Geneva.

4.7.4.1 Demographic survey

From the 26 students, 18 were males and 8 were females with an age of 21.03±1.95 (mean ±sd) years old in average. Specifically, 9, 6, 7, 2 and 2 students were in their 1st, 2nd, 3rd, 4th and 5th year of study respectively. All the students were francophone (Swiss and/or French nationality). In addition, all except one (96.15%) confirmed at least an intermediate level of English and 19 out of 26 (73.08%) confirmed a good to very good level of English.

4.7.4.2 Internet usage survey

All the students confirmed a daily use of the Internet and 25 out of 26 (96.15%) agreed to be well or very well experienced in information search on the Internet. However, only half of the students confirmed to make daily Internet searches in English, while for the other half the frequency were weekly, monthly or even yearly. For their general online searches, the great majority of students reported use of Google (25 out of 26 = 96.15%), while only one student selected Bing.

For online health searches, 14 out of 26 (53.85%) of the students seemed to be only occasional online health information seekers. On the other hand, 3 out of 26 students (11.54%) confirmed to search online health information concerning their own health or that of their relatives more than once a day. The frequency of online health search for the rest of the participants was weekly or monthly.

All the students except one (25 out of 26 = 96.15%) agreed to search online health information concerning at least one of the eight suggested health topics. Specifically, 20 out of these 25 (80%) searched in more than one health topic (mean number of topics searched ± sd = 2.53±1.83). "Specific disease or medical problem" and "Certain medical treatments or procedures" were the most sought health topics with 23 and 18 cases respectively.

4.7.4.3 Criteria of quality of health information of students of medicine

In general, when asked about their trust with regards to online health information, 21 out of 26 students (80.77%) stated that this was dependent on different factors and that the quality of online health information and results could be generally improved. In addition, 2 and 3 students were normally satisfied and unsatisfied respectively with the quality of the online health information and results.

Specifically, based on their open explanations, the most common criteria to trust online health information among students was the presence of references (14 out of 26 = 53.85%). Other important aspects were language/vocabulary, the way the information is presented/structured and information about the author (8 out of 26 = 30.77%). Seven participants also mentioned another trust factor as being whether the web site was a general health site or was one that provided information on a specific health topic (26.92%). Fewer participants identified presence of bias or advertisements as a quality evaluation criteria (4 out of 26 = 15.38%), while the same number paid attention as to whether the information explained on the web site corresponded to previous knowledge (15.38%). 4 out of 26 (15.38%) stated they did not trust forums, while two, on the contrary, were interested in checking other users' input (7.69%).

4.7.4.4 Analysis of results selection

In average, considering all the eight health scenarios, 25.12% of the participants selected their first, second and third choice (preference 1, 2 & 3. hereafter) from Khresmoi (table 2). The rest of the participants, 74.88%, preferred Google results instead. When looking specifically at preference 1, 2 and 3 for all the eight scenarios, the same pattern persisted (see averaged values from all scenarios, Table 6).

The proportion of participants that preferred results from Google was greater than that from Khresmoi for all the eight scenarios (see in Table 6 both the mean proportion of preferences 1, 2 & 3, but also the proportion of preferences 1, 2 & 3 independently, except for preference 3 scenario 6). However, the differences between the proportion of participants who preferred results from Google or Khresmoi varied between health scenarios (see differences in Table 6).

Table 6. Proportion of participants that selected their three most preferred results (preference 1, 2 & 3) from Khresmoi or Google, and the mean position of these results within the top ten list of results for each of the eight health scenarios. Mean values of proportion of participants and position of results are presented for each independent scenario by averaging values for preference 1,2 & 3 (mean 1, 2 & 3), and for all the eight scenarios by averaging first for preference 1,2 & 3 independently (mean 1, mean 2 and mean 3), and for preferences 1,2 & 3 all together (mean 1, 2 & 3).

Scenario	Number of participants	Preference category	Proportion of participants whose preference 1, 2 or 3 was from		Difference between proportion of Google and Khresmoi	Mean position of results selected as preference 1, 2 & 3	
			Khresmoi	Google		Khresmoi	Google
1	7	1	42.86	57.14	14.29	2.33	5.25
		2	14.29	85.71	71.43	7.00	4.50
		3	28.57	71.43	42.86	8.00	5.80
		Mean 1, 2 & 3	**28.57**	**71.43**	**42.86**	**5.78**	**5.18**
2	13	1	38.46	61.54	23.08	3.80	2.30
		2	46.15	53.85	7.69	3.67	2.14
		3	7.69	92.31	84.62	1.00	3.00
		Mean 1, 2 & 3	**30.77**	**69.23**	**38.46**	**2.82**	**2.48**
		1	33.33	66.67	33.33	8.67	6.67

3	9	2	22.22	77.78	55.56	5.00	5.43
		3	44.44	55.56	11.11	8.25	7.20
		Mean 1, 2 & 3	**33.33**	**66.67**	**33.33**	**7.31**	**6.43**
4	10	1	10.00	90.00	80.00	2.00	2.78
		2	0.00	100.00	100.00	2.00	3.50
		3	20.00	80.00	60.00	2.00	6.00
		Mean 1, 2 & 3	**10.00**	**90.00**	**80.00**	**2.00**	**4.09**
5	7	1	14.29	85.71	71.43	3.00	4.33
		2	14.29	85.71	71.43	7.00	7.33
		3	14.29	85.71	71.43	8.00	7.00
		Mean 1, 2 & 3	**14.29**	**85.71**	**71.43**	**6.00**	**6.22**
6	14	1	28.57	71.43	42.86	2.75	3.30
		2	42.86	57.14	14.29	2.33	3.88
		3	57.14	42.86	-14.29	3.88	2.83
		Mean 1, 2 & 3	**42.86**	**57.14**	**14.29**	**2.99**	**3.34**

7	6	1	33.33	66.67	33.33	4.50	1.00
		2	16.67	83.33	66.67	1.00	3.20
		3	33.33	66.67	33.33	8.00	6.75
		Mean 1, 2 & 3	**27.78**	**72.22**	**44.44**	**4.50**	**3.65**
8	5	1	0.00	100.00	100.00	-	3.60
		2	0.00	100.00	100.00	-	4.60
		3	40.00	60.00	20.00	3.50	2.67
		Mean 1, 2 & 3	**13.33**	**86.67**	**73.33**	**3.50**	**3.62**
1-8	71	Mean 1 (averaged from all scenarios)	25.11	74.89	49.79	3.86	3.65
		Mean 2 (averaged from all scenarios)	19.56	80.44	60.88	4.00	4.32
		Mean 3 (averaged from all scenarios)	30.68	69.32	38.63	5.33	5.16
		Mean 1, 2 & 3 (averaged from all scenarios)	**25.12**	**74.88**	**49.77**	**4.44**	**4.38**

Table 6 Proportion of participants that selected their three most preferred results (preference 1, 2 & 3) from Khresmoi or Google, and the mean position of these results within the top ten list of results for each of the eight health scenarios

Mean values of proportion of participants and position of results are presented for each independent scenario by averaging values for preference 1,2 & 3 (mean 1, 2 & 3), and for all the eight scenarios by averaging first for preference 1,2 & 3 independently (mean 1, mean 2 and mean 3), and for preferences 1,2 & 3 all together (mean 1, 2 & 3).

Considering first the averaged values between preferences 1, 2 & 3 of the proportion of participants who preferred results from Khresmoi or Google (see Mean Preference 1, 2 & 3 in table 2), we find that the lowest difference between Google and Khresmoi was of 14.29% for scenario 6. An intermediate category, included scenarios 1, 2, 3 and 7 where the differences between Google and Khresmoi ranged from 33.33% in scenario 3 to 44.44% in scenario 7. Finally, scenarios 4, 5 and 8 showed particularly high differences over 70%, indicating an important preference for Google in these three scenarios.

Concerning specifically preference 1, the differences between the proportion of participants who preferred Google and those who preferred Khresmoi was the lowest for scenario 1 and 2, 14.29% and 23.08% respectively. However, in scenario 4 and 8 the preferences were clearly in favour of Google with differences in the proportions over 80%. Specifically, 90% and 100% of the participants respectively for scenario 4 and 8 selected their preference 1 from Google (Table 6).

Considering all scenarios, on average 25.11% of participants selected their preference 1 from Khresmoi suggesting that these results were more relevant. However relevance may be defined differently for each participant, and often they are not able to explain what they exactly mean by relevance. Other reasons were:

- adequate vocabulary used (not too complex)

- trustworthy/"serious" image of the web site

4.7.4.5 Resources coverage: availability of preferred resources in the Khresmoi search

We have also taken into consideration which Google results participants preferred and have analysed whether these results existed in the list of the top 10 Khresmoi results, or in the Khresmoi index.

We have found that none of the Google preferred results (exact URL) can be found in the top 10 of the Khresmoi results, and only one preferred Google result domain (ranked the second) appeared in the top 10 Khresmoi results (ranked 7) for scenario 7. For the latter case, the exact URL was available in the Khresmoi corpus but did not pop up in the top 10 results. As for the domain preferred in Google search results, 6 out of 20 domains having been identified were crawled and indexed by Khresmoi (30%).

4.7.4.6 Trustworthiness of the resources preferred

Amongst the 70% of the URLs selected by the participants in Google but not present in Khresmoi, only 20% could satisfy the HONcode set of principles to comply with certain criteria of transparency and trustworthiness for online health and medical information. The other web sites could not have been certified by HON as none of them respected the minimum level of such criteria, e.g. non-moderated forums etc.

4.7.4.7 Ranking

Considering the averaged values for all the eight scenarios, the mean positions (rankings) of preferences 1, 2 & 3 were equivalent when comparing Khresmoi and Google and were in all cases over 3 (3.86 vs 3.65; 4.00 vs 4.32; 5.33 vs 5.16 for preference 1, 2 and 3 respectively, see Table 6).

In addition, these mean rankings of preference 1, 2 & 3 had an ascendant order. That is, the mean ranking of preference 1 was the lowest while that of preference 3 was the highest in both Khresmoi and Google (Table 6).

Interestingly, in scenario 1, 4, 5 and 6 the mean ranking of preference 1 had a lower position in Khresmoi than in Google. Mean ranking of preference 1 was particularly high in scenario 3 for Khresmoi (Table 6).

4.7.4.8 Coincident results as the indicator of relevancy

There were no coincident results of Khresmoi among participants for scenarios 1, 5 and 8, but there were for scenarios 2, 3, 4, 6 and 7 (Table 7). Considering these latter groups, there were a total of 11 most coincident results when including preference 1, 2 and 3 of participants, which seems low taking into account that there are 213 possibilities of coincidence (total number of scenarios run by participants (71) x number of results asked to select (3)). Specifically, scenario 6 had the largest number (five) of coincident results, while scenario 2 and 3 had two, and scenario 4 and 7 only had one.

Some results seemed more frequently chosen by the participants. Although not necessarily for the same preference 1, 2 or 3, seven out of 13 participants coincided in result 2 from the list of Khresmoi for scenario 2, 4 out of 9 participants coincided in result 8 from Khresmoi for scenario 3, and 4 out of 14 participants coincided in result 2 from Khresmoi for scenario 6.

Open-answer explanations by participants for the coinciding results again indicated that the main criteria is relevance. Other reasons were clarity and trustworthiness of information presented. Again, forums were regarded completely differently by various participants: some perceived them as an additional information source they appreciated to have an access to, while others as non-trustworthy sources.

Scenario	Total number of participants per scenario	Preference category	Most coincident result	
			Number of participants	Position within top ten list of Khresmoi
2	13	1	3	2
		2	4	2
3	9	1	2	8
		3	2	8
4	10	3	2	2
6	14	1	2	2
		2	3	1
		3	2	2, 3 & 6*
7	6	3	2	8

Table 7 Most coincident results among participants who selected preference 1,2 or 3 from the list of Khresmoi. In each case of coincident result, the number of participants who coincided in their choice is specified as well as the position of the coincident result

*In preference 3 of scenario 6, results in position 2, 3 & 6 coincided each of them in two participants.

4.7.5 Discussion:

It should be clearly noted that students of medicine only represent a small fraction of all online health information seekers. In this sense, this first set of evaluations should be considered as preliminary. We plan to broaden our sample in further evaluations to have better representation of the whole heterogeneity of online health information seekers.

Most of the participants reported being experienced Internet users with Google as a preferred search engine. Most sought medical topics concerned diseases and treatments, confirming the general tendency (Fox S, 2013), while very few of them search online health information on a daily basis.

Based on their answer regarding how they judge quality of the health information on the Internet, the most popular answer was presence of references. Clearly this alone is not sufficient to judge the quality of the information. Few specified that they check for possible bias or compare with previous knowledge, which demonstrates higher awareness and a more critical approach to evaluating the information found online. On the other hand, as shown in the study of Feufel (Feufel, 2012), once end users access a web site, quality concerns disappear despite the online navigation skills of the end user. Hence despite the fact that some participants listed certain criteria to evaluate the quality of online health information, it is possible that once they accessed the web sites, these were not taken into consideration at all or to the extent that they should have been taken and the participants were guided by some other criteria (relevancy, presentation of information, design etc.). In general, we have had an impression that the skills to critically evaluate the online health information need to be improved in this participant sample. We need to further investigate to confirm these findings. From other hand, it seems to be quite a difficult task to raise end-user awareness of such matters, when it becomes a personal responsibility, if even medical students were not always able to distinguish the trustworthy web sites from the ones in question. We suppose that by providing the end users with Khresmoi and

marketing it as a search engine with trustworthy content, a safe environment will be created and end users will be more protected from the possible harms fraudulent online health information can bring.

In general, about three quarters of participants preferred results from Google. This pattern of preference for Google results was confirmed for all eight different scenarios. However, the magnitude of the differences in the preferences between search engines varied between scenarios. In this sense, although still lower than Google, Khresmoi results were particularly preferred for scenario 6, which was about pregnancy and more precisely about taking a decision regarding amniocentesis. The scenarios where Google was strongly preferred by the participants were scenarios 4, 5 and 8 (food safety, drug safety and taking care of elderly persons respectively). In between (from 33.33% to 44.44% of preference towards Google comparing with Khresmoi) there were scenarios 1 (specific disease or medical problem), 2 (certain medical treatment or procedure), 3 (how to lose weight) and 7 (medical test results).

One of the possible reasons to explain the difference between scenarios where Google was largely preferred is coverage of certain topics in the database of HONcode certified web sites. Only 1 out 23 URLs and 5 out of 21 domains of most Google common results among participants do exist in Khresmoi index. The Khresmoi prototype index largely contains HONcode-certified web sites. About 8,300 websites comprising the HONcode-certified database have all undergone a rigorous certification process to ensure the web site's transparency and a certain ethical commitment of web site editors and content providers. However, as applying for the certification is a voluntary process, the database contains various kinds of web sites related to human health, but does not explicitly cover all health-related topics. Additionally, the HONcode certification process itself eliminates certain types of websites such as sites making various claims without proper evidence backing (various fat loss sites, miracle pain relief sites, etc.). However, many of these sites have a high popularity ranking with high clickability in Google, though they would not even appear in Khresmoi. There is no classification of the web sites in the database by health topic defined by Pew, however clearly the web sites corresponding to the scenarios where Google strongly overtook Khresmoi (4, 5, 8) are presented in minor quantities, while for the ones where Khresmoi was chosen in more cases (especially pregnancy), there are more web sites in the database and of better quality. Additionally, compared to the other scenarios, scenario 6 was most likely to be the most contradictory in terms of online information existing on the web regarding this topic. The fact that the participants preferred Google less significantly than in other cases shows that Khresmoi has a potential to be more appreciated and eventually preferred by the end-users when they research health topics for which many contradicting opinions exist on web. To make it possible for Khresmoi to be more preferred by end users, more resources of a high quality should be added into the database. These selected web sites should not necessarily be certified by HON as it is a voluntary process initiated by the web site's editors. However these websites would need to meet certain quality and transparency criteria. It is important to keep in mind that any domain-specific search engine will always have less coverage than a generalized one.

We have also seen that few participants selected the same results for the same scenarios, though there were some results, which were commonly chosen by the participants. Further exploration of this finding can possibly bring new insights into the concept of relevancy of search results. It seems that the concept of relevancy is highly subjective and differs from one user to another. Possible relationships with previous health knowledge of end users should be investigated.

In average, we could say that the position of the results selected by the participants were similar for Google and Khresmoi. When considering all scenarios together, the mean position of preference 1 was around 3-4, the mean position of preference 2 was around 4 and mean position of preference 3 was around 5, in both Google and Khresmoi. A first interesting point refers to the fact that these mean positions were always over three, suggesting that results 1 and 2 from both lists seemed not particularly relevant for the participants. We do not have a clear explanation for this since open answers were generally not very detailed and precise in this specific concern. One possibility could be that the first and second results of Google generally included sponsored links. However, this

possibility seems unclear for Khresmoi, since it does not prioritise certain web sites based on contractual agreements, so-called sponsored links (as the HONcode certification is free of charge). It is also possible that in our particular sample the participants in general tend to ignore the two top ranked search results. It seems to contradict the findings that users "tend to look at only the top-ranked retrieved results and are biased towards thinking the top one or two results are better than those beneath it simply by virtue of their position in the rank order" (Baeza-Yates, 2011). Further analysis of these specific search results is required to obtain a better understanding of this pattern. Further exploration of click-through data and logs is foreseen. A second interesting point refers to the fact that the mean positions of results for preference 1, 2 and 3 had an ascending order (the mean position of results selected as preference 1 < preference 2 < preference 3). This seems in accordance with the general tendency of assuming that lower positions means lower quality of results (Baeza-Yates, 2011).

When looking specifically at preference 1, the mean position of results selected was lower in Khresmoi for scenario 1, 4, 5 and 6, suggesting that the ranking of Khresmoi could be better than that of Google in certain health topics.

The study has numerous limitations. First of all, the sample of the participants selected does not represent the general population and is quite homogeneous. However, follow-up evaluations are foreseen to see whether there is a difference with other samples of the general population and whether these findings can be generalized. Also, despite the conclusive findings, it would be better to have a bigger sample of participants. Secondly, the participants could not formulate and reformulate the query and check more than the first ten results, hence observing and studying the "real-life" users' behaviour was limited, however this was out of scope of this blinded comparison. Also, sponsored links provided by Google were included and not specifically identified, while in real life participants could have paid attention to them and excluded them from their investigation. Thirdly, the tasks proposed were mostly exploratory (both learning and investigating), while end-users may also simply look up certain information (for example, disease definitions, the address of the pharmacy, a specific web site etc.). Fourthly, we used keywords queries, which are more common, but excluded the possibility of natural language queries. Fifthly, although the participants were not limited with time to complete the tasks, most of the participants came during their lunch break and potentially could spend more time to research on certain topics, especially if they had a personal experience with health scenarios proposed. Further exploration of data is foreseen to improve the follow-up stages of this study and to address some of the limitations, while other limitations will be addressed with full user tests.

4.7.6 Conclusions

In general we can conclude that the search results provided by Google have been preferred due to wider coverage of resources and lack of verification by the participants with regard to the trustworthiness of the web pages. In practice, quality and trustworthiness do not appear to be criteria for selecting web pages, which is contradictory to the stated position of the study participants. This also means that when the topic is more specific and/or sensitive or subject to possible bias, end users may have almost equal preference for Khresmoi at the current stage. Additionally, the test cohort in this study was not representative of the general population. To truly gauge the response of the general public, further studies would have to be conducted using a far more heterogeneous cohort representative of the general population. Results obtained in the current study demonstrated that despite claims to the contrary,

quality is not a priority for selection of online health pages amongst the medical students. The general public would likely have lower health and general literacy than a cohort of medical students, so it is likely that results obtained using a more heterogeneous and representative cohort would demonstrate an even lower prioritization of quality than was demonstrated in this study. The position of the most preferred results seems equivalent for both search engines, however further investigation in the concept of "relevance" is required. Hence, the following efforts regarding the prototype should be directed towards adding more high-quality resources and adapting the search engine taking into

account low awareness and critical evaluation skills of the end-users in respect to online health content. The next step in respect to the evaluation is to perform non-blind tests where participants are aware of the overall higher quality of Khresmoi results comparing to Google, and to ascertain whether it will change their mind and attitudes towards it.

4.8 Results of full user tests

4.8.1 Pilot tests in October 2012 – February 2013

4.8.1.1 Paris pilot tests

A first pilot user test for the general public was conducted in October 2012 during a meeting between HON, CUNI, and ELDA in Paris, France. The aim of this meeting was to present the user-centred evaluation process to CUNI and ELDA in order to prepare the evaluation for Czech and French participants. Both the protocol of the evaluation sessions for general public and the use of the Morae recording software have been discussed and tested, namely information sheet and consent form for the participants, protocol and check-list of the session, Morae configuration file with all tasks and questionnaires (demographic and Internet usage questionnaire, adapted SUS questionnaire, task questionnaires) and an evaluation matrix for the observer to fill in based on users feedback. A complete scenario were conducted:

- Introductory part (welcome, informed consent, demographic and background, internet usage questionnaires),
- Testing the search engine to get familiarised, performing predefined tasks,
- Feedback (informal feedback given during the session, answering SUS questionnaire with some additional questions specific to Khresmoi, some concluding remarks on what should be changed etc.)

After this session some general recommendations have been issued, such as ensuring stability of the interface (no change on the interface while evaluations are conducted, in order to prevent all risks of crash during the evaluation).

It was found that some questions and formulation of questions needed to be adjusted; participants should not feel they are evaluated on how to retrieve information but rather that they are the ones evaluating the system. It was decided to simplify the tasks.

At this stage, both interfaces of Khresmoi were planned to be evaluated by the general public : the ezDL prototype and Khresmoi for Everyone. As some components were not yet integrated into simple search, such as the translation system developed by CUNI, it was decided to add it before the evaluation starts.

It was also pointed out that the ezDL prototype evaluation by the general public should start by explaining how to use the “new” features such as the personal library and collaborative tools, which are not usually provided by general search engines.

After Paris test some improvements have been made in the protocol and the tasks.

4.8.1.2 Sofia pilot tests

At the second stage a series of pilot user tests was conducted during the Khresmoi annual meeting in November 2012 in Sofia, Bulgaria. A total of 8 people, both project partners and members of the Advisory Board, volunteered to evaluate the system :

- four of them did the test on Khresmoi for Everyone,
- two of them on the ezDL prototype,

- two of them did it on both interfaces..

Feedback received during these sessions can be split into following categories:

1) regarding the evaluation test setup

2) regarding the tasks

3) regarding the prototypes

4.8.1.2.1 Feedback regarding the evaluation test setup

After these sessions, some comments were made to improve the evaluation matrix for the observer :

- this table could be completed when analysing the recording after the session instead of during the recorded session.

- the document should be improved : question could be classified for a better use

In addition, some recommendations were made regarding other materials to be used and evaluation setup:

- beware of using a keyboard in the language of the participant
- have a mouse available
- ensure fast Internet connection
- use some external devices for the webcam and microphone if using laptops. Some of the participants during the pilot-evaluation in Sofia were moving the screen during the recording (in order to see correctly), which could result in a low-quality image in the recording.

4.8.1.2.2 Feedback regarding the tasks

Some observations were made in respect to the tasks and Morae configuration file:

- after the Paris tests tasks were simplified, and participants in Sofia tested both complex and simple tasks. It became clear that complex tasks were still too complex and created the impression that participants' knowledge and skills are being evaluated instead of the prototype. Also, it was noted that participants tended to copy-paste the words or even phrases from the tasks into the search bar. We concluded that the task formulation was too difficult, hence it was important to simplify both tasks and their formulation.

- all questionnaires should be checked for single/multiple answer and verified that all is correct

- questionnaires should be simplified (wording, questions themselves) - both demographic and Internet usage as well as extended SUS questionnaire. There was a suggestion to modify the SUS questionnaire as some questions were ambiguous. Eventually it was decided not to change the original SUS questions and modify the questions originally created for Khresmoi evaluations

- The Morae windows should be extendable. Eventually it was checked and it was not possible

- English needs to be improved for all materials given to the participants, however since there are no participants with English as a mother tongue, they should be simply translated into French and Czech and spell-checked.

- Tasks should be focused on the functionalities provided by Khresmoi, keeping still in mind that they should present "real-life" health scenarios.

- A tutorial is needed for both prototypes to improve the experience and evaluation results

4.8.1.2.3 Feedback regarding the prototypes

Feedback concerning ezDL prototype:

- The main problem is low relevancy of results, clearly there are no enough resources indexed, or they do not show up
- The ezDL Swing prototype is confusing for the user with several tabs; it is not clear how these tabs are interconnected
- In some cases spelling correction and query suggestions were not working
- After having typed the query, the window showing different libraries appears with the loading sign which was confusing for users
- Once results appeared, there are also certain categories on the left side of the tab. They are also confusing, and participants had to try several of them before finding the results; it is not necessary to display the categories for which there are no results available
- Sometimes some filters did not work, like a language (show by language) filter for example
- The horizontal scroll bar is annoying
- The results preview box covers the scroll bar; cannot mouse over and scroll down
- It is not clear how to use disambiguation toolbar
- In certain cases image results popped up while there is no request for the image, and these were not appropriate. Also they occupied the whole space of the tab
- The query search bar is too small and the query cannot be well read
- The drop down box with query suggestions overshadows the top results
- The tag cloud does not contain the query terms
- Query suggestions are not always helpful
- Some results are duplicated
- The Swing prototype is a separate Java application is not very convenient to use as users automatically tend to refer to browser tabs to find the one with the search engine
- Colour markers and lock signs are not very clear
- Excerpts are not informative in some cases
- Some results do not have titles, which creates confusion, as users do not know where to click.

Feedback concerning the Khresmoi for Everyone prototype:

- Low relevancy of results
- Some participants ignored using the filters, while others did use them: there was confusion between “by topic” and “filter by”, creates the impression that the first one is not filter.
- There is no “scientific article” in the topic cloud (was mentioned several times when the task was to find the article)
- PubMed was not available for the tasks about finding the scientific article
- Links already clicked have the same colour. It would be better to change this.
- For the task about BMI, participants often were redirected to the web sites providing different metric system (pounds instead of kg), which was confusing; the local resources should be prioritised.
- Participants who used translation (French --> English) were annoyed when the translation was no longer provided for the web site accessed

- It is frustrating to have very few results; it was suggested that the filter might be deactivated if there are extremely few results. In certain cases there were only 4 links while it says there are over 200 results, which created confusion.
- The word cloud does not contain the query search terms which was disturbing to the participants
- Geolocalisation is not currently possible, but would be useful and helpful
- Some participants appreciated definitions of the search terms below the search bar
- Some search results looked like spam
- Images in the middle of the page are disturbing in some participant's opinion
- The query suggestion is slow; they type faster than the suggestions appear
- There is a confusion with opening the result in a new tab or window: if it opens in the same window as HONsearch, then they need to restart the search

4.8.1.2.4 Some other observations regarding user tests setup

The initial idea to perform evaluation tests with several participations simultaneously was not a good one. As experienced in Sofia, it creates confusion for the participants, observers and the ones analysing the recordings. Hence, ideally each evaluation test should invite one participant and have one observer and one facilitator. It was also agreed that the best way to observe the participants was during the session, and not afterwards as contextual information is missing.

Only one of the participants in the pilot tests has typed the query in his mother tongue (French), there were some participants with German as a mother tongue, but they were querying in English (probably due to fluent English and the fact that all tasks were formulated in English). It created a certain bias as the general public is unlikely to query for unknown / unfamiliar topics in any language other than their mother tongue. It is also unlikely that the "average" representative of online health information seeker would query in English if his or her mother tongue was different.

After the Sofia tests were carried out further changes have been made regarding the protocol and tasks.

4.8.1.3 Geneva pilot tests

Further on, two more tests were performed in Geneva in February 2013 with two French-speaking librarians working in the library of Faculty of Medicine of the University of Geneva. One of their duties is to assist the patients contacting the library to find more information about their disease. It is difficult however to recruit these patients as they come anonymously and do not leave their contact details.

The following feedback has been received:

Regarding the overall organisation/perception:

- the perception of being evaluated persists
- it has been identified that the person staying close to the participants during the evaluation is perceived as a control if the tasks are well performed as if the participant is being evaluated.
- The person should remain in the same room in order to help when necessary but not close to the participant.

Regarding tasks and Morae:

- tasks were easy to handle and not stressful for the participants
- tasks description and questionnaires should be translated into the mother tongues of the participants

- adapted SUS questionnaire was still too long and difficult to understand
- Morae interface should be translated into the mother tongues of the participants --> have been checked with Morae technical support and it is not possible.

Regarding the *Khresmoi for Everyone* prototype:

- Disease section only works in English, hence cannot be evaluated with French and Czech speaking end users
- Used keyword classification, but noted that it was not clear (mixed words, the concept is not clear)
- Used forum filter
- Interface language changes after using the translation
- Results preview is not necessary
- Different background font for all filters to make them more distinguishable
- The idea of providing trustworthy web sites is not clear from the search engine presentation
- Web sites with restricted access should be marked as such
- The results should exclude pages under construction

4.8.1.4 Report on the work done by CUNI

CUNI is preparing general a public evaluation for Czech users, which is supposed to provide additional evaluation data, especially from the point of view of cross-language medical text and document retrieval. CUNI has performed initial tests (in cooperation with HON and ELDA, in Paris), after which the Czech general public evaluation methodology was decided upon. The methodology and tasks will be the same as for the other general public test(s). Also, the Morae software will be used for the actual testing. 5-10 people are assumed to be recruited for the initial test, to be performed in the May/June period in 2013. However, the following localization, translation and adaptation has been performed as part of the preparation:

- Preparation of localisation of the ezDL interface to Czech, and the actual localisation; this involved translation of all resources (menus, short texts, tab labels, error messages, etc.), of the tutorial (since the users will be allowed to "play" with the system at the beginning of the evaluation session). These resources are supposed to be reused for the web interface as well.

- Translation of the info leaflet for evaluation participants, and translation and adaptation of the consent form to obey local Czech human subject rules and the Czech law on personal data and information storage and manipulation (no ethical issues are involved, in any case).

- Translation of the Morae recording settings and tasks, to display online to the users while performing the evaluation tasks.

- Translation of the questionnaires for the users being evaluated as well as for the observers.

- Modification of the testing procedure due to the fact that the Morae evaluation recording software cannot be fully localised to Czech (two observers will be present: one will help the users to navigate the Morea software, while the other will be hidden and performing the observer's tasks proper, including notes etc.).

4.8.1.5 Conclusions

Overall, over the period from June 2012 till September 2012 the protocol for the "full user tests" was created, while from October 2012 till February 2013 it was modified and refined several times. We had to take into considerations certain limitations of the software chosen (Morae), for example, difficulty with task randomisation. We had to take into consideration the significant feedback received during the pilot tests regarding evaluation tests setup, protocol, tasks and questionnaires. After having

conducted three stages of pilots, the tasks and questionnaires were significantly simplified. The Geneva pilot tests also took into account the parallel evaluation being conducted in the same period of time – blind comparison, which also influenced the refinement of the questionnaires and tasks for the full user tests. It should be also noted that over the period from June 2012 till February 2013, following the Sofia evaluations, the decision was taken to exclude the ezDL prototype from the general public evaluations at least at this stage, as it would be too difficult and overwhelming to deal with it for the end users. Hence the HONsearch was rebranded as "Khresmoi for Everyone", while ezDL one was renamed "Khresmoi Professional".

Further steps to improve both prototypes have been listed in the results. The most important and crucial points remain the relevancy and increasing the index. It is clear that there is a lack of resources at this stage of both prototypes. The tools provided need to be better integrated in the interface.

4.8.2 Full user tests in May-July 2013

4.8.2.1 Evaluation tests set up

4.8.2.1.1 Prague

4.8.2.1.1.1 Pre-evaluation preparatory tasks

Prior to the evaluation, the following tasks have been performed, partially at the visit to Paris (with the help of HON and ELDA), partially at the later Sofia meeting, and partially in Prague:
(a) Localization of the Khresmoi for everyone interface to Czech (partially done by HON, the rest by CUNI)
(b) Installation of the Morae software to three notebooks: one of the observer (Zdenka Uresova), one to a separate Khresmoi-only notebook with a clean installation of the OS, for the test subjects to use, and a temporary demo version for the facilitator for checking and testing purposes (Jan Hajic)
(c) Localization (translation) of the Morae test configuration file (8.7), as provided by the Evaluation group in English, to Czech; this included the demographic questionnaire, the SUS questionnaire, and the three task descriptions and their short post-task questionnaires
(d) Watching several test sessions in Sofia conducted by HON, GAW, DCU and ELRA
(e) Recording a test session on ourselves to test the configuration
(f) Translation of the information sheet to be distributed to test subjects prior to the test (8.6)
(g) Localization of the informed consent form, to be distributed and signed by the test participants; the original has to be adapted to suit the local laws on privacy and ethical issues (8.6)
At CUNI, we have also localized (translated) all of the EzDL interface to Czech, but due to the switch to Khresmoi for everyone interface for the tests, it has not been used in the evaluation.

4.8.2.1.1.2 Recruitment

We used staff and researchers from several age groups and genders from the Faculty of Mathematics and Physics, Charles University in Prague, with various educational backgrounds. A total of 15 subjects were recruited. The youngest was 27 and the oldest 77 years old; education varied from vocational/technical school or college level to Ph.D. All of them were Czech native speakers, with several being highly fluent in English.

4.8.2.1.1.3 Location

All the recordings were made in the Charles University building at Malostranske nam. 25, 11800 Prague 1, Czech Republic; two of them in the office 420, 4th floor, the remaining 13 in the lab SU1 on the ground floor, where a better arrangement of the facilitator and the observer, and a quieter background were available.

4.8.2.1.1.4 Times and dates

All the recordings were made between May 23 and May 29, 2013. Typically, an hour and half was necessary to have the subject go through the forms, explain the initial setup, do the recording, and organize the paperwork and the machines (making file backups etc.) after the test.

4.8.2.1.1.5 Technical setup

A special "clean" notebook was acquired and used for the subjects themselves, with a fully upgraded version of Windows 7 Professional SP1. The notebook had been running only a virus/firewall software, MS Office and a few utilities for transferring files, plus the Morae Recorder Software. We provided users with a mouse to make navigation more convenient than with the notebook's integrated touchpad.

The observer used her own notebook with the Morae Observer software installed and connected to the subject´s notebook over an internal Eduroam WiFi network. The observer used headphones to listen to the recording being made while marking up the recording and making observation notes. She sat in a back corner of the room, not visible to the subject who was sitting in front of the room, facing the front wall (in the SU1 setup; the Rm420 setup was different, but the subject did not see the observer, either). The observer was able to view the subject from their position.

The facilitator sat next to the subject, and explained the project, the setup, the interface at the beginning and answered questions during the test. He also was in charge of distributing the info sheet and the consent form, and in charge of collecting and organizing signed consent forms. The observer also checked and prepared the subjects' notebook between recordings, transferred the observed files and made backups and solved upcoming technical issues. The facilitator used his own notebook for comments, but did not run the Morae software (although the demo had been installed) while conducting the evaluation and was not connected to the two other notebooks.

This setup worked well except the Morae Observer frequently "stalled" at the beginning of the recording, but resumed after a minute or so, which was not critical and the recording has not been affected. In one case, the audio at the observer software stopped working, but the observer sat close enough to still listen and make making relevant notes. However, the audio was recorded properly into the recorded file. In one case, the subject pressed a key combination that initiated a skipping of the demographic questionnaire, which was then filled in manually in the Manager part of the software.

At the post-evaluation stage, when re-analyzing the recordings in July and August (see below), a bug in the Morae software was discovered causing it to not re-play the recorded screen (only the face of the user in the small PIP window was visible). After some investigation, it was found that the bug was caused by an ill-behaved interaction of a recent Windows security update (KB2803821) and the Morae Manager software. A temporal removal of the update solved the problem (otherwise, it would be very difficult if at all possible to analyze the recordings in a synchronized way).

4.8.2.1.1.6 Organizational setup and staff

Two people were run the Prague tests: Jan Hajic (CUNI) as the technical person and organizer and Zdenka Uresova (CUNI) as a translator, localizer, recruiter, observer, and post-test evaluator. Milan Fucik has helped to acquire and install the special "clean" notebook for the subjects. Katerina Stuparicova handled necessary administrative tasks.

4.8.2.1.1.7 Post-evaluation process

First, signed consent forms and Info sheets were scanned and emailed to GAW/HON. Next, all the recordings were re-played and markers and comments edited by the observer (first pass).
All 15 subjects were rewarded by vouchers to buy books in one of the large bookstore chains in the Czech Republic (NeoLuxor), valued at EUR 12 each.

Later, summaries were written while re-playing the recordings of all subjects (second pass). Also, all the data from the questionnaires was exported from the Morae software to several .csv files as the software allowed, and manually re-entered to the master spreadsheet with all the data; headings have been re-extracted from the English configuration file and put into the master spreadsheet, and

also copied from the Czech configuration file to extract a complete record of what users have been seeing while they filled in questionnaires .

4.8.2.1.2 Geneva

4.8.2.1.2.1 Pre-evaluation preparatory tasks

Before evaluation tests started all necessary documents were translated into French, i.e. information sheet, consent form and Morae file configuration (8.6, 8,7). The translation was performed by HON. To adapt these documents to French settings, DCU contributed to double-checking and changing wording where necessary. A protocol for each evaluation session was defined (8.8).

4.8.2.1.2.2 Recruitment

User tests in Geneva were conducted in the beginning of June while the preparation started long ahead. Two main recruitment channels were:

- Emailing all Switzerland and France-based participants of the survey (D8.1.1) with the invitation to participate in the evaluation tests.

Out of all participants who left their email addresses to be updated about the study, we extracted 81 email addresses of those who indicated their countries of residence as France and Switzerland. An email was sent to them inviting them to participate in the evaluation tests. Eight people responded to the email; however they were mostly located in various cities across Switzerland and France, so it was difficult to organize the user tests from logistical point of view. Two persons out of eight were from Geneva and one of them tested the search engine. This person then advertised the test and we obtained another volunteer to test the search engine. Additionally, two persons were recruited by one of the eight persons that responded to our email and they participated in Paris evaluation tests.

- Emails sent by the library of the Faculty of Medicine of University of Geneva.

Previously in this deliverable we described the collaboration we had with the library for another type of user evaluation - blind comparison of search results. The same librarians sent out our invitation email to a group of patients who from time to time come to a library to research for their health condition. We got two responses, and conducted two evaluation tests.

Overall four people were recruited and conducted tests in Geneva, their feedback and experience is further analyzed in this deliverable.

4.8.2.1.2.3 Location

Test were conducted in the premises of Hopital Cantonal Universitaire de Geneve and at the library of the Faculty of Medicine of the University of Geneva located at Centre Medical Universitaire. We chosen these premises as they are conveniently located in the city center.

4.8.2.1.2.4 Times and dates

Four tests were conducted on the following days: 4th, 5th, 6th and 13th of July. Overall each evaluation session took about an hour.

4.8.2.1.2.5 Technical setup

Two laptops were used for the tests. One had Morae Recorder, Observer and Manager installed, while another only Observer. The first one was used by the participant along with the mouse and external microphone. Network varied depending on the location, it could be either hospital public WiFi network or University password protected network. Both the observer and facilitator were at the same room with the participants, but not visible to him/her (behind participant back).

4.8.2.1.2.6 Organizational setup and staff

Two people were running the tests: Rafael Ruiz de Castaneda (HON) as a facilitator and Natalia Pletneva (HON) as an observer.

Facilitator's role was to respond to participant questions during the session, while the observer was adding the markers of the test in real time using Morae observer. Both facilitator and observer welcomed the participant and presented the project, as well as thanked him/her at the end of the session.

Each session started with welcoming, short introduction, reading and signing the consent form, and a demo of the prototype. Then a recording started during which participants had to

1. respond to a demographic questionnaire
2. use a prototype on their own, performing a typical search they would have usually done
3. do three tasks (calculate BMI, find treatment options, find patients experiences about a drug); after each task they had to respond to a couple of questions
4. fill in expanded SUS questionnaire which included 10 standard questions and 15 specifically designed for Khresmoi questions

After the recording we asked the participants one more time to sum up their main impressions and feedback: what was positive and negative, helpful and distracting etc. After this, we thanked the participants for their time and participation. All participants were given "thank you" gift cards of 20chf (approximately 16 euro) for a local supermarket chain.

4.8.2.1.3 Paris

4.8.2.1.3.1 Pre-evaluation preparatory tasks

Prior to the evaluation, Morae software was installed on two notebooks (in addition to two laptops already used for the Geneva tests) and a 3G internet connection provided by a French telecommunication company was tested.

4.8.2.1.3.2 Recruitment

Thanks to the collaboration of Dr Patrice Degoulet, Director of the Medical Informatics division at the Hôpital Européen Georges-Pompidou (HEGP) (Georges Pompidou European Hospital) in Paris, two departments were identified to take part in Khresmoi patients' evaluation. Professor Jean-Jacques Altman (Service de Diabétologie Nutrition et Endocrinologie - department of diabetology, nutrition and endocrinology) and Dr. Postel-Vinay (Service de Médecine Vasculaire et Hypertension Arterielle - department of vascular medicine and arterial hypertension) assisted us with patients recruitment. There was a total of nine patients: six of them were from the hypertension department, three - from the diabetology department of the hospital. Two other participants were recruited via mail described in section 4.8.2.1.2.2. They represented one of the patient organizations. They agreed to come to the hospital for these evaluations.

4.8.2.1.3.3 Location

All the recordings were made in the hospital located at 19 rue Beaubourg, 75004 Paris, France.

4.8.2.1.3.4 Times and dates

All the recordings were made between July 12 and July 15, 2013. Typically, each evaluation session took about an hour to recruit a patient and perform the test.

4.8.2.1.3.5 Technical setup

Four notebooks with a Windows 7 version were used to do the tests. We provided participants with an additional mouse and headsets (the latter one was used by few participants). All laptops were connected to 3G network as there was no WiFi available in the hospital. This eventually affected the speed of the prototype and could lead to some problems with display and recording of the sessions.

4.8.2.1.3.6 Organizational setup and staff

Three people involved in running the Paris tests, Célia Boyer (HON), Samia Chahlal (intern at HON) and Jérémy Leixa (ELDA). As in the tests in Prague and Geneva, two people were running each

evaluation test. The observer used headphones to follow the evaluation session and was adding markers. The observer could see a participant. In the majority of cases, patients were staying in their beds. The facilitator sat next to a participant and was responsible for walking a participant through a whole session, as well as collecting documents for each session. A system engineer was appointed for the first day inorder to solve technical problems in case of Internet or Morae bugs. The experience from Paris tests set up was that it was sufficient to have only one person running the session who would combine roles of both facilitator and observer.

After the session, participants were thanked for their participation and given a Khresmoi pen. Two participants who came to the hospital only for these evaluations were given 20 euro each to cover their travel expenses.

All the recordings were further replayed, and markers added by Jérêmy Leixa and Priscille Schneller (ELDA). Overall eleven recordings were made, however due to specific hospital settings and unstable Internet connection, two participants were eventually excluded from the analysis: one due to problems with the recording (only half of the session was recorded) and the second due to the fact that he did not meet the inclusion criteria, i.e. at least occasional search for online health information.

4.8.2.2 Prototype

All participants tested the same prototype "Khresmoi for everyone", described in (D8.3, 2012) as "Khresmoi classic search" and its updated version in D8.5.1. As the evaluation took place in May-July 2013, the version being tested was close to the one described in (D8.5.1, 2013), however before adding an advanced search mode. Some changes made in the prototype were motivated by the results of the pilot tests as described in (D8.5.1, 2013), hence the prototypes tested during pilots and "real" tests are not exactly the same.

4.8.2.3 Results of the tests

4.8.2.3.1 Demographic questionnaire analysis

27 recordings were taken into account for the analysis, others were excluded if a significant part of the recording was missing, or when participants stated they had never sought online health information before. One of the 27 participants encountered a technical problem with Morae, and his answers to the demographic questionnaire were not recorded, however some of them were recovered by watching the video; we decided to keep this participant included in the sample, hence in some cases a sample is 26 persons indicated by N=26.

The participants were mostly from the Czech Republic (14) and France (9) while two were from Switzerland and 1 originally from Belgium, but living in Switzerland (N=26). There were 14 females and 13 males of various age groups: the most prevalent were aged 30-39 (37%) and 60-69 (22%).

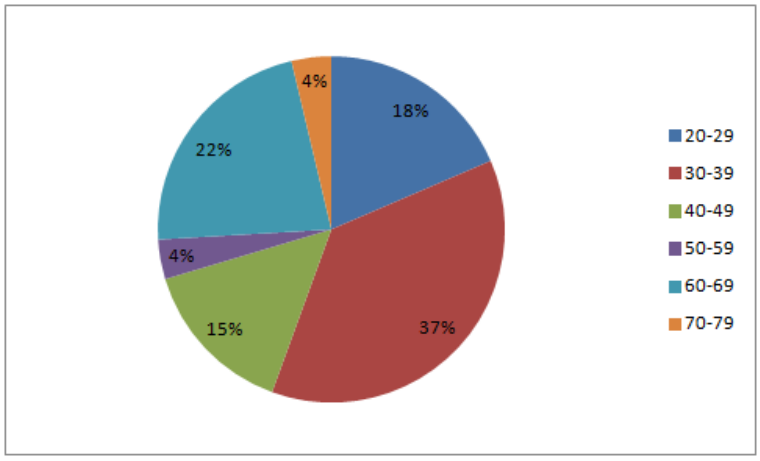


Figure 24 Participants age groups

Majority of participants hold a Master degree (41%), while all education levels were present:

High school	4%
Vocational/technical school	22%
Bachelor	11%
Master	41%
PhD	22%

Table 8 Level of education of participants

There was one participant currently enrolled at the University, whom we counted as Bachelor even if (s)he still had not finished his/her studies.

Most of the Czech participants worked in research, and some held administrative positions. Two of the Geneva participants received a health-related degree (not physicians) and are working in collaboration with a local hospital. The French participants occupations was very varied: marketing, teaching, bank, advocate, press editor etc. One participant indicated (s)he was retired.

4.8.2.3.1.1 Internet use

All participants were active online, 25 indicated they connected to the Internet on a daily basis, and two did so several times a week. 26 used PC or laptop and one indicated a smartphone as their principal device for connecting to the Internet (participants could only select one answer). 24 participants indicated they used the Internet for work or studies, while two mentioned that they did not work at the moment and used it for personal inquiries. One participant responded he did not use the Internet for work.

All participants also conducted online searches: 23 every day, and four several times a week. Everyone indicated Google as the main search engine being used (27), two Czech participants also mentioned Seznam.cz, one added Yahoo and another one added DuckDuckGo. Almost half of them (48%) said they were very confident with web search and considered themselves as expert users (N=27). 44% were slightly less confident,, they reported having problems in finding information from time to time. Only two (7%) judged their skills as average and reported often having problems with online search.

4.8.2.3.1.2 Languages

52% of the sample (N=27) stated their mother tongue as Czech, 41% as French. For the two remaining participants a mother tongue was Wolof (Senegal) and Russian (both communicated in French, so we further included them in a Francophone group). We asked participants about their level of English: 22 had at least average knowledge of English (N=26).

Basic	15%
Average	31%
Good	23%
Fluent	31%

Table 9 Level of English

37% reported searching for or reading online information in English on a daily basis, overall 85% did it at least several times a month (N=27):

Every Day	37%
Several times a week	15%
Once a week	19%
Several times a month	15%

Once a month	4%
Other	11%

Table 10 How often do you search for or read any information on the Internet in English?

Those who responded "Other" indicated that they did it very rarely, when there was no other option, or even almost never.

4.8.2.3.1.3 Online health search

As for online health search 42% of the participants did it once a month and 16% at least once a week (N=26).

Every Day	8%
Several times a week	4%
Once a week	4%
Several times a month	19%
Once a month	42%
Other	23%

Table 11 How often do you search for online health information regarding your or your family/friends health?

All participants gave preference to Google for looking for health content (N=27). Among the second and third choice were: seznam.cz (twice), PubMed, atlas.cz, databases (not clear which) and the web site CDS.

We also asked participants which types of online health information they were looking for (multiple choice, N=26). As seen from the below graph, all participants sought for information about a specific disease or medical problem.

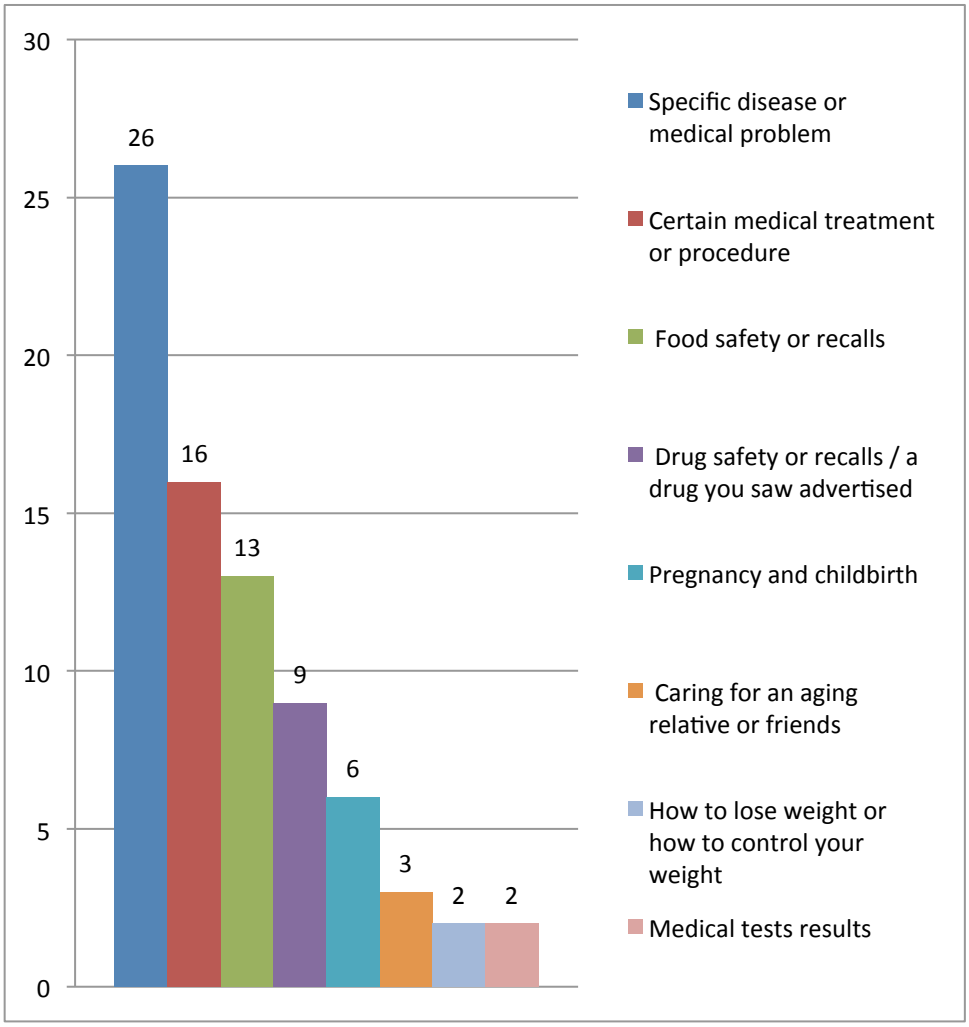


Figure 25 Which types of online health information you are looking for? (select all that apply)

Out of 26 respondents, nine (35%) reported having been diagnosed with various diseases, mostly (seven out of nine) for more than three years. Six out of nine participants reported that their Internet use to search for online health information did not changed after the diagnosis, while for two it increased and for one decreased. We also asked whether the participants encountered the situation of a family member or a friend being diagnosed with any chronic or life-threatening disease: 13 out of 26 reported having such experience, while the other half had not.

4.8.2.3.2 Task solving ability

Apart from the task which was free search engine use and was introduced in order for the participants to get accustomed with a search prototype, there were three other tasks. During "free search" participants were asked to perform a typical health search they would normally do.

The first one required participants to calculate their body mass index (BMI). We did not particularly insisted that they actually calculated their BMI as it could be a sensitive/private issue for some of them, hence finding a formula was sufficient. After this task we asked whether they were already familiar with BMI before and whether they calculated their BMI. 70% reported being familiar with BMI before making conducting the task, and 100% reported that they had calculated it (N=27). The average time to complete this task was 2.50 minutes.

The second task was concerned with finding treatment for liver cancer. After this task we asked whether someone close to the participants had being diagnosed with cancer and whether the participants were able to find the requested information. 92% reported that they had not experienced someone they knew and cared about being diagnosed with cancer. 81% reported having found

treatment options, hence completing the task (N=27). The average time to complete this task was 5.09 minutes.

Using Khresmoi for everyone was very helpful for this task. 80% of the participants who had no prior experience with liver cancer were able to solve the task, and all of those who had experience. Nobody who had prior experience with liver cancer was unable to solve the tasks, and 20% of those who did not were not able to find a satisfying answer.

The third task aimed at finding patients opinion on/experience with a medicine "Metformin". After completing the task they were asked whether they had previously searched for medication information on the Internet and whether they were able to find the requested information using the prototype. 74% reported having previously sought medication information, and 48% considered they completed the task (N=27). The average time to complete this task was 5.22 minutes.

We pose the question: does having previously searched for medication information on the Internet help with solving this task? Taking into account the data we have (27 participants), it does not. To investigate this, a large scale study is needed.

There was no significant difference in task success rate between Czech and Francophone participants.

4.8.2.3.3 Logs analysis

There was an unexpected bug with logs: they were not recorded during user tests, hence there was no additional log data to be collected, analyzed and presented in this section. Pages change and mouse click data recorded by Morae software did not yield interesting findings.

4.8.2.3.4 Users satisfaction

A specific questionnaire was developed and modified several times to measure users' satisfaction with the prototype. We used ten standard questions of the SUS questionnaire and also added 15 specific prototype-related questions. Participants had to grade each statement using the Likert scale from one to five, from strongly disagree to strongly agree accordingly. For the analysis we split all statements into "happy" (17) and "not happy" (8). "Happy" statements reflected positive user feedback, for such statements the higher the grade was, the more satisfied users were with the prototype and their experience. On the contrary, "not happy" statements reflected disappointment and dissatisfaction with the system; for such statements higher grade meant higher disappointment and dissatisfaction with the system and their experience.

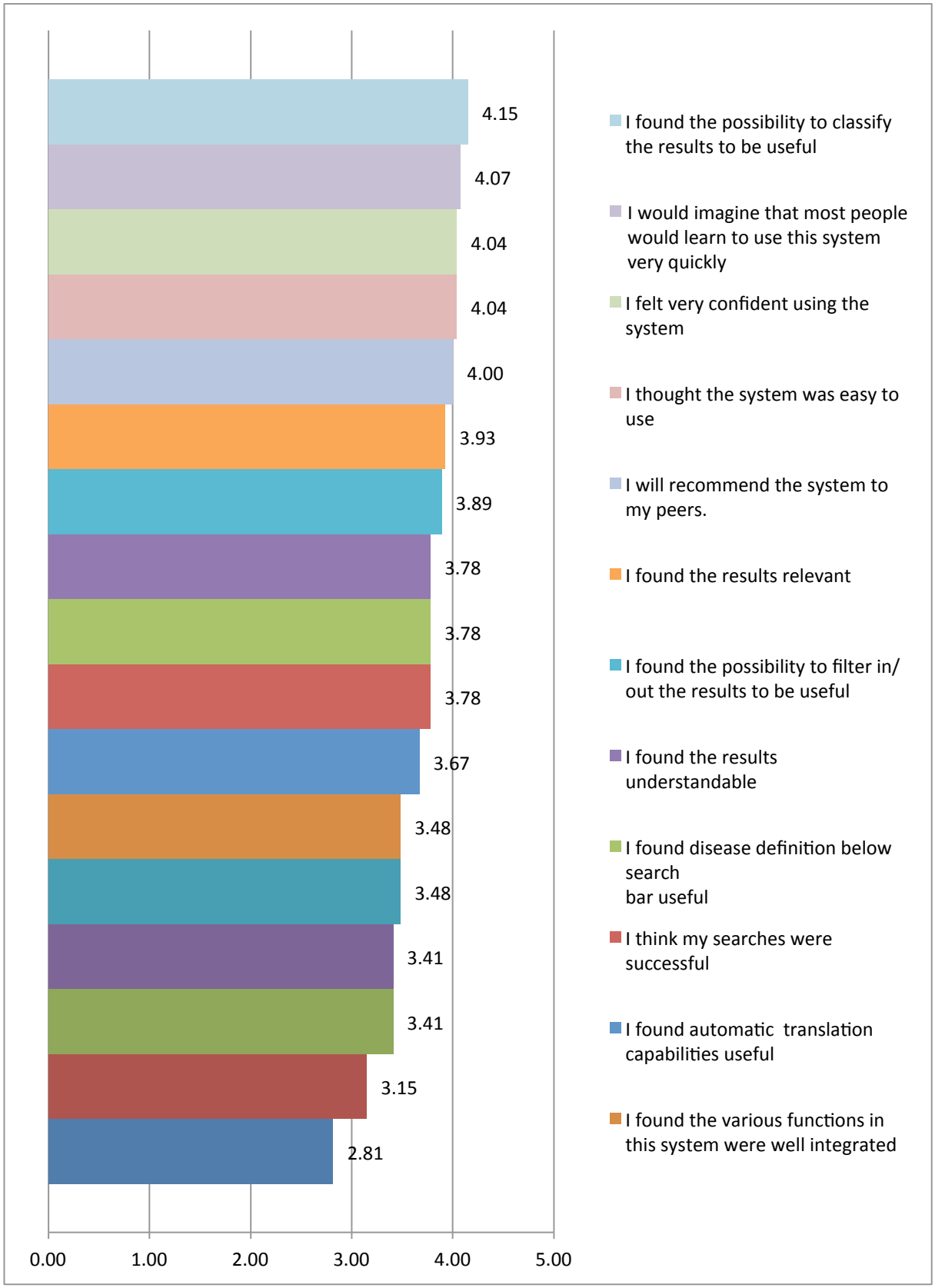


Figure 26 "Happy" statements from SUS questionnaire

As we can see from the figure above most of the "happy statements" responses fall between three and four, hence participants tended to agree. Only one statement concerning query formulation

was slightly less than 3. On the top of the list, there are several statements which received more than four points out of five (i.e. tending to "strongly agree"), i.e.the following system characteristics were most appreciated:

- results classification (4.15)
- quick learning how to use the system (4.07)
- feeling of confidence (4.04)
- easy to use (4.04)

Based on these findings, we can conclude that overall perception was positive, while there is a need for improvement of the specific features.

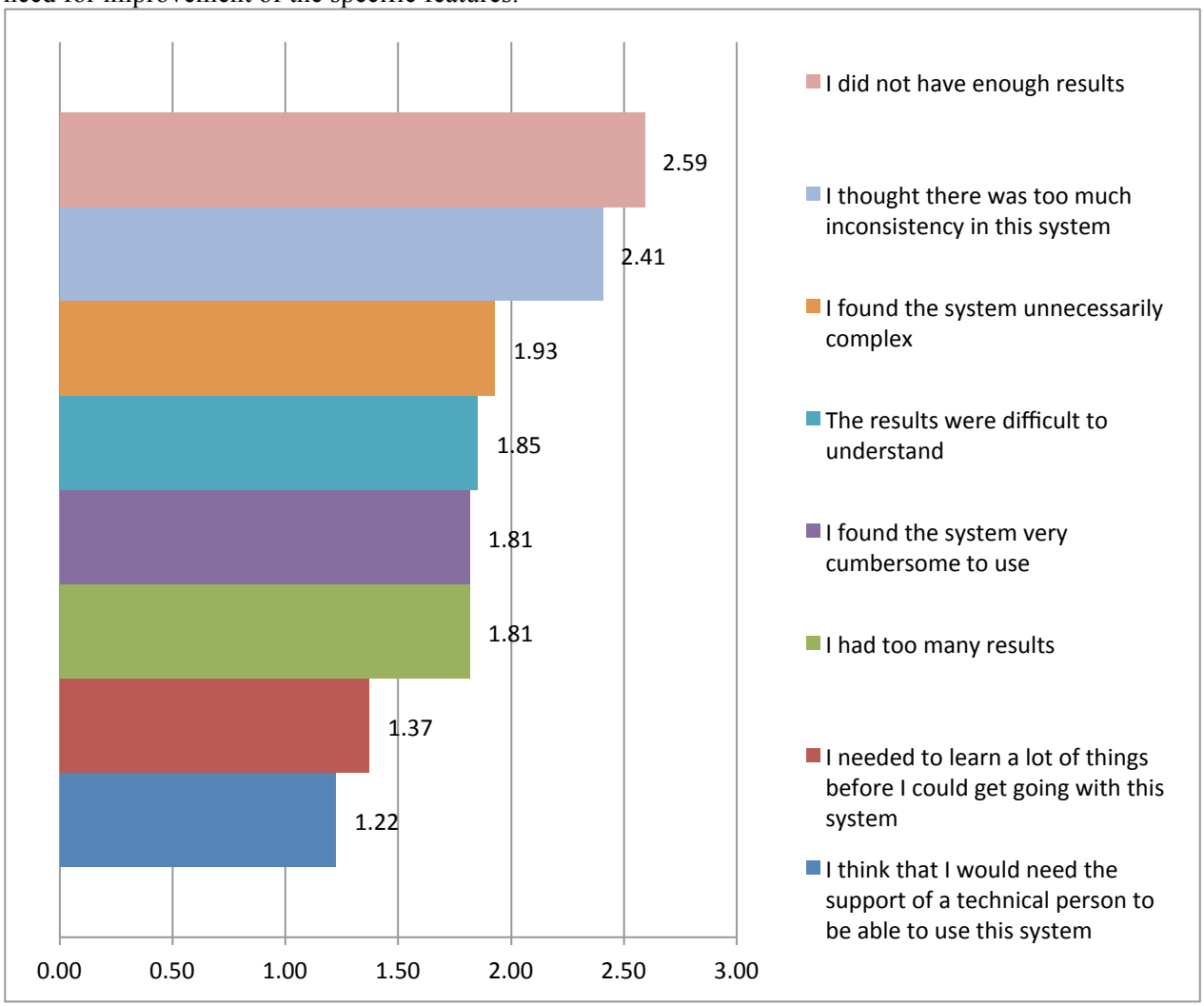


Figure 27 "Not happy " statements frmo SUS questionnaire

As for "not happy" statements, all of them received less than 3 out of 5 which indicates that on average participants disagreed with all of them.

We split the responses of Czech and Francophone participants and analyzed whether there was a difference of at least 20% between means and medians. All nine statements included in the table have a median difference of at least 20%, four out of them have mean difference of at least 20% (in **bold**). Positive numbers signify higher grades given by Francophone participants, and the negative correspond to higher grades given by Czech participants.

	Mean difference				Median difference			
	Mean		Difference		Median		Difference	
	FR	CZ	absolute	percentage	FR	CZ	absolute	percentage
I found the system unnecessarily complex	2.08	1.79	0.29	6%	2	1	1	**20%**
I think I was able to find an answer to the questions quickly	3.77	3.07	0.70	14%	4	3	1	**20%**
I did not have enough results	2.15	3.00	-0.85	-17%	2	3	-1	**-20%**
The system assisted me in my query formulation	3.46	2.21	1.25	**25%**	4	2	2	**40%**
I found automatic translation capabilities useful	3.00	4.29	-1.29	**-26%**	3	5	-2	**-40%**
I found the possibility to classify the results to be useful	3.85	4.43	-0.58	-12%	4	5	-1	**-20%**
I found the possibility to filter in/out the results to be useful	3.31	4.43	-1.12	**-22%**	3	5	-2	**-40%**
I think that having images among search results helps me to answer the question	3.69	3.29	0.41	8%	4	3	1	**20%**
I found disease definition below search bar useful	4.46	3.14	1.32	**26%**	5	3	2	**40%**

Table 12 Statements with at least 20% difference between mean and median answers of Francophone and Czech participants

From this table we can see that Francophone participants gave special preference to the tools which would help them to understand the topic, for example, disease definition, i.e. tools targeting a health domain. At the same time Czech users tend to appreciate more "common" search engine/IR tools such as automatic translation and filtering.

4.8.2.3.5 Overall users feedback

Due to the language barrier, this part of the results is split into two, depending on the language of the participant and evaluation: Czech-speaking and Francophone.

4.8.2.3.5.1 Czech evaluations (Prague):

This summary extracts the most frequent as well as special comments by the users and by the observer and facilitator while running and post-analyzing the tests.
The five most important observations from the Czech user tests can be listed as follows:

- All the users were able to use the system and follow the tasks
- Almost all users were successful in finding results in the first two simpler tasks, while half did not consider what they found in the third task as a success.
- The interface was found to be simple enough to use, but the translation and cloud features were confusing to a significant number of users, and several other features were mentioned as needing improvement.
- The quality of the snippet translation, independent of the problems encountered in (3), was deemed sufficient but must be improved.
- The availability of the contents was judged as poor by most users, especially in the Czech language.

We describe these points in more detail in the next section.

- System use and tasks

The users were able to use the system without apparent problems. Most of them, however, were not able to simultaneously comment on the things they were doing, but thanks to the fact that the sessions were fully recorded and the observer could make comments and mark the recording, this was not an issue with regard to the resulting analysis.

Some users complained though that the time for learning ("playing" with) the system before the tasks started was too short (they were allowed to try the system, ask questions to the facilitator and re-try for about 15-20 minutes each). It was observed that especially the language content feature, the topic filter feature, the definition feature and the word cloud feature would need more explanation time and then more time to try in various situations. The fact that the Czech content was very small compared to the English content, and that even in the English content, the topics used for filtering were only assigned to a few documents did not help the learning phase either. A contextual help feature would also help (and something like the Tutorial pages in the old ezDL interface would also be helpful, provided there more time to read it and go through it was provided to the test subjects).

- Success in finding results

All the users were able to find, and actually very quickly, the BMI calculator – most of them on Czech pages, since that one was already in and came on top. Similarly, most users (except for two) were satisfied with what they found for the task No. 2, treatment for liver cancer. It was observed though that for in-depth information on this topic, more than 5 minutes would be necessary in a real-life situation. Most of the results came from the translated pages from English, since the content on Czech was minimal for this task. Task 3 proved to be the most difficult, since it asked for other patients' opinions, typically found in users' forums and blogs. Most people used the filter feature, but apparently the classification of pages does not work very well for finding such web pages: in many cases, the pages were full of ads or not structured well – for example the users hit a page which allowed them to add user content several times, but how to get to the content added previously by other users was very confusing. Again, most of the content found was in English. Overall, 7 users out of the 14 taking part in the evaluation felt they did not find other users' opinions on the drug in question.

- The interface

The Khresmoi for everyone interface was found easy to use, users thought they would not need much expert help to use it, but they would need a little more time to try all the additional "non-Google-like" features to effectively use them during the test tasks.

Translation was judged as the most confusing feature. Users did not understand what the title meant, despite the fact that the Czech translation of the original button label "International" (which is completely inappropriate for what it actually means) was a little more telling (lit. "Pages translated from other languages").The label for the buttons meaning individual languages swept the users' thinking in several ways (some thinking they should use English, some thought the direction of the translation was the opposite, and all were confused by having "Czech" there as well, since the pages in Czech were actually not been translated at all). We present some suggestions for correcting the interface in the Conclusion section.

On the other hand, the "Filters" section was mostly clear to the users from the beginning; the problem here is rather in the content, not so much in the structure of the filter labels or the interface.
The "Keywords" (word cloud) feature was confusing and the users did not use it, but also felt they did not need it at all. It was unclear what this should do; those users who tried it did not see the usefulness of it since it was not clear what the selection actually does (at least not from the results displayed). Either some help, tutorial or other type of explanation is required.

The disease definition bar was not noticed by most users; in fact, it does not work in Czech, hence only the users who switched to English could see and judge this tool.
The query suggestion use has not been reported. Again, it probably stems from poor Czech content, since e.g. in the liver cancer task the only continuation offered for "rakovina" (lit. "cancer") was "rakovina nadledvinek" (adrenal gland tumor) despite the existence of tens of cancer types.

The images have not been used practically at all, perhaps they have not been very relevant for the three tasks as evaluated.

Other problems with the interface included the pop-up window showing the original text (for translated snippets) covering the snippet so it became unreadable, slow translation, HTML/SGML entities showing in the text (both translated and non-translated), wrong text encoding (on Czech pages – accented characters etc.), and several times reported confusion about the number of results shown on the top of the page and then the number of snippets on the page (it had to be explained by the facilitator that similar pages are being merged, and that they should be adding the number in orange parentheses instead of simply counting the snippets).

- Quality of Translation

Overall, people found the translation feature useful or very useful (apart from two –one said s/he has not concentrated on the translation, and the other one, a fluent speaker of English, would rather use everything – incl. the interface – in English). However, they all said that the translation could improve, and some commented on the (slow) speed, too. They would also like to have the pages translated right away (regardless of whether by Khresmoi MT or Google Translate etc.) when they click on the (already Khresmoi-translated) snippet shown on the results page. They thought this should be easy to add.

- Contents

Poor Czech content influenced, in our opinion and sometimes also commented by the users, several other test results as reported here; on the other hand, it allowed us to see the users exploit the translation feature (almost all did extensively, except for Task 1) and discover its problems.

We believe that the following features have been influenced by the lack of Czech content the most: results shown, query suggestion feature, keyword feature, and in part the filtering feature. It is not clear where the lack of the definition feature comes from – this might not be from the lack of contents, but from some shortcomings in the UMLS/MeSH definitions for Czech (but it could also be from some system or interface problem).

4.8.2.3.5.2 Francophone evaluations (Geneva and Paris)

Main findings are listed as following:

- All participants were able to use and interact with the prototype
- All participants succeeded in completing task 1 (calculate BMI), largely completed task 2 (find liver cancer treatment) and had difficulties with completing task 3 (patients' opinions about Metformin).
- They also reported lack of (relevant) results, but to a lesser extent than Czech participants; French-speaking participants could complete the tasks without automatic translation of results into English, though some were not satisfied with results in French and used automatic translation from English to find satisfactory results
- Participants faced several problems with the interface and gave some suggestions
- Overall impression was that Francophone participants used functionalities less than Czech ones
- A particular problem Francophone participants encountered was with accents while typing a query
- Having forums appeared to be a contradictory issue

Detailed findings are described below:

- Overall system use

As well as Czech participants, Francophone ones often needed more time to explore the system on their own (part entitled "free search engine use"). Typically we were asking them to perform a typical search regarding any health topic they would usually do. In the course of ten minutes dedicated to it, they got really absorbed by a search process and we had to stop them to keep the timeline for the session. One participant looked for local information about a health professional, which was unsuccessful as there is relatively little data in the index about this subject.

While testing, some participants ignored features and used the prototype as they have used Google, while others seemed to enjoy playing around with features and integrated tools. Overall, we noticed that if participants did not try tools during the first phase, they were unlikely to try them during completion of tasks. For example, instead of using forum filter, they would add "forum" in a search bar. Due to the log data missing, this could not be confirmed by log analysis.

France-based participants faced some difficulties with the keyboard as it varies from Switzerland to France which could affect their overall performance, especially in the beginning of the test.

A common problem for participants was "losing" a page with a search engine as once they clicked on a result they were redirected. They then had to either hit the "return" button, or if they had already closed a window, reopen it, sometimes with help from the facilitator. "Back" and "Forward" buttons did not seem to work well, returning a coding error, and as a consequence users had to start from the beginning.

In certain cases participants used a filter, and after typing a new query got the unfiltered results, i.e. previously selected filter was disabled, which was not convenient for the users. There was another case where after typing a new query, results of the previous query were retrieved which obviously did not meet users expectations.

Interestingly, a couple of participants typed sentences instead of keywords in a search bar. In a survey conducted in (D8.1.1, 2011) approximately 20% positively responded to the question whether they would like to ask a question in the same way they would have asked their physician.

Overall, participants often complained that the system was slow. It did not critically affect their experience with the prototype, but they clearly wished it would be faster.

- The interface

Overall participants were comfortable with the interface. Some mentioned that the page was a bit too noisy and confusing as there were many different filters and other functionalities. Two participants

suggested moving filters and other functionalities to the left side of the screen, so they do not draw all of the attention from results.

A common problem was the inability to identify the links already clicked on: after having clicked on a result, its colour did not change, hence after checking a web site, a participant could not figure out which one (s)he had previously clicked on.

Different tools and functionalities were used to a different extent:

- **Query suggestion** was rarely used by users, as they were faster typing and were not waiting for a drop down box with different suggestions. The ones who used it in some cases commented that the suggestions displayed were not useful. In other cases suggestions did not correct spelling errors. One participant commented that the box was too small to read. At least once abbreviation search did not work when participants sought "IMC" ("l'indice de masse corporelle" in French - "body mass index" in English). Another issue related and specific to Francophone users were the accents (é, à, ö etc). We noticed that some users were typing without paying attention to the accents, while others did type with accents. In a second case, when the page was reloaded, there was a coding error and users had to start from the beginning. Not surprisingly, this item had the lowest "grade" among all "happy statements" of the SUS questionnaire.
- **Preview of the results** was not used much, some commented that it is an interesting feature but it was not clear enough how to activate it and that it was taking too much time to load
- **Images** among search results were not hugely popular. Some participants did not notice them at all, while others questioned their relevance to a query. Only a couple of participants said they liked them, but some of the links they clicked were broken.
- **Translation tools**: there were two different things which confused some of the participants. The first one - localization, i.e. having an interface in a local language, and the second one - automatic translation of results in our case from English to French. Clearly, the latter feature is not self-explanatory. Some participants thought it served to filter out the results by their original language, similarly others thought it implied country restrictions. The title "international" was confusing. Not all participants used automatic translation: we observed a tendency that the higher the level of English was, the more likely these users were to use automatic translation. However, similar to Czech tests, we had one participant with fluent English (she was teaching English before retirement) who would rather use the whole prototype in English.
- **"Filters" menu**, or classification by the type of content was used by many users as it was available in French. However we had a few comments that users did not understand the logic behind this classification, also the titles were not comprehensible, for example, "les fournisseurs des soins" (healthcare providers). The confusion could happen because there are two different classification rationales behind this merged in one list: the first one type of content (health information, research etc) and the second one - format of content (forum and pages with video). A classification rationale should be crystal clear for the users and expressed in clear, understandable terms. For example, at least two participants said they wanted to clearly see which websites were coming from pharmaceutical companies inorder to filter them in/out. The most contradicting filter turned out to be "forums". Some participants expressed quite negative opinions about forums and their trustworthiness; they claimed they never check information on the forums as they do not perceive it to be reliable. Others were positive and enthusiastic about forums and said that they sometimes may go to read some information on the forums depending on the type of information they are interested in.
- **Keyword cloud** was barely used by participants. It seemed to be too complex to understand despite a tutorial given by a facilitator before the evaluation session.
- **Disease definition** below a search bar was one of the tools available only in English (or when translation from English is activated). Many of Francophone participants mentioned they would find having it in French very helpful.
- **Disease section** is another type of classification/filter. It is available in English only or when translated from English. Many participants also thought it was a good feature and worth developing it in French. Instead many had to type "effets secondaires" (side effects) in a search bar as such a filter was not available in French.

- Contents / Resources coverage

Although completion rate for the tasks was quite high (with the exception of the third one), participants commented on having few, insufficient number of results. First of all, there was often a confusion between a number of results (for example, 28) and number of links displayed in a results list (for example two): users had an impression that "missing" 26 results were hidden which was not the case. Secondly, going to a second page, there were not results in some cases which frustrated users. Similarly, when using some filters in some cases the participants were getting no results. Another important problem, when talking about lack of results, many of the users actually meant lack of relevant results, so they could not quickly get a satisfactory answer to their query. Further on, some results contained broken links which need to be remedied. The next problem was that the titles of results were not always clear, and snippet was not regarded as relevant to a query entered. The last, but not the least, some participants expressed their concerns about ranking, i.e. they could not understand the order of results in the list. Overall, Francophone participants had access to more resources in French comparing to Czech participants as the database of HONcode certified web sites (which is the baseline of Khresmoi for everyone) traditionally has had more web sites in French than in Czech. Currently it offers most of its resources in English. A separate list of resources to be added to the index to enhance variability and coverage of topics and languages required has been created.

- Suggestions from the users

During the evaluation sessions few users expressed their "wishes" to remedy the problems or inconveniences encountered:

- to add a "drug catalogue" so all drug information is all in one place and trustworthy
- to add an advanced search option to exclude certain terms, for example, a user was not interested in reading the articles about pregnant women which had the highest ranking for a given topic
- to add MOOC (massive online open courses)
- to add Question-Answering service with a real physician as an alternative to forums

4.8.2.3.6 Answering research questions and hypothesis

In the beginning of this chapter we list main goals of the general public evaluation tests, main questions to be answered and hypothesis. We attempt to confront the hypothesis, respond to the questions and judge whether we have achieved the goals.

The first hypothesis was regarding the usability of the system. After conducting user tests we can state that the system was helpful and some tools were especially helpful to solve the proposed tasks.

The second hypothesis stated that user satisfaction depended on certain variables. We assumed that disease knowledgeable users would type medical vs. layman terms, look for more detailed information in more complex resources and appreciate collaborative platforms such as forums etc. Due to a small sample we could not find any correlation and prove this tendency specifically, however based on observations we tend to think that this correlation does exist, except for forums, as in our case the more knowledgeable would rather avoid forums. We also assumed that internet experienced users would look for more functionalities. Again, due to a small sample we could not prove any statistically valid correlation; however, it seems to be the case and definitely requires more investigation. Another assumption was that both disease knowledgeable and web savvy users would be interested in trying and using advanced search functionalities and semantic search. Out of all participants, we could only for sure name one person (from Geneva) who would satisfy both of these criteria, and she was definitely eager to have more complex functionalities. Some of her comments could be satisfied only by a semantic search and we will contact her again to test the next versions of the prototype.

The third hypothesis suggested that more medically knowledgeable users will appreciate a higher quality of Khresmoi index over a general search engine. We started testing this hypothesis in another evaluation we performed, i.e. blind comparison of search results, however it turned out that in a blind comparison medical students preferred Google over Khresmoi. The next step is to conduct a randomized study where participants in some cases will know the difference between lists of the results, i.e. non-blind, and in other cases - not know, i.e. blind. These future findings will allow us to confirm or disregard this hypothesis.

The fourth hypothesis stated that after having learnt about trustworthiness of Khresmoi, users will tend to turn to it. This will also be studied along with the third hypothesis.

In the previous paragraphs we already partially answered research questions, i.e. questions 1, 3 and 5. It remains difficult to respond to question 2 regarding correlation between user profile and search outcome. With the number of participants we had (27) we cannot draw any statistically valid conclusion, we would simply need more users to test these hypotheses and respond to this question. However, this implies a considerable amount of extra work, as well as organizational difficulties, as it is very difficult to recruit participants for such studies. On the other hand, if the main goal is to collect user feedback to improve the prototype, we do not need to run tests with so many people - probably longer sessions with more varying tasks, but fewer users would eventually provide more deep understanding of users' behaviour and their usage of the prototype. Due to the failure to record logs during the general public evaluation we cannot respond to question 4, regarding studying user behaviour.

4.8.2.4 Conclusions and further steps for the prototype development

In this section we draw overall conclusions from the "full user tests" conducted over the period from May to July 2013 in Prague, Geneva and Paris. It has to be noted that the user (general public) tests have been an important experience for the CUNI/Prague team; despite being interested the most in the translation feature and its acceptance by the users, the Prague user tests contributed to finding other issues as well, mostly in the localization area.

Overall, comparing with the pilot tests which took place from October 2012 to February 2013 we can see a substantial progress regarding evaluation setup, protocol and prototype. Due to the fact that we simplified the tasks and protocol, users felt less stressed compared to during the pilots. Also, some changes were made to improve the prototype and overall satisfaction was increased.

In our tests we had quite a representative sample of online health information seekers, i.e. Internet users of both genders, all ages, education levels and professional backgrounds. They also varied in their web search experience and skills and personal health experience. There were still quite a lot of very educated people (a high percentage of participants with Master degree or PhD), however, we should not forget that most online health seekers tend to have higher education compared to the average web surfer. At least occasional online health information search was an important inclusion criterion,that is why eventually we had to exclude a few participants who did not meet the criteria.

Overall, participants were positive about the prototype, they would not need the help of technical person to use it.

Having analyzed users' feedback and experience we propose the following steps be taken in the near future, in any case before the final evaluation of the system takes place towards the end of the project in 2014:

- **Classification and filtering in/out** were highly appreciated by the users; however their presentation in the interface has to be clearer in terms of functionality and presentation. We propose to change the wording and adapt in each "active" interface language.
- **Disease section** needs to be added for French and Czech websites, and then displayed in the appropriate language based on the interface language selected or selectable by the user.
- **Keywords cloud** remained unclear for the users, hence we propose to remove it or at least hide it. More reflection is needed on how to present this filter in a clear and understandable way.
- For **all filters** we propose to have a checkbox, allowing checking or unchecking of various options simultaneously. Especially in the case of forums, it could serve to both groups of users

- the ones interested in reading forums can simply tick the box, while the ones not interested would tick all other boxes but forums. All documents in all languages should be (correctly) classified so that filters when applied get a reasonable number of results.

- **Definitions in French and Czech** below a search bar were regarded as interesting and potentially useful features which should be developed.
- **Query assistance** should be dramatically improved, especially taking into account problems with French accents. The same holds for Czech, where the problem is probably lack of content from which the query suggestions are derived. Accent problem is only presumed, since not enough data has been collected from the users to be able to look at the problem.
- **Images** were not regarded as particularly useful, however this can be explained by the fact that there was no specific task proposed to test this functionality (due to time restrictions we could not propose tasks for each functionality). It is clear that relevance of images should be increased if they are kept in the interface. Further investigation is required.
- **Technical bugs** detailed in the results section should be fixed.
- **Index** has to be expanded, especially with Czech resources. Already existing indexed documents should be cleaned up to decrease the number of broken links. Annotations should be double-checked and verified to ensure filters give expected results. If no results are available, ideally a filter should be disabled. It would be desirable to add more local healthcare services information in the index as directories of hospitals, pharmacies, physicians, paramedical professionals etc. Blind comparison of search results also gave some insights into which topics should have wider coverage in the French index of Khresmoi for everyone.
- **Automatic translation** proved to be an important feature for the participants with no fluent English, however its quality should be improved and presentation simplified. If the user clicks on a snippet which has been translated from language X to his-her selected language, the same translation should be performed automatically from language X to the user's language on the page where the user is redirected (presumably by Google Translate at the moment, since full page translation is out of the scope of Khresmoi). Other engines can be used too, but Google Translate is currently the best for pairs with Czech and German (WMT papers, 2013).
- **The interface for non-English users** should be improved, to make it clearer what the "Translation" feature actually does, and to make clearer the difference between the interface language selection in the upper right-hand corner and the "International" (i.e., "Pages translated from other languages") selection section. We suggest to make the available languages selectable as a set (checkbox next to each language), move the original language out of the list (but make it (de)selectable too), and allow for specific labels (inflections change non-English word for the particular languages, e.g. for pages translated from English it is "z angličtiny" and not "anglicky" or "angličtina", which is otherwise correct as the language of the interface).
- **System speed** should be increased

Regarding the performance of the evaluation itself, we recommend giving the users (test subjects) more time to explore all the features of the system in the initial phase of the test, to be more proficient when working on the tasks, which are (rightly so) time-restricted and thus do not allow for additional exploration.

Although not confirmed by a statistical analysis due to the small sample of participants, we had an impression that Czech participants used more and gave preference to common information retrieval functionalities, while Francophone users tended to appreciate more tools assisting them with simplifying and understanding of medical information.

Based on the above, we can conclude that at the current stage of prototype development we achieved the following goals: evaluating search success and success in using the tools, evaluating possible benefits of Khresmoi over Google. However we failed to study users' behaviour, which should be remedied in following evaluations.

5 Conclusions and future steps

In this deliverable the report on the user tests conducted has been presented. All evaluation tests made provided a significant feedback on further improvements of the prototypes. The most crucial things to do at this stage are to increase the index, to improve relevancy, and clarify filter functionalities. The next steps are directed towards prototype improvements and further tests. In the case of the general public, further steps of blind and non-blind comparison are foreseen, as well as follow up "full user tests". For the physicians, a lightweight method for on-going tests needs to be established. One of the goals of continuing the evaluation tests is also to increase the sample of participants, especially when it does not directly involve significant human resources to be spent (as full user tests do). Further tests will allow us to see whether the changes made in the prototypes yield more benefit for end users.

6 References

(Beaze-Yates, 2011) Modern Information retrieval. The concepts and technology behind search. Baeza-Yates, Ribeiro-Neto. P22-25

(D1.3, 2012) D1.3, Report on results of the WP1 first evaluation phase, Khresmoi project deliverable, August 2012, updated October 2012, http://www.khresmoi.eu/assets/Deliverables/WP1/KhresmoiD13V2.pdf

(D2.3, 2012), D2.3, Report on results of the WP2 first evaluation phase, Khresmoi project deliverable, August 2012, http://www.khresmoi.eu/assets/Deliverables/WP4/KhresmoiD23.pdf

(D3.2, 2012) D3.2, Report on results of the WP3 first evaluation phase, Khresmoi project deliverable, August 2012, http://www.khresmoi.eu/assets/Deliverables/WP4/KhresmoiD32.pdf

(D4.3, 2012), D4.3, Report on results of the WP4 first evaluation phase, Khresmoi project deliverable, August 2012, http://www.khresmoi.eu/assets/Deliverables/WP4/KhresmoiD43.pdf

(D5.3, 2012), D5.3, Scalability and performance evaluation report, Khresmoi project deliverable, August 2012, http://www.khresmoi.eu/assets/Deliverables/WP5/KhresmoiD53.pdf

(D8.3, 2012), D8.3, Prototype of a first search system for intensive tests, Khresmoi project deliverable, August 2012, http://www.khresmoi.eu/assets/Deliverables/WP8/KhresmoiD83.pdf

(D8.2, 2012), D8.2, Use case definition including concrete data requirements, Khresmoi project deliverable, February 2012, http://www.khresmoi.eu/assets/Deliverables/WP8/KhresmoiD82.pdf

(D8.1.2, 2011), D8.1.2, Requirements document for health professional search, Khresmoi project deliverable, August 2011, http://www.khresmoi.eu/assets/Deliverables/WP8/KhresmoiD812.pdf

(D8.1.1, 2011), D8.1.1, Requirements document for the general public health search, Khresmoi project deliverable, May 2011, http://www.khresmoi.eu/assets/Deliverables/WP8/KhresmoiD812.pdf

(D8.5.1, 2013), D8.5.1, Intermediate prototype of a second search system, Khresmoi project deliverable, August 2013

(ezDL, 2012) Thomas Beckers, Sebastian Dungs, Norbert Fuhr, Matthias Jordan, and Sascha Kriewel. ezDL: An interactive search and evaluation system. In SIGIR 2012, Workshop on Open Source Information Retrieval (OSIR 2012), August 2012.

(Feufel, 2012) Feufel MA, Stahl SF: What do Web-Use Skill Differences Imply for Online Health Information Searches? J Med Internet Res 2012;14(3):e87 URL: http://www.jmir.org/2012/3/e87/, doi: 10.2196/jmir.2051, PMID: 22695686

(Fox, 2013) Fox S, Duggan M. 2013. Health online 2013. Pew Internet: Pew Internet & American life project. Retrieved from: http://pewInternet.org/Reports/2013/Health-online.aspx

(Higgins, 2011) Higgins O, Sixsmith J, Barry MM, Domegan C. 2011. A literature review on health information seeking behaviour on the web: a health consumer and health professional perspective. Stockholm: ECDC.

(Lenzerini, 2001) Lenzerini, M. (2002): Data integration: a theoretical perspective. In: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, New York, 233–246.

(Morae, 2013) Morae usability testing software, version 3.3.2, 2013, htt://www.techsmith.com/morae.html

(Pletneva, 2011) Pletneva N, Cruchet S, Simonet MA, Kajiwara M, Boyer C. Results of the 10 HON survey on health and medical Internet use. Stud Health Technol Inform (2011), PMID 21893717

(Pletneva, 2012) Pletneva N, Vargas A., Kalogianni K., Boyer C. Online health information search: what struggles and empowers the users? Results of an online survey. Stud Health Technol Inform (2012), PMID 22874311

(Sheth, 1998) Sheth A 1998, Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics, in Interoperating Geographic Information Systems, M. F. Goodchild, M. J. Egenhofer, R. Fegeas, and C. A. Kottman (eds) Kluwer Publishers.

(SUS, 1996) Brooke, J.: "SUS: a "quick and dirty" usability scale". In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland. Usability Evaluation in Industry. London: Taylor and Francis, 1996

(WMT papers, 2013) Proceedings of the Eighth Workshop on Statistical Machine Translation, 2013, http://aclweb.org/anthology//sigmt.html#2013_0

7 Appendix

7.1 Physicians

7.1.1 Demographics

7.1.1.1 Demographics questionnaire

Demographics

All information collected is used exclusively for the purpose of the study and will be kept confidential. We do not collect personally identifiable information without your consent.

1. Age

Comments:

2. Gender

( ) Male

( ) Female

3. Mothertongue

( ) German

( ) English

( ) Italian

( ) Spanish

( ) Dutch

Comments:

4. Country of origin

( ) Austria

( ) Germany

( ) Italy

( ) Spain

( ) Netherlands

Comments:

5. Medical specialization

Comments:

6. Current occupation

( ) Physician in training

( ) General Practicioner

( ) Specialist

( ) Unemployed

( ) Retired

( ) Other (please specify)

Comments:

7. Highest completed academic qualification

( ) Medical degree (Dr. med.)

( ) Medical degree and PhD

( ) Medical degree and professorship (Prof.)

( ) Other (please specify)

Comments:

8. What year did you complete your medical degree?

Comments:

9. How good do you evaluate your own skills in using the internet to look for information?

( ) Very good

( ) Average

( ) Below average

10. How good do you evaluate your own skills of understanding medical English?

( ) Very good

( ) Average

( ) Below Average

11. How often do you use the Internet to search for medical information?

( ) Always

( ) Often

( ) Sometimes

( ) Rarely

( ) Never

12. Where do you currently work? (please leave blank if retired or unemployed) - City/ Country

Comments:

7.1.1.2 Demographics answers

Age	
Recording	Comment
physA	32
physB	28
physC	71
physD	25
physE	73
physF	25
physG	30
physH	29
physI	26
physJ	26
physK	26
physL	25
physM	25
physN	37

Gender		
Recording	Male	Female
physA	X	
physB	X	
physC	X	
physD	X	
physE	X	
physF		X
physG	X	
physH		X
physI	X	
physJ		X
physK	X	
physL		X
physM		X
physN		X

Mothertongue						
User ID	German	English	Italian	Spanish	Dutch	Comment
physA	X					
physB	X					
physC	X					
physD	X					
physE	X					
physF	X					
physG	X					
physH	X					
physI	X					
physJ	X					
physK	X					
physL	X					
physM	X					
physN	X					

Country of origin						
User ID	Austria	Germany	Italy	Spain	Netherlands	Comment
physA		X				
physB		X				
physC	X					
physD	X					
physE	X					
physF	X					
physG		X				
physH		X				
physI	X					
physJ	X					
physK	X					
physL	X					
physM		X				
physN	X					

Medical specialization	
User ID	Comment
physA	none
physB	turnusarzt
physC	internal medicine, gastroenterology and hepatology
physD	Dermatology
physE	Hematooncologist

physF	General Practitioner
physG	dr. med. univ.
physH	Residant
physI	Training in Radiology
physJ	resident
physK	
physL	Radiology
physM	Neuro-Oncology
physN	Allgemeinmedizin, Arbeitsmedizin

Current occupation							
User ID	Physician in training	General Practitioner	Specialist	Unemployed	Retired	Other (please specify)	Comment
physA				X			
physB						X	turnusarzt
physC					X		
physD	X						
physE						X	Physician, Research
physF	X						
physG	X						
physH	X						
physI	X						
physJ	X						
physK				X			
physL	X						
physM	X						
physN		X					

Highest completed academic qualification					
User ID	Medical degree (Dr. med.)	Medical degree and PhD	Medical degree and professorship (Prof.)	Other (please specify)	Comment
physA	X				
physB	X				
physC	X				
physD	X				
physE			X		

physF	X				
physG				X	dr. med. univ.
physH	X				
physI	X				
physJ	X				
physK	X				
physL	X				
physM	X				
physN	X				

What year did you complete your medical degree?	
User ID	Comment
physA	2012
physB	2012
physC	1966
physD	2012
physE	1964
physF	2012
physG	2013
physH	2013
physI	2011
physJ	2012
physK	2012
physL	2012
physM	2011
physN	2002

How good do you evaluate your own skills in using the internet to look for information?			
User ID	Very good	Average	Below average
physA	X		
physB		X	
physC		X	
physD	X		
physE		X	
physF	X		
physG	X		
physH		X	

physI	X		
physJ	X		
physK	X		
physL		X	
physM		X	
physN		X	

How good do you evaluate your own skills of understanding medical English?			
User ID	Very good	Average	Below Average
physA		X	
physB		X	
physC	X		
physD		X	
physE	X		
physF	X		
physG	X		
physH		X	
physI	X		
physJ		X	
physK	X		
physL		X	
physM		X	
physN			X

How often do you use the Internet to search for medical information?					
User ID	Always	Often	Sometimes	Rarely	Never
physA		X			
physB		X			
physC	X				
physD		X			
physE		X			
physF		X			
physG	X				
physH		X			
physI		X			
physJ		X			
physK		X			
physL		X			

physM	X				
physN			X		

Where do you currently work? (please leave blank if retired or unemployed) - City/ Country	
User ID	Comment
physA	
physB	Stockerau, Austria
physC	
physD	Vienna/Austria
physE	Rudolfinerhaus Hospital
physF	Mödling / Austria
physG	
physH	Medical University Vienna
physI	Krems/Austria
physJ	Wiener Neustadt/ Niederösterreich
physK	
physL	Vienna general hospital
physM	Medical University of Vienna, Vienna, Austria
physN	Vienna / Austria

7.1.2 Tasks

7.1.2.1 Task questionnaires

0 Atrial Fibrillation

Is it ok for a 69-year-old women with a history of atrial fibrillation and cardioversion to stop anticoagulation due to recent rhythm stability? Case scenario: A 69 year old women, diagnosed 4 years ago with atrial fibrillation has successfully received cardio version. That time she felt elevated heart rate and palpitations and is taking oral anticoagulants. Since then she is symptom free. She is health conscious and regularly measures her heart rate, which seems ok. She is otherwise healthy, her heart has a normal structure, only the left ventricle shows a moderate enlargement. She wants to stop oral anticoagulants.

From your knowledge: Is it ok for her to STOP taking oral anticoagulants?

( ) Yes

( ) No

( ) I don't know / I require further information to answer this question

1 Diabetes and malignant illnesses

From your knowledge: Does a 40-year-old female patient diagnosed with type 2 Diabetes have an elevated risk of getting malignant illnesses?

( ) Yes

( ) No

( ) I don't know / I require further information to answer this question

2 Diagnosis of an X-Ray

Please examine the chest X-ray of a patient with shortness of breath. FROM YOUR KNOWLEDGE: What diagnosis would you make? Pneumonia or atelectasis?

( ) Pneumonia

( ) Atelectasis

( ) I don't know / I require further information to answer this question

Task 1.4 Scientific Task

Imagine you have to prepare yourself for a scientific lecture about lung ultrasound. One of the key questions is the sensitivity of lung ultrasound in diagnosis of pneumothorax. You know that it is already a well-used diagnostic tool in some hospitals, but not yet considered by the majority of physicians. In a recent meeting you have heard that the sensitivity is high and you want to underline this in your presentation with a couple of scientific resources supporting that issue and give a concrete value.

( ) I don't know / I require further information to answer this question

( ) Answer:

7.1.2.2 Task answer sheets

Answer, Supporting Websites & Feedback - Task 1

Cite at least 3 websites (or more until you are confident in your answer) that you consider supportive. Please provide feedback on the search system.

1. KHRESMOI SUPPORT: Please use KHRESMOI to find the (evidence to support your) answer and cite at least 3 websites (or more until you are confident in your answer) that you consider supportive.

( ) I could not find the answer using KHRESMOI

( ) Answer:

Comments:

2. Supporting Website - URL 1

Comments:

3. Supporting Website - URL 2

Comments:

4. Supporting Website - URL 3

Comments:

5. What do you like about the system? What functionality/aspect helped you or did you like?

Comments:

6. What did you dislike about the system? Did anything annoy you or hold you up in the search?

Comments:

7. What functionality/aspect did you miss that you think would have helped you in your search and retrieval of the answer?

Comments:

Answer, Supporting Websites & Feedback - Task 2

Cite at least 3 websites (or more until you are confident in your answer) that you consider supportive. Please provide feedback on the search system.

1. KHRESMOI SUPPORT: Please use KHRESMOI to find the (evidence to support your) answer and cite at least 3 websites (or more until you are confident in your answer) that you consider supportive.

( ) I could not find the answer using KHRESMOI

( ) Answer:

Comments:

2. Supporting Website - URL 1

Comments:

3. Supporting Website - URL 2

Comments:

4. Supporting Website - URL 3

Comments:

5. What do you like about the system? What functionality/aspect helped you or did you like?

Comments:

6. What did you dislike about the system? Did anything annoy you or hold you up in the search?

Comments:

7. What functionality/aspect did you miss that you think would have helped you in your search and retrieval of the answer?

Comments:

Answer, Supporting Websites & Feedback - Task 3

Cite 1 website that you consider supportive. Please provide feedback on the search system.

1. KHRESMOI SUPPORT: Please use KHRESMOI to find the (evidence to support your) answer and cite at least 1 website (or more until you are confident in your answer) that you consider supportive (similar images).

( ) Pneumonia

( ) Atelectasis

( ) I could not find the answer using KHRESMOI

2. Supporting Website - URL 1

Comments:

3. What do you like about the system? What functionality/aspect helped you or did you like?

Comments:

4. What did you dislike about the system? Did anything annoy you or hold you up in the search?

Comments:

5. What functionality/aspect did you miss that you think would have helped you in your search and retrieval of the answer?

Comments:

Answer, Supporting Websites & Feedback - Task 4

Cite at least 3 websites (or more until you are confident in your answer) that you consider supportive. Please provide feedback on the search system.

1. KHRESMOI SUPPORT: Please use KHRESMOI to find the (evidence to support your) answer and cite at least 3 websites (or more until you are confident in your answer) that you consider supportive.

( ) I could not find the answer using KHRESMOI

( ) Answer:

Comments:

2. Supporting Website - URL 1

Comments:

3. Supporting Website - URL 2

Comments:

4. Supporting Website - URL 3

Comments:

5. What do you like about the system? What functionality/aspect helped you or did you like?

Comments:

6. What did you dislike about the system? Did anything annoy you or hold you up in the search?

Comments:

7. What functionality/aspect did you miss that you think would have helped you in your search and retrieval of the answer?

Comments:

7.1.2.3 Task Answers

7.1.2.3.1 Task 1

What do you like about the system? What functionality/aspect helped you or did you like?	
User ID	Comment
physA	The user interface has a clear overview. The search was very fast.
physB	aufteilung links in unterpunkte (patient information, research, drugs...)
physC	
physD	Filtering, Quick overview with highlighted keywords
physE	good presentation in the interface
physF	i liked that i found overview webpages that gave me the information about the score but also high level web pages with detailed information. unfortunately access to the high level web page was denied
physG	
physH	Artikel decken viele Bereiche ab. Zugang zu der Thematik aus verschiedenen Perspektiven möglich. Die Seite ist übersichtlich und gut strukturiert. Die persönliche Bibliothek ist hilfreich.
physI	No advertisements, good percentage of scientific material.
physJ	I liked that I can choose subgroupes. Might help to specify the search and to be faster
physK	
physL	Translation help
physM	Would be helpfull to have the subgroup for therapy
physN	

What did you dislike about the system? Did anything annoy you or hold you up in the search?	
User ID	Comment
physA	The results were confusing respectively unclear. In some cases, I had no evidence, whether the result page had a scientific background or not.
physB	das anklicken der aufteilung sollte im separaten fenster erfolgen, eine detailgenauere suche ergab jeweils keine treffer
physC	unspecificity of response to specific questions
physD	System a little slow (not search but GUI).
physE	to many graphs
physF	during the time given I had problems filtering the information and links. it took me too long to find orientation in the sources given. scrolling down in the middle window opens the preview of the source, but this makes it impossible to push the down button.
physG	
physH	Nein
physI	Very slow response to cursor inputs; lack of horizontal and vertical resolution takes a lot of scrolling.
physJ	I normally need more time to get to know a program, so maybe I will get used to it soon and will like it more later. I think for a quick search there is too much information at once
physK	It would have been easier to just use pubmed. pubmed is less cluttered. source is very clear in pubmed.
physL	Verlässlichkeit der Quellen z.B Forum- unübersichtlich
physM	Maybe I made a spelling mistake, would be great to have a 'did you mean ....' option
physN	

What functionality/aspect did you miss that you think would have helped you in your search and retrieval of the answer?	
User ID	Comment
physA	How actual the website/ publication?
physB	to get the publication year more clearly, you should see that at the beginning of the page on the left side.
physC	
physD	Questionable if feasible: Better access to external sources (like uptodate.com) Better links to medical societies, and their guidelines
physE	more queries
physF	clear marking of the sort of source
physG	
physH	Sortierungsmöglichkeiten in der Bibliothek selbst
physI	Negative filter option is missing. No 'tabbed browsing'.

physJ	
physK	
physL	export the source in endnote
physM	
physN	all items - German translation, search in German

7.1.2.3.2 Task 2

What do you like about the system? What functionality/aspect helped you or did you like?	
User ID	Comment
physA	
physB	same as before
physC	
physD	
physE	the rapid finding of answers
physF	
physG	
physH	Unterteilung der Quellen in verschiedene Kategorien selektieren vorab die relevanten Studien
physI	
physJ	I liked that there was the right undergroup so that this time the search was quick. Now I think that it is better to search with one topic first, before you get more specific in the question
physK	sub category 'medsearch' delivers only scientific articles
physL	
physM	That I did find the answer in a reasonable amount of time
physN	I found a source for my thought answer.

What did you dislike about the system? Did anything annoy you or hold you up in the search?	
User ID	Comment
physA	I wasn't able to mark my words in the searching window. Picture-results are, in my opinion, useless when searching for complex problems.
physB	die einzelne wortsuche ist schwierig, mit bindestrich oder anführungszeichen damit das system artikel sucht in denen es um 1. typ 2 diabetes und 2. maligne erkrankungen geht
physC	
physD	Too much results unrelated to the question. Especially for lay people too complicated.
physE	no direct answer to my question
physF	entering specific searching phrases only let to images. by entering only

	the main search items like 'diabetes cancer prevalence' led to a bigger amount of links but those weren't helpful for the question, but too specific as for example gen expressions or drinking water guidelines. I had troubles deleting the search items.
physG	
physH	Erläuterungen zu den Quellen sind teilweise zu wenig, vor allem bei Bildern
physI	Mouseover-event in search results (popup of the whole title) is slow and gets in your way while scrolling. editing your search only works by removing every single letter.
physJ	
physK	you should be referred to an original sized image when you click on it
physL	Verlässlichkeit der Quellen? in one source diabetes was only another link at the end of the homepage and so the article was not helpful
physM	filter system could be better
physN	

What functionality/aspect did you miss that you think would have helped you in your search and retrieval of the answer?	
User ID	Comment
physA	I had the impression, that the search engine was just collecting sites with the words 'diabetes' and 'malignancy' in it. But I wasn't able to find a true correlation of these two conditions.
physB	
physC	Compared with PubMed, this system is rather unspecific and in terms of scientific research not equally satisfying
physD	
physE	
physF	
physG	
physH	Die Fragestellung ist nicht ganz klar
physI	
physJ	
physK	
physL	
physM	I expected other information in the subgroup of disease biology
physN	Same answer like before

7.1.2.3.3 Task 3

What do you like about the system? What functionality/aspect	

helped you or did you like?	
User ID	Comment
physA	
physB	die möglichkeit der speziellen bildersuche als unterpunkt
physC	
physD	Many different pictures of the desired topic
physE	x-rays
physF	
physG	
physH	
physI	
physJ	
physK	
physL	in this case it was helpful to have many images and to see a miniature of it before opening it
physM	Pictures subgroup was very helpful
physN	picture search

What did you dislike about the system? Did anything annoy you or hold you up in the search?	
User ID	Comment
physA	The user interface was very slow, especially when I tried to scroll down or up.
physB	unklare aufteilung mit radiologie bei den unterpunkten. unwissenheit darüber ob das system auch auf lehrbuchartige quellen zugreift die für mich bei der dieser suche erste wahl sein würden
physC	
physD	
physE	no description of x--ray pictures
physF	trying to find description about what the different diseases look like in the chest x ray did not result in satisfying answers (similar problems as before, the results were either to unspecific or, if the question was more specific, the results were only images. I could find some images of both diseases but not enough to have a sufficient idea about the presentation in the x ray. nevertheless I changed my answer because of the pictures seen.
physG	
physH	
physI	Very few search results. Images should be preferably ones with a better quality.
physJ	
physK	
physL	That it is not possible to mark the words to delete them in the search

physM	to many item appear for a search
physN	not the whole pictures to see

What functionality/aspect did you miss that you think would have helped you in your search and retrieval of the answer?	
User ID	Comment
physA	Easier function, to hide unused features in the user interface, for having more space for the temporary more important things.
physB	unterteilung in wissenschaftliche quellen und Lehrbücher
physC	It would be helpful to have the possibility to take a chest-X-ray into the system and then to start a search to compare with similar morphologies and interpretations
physD	Maybe access to a known and tagged database of radiologic images.
physE	
physF	the possibility of selecting that only articles are shown were my search items are in the title of the article
physG	
physH	
physI	If you filter your search results, you should not be able to see the unfiltered results just by scrolling.
physJ	
physK	
physL	
physM	I would like to find the answer under the diagnosis subgroup
physN	as at question 1

7.1.2.3.4 Task 4

What do you like about the system? What functionality/aspect helped you or did you like?	
User ID	Comment
physA	
physB	bei dieser fragestellung war die aufteilung in grobe suche oben und genauere suche unten hilfreich, weil auch klassisch wissenschaftliche fragestellung
physC	
physD	
physE	rapid answer, ultrasound picture
physF	it was the first question I could answer, it was the right combination of overview webpages and scientific articles.
physG	
physH	

physI	good results
physJ	liked the possibility to compare x-rays
physK	
physL	in this case the system was very helpful
physM	I did know what I was looking for (research article) and I found a matching article for my question
physN	

What did you dislike about the system? Did anything annoy you or hold you up in the search?	
User ID	Comment
physA	
physB	wiederum keine Jahresanzahlangaben bei den wissenschaftlichen artikeln. angegebener artikel stammt aus jahr 2006, man hätte es durch noch aktuellere studien unterstützen sollen
physC	
physD	Too many results not containing (all of) the original search terms, at times completely unrelated.
physE	
physF	
physG	
physH	
physI	annoyingly slow
physJ	
physK	Feels like not directly researching the source, with pubmed you have the feeling that you get straight to the journal article. In contrast, the search engine feels like there is on more step to go
physL	-
physM	Many items/result did not fit my questions e.g. articles about breast cancer case report
physN	

What functionality/aspect did you miss that you think would have helped you in your search and retrieval of the answer?	
User ID	Comment
physA	
physB	angabe der jahreszahlen tabellen aus studien in der vorschau
physC	
physD	
physE	

physF	it would be nice to have pubmed reviews dealing with the question listed on the first position. matching of the search terms with the titles of the articles.
physG	
physH	
physI	Export should be available as PDF directly into Email HTML only as a link, not as the entire web page
physJ	
physK	
physL	
physM	
physN	

7.1.3 User satisfaction

7.1.3.1 User satisfaction questionnaire

How did you like Khresmoi?

In the following part we are interested in your feedback. Please answer the questions as honestly as possible. If you don't understand a question please ask the researcher for further clarification.

1. It was easy to store websites in the Personal Library (tray)
strongly agree 1 2 3 4 5 strongly disagree

2. I would find the personal library (tray) a helpful tool for my work.
strongly agree 1 2 3 4 5 strongly disagree

3. The Tags function is a helpful tool for my work.
strongly agree 1 2 3 4 5 strongly disagree

4. It was easy to answer to tasks using the search system.
strongly agree 1 2 3 4 5 strongly disagree

5. I understood the KHRESMOI system without further training.
strongly agree 1 2 3 4 5 strongly disagree

6. I would use KHRESMOI again to obtain medical information.
strongly agree 1 2 3 4 5 strongly disagree

7. Finding information took me more time than usual.
strongly agree 1 2 3 4 5 strongly disagree

8. The types of resources offered were what I looked for
strongly agree 1 2 3 4 5 strongly disagree

9. The organization of information on the system screens is clear
strongly agree 1 2 3 4 5 strongly disagree

10. The information is effective in helping me complete the tasks and scenarios
strongly agree 1 2 3 4 5 strongly disagree

11. Whenever I make a mistake using the system, I recover easily and quickly
strongly agree 1 2 3 4 5 strongly disagree

12. I think that I would like to use this system frequently
strongly agree 1 2 3 4 5 strongly disagree

13. I found the system unnecessarily complex.
strongly agree 1 2 3 4 5 strongly disagree

14. I thought the system was easy to use.
strongly agree 1 2 3 4 5 strongly disagree

15. I think that I would need the support of a technical person to be able to use this system
strongly agree 1 2 3 4 5 strongly disagree

16. I found the various functions in this system were well integrated.
strongly agree 1 2 3 4 5 strongly disagree

17. I thought there was too much inconsistency in this system.
strongly agree 1 2 3 4 5 strongly disagree

18. I would imagine that most people would learn to use this system very quickly
strongly agree 1 2 3 4 5 strongly disagree

19. I found the system very awkward to use
strongly agree 1 2 3 4 5 strongly disagree

20. I felt very confident using the System
strongly agree 1 2 3 4 5 strongly disagree

21. How did you like KHRESMOI? Any improvement suggestion?
Comments:

7.1.3.2 Answers user satisfaction questionnaire

It was easy to store websites in the Personal Library (tray)							
User ID	strongly agree	1	2	3	4	5	strongly disagree
physA		X					
physB			X				
physC		X					
physD		X					
physE		X					
physF		X					
physG							
physH		X					
physI		X					
physJ		X					
physK		X					
physL		X					
physM		X					
physN			X				

I would find the personal library (tray) a helpful tool for my work.							
User ID	strongly agree	1	2	3	4	5	strongly disagree
physA		X					
physB			X				
physC			X				
physD					X		
physE			X				

physF		X					
physG							
physH		X					
physI			X				
physJ		X					
physK		X					
physL			X				
physM		X					
physN		X					

The Tags function is a helpful tool for my work.							
User ID	strongly agree	1	2	3	4	5	strongly disagree
physA				X			
physB				X			
physC							
physD					X		
physE			X				
physF				X			
physG							
physH		X					
physI							
physJ							
physK			X				
physL			X				
physM		X					
physN			X				

It was easy to answer to tasks using the search system.							
User ID	strongly agree	1	2	3	4	5	strongly disagree

physA					X		
physB				X			
physC			X				
physD						X	
physE			X				
physF					X		
physG							
physH		X					
physI					X		
physJ				X			
physK					X		
physL				X			
physM				X			
physN				X			

I understood the KHRESMOI system without further training.							
User ID	strongly agree	1	2	3	4	5	strongly disagree
physA			X				
physB			X				
physC			X				
physD		X					
physE		X					
physF			X				
physG							
physH			X				
physI		X					
physJ			X				
physK		X					
physL			X				
physM			X				
physN				X			

I would use KHRESMOI again to obtain							

medical information.							
User ID	strongly agree	1	2	3	4	5	strongly disagree
physA					X		
physB				X			
physC				X			
physD						X	
physE		X					
physF				X			
physG							
physH		X					
physI				X			
physJ			X				
physK				X			
physL				X			
physM					X		
physN			X				

Finding information took me more time than usual.							
User ID	strongly agree	1	2	3	4	5	strongly disagree
physA			X				
physB			X				
physC			X				
physD		X					
physE				X			
physF			X				
physG							
physH						X	
physI		X					
physJ				X			
physK		X					
physL			X				
physM				X			
physN			X				

The types of resources							

offered were what I looked for							
User ID	strongly agree	1	2	3	4	5	strongly disagree
physA					X		
physB				X			
physC				X			
physD					X		
physE				X			
physF					X		
physG							
physH			X				
physI		X					
physJ				X			
physK					X		
physL				X			
physM			X				
physN				X			

The organization of information on the system screens is clear							
User ID	strongly agree	1	2	3	4	5	strongly disagree
physA					X		
physB		X					
physC				X			
physD				X			
physE		X					
physF			X				
physG							
physH		X					
physI			X				
physJ		X					
physK			X				
physL			X				
physM					X		

physN			X				

The information is effective in helping me complete the tasks and scenarios							
User ID	strongly agree	1	2	3	4	5	strongly disagree
physA				X			
physB			X				
physC				X			
physD					X		
physE			X				
physF					X		
physG							
physH							
physI				X			
physJ				X			
physK					X		
physL			X				
physM					X		
physN				X			

Whenever I make a mistake using the system, I recover easily and quickly							
User ID	strongly agree	1	2	3	4	5	strongly disagree
physA			X				
physB		X					
physC				X			
physD		X					
physE			X				
physF				X			
physG							

physH		X					
physI		X					
physJ			X				
physK		X					
physL			X				
physM		X					
physN				X			

I think that I would like to use this system frequently							
User ID	strongly agree	1	2	3	4	5	strongly disagree
physA					X		
physB				X			
physC			X				
physD						X	
physE			X				
physF				X			
physG							
physH		X					
physI					X		
physJ			X				
physK					X		
physL				X			
physM			X				
physN				X			

I found the system unnecessarily complex.							
User ID	strongly agree	1	2	3	4	5	strongly disagree
physA					X		
physB					X		
physC					X		
physD					X		
physE					X		
physF					X		

physG							
physH						X	
physI				X			
physJ					X		
physK		X					
physL					X		
physM					X		
physN						X	

I thought the system was easy to use.							
User ID	strongly agree	1	2	3	4	5	strongly disagree
physA				X			
physB			X				
physC			X				
physD		X					
physE			X				
physF			X				
physG							
physH		X					
physI		X					
physJ			X				
physK				X			
physL			X				
physM			X				
physN			X				

I think that I would need the support of a technical person to be able to use this system							

User ID	strongly agree	1	2	3	4	5	strongly disagree
physA					X		
physB						X	
physC					X		
physD						X	
physE						X	
physF						X	
physG							
physH						X	
physI						X	
physJ						X	
physK						X	
physL					X		
physM					X		
physN						X	

I found the various functions in this system were well integrated.							
User ID	strongly agree	1	2	3	4	5	strongly disagree
physA					X		
physB			X				
physC			X				
physD			X				
physE			X				
physF			X				
physG							
physH		X					
physI				X			
physJ		X					
physK				X			
physL			X				
physM			X				
physN			X				

I thought there was too much inconsistency in this system.							
User ID	strongly agree	1	2	3	4	5	strongly disagree
physA				X			
physB				X			
physC					X		
physD						X	
physE				X			
physF				X			
physG							
physH						X	
physI			X				
physJ					X		
physK			X				
physL					X		
physM		X					
physN					X		

I would imagine that most people would learn to use this system very quickly							
User ID	strongly agree	1	2	3	4	5	strongly disagree
physA				X			
physB		X					
physC		X					
physD		X					
physE			X				
physF		X					
physG							
physH		X					

physI			X				
physJ			X				
physK		X					
physL			X				
physM		X					
physN			X				

I found the system very awkward to use							
User ID	strongly agree	1	2	3	4	5	strongly disagree
physA					X		
physB						X	
physC						X	
physD						X	
physE					X		
physF						X	
physG							
physH						X	
physI					X		
physJ						X	
physK					X		
physL					X		
physM				X			
physN						X	

I felt very confident using the System							
User ID	strongly agree	1	2	3	4	5	strongly disagree
physA			X				
physB			X				
physC			X				
physD			X				
physE				X			

physF			X				
physG							
physH		X					
physI			X				
physJ				X			
physK		X					
physL				X			
physM		X					
physN				X			

How did you like KHRESMOI? Any improvement suggestion?	
User ID	Comment
physA	
physB	verweis auf die mündlichen feedbacks/Verbesserungsvorschläge während der bearbeitung der aufgaben
physC	Interesting new tool
physD	Use/implementation of big, reliable sources.
physE	help to find the right choice of queries
physF	
physG	
physH	very helpful
physI	I did like the quality of the results. It was way to slow. You need a very high resolution on your screen => mobile use does not seem very likely. I'd like more search results + filtering options. The tags don't really work - maybe you should reduce those to 'Images' + 'Scientific' and only one or two others. You could also remove it alltogether and make the tags only available by user choice. I would only use Khresmoi if it became way faster on mouse/keyboard inputs and it found a lot more search results.
physJ	I would not use it for a quick search. For more detailed information I think it is very helpful.
physK	
physL	add articles of the library to endnote
physM	
physN	je nach zielgruppe, wenn für niedergelassene allgemeinmediziner suche und ergebnisliste auf deutsch.

7.1.4 Pilot test protocols

7.1.4.1 Pilot test 1

Preparation:

Laptop Acer Aspire 5830TG, 15.6" used with Morae Recorder, ezDL swing prototype, visitor account.

Demographics:

Age: 27, medical student, 6th year

Occupation: student

Level of Medical English: average

Level of Computer skill: high

Procedure:

In the first 10 minutes the user was informed on the purpose of the study, was given the consent sheet and introduced to a demo search in the swing prototype interface. Following Morae recorder was started (audio and screen recording) and the autopilot asked to enter demographic data.

After that 4 tasks were given to the user (task 1,3,10,12) to solve in 40 minutes. Each task was started manually by the researcher. After each task the user was asked to provide feedback within the auto pilot of Morae as well as verbally on what he found helpful (positive remarks), what resources and what attributes he missed or were helpful/important to him and what he disliked/prevented him to obtain the answer (negative remarks). Assistance was given during tasks when required. Free feedback during tasks was reported handwritten.

Queries, reported thoughts and any additional feedback were reported.

In addition, it was reported whether the user was able to solve the task and if he was satisfied. The search process was documented. In order to obtain as much feedback as possible from this pilot user task, the procedure was kept informal, allowing the user to ask questions at any point.

The user was debriefed after each task and at the end of the study.

After the completion of four tasks, the user was provided with SUS items and asked for their final feedback/suggestion on the improvement of the search system. The whole experiment took about 60 minutes.

TASKS:

Task 1: Atrial Fibrillation:

Initial knowledge: The student provided an incorrect answer.

First Query: cardioversion AND anticoagulation

Second Query: /

Important attributes reported:

“show by decade” and “show by target audience” is not relevant to the user.

Search results sometimes doubled.

Barriers:

Some information provided was not accessible (UpToDate – she would have needed to sign in), resources were shown twice, Query words were not reflected in the links provided, making it hard to select relevant sources.

Time taken: 5 minutes

The search process: First query provided results where the user scrolls down to find valuable resources. Same results appear twice (see screenshot).

Useful information is found in a resource provided by “Uptodate”. After clicking on the resource provided by UpToDate the user is confronted with the barrier of inaccessibility, since only registered users can access the database.

Moving back to the prototype the user is having a look at other resources from the query and is of the opinion that the excerpt isn’t meaningful because it starts in the mid of the sentence. Often there is no reference to the results.

Though of high-speed Internet connection, java ezDL searches slowly, changing the size of different windows (results, details, etc.) is slow.

Option “show by decade” and “show by target audience” is not relevant to the user. After 5 minutes of searching the user wants search the answer in a book and cancels the task frustrated.

Interface is confusing: Two many icons, too little colours. The user doesn’t know where to look first, too many windows. User refers to easy search structure, e.g. Pubmed.

Final remarks:

Positive:

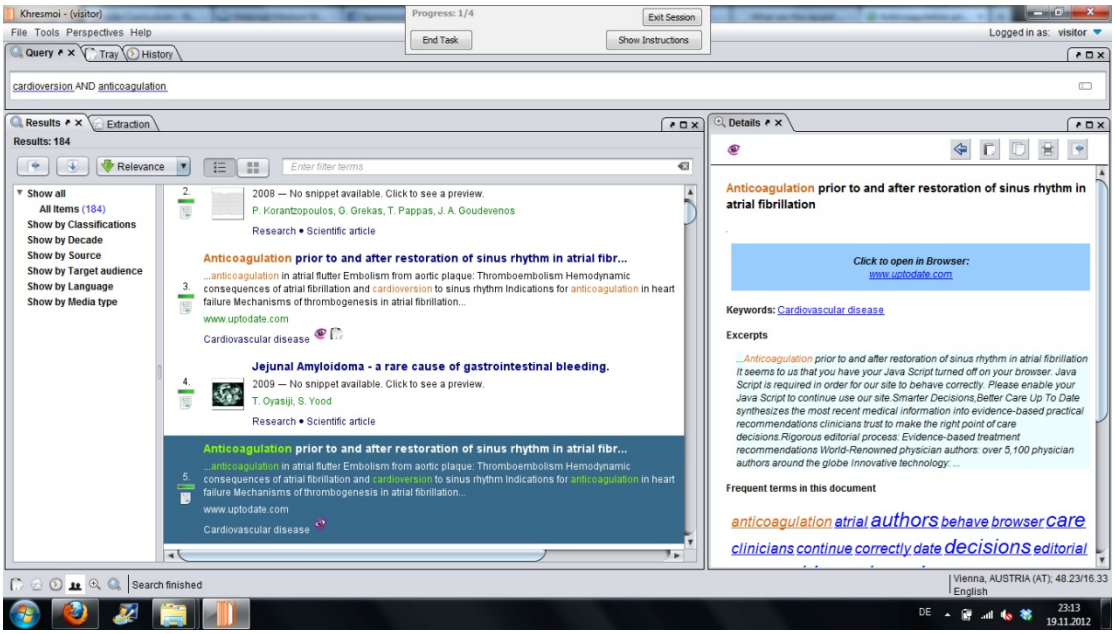
Was able to understand the interface.

Negative:

Interface is confusing. Potentially verifying answer was not accessible.

Excerpt seems not be meaningful because it starts in the mid of the sentence with no reference to the search results. Insufficient guidance in link on whether the answer is contained in website.

User could not solve the task.



Task 2: Diabetes and malignant illnesses

Initial knowledge: The student didn't know the answer

First Query: diabetes II

Second Query: diabetes II and malignant disease

Third Query: see Morae recording

Important attributes reported:

Translation tool offered nonsense word (malign -> Malibu, see screenshot)

Barriers: Inappropriate resources provided

Time taken: 4 minutes

The search process:

The user expects a fast answer related to wide information about diabetes II. After the first query, information in the results is too general. Adding "malignant disease" the user is offered a weird translation (malign -> Malibu). Displaying results do not show diabetes and malignant disease. Either the first part oft he query or only the second part is shown. Going on with the search ezDL crashes and needs to be restarted. After 4 minutes of search, the user wants to stop again and look up some information in books of internal medicine.

Simple query needs to be solved fast. If a book is not accessible the user would change to Google. The user could not solve the task using Khresmoi and was frustrated again.

Final remarks:

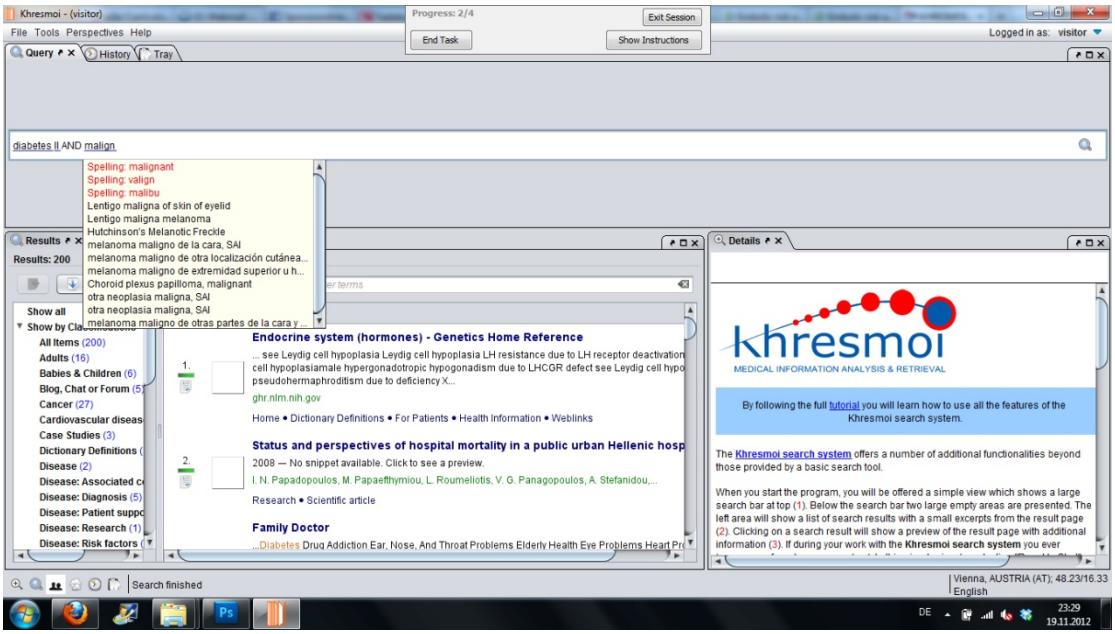
Positive: /

Negative:

Only parts of the query were displayed in the results.

ezDL crashed down during search.

Translation offered the user weird options (malign -> Malibu).



Task 3: Diagnosis of an X-ray

Initial knowledge: The student does not know the answer

First Query: x ray atelectasis

Second Query: x ray cardiac stasis

Important attributes reported:

Barriers:

Not enough picture material available.

Search process:

Interface mode is changed to image search perspective. Image in the task is of low quality (can be both answers from the small image).

User wants to see the picture the whole time during her search to compare directly. Query with “cardiac stasis” reveals no images with x-ray and cardiac stasis. Instead of expected results, websites with content of metoclopramide are shown. Frustrated of this search, the user changes query to “x-ray and atelectasis”. Also in these results no comparable image can be found.

With the image perspective in the interface, sometimes the title is not shown by scrolling the mouse over the result.

Khresmoi didn’t help to solve the question.

Final remarks:

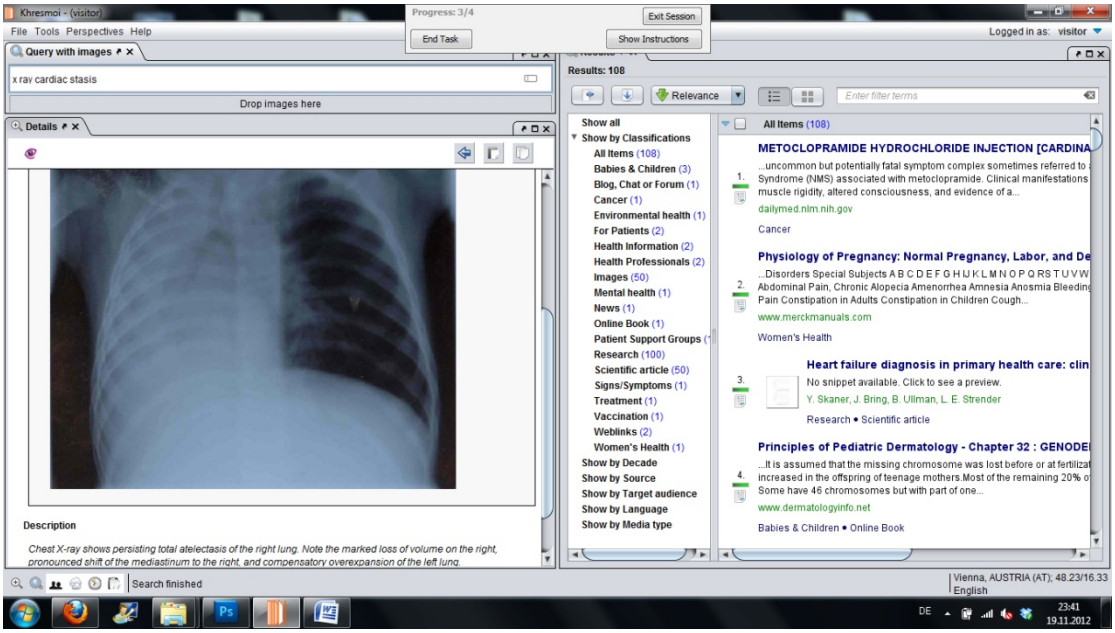
Positive: /

Negative:

Image quality on the paper is too low.

Query and results don’t match each other (metoclopramide <-> cardiac stasis).

With the image perspective in the interface, sometimes the title is not shown by scrolling the mouse over the result.



Task 4: Scientific Task

Initial knowledge: The student does not know the answer.

First Query: progesterone AND breast cancer

Second Query: see Morae recording

Important attributes reported:

Barriers:

Search process:

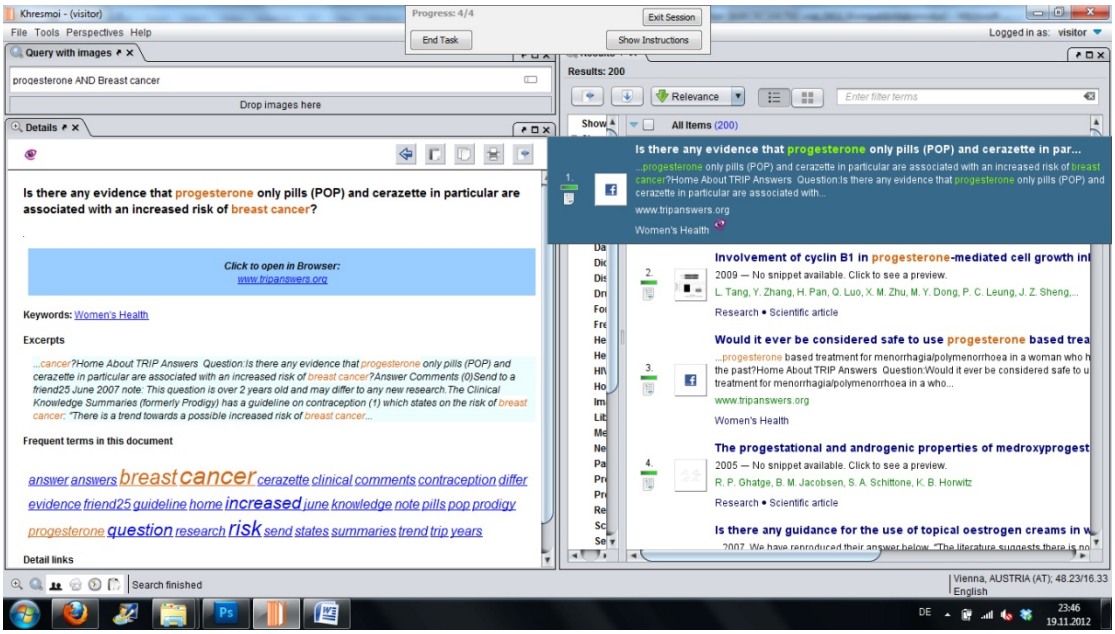
The first query is “progesterone breast cancer”. Various scientific results are displayed where the user cannot find hints to solve the task. After changing the query to a quite similar content other results appear but do not help the user to solve the task again. For usage of the interface she has to export content of a certain website with the export function. Uncomfortable possibilities of export formats are displayed. Copy and paste is easier and faster.

The User couldn’t find the answer but states again the overloaded interface (word cloud is not necessary) but states the export function useful if it will be improved to share the information with a colleague.

Final remarks:

Positive: /

Negative: User could not solve the task



Overall remarks:

Feedback/Suggestion made by the user:

User test setup:

User tasks need to be accessible the whole time – extra written form.

Question 4, demographics: Country? Which country is meant – country of origin or country to live in?

Question 19, overall feedback: Explain a word in a feedback is strange, just use the popular one. Question 1 – 3: What is the personal library, what is the tag function? User was logged in with a visitor account. Personal library and tag function weren't accessible.

User tasks, autopilot and Morae Recorder functioned well and easy structured. There is a difficulty in inserting cited websites, while a survey with Morae is running. Needs to be solved by manual insertion of the researcher.

Interface: Confusing overloaded interface, elimination of word cloud, slow interface

The Interface is confusing and overloaded with too many windows. Two many icons, too little colours. The user doesn't know where to look first, too many windows. User refers to easy search structure, e.g. Pubmed: Show one query row in the middle and have options to edit on the left and right in a column.

Insufficient guidance in the link on whether the answer is contained in website.

ezDL crashed during search. Prototype works quite slowly during search and optimizing the interface by changing size of windows.

Resources: Query and results do not match in different cases. Excerpts sometimes not meaningful. The link of specificity of query and provided information is important.

Excerpt seems not be meaningful because it starts in the mid of the sentence with no reference to the search results. Query and results don't match each other (metoclopramide <-> cardiac stasis). Only parts of the query were displayed in the results.

Tools: No special tools used from the user

Translation offered the user weird options in the query (malign -> Malibu).

Tray was used efficiently, but placed uncomfortable on the interface.

Rating:

Rating and answering final feedback questions of the user test needed 10 minutes as expected. Rating was generally put low because of frustration during search with the tasks. The user would not use Khresmoi if it stays like that. :

7.1.4.2 Pilot test 2

Preparation:

Laptop Acer Aspire 5830TG, 15.6" used with Morae Recorder, ezDL swing prototype, visitor account.

Demographics:

Age: 30, resident

Occupation: unemployed

Level of Medical English: average

Level of Computer skill: high

Procedure:

In the first 10 minutes the user was informed on the purpose of the study, was given the consent sheet and introduced to a demo search in the swing prototype interface. Following Morae recorder was started (video, audio and screen recording) and the autopilot asked to enter demographic data.

After that 4 tasks were given to the user (task 1,3,10,12) to solve in 40 minutes. Each task was started manually by the researcher. After each task the user was asked to provide feedback within the auto pilot of Morae as well as verbally on what he found helpful (positive remarks), what resources and what attributes he missed or were helpful/important to him and what he disliked/prevented him to obtain the answer (negative remarks). Assistance was given during tasks when required. Free feedback during tasks was reported handwritten.

Queries, reported thoughts and any additional feedback were reported.

In addition, it was reported whether the user was able to solve the task and if he was satisfied. The search process was documented. In order to obtain as much feedback as possible from this pilot user task, the procedure was kept informal, allowing the user to ask questions at any point.

The user was debriefed after each task and at the end of the study.

After the completion of four tasks, the user was provided with SUS items and asked for their final feedback/suggestion on the improvement of the search system. The whole experiment took about 60 minutes.

TASKS:

Task 1: Atrial Fibrillation:

Initial knowledge: The user provided a correct answer.

First Query: atrial fibrillation

Second Query: atrial fibrillation + extract function “anticoagulation”

Important attributes reported:

Slow search, definition of ranking by relevance? Translation tool missing in the extract function

Barriers:

No relevant information to the task could be found. User was not sure whether his search goes wrong or the interface confuses him.

Time taken: 14 minutes

The search process:

First query (atrial fibrillation) directly confronts the user with the translation tool. Because of the wrong spelling a correct wording is inserted, but changes the query to a doubled wording (see screenshots 1&2)

The search appears to be quite slowly and needs several seconds though there is a high-speed Internet connection. The user discovers the extract function and types in “anticoagulation”. Following no proper resource is found. The user tries to rank his results by relevance but is unsure about the meaning of that ranking. What is the definition of relevance?

He uses again the extract function and is confronted with no results but only sees a blank screen. The user would be happy about a message like “no results available” (see screenshot 3). Furthermore he mocks about no translation tool within the extract function while he is again not sure about a correct spelling.

The user puts one citation of a website into the tray.

Finally the user is blocked in the interface, extends the query but looses orientation. Restarting with a new query the user is looking for guidelines but fails to find them. In consequence could not solve the task

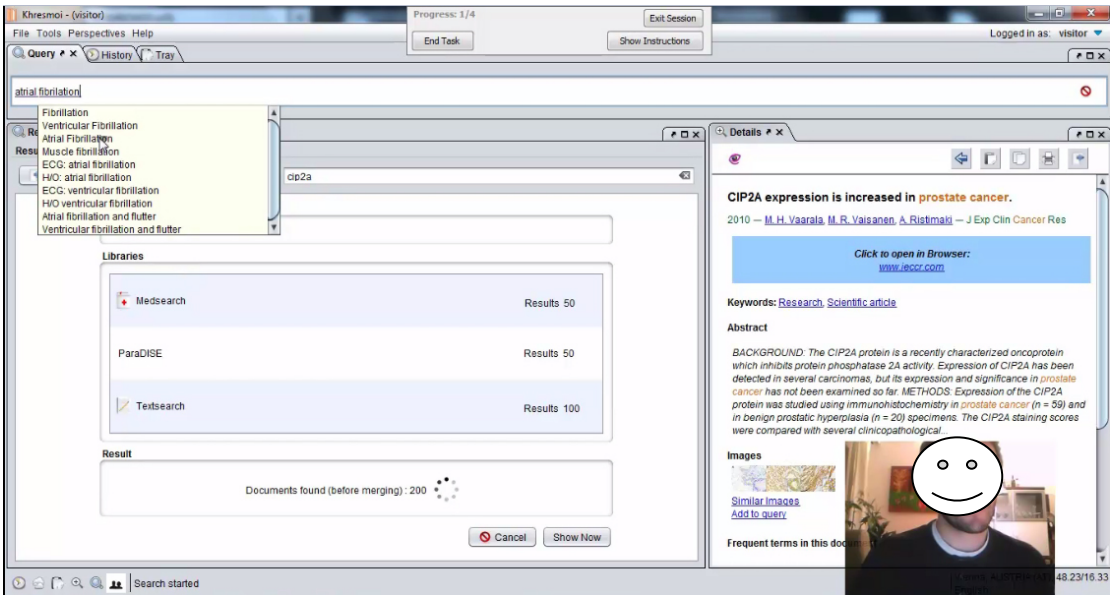
Final remarks:

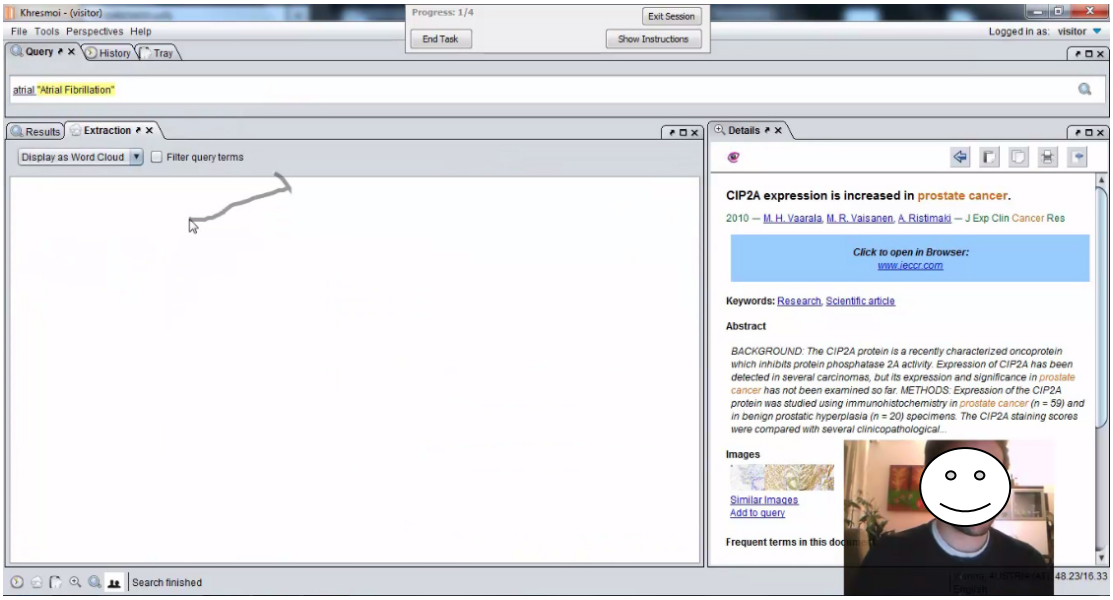
Positive:

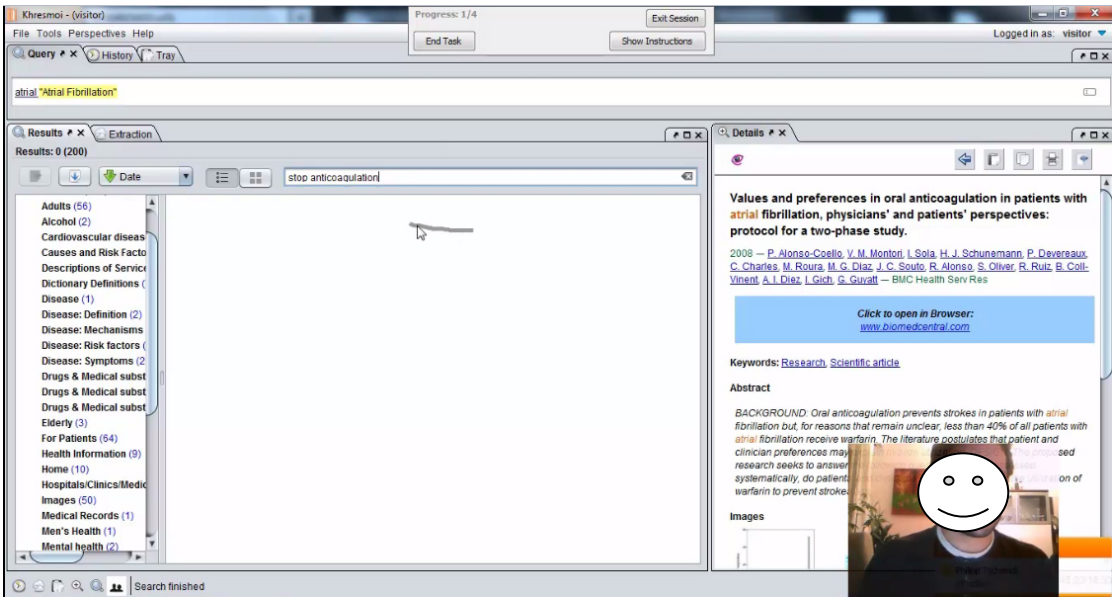
Excerpts in the details window help and give the impression of a summary of the chose website. Translation tool not only proposes the right spelling, it also gives options for word combinations.

Negative:

Interface is confusing. No translation tool in the extract function. User wants to have a short tutorial of query usage. The user could not solve the task


Khresmoi - (visitor)
File Tools Perspectives Help
Progress: 1/4
End Task
Exit Session
Show Instructions
Logged in as: visitor
Query
History
Tray
atrial fibrillation
Fibrillation
Ventricular Fibrillation
Atrial Fibrillation
Muscle fibrillation
ECG: atrial fibrillation
H/O: atrial fibrillation
ECG: ventricular fibrillation
H/O ventricular fibrillation
Atrial fibrillation and flutter
Ventricular fibrillation and flutter
cip2a
Libraries
Medsearch
Results 50
ParaDISE
Results 50
Textsearch
Results 100
Result
Documents found (before merging) : 200
Cancel
Show Now
Search started
Details
CIP2A expression is increased in prostate cancer.
2010 — M. H. Vaarala, M. R. Vaisanen, A. Ristimaki — J Exp Clin Cancer Res
Click to open in Browser:
www.jeccr.com
Keywords: Research, Scientific article
Abstract
BACKGROUND: The CIP2A protein is a recently characterized oncoprotein which inhibits protein phosphatase 2A activity. Expression of CIP2A has been detected in several carcinomas, but its expression and significance in prostate cancer has not been examined so far. METHODS: Expression of the CIP2A protein was studied using immunohistochemistry in prostate cancer (n = 59) and in benign prostatic hyperplasia (n = 20) specimens. The CIP2A staining scores were compared with several clinicopathological...
Images
Similar Images
Add to query
Frequent terms in this doc
48.23/16.33


Khresmoi - (visitor)
File Tools Perspectives Help
Progress: 1/4
End Task
Exit Session
Show Instructions
Logged in as: visitor
Query
History
Tray
atrial "Atrial Fibrillation"
Results
Extraction
Display as Word Cloud
Filter query terms
Search finished
Details
CIP2A expression is increased in prostate cancer.
2010 — M. H. Vaarala, M. R. Vaisanen, A. Ristimaki — J Exp Clin Cancer Res
Click to open in Browser:
www.jeccr.com
Keywords: Research, Scientific article
Abstract
BACKGROUND: The CIP2A protein is a recently characterized oncoprotein which inhibits protein phosphatase 2A activity. Expression of CIP2A has been detected in several carcinomas, but its expression and significance in prostate cancer has not been examined so far. METHODS: Expression of the CIP2A protein was studied using immunohistochemistry in prostate cancer (n = 59) and in benign prostatic hyperplasia (n = 20) specimens. The CIP2A staining scores were compared with several clinicopathological...
Images
Similar Images
Add to query
Frequent terms in this doc
48.23/16.33



Task 2: Diabetes and malignant illnesses

Initial knowledge: The user provided an incorrect answer

First Query: diabetes mellitus, type 2, risk of developing cancer

Second Query: diabetes mellitus, type 2, risk cancer

Third Query: diabetes mellitus, type 2, cancer

Important attributes reported:

Various queries where no results are displayed.

Barriers: No relevant information to the task could be found. User was not sure whether his search goes wrong or the interface confuses him. Limited choice of options – only scientific results.

Time taken: 10 minutes

The search process:

Directly from the beginning the user works with the translation tool. A long query is produced. He is confused about the displayed query, results and details that were not cleared. Therefore these three components do not match together (see screenshot 1). During the on going the search the user is actively asked for likes and dislikes. Again no relevant resources can be found.

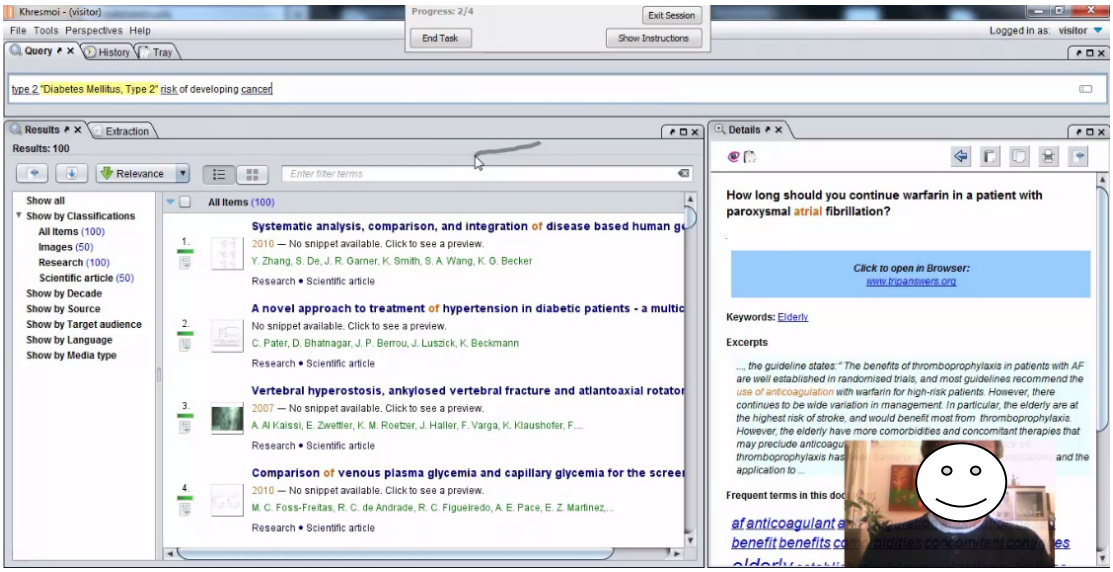
Only scientific results are displayed which gives the user a limited choice of options (see screenshot 2). By using the extract function with the word “malignancies” no results are found. The user feels helpless and tries again with “cancer”. Again no results are found. Now the user is confused of the interface. What is searched and displayed? All results or focused material? The user could not solve the task using Khresmoi and was frustrated.

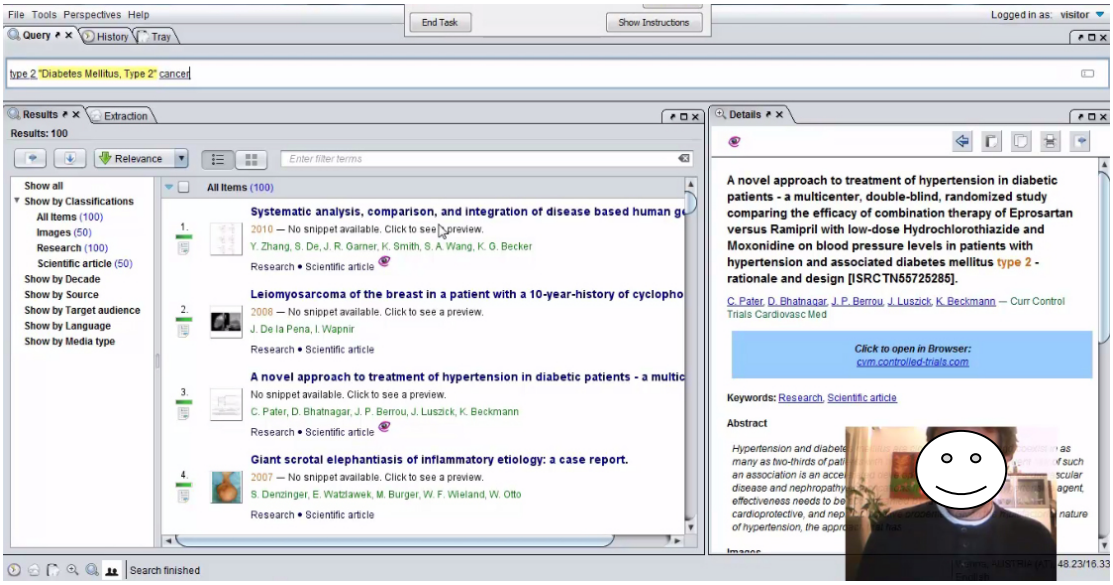
Final remarks:

Positive: Short overview in the excerpt is useful again

Negative: Interface is not user friendly. Three fields overstrain the user while is used to single one. To him it is a question of adaption.

User is not sure whether he failed or the interface with its search.


Khresmoi - (visitor)
File Tools Perspectives Help
Progress: 2/4
End Task
Exit Session
Show Instructions
Logged in as: visitor
Query
History
Tray
type 2 "Diabetes Mellitus, Type 2" risk of developing cancer
Results
Extraction
Results: 100
Relevance
Enter filter terms
Show all
Show by Classifications
All Items (100)
Images (50)
Research (100)
Scientific article (50)
Show by Decade
Show by Source
Show by Target audience
Show by Language
Show by Media type
All Items (100)
Systematic analysis, comparison, and integration of disease based human g
2010 — No snippet available. Click to see a preview.
Y. Zhang, S. De, J. R. Garner, K. Smith, S. A. Wang, K. G. Becker
Research • Scientific article
A novel approach to treatment of hypertension in diabetic patients - a multic
No snippet available. Click to see a preview.
C. Pater, D. Bhatnagar, J. P. Berrou, J. Luszick, K. Beckmann
Research • Scientific article
Vertebral hyperostosis, ankylosed vertebral fracture and atlantoaxial rotator
2007 — No snippet available. Click to see a preview.
A. Al Kaissi, E. Zwettler, K. M. Roetzer, J. Haller, F. Varga, K. Klaushofer, F....
Research • Scientific article
Comparison of venous plasma glycemia and capillary glycemia for the scree
2010 — No snippet available. Click to see a preview.
M. C. Foss-Freitas, R. C. de Andrade, R. C. Figueiredo, A. E. Pace, E. Z. Martinez,...
Research • Scientific article
Details
How long should you continue warfarin in a patient with paroxysmal atrial fibrillation?
Click to open in Browser:
www.tripanswers.org
Keywords: Elderly
Excerpts
... the guideline states: " The benefits of thromboprophylaxis in patients with AF are well established in randomised trials, and most guidelines recommend the use of anticoagulation with warfarin for high-risk patients. However, there continues to be wide variation in management. In particular, the elderly are at the highest risk of stroke, and would benefit most from thromboprophylaxis. However, the elderly have more comorbidities and concomitant therapies that may preclude anticoagu
thromboprophylaxis has
application to ...
and the
Frequent terms in this doc
af anticoagulant a
benefit benefits co
es


File Tools Perspectives Help
End Task
Show Instructions
Logged in as: visitor
Query
History
Tray
type 2 "Diabetes Mellitus, Type 2" cancer
Results
Extraction
Results: 100
Relevance
Enter filter terms
Show all
Show by Classifications
All Items (100)
Images (50)
Research (100)
Scientific article (50)
Show by Decade
Show by Source
Show by Target audience
Show by Language
Show by Media type
All Items (100)
Systematic analysis, comparison, and integration of disease based human g
2010 — No snippet available. Click to see a preview.
Y. Zhang, S. De, J. R. Garner, K. Smith, S. A. Wang, K. G. Becker
Research • Scientific article
Leiomyosarcoma of the breast in a patient with a 10-year-history of cyclopho
2008 — No snippet available. Click to see a preview.
J. De la Pena, I. Wapnir
Research • Scientific article
A novel approach to treatment of hypertension in diabetic patients - a multic
No snippet available. Click to see a preview.
C. Pater, D. Bhatnagar, J. P. Berrou, J. Luszick, K. Beckmann
Research • Scientific article
Giant scrotal elephantiasis of inflammatory etiology: a case report.
2007 — No snippet available. Click to see a preview.
S. Denzinger, E. Watzlawek, M. Burger, W. F. Wieland, W. Otto
Research • Scientific article
Search finished
Details
A novel approach to treatment of hypertension in diabetic patients - a multicenter, double-blind, randomized study comparing the efficacy of combination therapy of Eprosartan versus Ramipril with low-dose Hydrochlorothiazide and Moxonidine on blood pressure levels in patients with hypertension and associated diabetes mellitus type 2 - rationale and design [ISRCTN55725285].
C. Pater, D. Bhatnagar, J. P. Berrou, J. Luszick, K. Beckmann — Curr Control Trials Cardiovasc Med
Click to open in Browser:
cvm.controlled-trials.com
Keywords: Research, Scientific article
Abstract
Hypertension and diabete
many as two-thirds of pati
an association is an acce
disease and nephropathy
effectiveness needs to be
cardioprotective, and nep
of hypertension, the appr
as
if such
scular
agent,
nature

Task 3: Diagnosis of an X-ray

Initial knowledge: The user didn't know the answer.

First Query: x ray breast shortness of breath

Second Query: x ray cardiac stasis

Important attributes reported:

Barriers:

Picture with low quality – user test setup. No reasonable results to solve the task. Crash down of the interface.

Time taken: 13 minutes

Search process:

Directly after typing the first query the interface crashes down "Connection to backend failed" (see screenshot 1). The high-speed Internet connection is still available and the search process is broken. After restarting the user again enters his first query. The interface mode is manually changed to the image search perspective.

The user searches his results and sees findings not correlated to the query (x-ray – mr angio) (see screenshot 2). By scrolling down the first x-ray images appear. Five images are put into the tray. The user searches for a summary or excerpt to the pictures but cannot find anything. He hopes to get a better ranking or clearer results the extract function "cardiac stasis". No results are displayed. The same appears with "athelectase", where the user doesn't know the correct spelling but had liked to have the translation tool for help. He changes to the query to work with the translation tool. The correct spelling is finally identified. Scrolling the mouse over the results displays flickering extension of the content, which is uncomfortable for the user.

After five relevant resources are put into the tray the user changes the window to the tray overview. He needs quite a time to adapt the window size and wants to have an automatic size adaption or a short overlap of the whole information inside to compare to the correlating pictures.

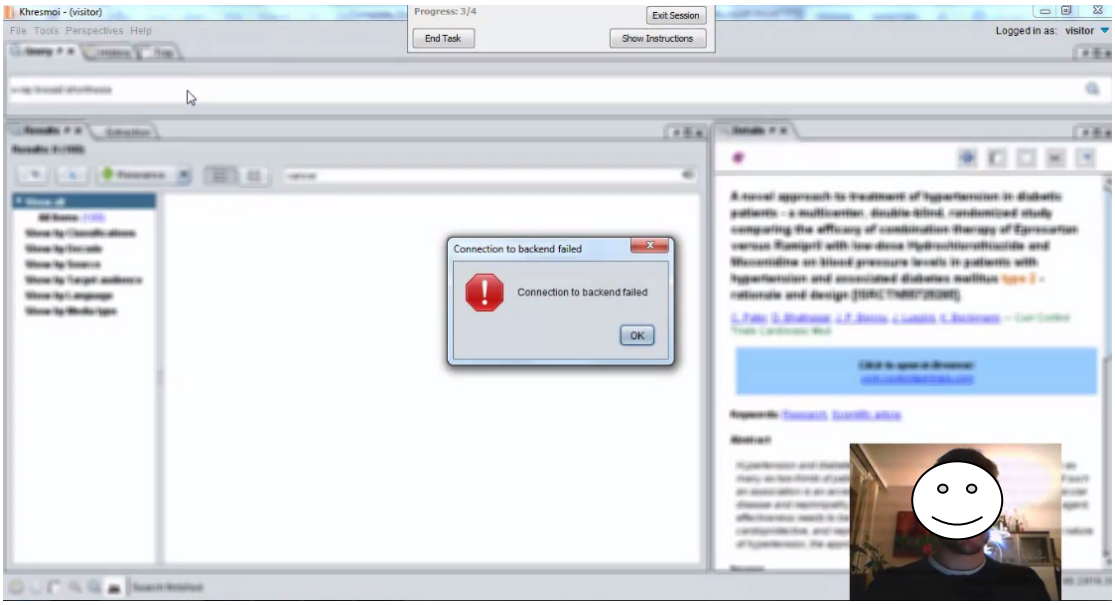
Khresmoi didn't help the user to solve the question.

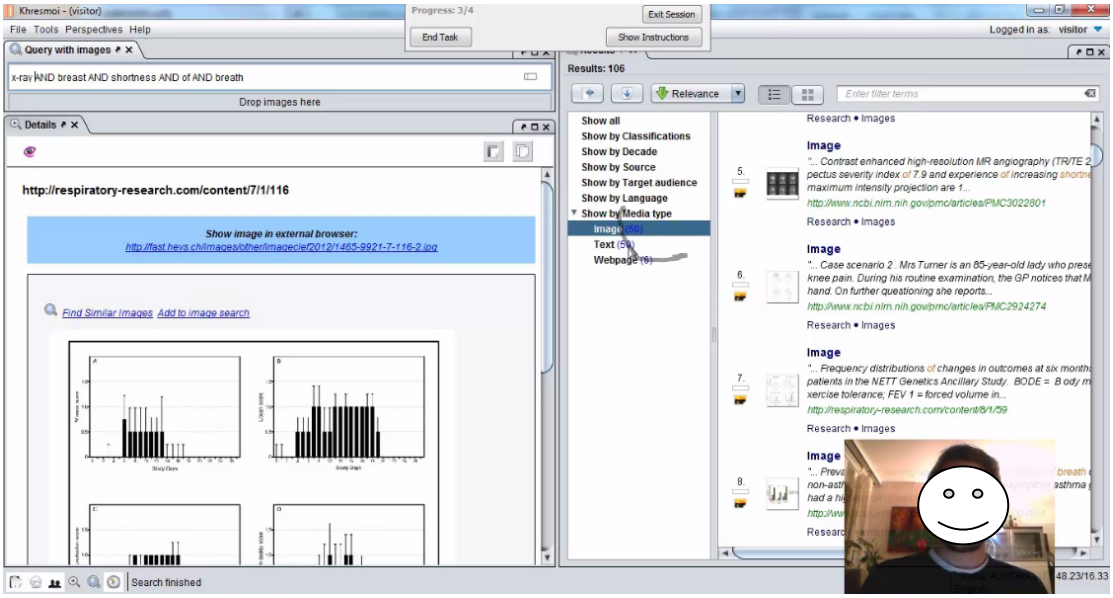
Final remarks:

Positive: /

Negative:

Image quality is too low. Picture description often missing. Uncomfortable interface with adapting field sizes.


Khresmoi - (visitor)
File Tools Perspectives Help
Progress: 3/4
End Task
Exit Session
Show Instructions
Logged in as: visitor
Connection to backend failed
Connection to backend failed
OK


Khresmoi - (visitor)
File Tools Perspectives Help
Progress: 3/4
End Task
Exit Session
Show Instructions
Logged in as: visitor
Query with images
x-ray AND breast AND shortness AND of AND breath
Drop images here
Details
http://respiratory-research.com/content/7/1/116
Show image in external browser:
http://fast.hevs.ch/images/otherImage/clef2012/1465-9921-7-116-2.jpg
Find Similar Images Add to image search
Results: 106
Relevance
Enter filter terms
Show all
Show by Classifications
Show by Decade
Show by Source
Show by Target audience
Show by Language
Show by Media type
Image (50)
Text (50)
Webpage (6)
Image
Research • Images
Search finished

Task 4: Scientific Task

Initial knowledge: The user does not know the answer.

First Query: artificial progesterone breast cancer

Second Query: artificial progesterone therapy breast cancer

Important attributes reported: /

Barriers:

Time taken: 11 minutes

Search process:

Answer option is missing in the task. User test setup will be completed

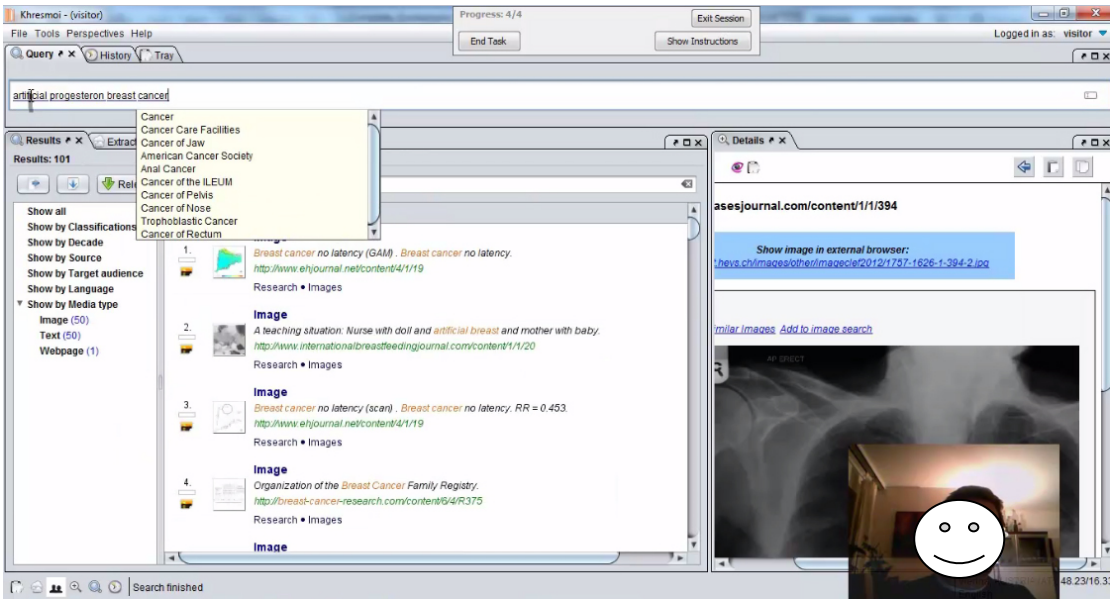
Print newspaper article so it can be seen every time, over pc or print

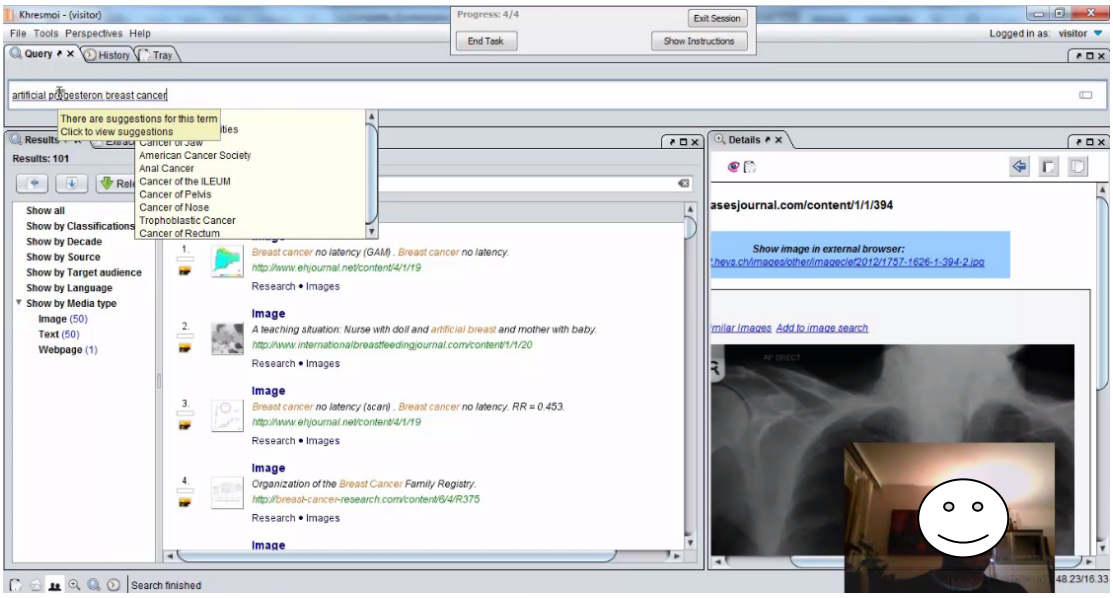
The user wants to use the query suggestions but they are just displayed for the last word, even if he clicks on the first or second word (see screenshot 1&2). Sometimes the suggestions appear automatically, sometimes they have are displayed by manual clicking on them. The search needs 23 seconds to display results, which is too slow to the user compared with other existing search engines. The results should directly be sorted by "show all" and not by the last chosen choice which was "show by media type" for the image search of task 3 for the first time the user tries out "show by" and selects Medsearch. He identifies a source with which he believes to solve the task and wants to share this source with a colleague. Therefore the export function is involved. Saving options deliver unuseful datatypes. The word cloud is identified as not necessary and gives too much information. The user couldn't find the answer to the task. Use the export function to share the information with a colleague.

Final remarks:

Positive: Existence of an export function

Negative: slow search, "moody" selection of word suggestions in the query





Overall remarks:

Feedback/Suggestion made by the user:

User test setup: Various small improvements identified

60 minutes is enough duration for the user test with 4 tasks and as well pre as post test information. Demographics – question 4: “What country do you mean? Country of origin or the country I actually live?” is identified as country of origin.

The finger pad may be a hint for the test, so the upcoming user test will be done with an additional mouse. Citation of websites is difficult during the autopilot of Morae Recorder. Copy and paste by the observer was the best solution onsite.

The quality of the x-ray picture needs to be improved. Furthermore the pictures need to be shown all the time, either in print version or on another computer display. The tasks of the autopilot should be read loudly. The font is not changeable in Morae and is displayed quite small.

In the overall feedback “Personal library” and “tag function” are unclear options.

The user was logged in with a visitor account. Personal library and tag function weren’t accessible.

Interface: Uncomfortable overloaded interface, no translation tool in the extract function, elimination of word cloud, slow interface, translation tool useful

Similar to Pilot 3 the Interface is confusing and overloaded with too many windows. This is uncomfortable to the user. He wishes to have one single window displayed with the option to change fast between the certain windows of the interface. He uses a Mac and is adapted to its design and comfortable setting. Furthermore the extract function misses a translation tool comparable to the query. The word cloud is unnecessary information.

During the test the search engine appeared to work quite slow up to 23 seconds for one search.

The translation tool as itself with its suggestions is quite helpful to the user.

Resources: Query, results and details sometimes do not match, resources not satisfying for solving the tasks

Query, results and details sometimes to do not match. This might be because of the earlier task and its results. The user was unsure whether he or the interface with the ranking and resources failed to solve the task.

None of the tasks could be solved or provided information to solve the tasks.

Tools: Tray and export function used

The Tray function was used in every task. The first two tasks to begin slowly and put one reliable resource into the tray. In the third task the user was instructed to put 5 resources in the tray and choose the best at the end of the task for solving. In task 4 the user asked to export his favourite website to share the content with a colleague. Uncomfortable formats to save the content appeared.

Rating:

Rating and answering final feedback questions of the user test needed 10 minutes as expected. Rating was generally put low because of frustration during search with the tasks. The user feels antipathy against Khresmoi at the current status.

7.1.4.3 Pilot test 3

Protocol of the fifth pilot user test:

Preparation:

Laptop Acer Aspire 5830TG, 15.6” used with Morae Recorder, ezDL swing prototype, visitor account.

Demographics:

Age: 30, Medical student, 6th year

Occupation: Medical student

Level of Medical English: average

Level of Computer skill: average

Procedure:

In the first 10 minutes the user was informed on the purpose of the study, was given the consent sheet and introduced to a demo search in the swing prototype interface. Following Morae recorder was started (audio and screen recording) and the autopilot asked to enter demographic data.

After that 4 tasks were given to the user (task 1,3,10,12) to solve in 40 minutes. Each task was started manually by the researcher. After each task the user was asked to provide feedback within the auto pilot of Morae as well as verbally on what he found helpful (positive remarks), what resources and what attributes he missed or were helpful/important to him and what he disliked/prevented him to obtain the answer (negative remarks). Assistance was given during tasks when required. Free feedback during tasks was reported handwritten.

Queries, reported thoughts and any additional feedback were reported.

In addition, it was reported whether the user was able to solve the task and if he was satisfied. The search process was documented. In order to obtain as much feedback as possible from this pilot user task, the procedure was kept informal, allowing the user to ask questions at any point.

The user was debriefed after each task and at the end of the study.

After the completion of four tasks, the user was provided with SUS items and asked for their final feedback/suggestion on the improvement of the search system. The whole experiment took about 60 minutes.

TASKS:

Task 1: Atrial Fibrillation

Initial knowledge: The student provided a correct answer.

First Query: stop anticoagulation after successful cardioversion

Second Query: anticoagulation after cardioversion

Important attributes reported:

Barriers:

Some information provided was not accessible (UpToDate – she would have needed to sign in), resources were shown twice.

Time taken: 16 minutes

The search process:

Directly after starting to read the task the user identifies a spelling mistake. The sentence is marked and corrected. In the task she also states, that the question is not clear enough: When was the cardioversion? Recently or in the past? The user proposes recently. Beginning to work with the interface she learns the fields and its options quite fast. She understands the query, the search, "what is displayed green", etc. After the first query the user also uses the "show by" function quite efficiently and chooses "show by classification". This gives her a valuable resource (heartdisease.about.com). The second resource is not accessible completely (UpToDate). After the pilot tests this is a known problem, so the user is advised to move on and find other resources to solve the task. During her search doubled resources appear (see screenshot 1 & 3). The user could not solve the task using Khresmoi but found quite valuable resources by scrolling often downwards.

Final remarks:

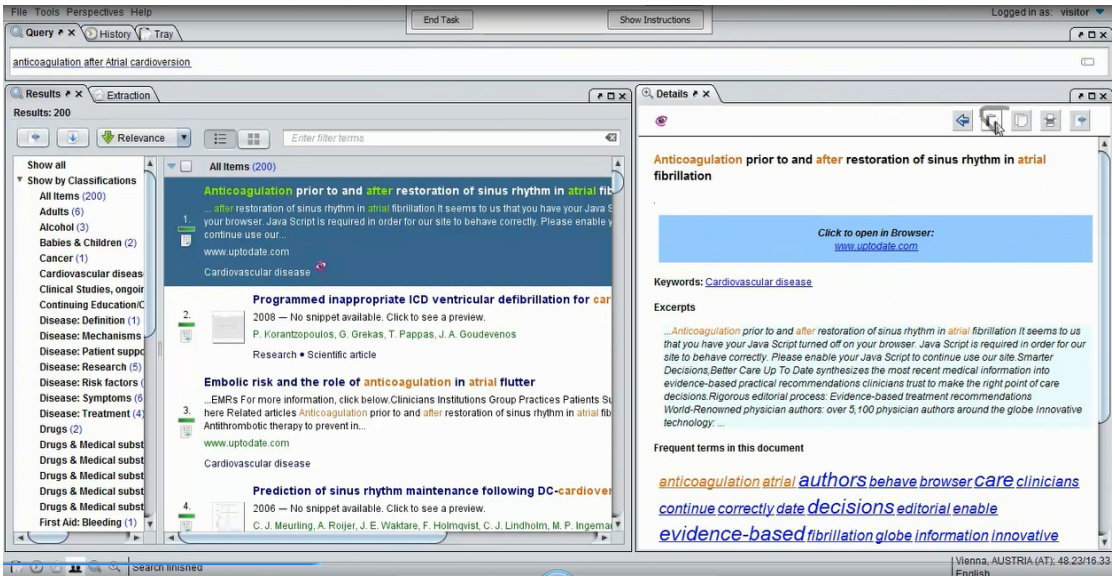
Positive:

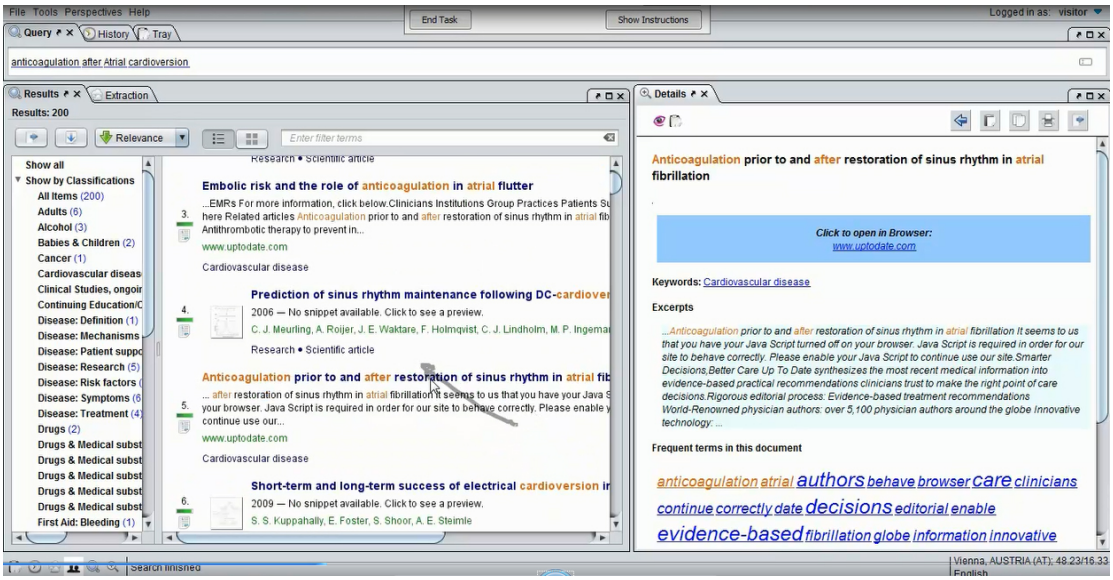
Sort function and translation tools are useful

Negative:

Question not precise enough.

Doubled results, not accessible resources (UpToDate)


File Tools Perspectives Help
End Task
Show Instructions
Logged in as: visitor
Query
History
Tray
anticoagulation after Atrial cardioversion
Results
Extraction
Results: 200
Relevance
Enter filter terms
Show all
Show by Classifications
All Items (200)
Adults (6)
Alcohol (3)
Babies & Children (2)
Cancer (1)
Cardiovascular disease
Clinical Studies, ongoir
Continuing Education/C
Disease: Definition (1)
Disease: Mechanisms
Disease: Patient suppo
Disease: Research (5)
Disease: Risk factors
Disease: Symptoms (6
Disease: Treatment (4
Drugs (2)
Drugs & Medical subst
First Aid: Bleeding (1)
All Items (200)
Anticoagulation prior to and after restoration of sinus rhythm in atrial fib
... after restoration of sinus rhythm in atrial fibrillation It seems to us that you have your Java S
your browser. Java Script is required in order for our site to behave correctly. Please enable y
continue use our...
www.uptodate.com
Cardiovascular disease
Programmed inappropriate ICD ventricular defibrillation for car
2008 — No snippet available. Click to see a preview.
P. Korantzopoulos, G. Grekas, T. Pappas, J. A. Goudevenos
Research • Scientific article
Embolic risk and the role of anticoagulation in atrial flutter
...EMRs For more information, click below.Clinicians Institutions Group Practices Patients Su
here Related articles Anticoagulation prior to and after restoration of sinus rhythm in atrial fib
Antithrombotic therapy to prevent in...
www.uptodate.com
Cardiovascular disease
Prediction of sinus rhythm maintenance following DC-cardiover
2006 — No snippet available. Click to see a preview.
C. J. Meurling, A. Roijer, J. E. Waktare, F. Holmqvist, C. J. Lindholm, M. P. Ingemar
Details
Anticoagulation prior to and after restoration of sinus rhythm in atrial fibrillation
Click to open in Browser:
www.uptodate.com
Keywords: Cardiovascular disease
Excerpts
...Anticoagulation prior to and after restoration of sinus rhythm in atrial fibrillation It seems to us that you have your Java Script turned off on your browser. Java Script is required in order for our site to behave correctly. Please enable your Java Script to continue use our site.Smarter Decisions,Better Care Up To Date synthesizes the most recent medical information into evidence-based practical recommendations clinicians trust to make the right point of care decisions.Rigorous editorial process: Evidence-based treatment recommendations World-Renowned physician authors: over 5,100 physician authors around the globe Innovative technology: ...
Frequent terms in this document
anticoagulation atrial authors behave browser care clinicians
continue correctly date decisions editorial enable
evidence-based fibrillation globe information innovative
Search finished
Vienna, AUSTRIA (AT); 48.23/16.33
English


File Tools Perspectives Help
End Task
Show Instructions
Logged in as: visitor
Query
History
Tray
anticoagulation after Atrial cardioversion
Results
Extraction
Results: 200
Relevance
Enter filter terms
Show all
Show by Classifications
All Items (200)
Adults (6)
Alcohol (3)
Babies & Children (2)
Cancer (1)
Cardiovascular disease
Clinical Studies, ongoir
Continuing Education/C
Disease: Definition (1)
Disease: Mechanisms
Disease: Patient suppo
Disease: Research (5)
Disease: Risk factors
Disease: Symptoms (6
Disease: Treatment (4
Drugs (2)
Drugs & Medical subst
First Aid: Bleeding (1)
Research • Scientific article
Embolic risk and the role of anticoagulation in atrial flutter
...EMRs For more information, click below.Clinicians Institutions Group Practices Patients Su
here Related articles Anticoagulation prior to and after restoration of sinus rhythm in atrial fib
Antithrombotic therapy to prevent in...
www.uptodate.com
Cardiovascular disease
Prediction of sinus rhythm maintenance following DC-cardiover
2006 — No snippet available. Click to see a preview.
C. J. Meurling, A. Roijer, J. E. Waktare, F. Holmqvist, C. J. Lindholm, M. P. Ingemar
Research • Scientific article
Anticoagulation prior to and after restoration of sinus rhythm in atrial fib
... after restoration of sinus rhythm in atrial fibrillation It seems to us that you have your Java S
your browser. Java Script is required in order for our site to behave correctly. Please enable y
continue use our...
www.uptodate.com
Cardiovascular disease
Short-term and long-term success of electrical cardioversion in
2009 — No snippet available. Click to see a preview.
S. S. Kuppahally, E. Foster, S. Shoor, A. E. Steimle
Details
Anticoagulation prior to and after restoration of sinus rhythm in atrial fibrillation
Click to open in Browser:
www.uptodate.com
Keywords: Cardiovascular disease
Excerpts
...Anticoagulation prior to and after restoration of sinus rhythm in atrial fibrillation It seems to us that you have your Java Script turned off on your browser. Java Script is required in order for our site to behave correctly. Please enable your Java Script to continue use our site.Smarter Decisions,Better Care Up To Date synthesizes the most recent medical information into evidence-based practical recommendations clinicians trust to make the right point of care decisions.Rigorous editorial process: Evidence-based treatment recommendations World-Renowned physician authors: over 5,100 physician authors around the globe Innovative technology: ...
Frequent terms in this document
anticoagulation atrial authors behave browser care clinicians
continue correctly date decisions editorial enable
evidence-based fibrillation globe information innovative
Search finished
Vienna, AUSTRIA (AT); 48.23/16.33
English

Task 2: Diabetes and malignant illnesses

Initial knowledge: The student didn't know the answer

First Query: Type 2 diabetes malignancies

Second Query: Type 2 diabetes risk of malignancies

Third Query: Type 2 diabetes risk of cancer

Important attributes reported:

Translation tool useful and clear, tag function and query mixed up

Barriers: /

Time taken: 4 minutes

The search process:

The user was quite fast and confident with the setting now though the tag function and the query are mixed up in the beginning. After the first query a valuable resource was displayed where a first hint to the answer was given. She supports the sort function "show by" and often used it. Furthermore direct displaying of the abstract is useful and fast for the search. She states, that small words from the query like "of" are taken to display results. The other words are not included, which does not help the user to find valuable resources. After clicking through some more resources the user is sure to have the answer to the task and ends it quite fast. Unfortunately she provides a wrong answer to say that there is no elevated risk of malignancies in diabetes type II.

Final remarks:

Positive:

Function "Show by", Abstract directly displayed

Translation tool where the spelling is clearly marked with red colour

Negative:

Wrong answer provided

Task 3: Diagnosis of an X-ray

Initial knowledge: The student provides an incorrect answer

First Query: atelectasis x ray

Second Query: /

Important attributes reported: /

Barriers:

Picture with low quality – user test setup (known problem)

Time taken: 13 minutes

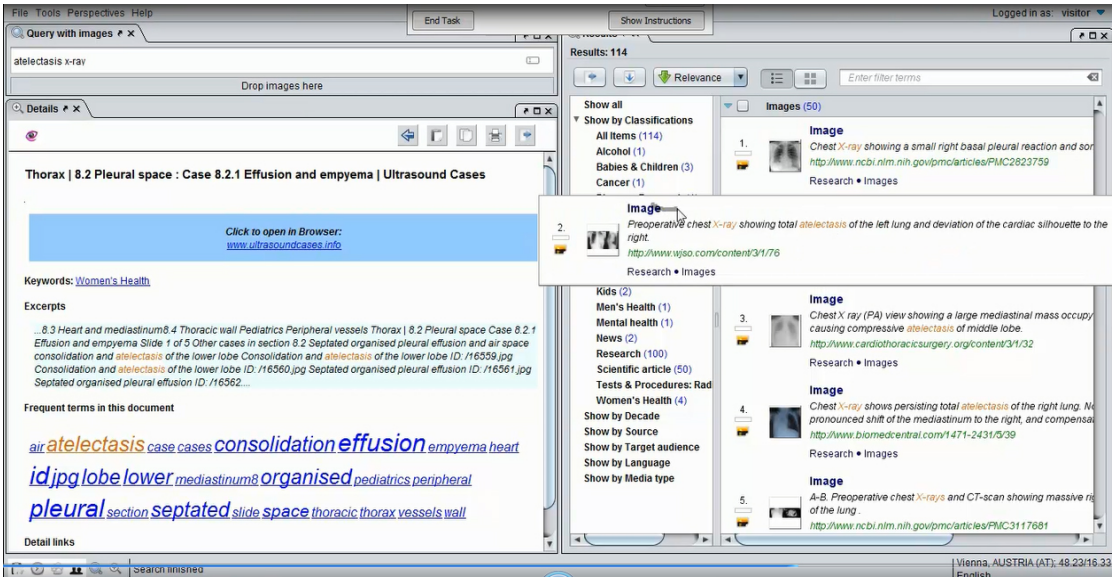
Search process:

The interface is changed to the image perspective. The first query “atelectasis x ray” displays several image results. The user needs some time to identify the picture section and that the interface automatically scrolls down to this part of the results. She wants to open the resource by scrolling over and is not linked to the result. Maybe this problem just appears in the image perspective, so the view is changed to the normal perspective because of easier handling and experience from the recent tasks. After a couple of minutes of search the user provides an incorrect answer (atelectasis). She wishes to have easier literature like “atelectasis, x-ray, and signs”. This search is quite easier with Google. Only scientific literature is blocking the task solution with Khresmoi

Final remarks:

Positive: /

Negative: Easier literature missing, task does not need scientific resources



Task 4: Scientific Task (was not included because of time overlapping)

Initial knowledge:

First Query:

Second Query:

Important attributes reported:

Overall remarks:

Easier literature is missing. Not every task needs scientific material. Wikipedia and schoolbooks need to be integrated. The advantage of this search engine lies in recruiting only medical literature (guidelines, “easier” and scientific literature) in contrast to Google where every resource sometimes with low quality is displayed.

Feedback/Suggestion made by the user:

User test setup:

User tasks need to be accessible the whole time – extra written form will be provided (was provided only by a second computer. First task needs to be more precisely. Old problems again identified and solved:

Question 4, demographics: Country? Which country is meant – country of origin or country to live in?

Question 19, overall feedback: Explain a word in a feedback is strange, just use the popular one. Question 1 – 3: What is the personal library, what is the tag function? User was logged in with a visitor account. Personal library and tag function weren’t accessible.

User tasks, autopilot and Morae Recorder functioned well and easy structured. There is a difficulty in inserting cited websites, while a survey with Morae is running. Needs to be solved by manual insertion of the researcher.

Interface: Translation tool and sort function useful

Directly from the beginning the user rated the sort function and translation tools as useful and positive. “Show by” was often used and helpful. Furthermore direct displaying of the abstract promoted the search process.

The spelling is clearly marked with red colour in the translation tool.

Word cloud not necessary, interface overloaded

Resources: Not accessible resources (UpToDate), doubled results

During one task doubled resources appeared. In another task not accessible resources were identified to solve the task (UpToDate). Unfortunately the abstract to these resources couldn’t help in solving the tasks. Two tasks were solved, but with a wrong answer. Consequently the resource supports the user in failing most of the tasks.

Tools: Translation tool and sort tool useful

See above

Rating:

Rating and answering final feedback questions of the user test needed 10 minutes as expected. Rating was generally put average or low because of frustration during search with the tasks.

7.2 Consent form for the general public evaluations (blind comparison) (French)

Formulaire de consentement

Je, soussigné, Mme/M. _____, donne par la présente mon consentement à ma participation à une étude scientifique menée dans le cadre d'un projet de l'UE à laquelle la Fondation Health On the Net participe. Cette étude vise à évaluer la qualité des résultats /sites web de différents moteurs de recherche dans le cadre de la recherche sur internet d'informations au sujet de la santé, de manière « transparente » et « fiables ».

J'ai été informé sur le contenu, le but et la portée de cette étude. J'ai eu suffisamment de temps pour réfléchir au sujet de ma participation. Je certifie, que si j'avais des questions, des réponses satisfaisantes y ont été apportées. Je comprends que ma participation est volontaire et peut être résiliée à tout moment, sans aucune raison particulière. Je comprends que ma participation à cette étude n'influence pas mon état de santé, et dès lors, qu'aucune couverture d'assurance ne m'est fournie.

Le test dure environ 45 minutes. Après avoir été familiarisé à une situation sur la santé, je vais devoir évaluer comparativement les résultats /sites web de deux moteurs de recherche en répondant à six questions.

Les chercheurs principaux, ainsi que tous les membres du projet concernés, s'engagent à n'utiliser les données recueillies que sous une forme anonyme. Ils sont tenus de traiter les données et observations de manière confidentielle. Aucune donnée personnelle identifiable n'est transmise à des tiers ou commercialisées.

Je donne ma permission aux membres du projet en question d'avoir accès à mes données anonymes recueillies lors de l'étude. Je suis d'accord avec le fait que ces données seront utilisées dans le cadre de la recherche pour améliorer la recherche sur internet d'informations au sujet de la santé.

_____ _____

*Nom du **participant** en MAJUSCULES* *Nom de l'**organisation** et **chercheur***

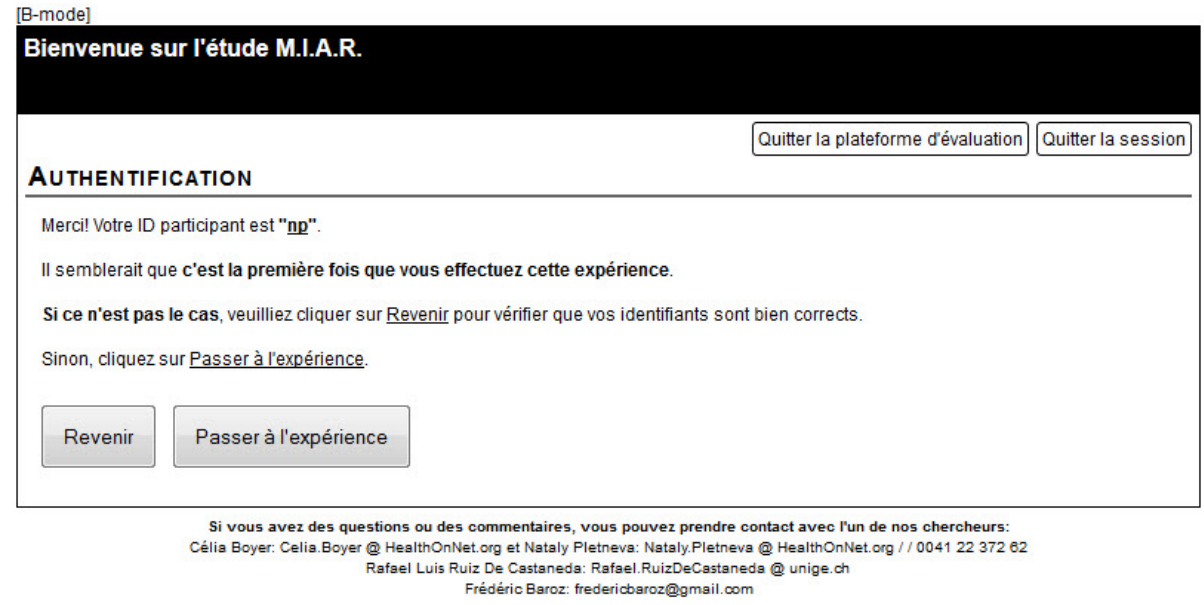
_____ _____

Date et signature *Date et signature*

Ce formulaire doit être signé en deux exemplaires. Le premier est retenu par le participant, et le second par le chercheur. En cas de questions, s'il-vous-plaît, contactez les chercheurs principaux:

Celia.Boyer@HealthOnNet.org, Rafael.RuizDeCastaneda@UniGe.ch et Nataly.Pletneva@HealthOnNet.org / / 0041 22 372 62 50

7.3 Screenshots of the blind comparison platform



[B-mode] - [Participant ID: "np"] - [0 scénarios terminés]

Bienvenue sur l'étude M.I.A.R.

Quitter la plateforme d'évaluation | Quitter la session

ETAPE 1

Le scénario qui vous a été attribué aléatoirement est:

#6:Grossesse et accouchement

La boîte ci-dessous contient les détails du scénario et restera accessible durant tout le test.

SCÉNARIO

Grossesse et accouchement

"Imaginez que vous êtes une femme de 38 ans,et enceinte depuis quelques semaines maintenant.

Vous avez récemment consulté votre gynécologue. Comme vous avez déjà 38 ans, ce dernier vous informe que vous courrez un risque considérable de mettre au monde un bébé atteint du Syndrome de Down (trisomie 21). En conséquence, il propose de réaliser une amniocentèse qui permettra de répondre à la question et vous amènera éventuellement à discuter d'une interruption de grossesse.

Vous savez que le syndrome de Down est une affection très handicapante pour l'enfant et lourde de conséquences pour sa famille et vous. Il est clair pour vous que cela changerait toute la vie familiale que vous aviez imaginé. Vous voudriez effectuer l'amniocentèse mais juste avant de donner votre consentement, le médecin se montre très clair sur les risques qui y sont liés (mort in utero, fausse couche)..

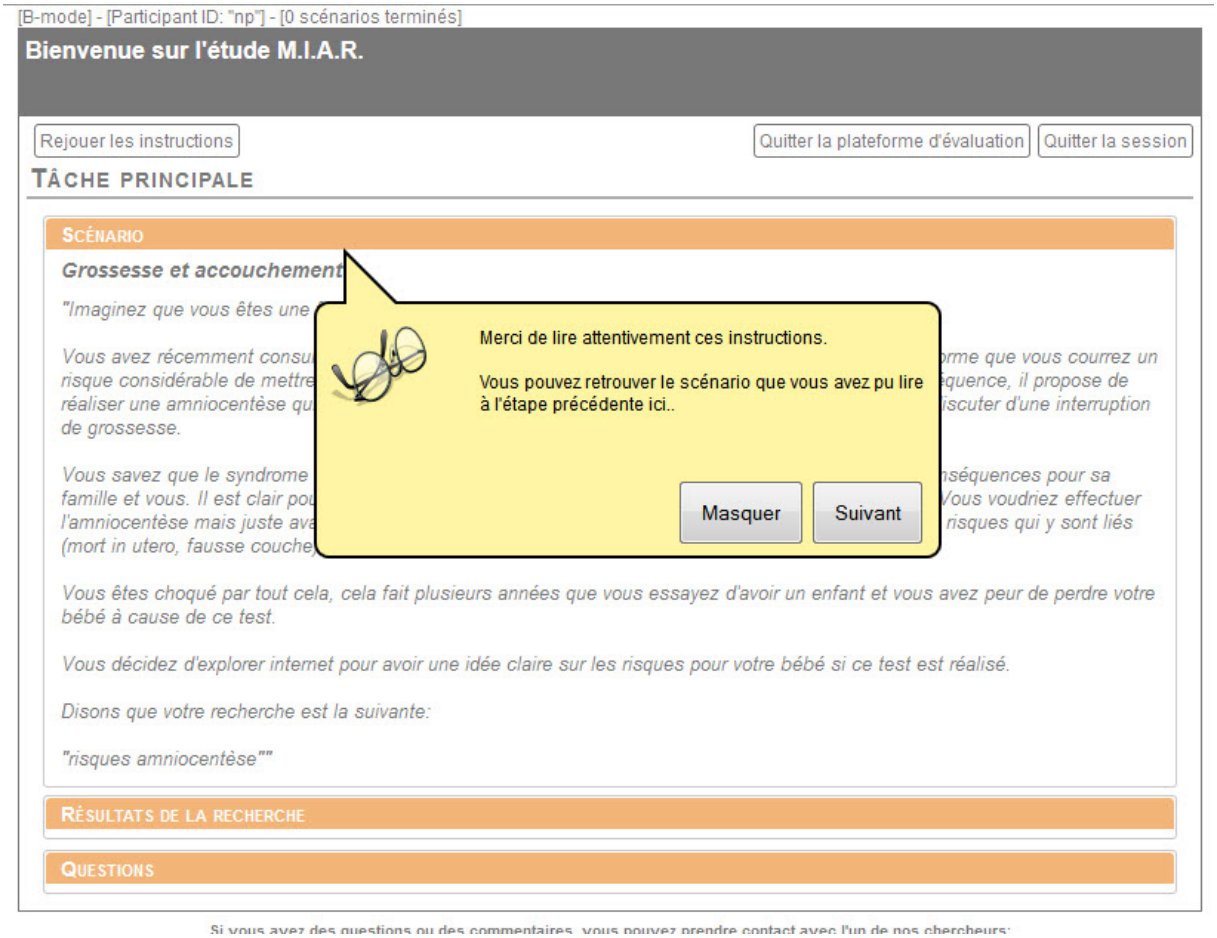
Vous êtes choqué par tout cela, cela fait plusieurs années que vous essayez d'avoir un enfant et vous avez peur de perdre votre bébé à cause de ce test.

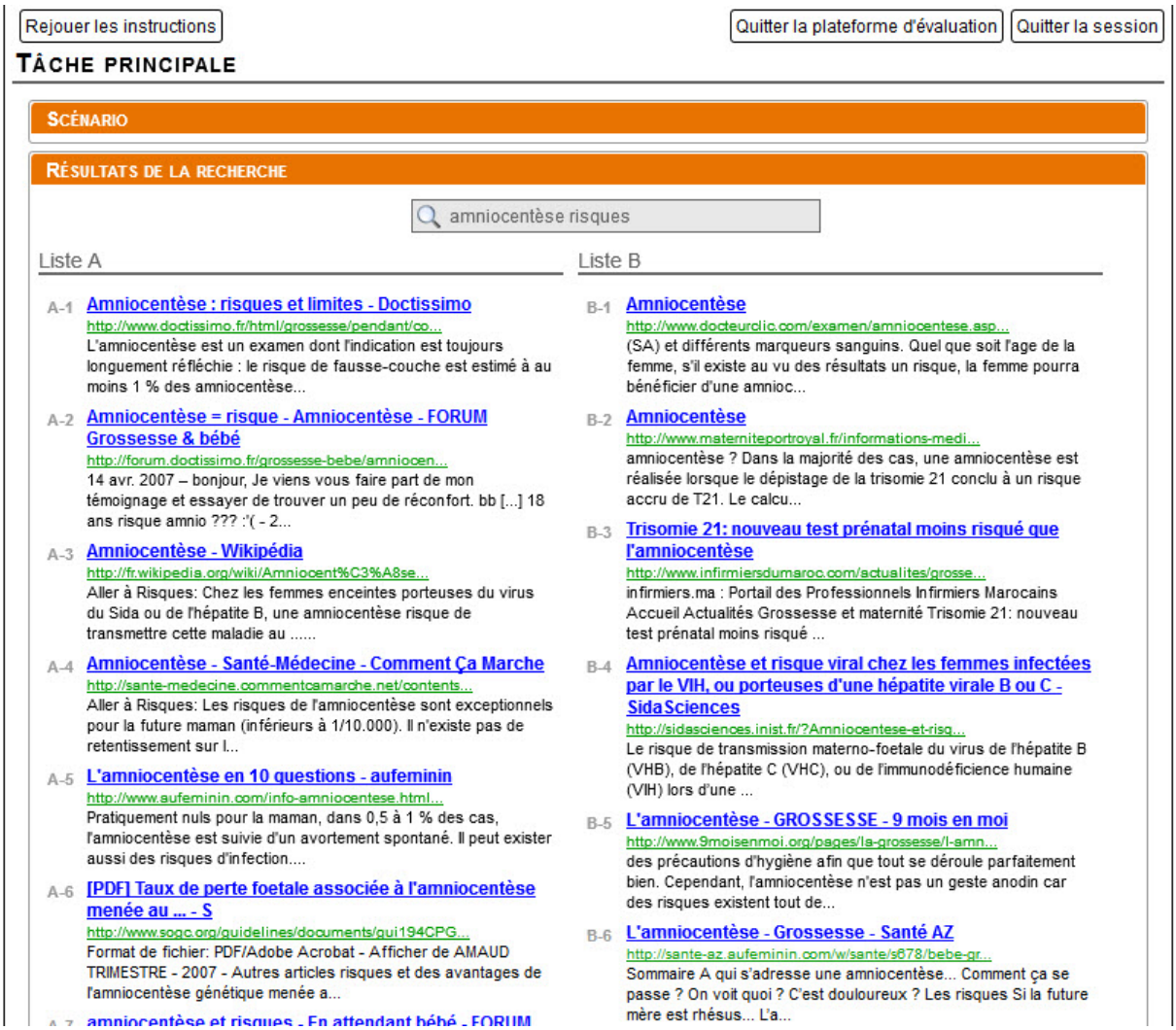
Vous décidez d'explorer internet pour avoir une idée claire sur les risques pour votre bébé si ce test est réalisé.

Disons que votre recherche est la suivante:

"risques amniocentèse""

Suivant


[B-mode] - [Participant ID: "np"] - [0 scénarios terminés]
Bienvenue sur l'étude M.I.A.R.
Rejouer les instructions
Quitter la plateforme d'évaluation
Quitter la session
TÂCHE PRINCIPALE
SCÉNARIO
Grossesse et accouchement
"Imaginez que vous êtes une
Vous avez récemment consu
risque considérable de mettre
réaliser une amniocentèse qu
de grossesse.
Vous savez que le syndrome
famille et vous. Il est clair pou
l'amniocentèse mais juste ava
(mort in utero, fausse couche)
Merci de lire attentivement ces instructions.
Vous pouvez retrouver le scénario que vous avez pu lire à l'étape précédente ici..
Masquer
Suivant
Vous êtes choqué par tout cela, cela fait plusieurs années que vous essayez d'avoir un enfant et vous avez peur de perdre votre bébé à cause de ce test.
Vous décidez d'explorer internet pour avoir une idée claire sur les risques pour votre bébé si ce test est réalisé.
Disons que votre recherche est la suivante:
"risques amniocentèse""
RÉSULTATS DE LA RECHERCHE
QUESTIONS



7.4 Demographic and Internet usage questionnaires for the blind comparison evaluation of the general public (French)

Origine et langue

1. Quel est votre pays d'origine ? Si vous avez une multi-nationalité, entrez les nationalités concernées.
2. Quelle est la langue avec laquelle vous vous sentez le plus à l'aise?
3. Comment maîtrisez vous l'anglais?

- Rien
- En dessous de la moyenne
- Moyennement
- Bien
- Très bien / excellent

Données personnelles

4. Genre

- Un home
- Une femme

5. Quel âge avez-vous ?

Education

6. Quel est votre niveau d'éducation maximal?

- Ecole élémentaire
- Etudes secondaires
- Ecole technique (2 ans)
- Degré universitaire (Bachelor)
- Degré universitaire (Master)
- Postgrade universitaire (doctorat)
- Autre, précisez

7. Quel est votre domaine professionnel et votre fonction?

8. Si vous êtes (ou avez été) étudiant universitaire, quelle discipline étudi(i)ez-vous? Sinon, saisissez "N/A".

9. Si vous êtes toujours étudiant universitaire, en quelle année êtes vous actuellement? Saisissez "N/A" si cela ne vous concerne pas.

Questionnaire sur votre utilisation d'internet

Remplissez svp ce questionnaire sur votre utilisation d'internet en rapport à la santé. Cliquez sur Envoyer en bas du formulaire pour continuer.

Si des questions à réponse libre ne sont pas applicables dans votre cas, SVP saisissez N/A.

Votre configuration

10. Quels dispositifs utilisez-vous habituellement pour vous connecter à internet ? Sélectionnez une ou plusieurs réponses.

- Ordinateur fixe / portable
- Tablettes électroniques
- Un téléphone mobile /smartphone
- Autre, précisez:

Votre utilisation d'internet

Note:
Nous utilisons le terme "moteur de recherche" au sens large: il peut s'agir de n'importe quel outil sur page web muni d'un champ texte et fournissant une liste de résultats selon les termes employés.

11. A quelle fréquence vous connectez-vous à internet ?

- Plusieurs fois par jour

- Environ une fois par jour
- Plusieurs fois par semaine
- Environ une fois par semaine
- Plusieurs fois par mois
- Environ une fois par mois
- Quelques fois par année et moins
- Autre, précisez:

12. A quelle fréquence utilisez-vous internet pour rechercher des informations sur des moteurs de recherche?

- Plusieurs fois par jour
- Environ une fois par jour
- Plusieurs fois par semaine
- Environ une fois par semaine
- Plusieurs fois par mois
- Environ une fois par mois
- Quelques fois par année et moins
- Autre, précisez:

13. Vous sentez-vous à l'aise lorsque vous faîtes des recherches sur internet?

- Tout à fait à l'aise, je suis un utilisateur expérimenté.
- Bon niveau, mais j'ai tout de même parfois des problèmes pour trouver les informations que je recherche.
- Moyennement, je n'arrive pas toujours à trouver ce que je recherche.
- Pas vraiment, je suis un utilisateur peu expérimenté.

14. A quelle fréquence cherchez-vous des informations en anglais sur internet?

- Plusieurs fois par jour
- Environ une fois par jour
- Plusieurs fois par semaine
- Environ une fois par semaine
- Plusieurs fois par mois
- Environ une fois par mois
- Quelques fois par année et moins
- Autre, précisez:

15. Quels "moteurs de recherche" (voir note ci dessus) utilisez-vous lorsque vous faîtes des recherches en général sur internet? Sélectionnez une ou plusieurs réponses

- Google
- Bing
- Yahoo
- Ask
- Autre, précisez:
- Non, je ne cherche jamais d'informations sur des moteurs de recherche

16. A quelle fréquence utilisez-vous internet afin de trouver des informations sur votre santé ou la santé de vos proches?

- Plusieurs fois par jour
- Environ une fois par jour
- Plusieurs fois par semaine
- Environ une fois par semaine
- Plusieurs fois par mois
- Environ une fois par mois
- Quelques fois par années et moins
- Autre, précisez:

17. Est-ce que vous recherchez habituellement des informations concernant les sujets de santé suivants? Sélectionnez une ou plusieurs réponses.

- Maladies ou problèmes de santé en particulier
- Traitements médicaux ou procédures
- Comment perdre ou controller son poids
- Risques de santé liés a l'alimentation
- Sécurité d'usage des medicaments / medicaments dont vous avez vu une publicité
- Résultats de test médicaux
- Soins apportés à un proche ou ami âgé
- Grossesse et naissance
- Autre, précisez:
- Non, je ne cherche jamais d'informations sur la santé

18. Faîtes vous confiance aux information sur la santé que vous trouvez sur internet?

- Normalement oui, je suis satisfait avec la qualité de l'information et les résultats trouvés.
- Cela dépend, parfois oui, parfois non, la qualité de l'information et les résultats de santé devraient être améliorés.
- Généralement non, je ne suis pas satisfait avec l'information et les résultats que je trouve.

19. Quels sont les critères que vous avez pour déterminer si vous faîtes confiance à l'information de santé sur un site internet ? Svp justifiez votre réponse:

20. Avez-vous, vous ou un de vos proches, un ou plusieurs problèmes de santé suffisament important(s), pour que vous ayez dû faire des recherches sur internet à son (ses) sujet(s)? SVP décrivez brièvement ce(s) problème(s).
Par exemple: Proche: cancer du colon, Moi: fracture de hanche

Ces données sont bien entendu anonymisées

7.5 Health scenarios for the blinded comparison

7.5.1 Specific disease or medical problem (Gout):

Imagine that you are a man of 60 years old that has been diagnosed with gout.

Gout is a form of arthritis, a form of joint disorder that involves inflammation of one or more joints. Symptoms of gout are characterized by sudden and severe episodes of pain tenderness, redness, stiffness and swelling of affected joints. It is so painful that some patients report that gout is as severe as a fracture of a long bone.

Your physician has told you that choosing a special diet is critical to limit the development of this disease and you decide to dramatically change your diet in order to reduce the severe pains you are facing.

You decide to make a search on the Internet to find a specific information about a diet that could help you with your problem. You know that some diets can have negative implications on health and therefore you will try to be particularly careful with the information found on the Internet.

Let's say that your search is:

Query: goutte régime

7.5.2 Certain medical treatment or procedure (ADHD Medication)

Imagine that you are mother/father of a kid of 7 years old who has been diagnosed with "Attention deficit hyperactivity disorder" (ADHD). Your kid shows high levels inattention, both at home and school, hyperactivity and impulsivity, all typical symptoms of ADHD.

The doctor who diagnosed your kid suggested treating this health problem by administrating a stimulant drugs. This seems very confusing for you, even contradictory, "why should I give stimulants to my kid that is already hyperactive?"

As medical consultation time is very limited, you do not have time to ask all questions you want, hence you decide to go online to have a better understanding on why your kid should take stimulants. You decide to look for the effects of these drugs on people with ADHD.

Let's say that your search is:

Query: TDAH médicament stimulants effects

7.5.3 How to lose weight or how to control your weight

Imagine you are student at the University. Since you left home and entered University, you spend most of your time working on yours studies: attending lectures, revising literature and reading, writing up reports and working on your class notes, preparing exams and presentations etc.

Unfortunately, you do not have enough time to practice as much sport as you would like to and your alimentation has also become quite unbalanced, you are not a very skilled cook and do not particularly enjoy spending too much time preparing your meals.

As a result of this type of hard-working student's life-style, you have gained considerable weight during the last year and feel embarrassed about your situation and want to take better care of your own health from now on.

You decide to explore the Internet using a search engine to find a possible method that could help you lose some weight rapidly and safely.

Let's say that your search is:

Query: Perdre poids rapidement

7.5.4 Food safety or recalls

You have always consumed pasteurized milk, however you have recently read an article in a magazine about the benefits of raw milk consumption. You are not completely convinced about this information as the magazine was not a specialized health magazine, but one of those referring to general lifestyle aspects.

Soon, you will be leaving for your summer vacations in a farm in the mountains and will have the opportunity to drink raw milk as the farmer milks his cows every day.

You are confused: should you drink this raw milk from the farm as the magazine suggests or alternatively go to a close supermarket and get pasteurized milk, as you generally do in the city? What are the risks of drinking raw milk?

To solve your doubts and to make sure that you do the best for your health, you decide to search on the Internet to find out the risks of consuming raw milk and how to avoid them.

Let's say that your search is:

Query: lait cru risques

7.5.5 Drug safety or recalls / a drug you saw advertised

Imagine you are constantly having headaches but do not want to visit a physician to consult your problem. Instead, you keep buying and taking analgesic drugs which do not require medical prescription.

You are worried because the problem persists and you keep taking more and more drugs. You decide to search online to find the information about the risks of taking too many analgesics (pain killers / anti-inflammatory).

Let's say that your search is:

Query: risques trop antidouleurs

7.5.6 Pregnancy and childbirth

Imagine that you are a 38 years old woman that is pregnant for a couple of weeks now.

You recently consulted your gynecologist and, as you are already 38, he suggested that you have a considerable risk to have a baby with Down's syndrome. Accordingly, he suggested to do an amniosynthesis to check for the possibility of the baby of having Down's syndrome and eventually discuss of abortion.

Although you are interested in knowing if your baby could have Down's syndrome, the doctor is very clear about this test and warns you about some risks of harming the baby which could end in its death.

You are very shocked about all this, you have been trying to have a baby for several years and do not want to take the risk of losing your baby due to this test.

You decide to check on the Internet to have a clear idea about the risks for your baby of doing this test.

Let's say that your search is:

Query: amniocentèse risques

7.5.7 Medical test results

Last Saturday, you met a boy/girl in a party and after a couple of drinks, went back home with him/her.

In the morning, the boy/girl had already left when you woke up. Besides having a terrible headache, and not remembering anything from last night, you notice the sealed box of condoms on the table. You think that you had sex with this unknown boy/girl without any protection!

You start panicking because you remember now that the appearance of the boy/girl was not very good. In fact, later on the day, a friend of you warns you that he heard a rumor about this boy/girl suggesting that he may have HIV.

You urgently need help, you do not know what to do or where to go. You decide to do an urgent search on the Internet to find help and support in your city.

Let's say that your search is:

Query: Sida aide Gèneve

7.5.8 Caring for an aging relative or friends

You are a 54 years-old important businessman. You are extremely busy with meetings and travels around the world and have not much time to take care of your mother, who is already 88 years old.

Your father died some years ago and you have no brothers, thus your mother is always alone at home. She suffers from an early form of dementia and she is becoming unable to take care of herself: she does not take her medication correctly, she is not able to cook properly…

You are really worried about this situation and want to find rapid solutions. Your mother is firmly opposed to moving to a nursing home so you have to find alternative possibilities.

You decide to go online and search the Internet for a solution. Could you maybe hire a specialized nurse that would come to her house regularly to check her medication and take care of her?

Let's say that your search is:

Query: aide personne-agée maison

7.6 Information sheets and consent forms for participants of full user tests (Czech and French)

7.6.1 Informační leták – Evaluace v rámci projektu Khresmoi

Tento dokument Vám poskytne podrobnější informace o projektu KHRESMOI (výzkumný projekt EU). Zároveň se dozvíte, čím je Vaše účast v této evaluační studii důležitá.

Proč KHRESMOI?

Možná jste sami zaznamenali, že široká veřejnost k vyhledávání informací o zdraví stále častěji využívá internet. Medicína se evidentně zcela nepochybně stává jeho součástí. Velké množství dostupných informací však vede k tomu, že najít informace, které by byly relevantní a přesné, je pro laika stále velice obtížné. Samotné vyhledání informací totiž nestačí. Je nutné zvážit nejen to, zda je vyhledaný dokument relevantní a obsahově správný, ale také například i to, zda je vydavatel dokumentu dostatečně věrohodný.

Přístup ke kvalitním informacím o zdraví považujeme za důležitý krok ve snaze dosáhnout „rovnosti ve zdraví“. V souvislosti s tímto cílem je náš projekt financován Evropskou komisí. Rozepsaná zkratka KHRESMOI zní Knowledge Helper for Medical and Other Information Users (v překladu „Pomocník pro uživatele internetu hledající informace nejen o zdraví“). Naším cílem je v průběhu 4 let vytvořit vyhledávač, který bude uživatelsky maximálně vstřícný a informace o zdraví na internetu zpřístupní následujícím skupinám uživatelů: veřejnosti, lékařům a odborníkům, rentgenologům. Pro zajištění spolehlivého a důvěryhodného prostředí využívá KHRESMOI moderní technologie a různorodé zdroje, které mu umožňují zpřístupnit nejen domácí, ale také zahraniční zdroje, a to v příslušném překladu.

Vaše názory a zkušenosti nám mohou pomoci vyhledávač KHRESMOI zdokonalit. Tím, že se s námi podělíte o své zkušenosti, pomůžete dalším lidem získat přístup k relevantním, důvěryhodným informacím o zdraví, a to bez ohledu na jejich rodný jazyk, na jejich zkušenosti s online informacemi, na jejich vzdělání nebo sociálně-ekonomické zázemí.

Na níže uvedeném obrázku vidíte, že KHRESMOI využívá údaje z rozličných zdrojů. Mezi tyto zdroje patří texty (například online časopisy, knihy, důvěryhodné internetové stránky) i obrázky. Na základě Vašich podnětů z vyhledávání v současné testovací verzi budeme moci identifikovat prvky, které je třeba zdokonalit.

Vaše účast

Po dvou letech výzkumu byl v rámci projektu Khresmoi vyvinut prototyp vyhledávače. Chtěli bychom, abyste otestoval náš nový vyhledávač informací o zdraví na internetu. Očekáváme, že test bude trvat přibližně jednu hodinu.

Účastí v této hodnotící studii přispíváte k tomu, aby všichni získali lepší přístup k relevantním a důvěryhodným informacím o zdraví. Všechna data a postřehy, které studií získáme, budou považovány za důvěrné informace. Účast v této studii je dobrovolná a můžete ji kdykoli bez uvedení důvodu přerušit. Účast v této studii nemá vliv na Váš zdravotní stav. Výsledky této studie mohou být prezentovány ve vědeckých publikacích. Získaná data budou s maximální diskrétností používána pouze jako data anonymní. Všechna získaná data a pozorování budou považována za důvěrná. Děkujeme Vám za spolupráci!

7.6.2 Účastnická smlouva a souhlas se zpracováním osobních údajů

Já, _____, souhlasím s účastí ve vědecké studii, která je součástí projektu KHRESMOI, podporovaného Evropskou unií a organizovaného nadací Health on the Net. Potvrzuji, že jsem byl(a) informován(a) o obsahu, účelu a důležitosti této studie a že jsem dostal(a) dostatečné množství času na rozvážení si své účasti ve studii. Veškeré mé otázky byly zodpovězeny a jsem srozuměn(a) s tím, že má účast ve studii je dobrovolná a mohu ji kdykoliv bez udání důvodu ukončit.

Dále jsem srozuměn(a) s tím, že tato studie nemá žádný vliv na mé zdraví. Vzhledem k tomu, že v souvislosti s touto studií pro mě neexistují žádná zdravotní rizika, souhlasím s tím, že se na mou účast ve studii nevztahuje pojistná smlouva.

Test v rámci studie bude trvat maximálně jednu hodinu. Bude tvořen dvěma dotazníky a několika úlohami zpracovávanými a nahrávanými na počítači. Tyto úlohy budou založeny na vyhledávání specifických otázek pomocí vyhledávacího nástroje, vyvinutého pro tento účel.

Hlavní řešitel a rovněž všichni spoluřešitelé studie se zavazují, že data, která byla získána v rámci studie, budou s maximální diskrétností používat pouze jako data anonymní. Hlavní řešitel a spoluřešitelé jsou povinni považovat získaná data a pozorování za důvěrná, data nesmí žádným způsobem poskytnut třetí straně.

Souhlasím s tím, aby řešitelé projektu KHRESMOI měli přístup k mým soukromým datům, která byla během studie získána a která mají anonymní formu. Souhlasím s použitím svých anonymizovaných dat pro schválený vědecký výzkum, jenž má napomoci základnímu výzkumu při získávání lékařských informací. Beru na vědomí, že žádná osobní data, která mě jako účastníka mohou osobně identifikovat, nebudou uchovávána, a to ani dočasně. Tato data tedy nepodléhají ustanovení Zákona 101/2000 Sb. O ochraně osobních údajů, s výjimkou autorizovaných dat spadajících do případné Dohody o provedení práce a administrativy s ní spojené.

Děkujeme za spolupráci!

_____	__ Univerzita Karlova v Praze / Jan Hajič ___
Jméno účastníka (hůlkovým písmem)	*Název organizace a jméno odpovědné osoby*
_____	_____
Datum a podpis	*Datum a podpis*

Tento formulář musí být podepsán ve dvou kopiích. Jednu si ponechá účastník studie, druhou odpovědná osoba.

V případě otázek kontaktujte prosím hlavní řešitele:

Celia.Boyer@HealthOnNet.org and Nataly.Pletneva@HealthOnNet.org // 0041 22 372 62 50, popřípadě hajic@ufal.mff.cuni.cz and uresova@ufal.mff.cuni.cz //00420 221 914 361

Více informací o projektu: www.khresmoi.eu

7.6.3 Feuille d'information

Ce document a été créé de sorte à vous fournir de plus amples informations sur le projet Européen Khresmoi.

Pourquoi Khresmoi?

Comme vous l'avez peut-être remarqué, les citoyens ont de plus en plus recours aux richesses du Web, lorsqu'il s'agit de rechercher des informations médicales. Plus généralement, les technologies Internet sont inéluctablement intégrées de plus en plus profondément dans nos systèmes de santé contemporains. Pourtant, la recherche précise et efficace d'informations reste une tâche difficile pour le néophyte, notamment, en raison du nombre important d'informations disponibles sur le web. Une fois l'information trouvée, il faut encore que l'utilisateur puisse juger de sa crédibilité, de sa pertinence et de sa justesse.

Nous reconnaissons que de fournir un accès de qualité pour tous aux informations de santé dignes de confiance est un pas important vers l'égalité de l'accès aux soins. C'est dans ce cadre que s'inscrit le projet KHRESMOI, projet financé par la commission européenne. KHRESMOI est l'acronyme utilisé pour *Knowledge Helper for Medical and Other Information user*. Pour un développement de quatre ans, le projet vise à proposer un environnement ergonomique pour la recherche de contenu en ligne traitant de la santé et destiné à divers groupes d'utilisateurs: la population générale, le médecin de famille et le radiologue. KHRESMOI utilise des technologies modernes et des ressources variées dans le but de fournir à l'utilisateur un environnement fiable, consistant en un support pour la recherche d'informations locales, tout comme pour la recherche d'informations internationales, au travers, notamment, de ses services de traductions.

Grâce à votre expérience sur notre système et à vos précieux retours, vous contribuez à rendre l'information de santé en ligne plus pertinentes et digne d'une confiance méritée, pour tous, quelque soit la langue, l'expérience sur un moteur de recherche, ou le niveau socio-économique des utilisateurs.

Khresmoi combine des sources textuelles (journaux en ligne, livres numériques, et sites Internet certifiés) à des images. Grâce à votre expérience sur l'outil, nous serons capables d'identifier les éléments qui nécessitent des modifications avant de les intégrer dans le prototype suivant.

Votre participation

Le prototype présenté ici est le fruit de deux ans et demi de travail. Nous vous invitons à l'utiliser en faisant des recherches sur des informations de santé. Une session dure environ une heure.

En participant à cette évaluation, vous contribuez à l'amélioration de l'accès aux informations de santé en ligne pertinentes et dignes de confiance, pour chacun. Toutes les données et observations collectées au cours du test sont traitées en tant que données confidentielles. La participation se fait sur une base volontaire et peut se terminer à n'importe quel moment, sans avancer de raison particulière. La participation n'influence, ni n'affecte l'état de santé du participant. Les résultats de l'expérience scientifique pourront être présentés dans des publications scientifiques. Les données collectées ne seront utilisées que sous une forme anonyme et dans un parfait respect de l'identité du participant. Nous vous remercions encore pour votre participation.

7.6.4 Formulaire de consentement

Je, soussigné, Mme/M. _____, donne par la présente mon consentement à ma participation à une étude scientifique menée dans le cadre du projet de l'UE *Khresmoi* à laquelle la *Fondation Health On the Net* participe.

J'ai été informé sur le contenu, le but et la portée de cette étude. J'ai eu suffisamment de temps pour réfléchir au sujet de ma participation. Je certifie, que si j'avais des questions, des réponses satisfaisantes y ont été apportées. Je comprends que ma participation est volontaire et peut être résiliée à tout moment, sans aucune raison particulière. Je comprends que ma participation à cette étude n'influence pas mon état de santé, et dès lors, qu'aucune couverture d'assurance ne m'est fournie.

Le test dure environ une heure. Il se compose de deux questionnaires et d'un certain nombre de tâches consistant en la recherche de sujets spécifiques sur la santé, grâce un prototype de moteur de recherche. Les chercheurs principaux, ainsi que tous les membres du projet concernés, s'engagent à n'utiliser les données recueillies que sous une forme anonyme. Ils sont tenus de traiter les données et observations de manière confidentielle. Aucune donnée personnelle identifiable n'est transmise à des tiers ou commercialisée.

Je donne ma permission aux membres du projet en question d'avoir accès à mes données anonymes recueillies lors de l'étude. Je suis d'accord avec le fait que ces données seront utilisées dans le cadre de la recherche pour améliorer l'accès à des informations sur internet, au sujet de la santé, dignes de confiance.

_____	_____
*Nom du **participant** en MAJUSCULES*	*Nom de l'**organisation** et **chercheur***
_____	_____
Date et signature	*Date et signature*

Ce formulaire doit être signé en deux exemplaires. Le premier est retenu par le participant, et le second par le chercheur. En cas de questions, s'il-vous-plaît, contactez les chercheurs principaux:

Celia.Boyer@HealthOnNet.org, Rafael.RuizDeCastaneda@UniGe.ch et Nataly.Pletneva@HealthOnNet.org / / 0041 22 372 88 86

7.7 Configuration for Morae file (all questionnaire and tasks proposed to participants, all in English)

Study name: Khresmoi prototype usability study 2012¨

Study description: Khresmoi is a 4 year EU funded project aiming at improving online search of medical information. In this study we present you the intermediate outcome of 2 years of our work. There is still a lot of work to be done and we need your feedback to better understand what you would like to see in a final product: what should be changed, removed, improved, added? Please, fill free to provide us with any thoughts and reflections you may have. We highly appreciate your help and insights!

Study instructions: We thank you for agreeing to participate in our study. You will have to fill **initial questionnaire**, then you will have some time to see and play around with the search engine prototype. Further on we will ask you to perform 3 **search tasks** using a search engine. Afterwards you will be asked to answer another **questionnaire about your search experience**. The whole test will take **about an hour**.

Each task describes a **situation and poses a question** which you should try to answer. Try to perform the task as if you were actually doing this task for yourself.

Please, remember we are NOT evaluating you, your knowledge or skills, we evaluate the system.

Please, **talk out loud** while you work and keep a running dialog of your thoughts, actions, expectations, and reactions.

7.7.1 Demographic and background questionnaire:

1. Your country:
    a. Austria
    b. Czech Republic
    c. Ireland
    d. France
    e. Germany
    f. Switzerland
    g. UK
    h. Other (please, specify)
2. Your gender:
    a. Male
    b. Female
3. Your age
    a. Below 20

b. 20-29
c. 30-39
d. 40-49
e. 50-59
f. 60-69
g. 70-79
h. 80 and more

4. What is your highest level of education?
    a. Elementary school
    b. High school
    c. Vocational/technical school
    d. University graduate (Bachelor)
    e. Master degree (MSc, MA, MBA etc)
    f. PhD
    g. Other
5. In which area do you work ? Please, specify the area and your function?
6. How often do you connect to the Internet?
    a. Every day
    b. Several times a week
    c. Once a week
    d. Several times a month
    e. Once a month
    f. Other
7. Which device do you mostly use to connect to the Internet?
    a. PC/laptop
    b. Tablet
    c. Mobile phone/smartphone
8. Do you use Internet for work/studies?
    a. Yes
    b. No
    c. I do not work/study at the moment, I use Internet only for my personal inquiries
9. How often do you search online?
    a. Every day
    b. Several times a week
    c. Once a week
    d. Several times a month

e. Once a month
f. Other

10. Which search engine(s) do you use while searching on the Internet? (select all that apply)
    a. Google
    b. Bing
    c. Yahoo
    d. Ask
    e. Other
11. How confident do you feel searching for the information on the Internet?
    a. Not confident, I am a new-comer
    b. Average, I am often not able to find what I search for
    c. Overall good, I have problems in finding information only from time to time
    d. Very confident, I am an experienced user
12. What is your mother tongue?
    a. Czech
    b. English (go directly to the question 14)
    c. French
    d. German
    e. Spanish
    f. Other
13. What is your level of English?
    a. No knowledge of English
    b. Basic knowledge of English
    c. Average knowledge of English
    d. Good knowledge of English
    e. Fluent English
14. How often do you search for or read any information on the Internet in English?
    a. Every day
    b. Several times a week
    c. Once a week
    d. Several times a month
    e. Once a month
    f. Other
15. How often do you search for online health information regarding your or your family/friends health?
    a. Every day
    b. Several times a week

c. Once a week
d. Several times a month
e. Once a month
f. Other

16. Which search engine(s) do you use while searching for online health information? (select all that apply)
    a. Google
    b. Bing
    c. Yahoo
    d. Ask
    e. Other
17. Which types of online health information you are looking for? (select all that apply)
    a. Specific disease or medical problem
    b. Certain medical treatment or procedure
    c. How to lose weight or how to control your weight
    d. Food safety or recalls
    e. Drug safety or recalls / a drug you saw advertised
    f. Medical tests results
    g. Caring for an aging relative or friends
    h. Pregnancy and childbirth
    i. Other
18. Have you ever been diagnosed with any chronic or/and life-threatening disease?
    a. No (go directly to the question 21)
    b. Yes (write which one in a comment field below)
19. For how long have you been diagnosed?
    a. Few weeks
    b. Few months up to a year
    c. From one to three years
    d. More than three years
20. After being diagnosed, have you changed your Internet use to search for online health information comparing with “before diagnosis”?
    a. Internet use has increased after diagnosis
    b. Internet use decreased after diagnosis
    c. Internet use has not changed after diagnosis
21. Is someone you know and care for has ever been diagnosed with any chronic or/and life threatening disease?
    a. No

b. Yes (write which one(s) in a comment field below)

22. Do you think you understand well your health situation?

5 point likert skale from very bad to very good

7.7.2 Free search engine use

3 to 5 minutes

7.7.3 Task 1: Body mass index

Imagine, you (or someone you know) are concerned with your weight and ask an opinion of your physician. He/she suggests you to check your body mass index.

Please, use Khresmoi to calculate your body mass index (formula or calculator).

Task 1 completion survey:

1) Have you already been familiar with body mass index before doing the task?
   a. Yes
   b. No
2) Have you been able to calculate your body mass index using Khresmoi?
   a. Yes
   b. No

7.7.4 Task 2: Liver cancer

Imagine the situation you or someone you know is diagnosed with liver cancer. Use Khresmoi to find information about possible treatment options.

Task 2 completion survey:

1) Have you (or someone you know) been diagnosed with liver cancer?
   a. Yes
   b. No
2) Have you been able to find information about treatment options for liver cancer?
   a. Yes
   b. No

7.7.5 Task 3: Diabetes medication

Imagine you or someone you know have been diagnosed with diabetes a while ago and have been prescribed Metformin. Before you start taking the medication you are curious to know what other patients think about this medication. Use Khresmoi to find such kind of information.

Task 3 completion survey

1) Have you previously searched for medication information on the Internet?

a. Yes
b. No

2) Have you been able to find and read other patients experiences with Metformin?
   a. Yes
   b. No

7.7.6 SUS questionnaire with additions (questions 11-24) for the general public

1. I think that I would like to use this system frequently
2. I found the system unnecessarily complex
3. I thought the system was easy to use
4. I think that I would need the support of a technical person to be able to use this system
5. I found the various functions in this system were well integrated
6. I thought there was too much inconsistency in this system
7. I would imagine that most people would learn to use this system very quickly
8. I found the system very cumbersome to use
9. I felt very confident using the system
10. I needed to learn a lot of things before I could get going with this system
11. I think I was able to find an answer to the questions quickly
12. I think my searches were successful
13. I did not have enough results
14. I found the results understandable
15. I found the results relevant
16. I perceived the results to be of better quality than those of other search engines.
17. The results were difficult to understand
18. The system assisted me in my query formulation
19. I had too many results
20. I found automatic translation capabilities useful
21. I found the possibility to classify the results to be useful
22. I found the possibility to filter in/out the results to be useful
23. I think that having images among search results helps me to answer the question
24. I found disease definition below search bar useful
25. I will recommend the system to my peers.

7.8 Evaluation session check list

PART 1 - PREPARATION:

1) General preparation

1. Recruitment of participants (10-15 in total): patients, carers (parents of children) and general public of all ages, education level, languages skills if possible; **inclusion criteria – using the Internet and search engines for health topics at least sometimes**
2. Find an affordable place to conduct the study
3. Set time and dates
4. Inform participants about the time and place and how to get there (by phone/email)
5. Cost estimate: Are travel expenses covered? Little gift offered? Beverages and snacks?
6. All participants test Khresmoi for everyone http://khresmoi.honservices.org/hon-search/

- Internal documents should not to be translated
- All information for the participants should be translated into their mother tongue: information sheet, participant consent, Morae file configuration

2) Preparation of the materials and documents:

- Make sure the **room** is available during the whole testing session
- Check the **internet connection**: WiFi (check password and login!) or cable
- Arrange **beverages and snacks** (pay attention if patients with particular dietary requirements!): water, juice, cookies, etc.
- **Timer**
- **Laptop** and **adaptor**: the laptop should have a **camera** and **microphone** (for a higher quality of sound, use of an external microphone is preferred)
- **Stickers/labels** and **files** for paper documents(one for every participant)
- Little **gift** for a participant (gift voucher, Khresmoi T-short)
- Prepare the following **documents** for every participant**:**
    - Assign a **ID** to each participant (investigator initials_day_month_year_number of participant, i.e. ZU_210513_1)
    - 1 x paper or e-format table "**List of participants**" with participant ID, date and observer initials
    - 1 x **Information sheet** in paper
    - 2 x **Informed consent** in paper
- Prepare a **laptop** for every participant:
    - **Morae** Recorder should be installed and tested
    - Open the file in Morae.

The file runs in AUTOPILOT mode after clicking "Record" and provides a brief introduction to the session, demographic and background questionnaire, tasks and adapted SUS (System Usability Scale) usability questionnaire.
Reference document: http://assets.techsmith.com/docs/pdf-morae/SW-Usability-Testing-with-Morae.pdf

- **Open the search engine prototype**:
  - http://khresmoi.honservices.org/hon-search/
- **Change the language of the interface if needed** (this should be the mother tongue of the participant)

- Prepare a **laptop** for the observer: an observer should get connected to Morae Observer from another laptop by introducing the IP address of the PC where Recorder will be played.

PART 2 – SESSION → total estimated time **1h**

1. **Welcome** the participants and introduce the session → estimated time **3 to 5 minutes**
2. Handing out the **documents** → estimated time **5 to 10 minutes**
   a. The information sheet should be read by every participant
   b. Then the consent form should be read and signed: 2 exemplars signed by the participant and the observer/facilitator; one retained by the participant, the other by Khresmoi
   c. Any questions? Make sure all questions are answered before the session starts
3. A **real-life demo of the search engine 3 to 5 minutes**: give an example of health situation:

"Imagine someone you know has developed pollen allergy (seasonal allergy to blossoming plants). You want to get more information what it is and how to deal with it. We now go to Khresmoi and search for some information about it."

Start typing "pollen allergy" slowly to show that there are query suggestions. You may want to misspell it to show that system would correct it. Once the results appear, demonstrate the the functionalities of the interface, i.e.

- Translation tool
- Filtering (for example, forum to know other patients experience, or "products and services")
- Key-word cloud (explain that size of letters corresponds to frequency of the term in the search results)
- Images if any

Please, test the query in your language with selected interface language and change it if needed. You can try with any other disease or health condition (obesity, headache etc).

4. **Starting the recording** session by running Morae recording configuration file by pressing the red button
5. Make sure:

   ☐ The user **reads the user ID out loud** when the recording starts

   ☐ **The user/researcher saves the Morae file with the written user ID**

6. Participant starts with reading out loud their user ID
7. Participant plays around with the prototype to familiarize with the search engine → estimated time **3 to 5 minutes**
8. Participant starts filling out the demographic and background questionnaire → estimated time **5 to 10 minutes**
9. **ASK A PARTICIPANT TO SPEAK UP while using the prototype and to be explicit about their feedback:**
   "Please, talk out loud while using the search engine, while explicitly describing and expressing your thoughts, actions, expectations and reactions. You can also ask questions. Please name loudly the most useful website for every task."
10. Participant starts with the 3 tasks (included in Morae configuration file) →estimated time in total **15 to 20 minutes**
    a. **5 min per task**: in case a participant can answer faster, move on to the next task.
    b. In case a participant cannot find an answer **within 5 minutes**, leave the task and move on to the next one. **The observer should keep track of time and should ask the participant to move on to a next task if the 5 minutes are over.**
    c. Remind the participant to **speak out loud and name the web site** that was most useful (up to three) in answering the question
    d. Observer does as many notes as possible by adding the markers, i.e. "task 2 started", "question", "error" describing the issues (detailed description may be added after the session)
11. Participant fills out of the adapted SUS usability questionnaire (included in Morae configuration file) → estimated time **10 minutes**
12. Discussion based on observations and general feedback →estimated time **5 to 10 minutes** depending on the amount of feedback
13. Ask for and collect email addresses/postal addresses/phone numbers in case a participant wants to be informed about the results of the study → estimated time **2 minutes**
14. Thanking the participant for their participation with a little gift/good bye → estimated time **4-5 minutes**

*Time schedule: total estimation **55 min – 1h20 min***

PART 3 - AFTER THE SESSION:

- Save all files using the participant ID (i.e. ZU_210513_1) on both laptops (Recorder and Observer)
- Open the file in Morae Manager, ensure all markers are correctly displayed, add some if missing and expands the description in English
- Write a short summary of the session, main findings and observation for each participant ID
- Send the recordings to HON, they will then be uploaded on wiki
- Collect, save and scan signed consent forms
- Send list of participants and scanned version of the original informed consent to HON copying TUW (Veronika).

- Each investigator group keep the original informed consent and are responsible for keeping it at least till the end of the project.