Grant Agreement Number: 257528

KHRESMOI

www.khresmoi.eu

Report on the extensive tests with the final search system

Deliverable number	*D10.3*
Dissemination level	*Public*
Delivery date	*August 2014*
Status	*Final*
Author(s)	*Célia Boyer, Jan Hajič, Allan Hanbury, Marlene Kritz, Natalia Pletneva, Priscille Schneller, Veronika Stefanov, Zdeňka Urešova*

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 document summarizes in detail the results of extensive user evaluation conducted with the final search system, KPro and K4E, in the final year of the project.

For KPro two rounds of user evaluation were carried out. After the first round of evaluation in November 2013 extensive user feedback directed subsequent prototype advancement. Final user tests, carried out from May-July 2014, revealed subsequent improvements in efficiency, effectiveness and usability of the system. The KPro search system is in its current state most useful for physicians in training, hospital clinicians and research physicians. Furthermore, the personal library, search facets and translation features have been classified as most useful and provide a good foundation for future exploitation and advancement of the KPro search system.

All evaluation tests done on K4E provided a significant feedback on the acceptance of the system and few fine tune improvements of the prototypes. The most crucial things to do at this stage are to increase the coverage of resources in Czech. The next steps are directed towards the improvement of the automatic detection of the readability and trust and prototype improvements of the interface. We realise how important and crucial real user tests are. It is also important to have users "who do not know you or the service to test" in order to have fresh comments. In that aspect the George Pompidou Hospital location was a fantastic place as we had real patients which could be IT persons, physicians, nurses, students. Overall, we can conclude that presenting K4E as a search engine offering trustworthy web sites did affect the choice of participants, as shown while conducting the blind / non-blind user tests.

Table of Contents

List of Figures

List of Abbreviations

CUNI	Charles University in Prague
Del.	Deliverable
ELDA	Evaluations and Language resources Distribution Agency
GAW	Gesellschaft der Ärzte in Wien (Society of physicians in Vienna)
GP	General practitioner
HON	Health on the Net Foundation
HONcode	Health On the Net Code of Conduct
KPro	Khresmoi Professional search system
K4E	Khresmoi for Everyone search system
SUS	System Usability Scale
TUW	Technical University Vienna
UDE	University Duisburg-Essen
Y4	Year 4 of the Khresmoi Project

1 Executive Summary

The following report has been written to fulfil the requirements for **Deliverable 10.3 "**Report on the extensive tests with the final search system". A detailed report on the results of the user evaluation of "Khresmoi for Everyone" and "Khresmoi Professional" is provided.

1.1 Khresmoi Professional

In year 4 of the Khresmoi Project, a total of 84 physicians provided extensive insight into the usability, effectiveness, efficiency and exploitation potential of Khresmoi Professional. In the first round of user tests, conducted in November 2013, detailed user feedback was obtained and implemented into subsequent prototype development. The second round of tests, done in May-July 2014, included an online field evaluation and provided insight into the exploitation potential of the final search system. In all Y4 evaluations physicians were asked to do a "free browsing task" where they used KPro to find information relevant to a question that they had formulated themselves.

Perceived efficiency, effectiveness and usability varied with respect to time of evaluation (November 2013 vs. May-July2014), occupational group and age group. A central problem affecting system effectiveness in the first round of user tests was that KPro had difficulties identifying complex queries. In terms of efficiency, slow loading time and inefficient ranking prevailed. Furthermore, system complexity, size of interface icons/search bars and problems with navigation created some usability issues. Successful implementation of user-feedback was documented by substantial improvements in KPro on all dimensions. The proportion of physicians retrieving relevant answers in the final user tests doubled and more than half were able to find the answer to a question when searching with the final KPro search system. Effectiveness was primarily associated with occupational group. Physicians in training and hospital clinicians were most, and general practitioners least, successful at finding the information they searched for using KPro.

Usability improved on average by 5 %. Many users were impressed with the final search system for being intuitive in handling, showing clear structure. Only a quarter instead of the initial 47%, regarded the final KPro search system as time consuming. The age of physicians was inversely associated with the magnitude of both perceived usability and efficiency. Older users were more likely to perceive the search with KPro as inefficient and be overwhelmed by the number of tools and features available. In addition to age, self-employed physicians were the most likely group to perceive the system as time-consuming and suggested the integration of a « quick search » platform.

More than half of the physicians evaluating the final KPro search system could imagine using the system on a regular basis. Hospital clinicians and physicians in training reported the highest and general practitioners the lowest likelihood of future use. Multilingual, unbiased access, the personal library, export function, search facets and summary translation were perceived as the most useful aspects of KPro. Research physicians and hospital clinicians expressed strong interest for the « scientific articles » filter and physicians in training rated the « definition » and « online education » restriction as useful while general practitioners were drawn to the "language" faceting options. System accessibility, incompatibility with IOS devices and relevance of search results were the biggest barriers to system use.

Further emphasis should be devoted to the ranking strategy of KPro, implementation of similar queries, resource accessibility, interface simplification and navigation, content expansion within the faceting features, tools implementation in the KPro web browser, system accessibility and the development of KPro mobile applications.

1.2 General public

The general public user tests reported in this deliverable consist of two parts: A blind and a non-blind comparison of search results between Khresmoi for Everyone and Google, in which the preferences of a total of 22 users completing the online survey were evaluated during December 2013 and February 2014, as well as the results of the large scale user tests conducted in the period from May 2014 to June 2014.

These tests conducted on a larger population (N=63 for the general public user tests) collected feedback from targeted user groups to understand the level of user acceptance and provide final tuning recommendations for work package 8 and others before the online launch of the KHRESMOI search systems.

The blind / non-blind user tests show that presenting K4E as a search engine offering trustworthy health web sites did affect the choice of participants.

Full user tests were further conducted with a population of the general public in Prague, Geneva and Paris. All evaluation tests made provided a significant feedback on the acceptance of the K4E. A few improvements and fine tuning of the prototypes are still needed on the interface level before the search engine is launched online.

Currently, the priority is to increase the coverage of resources in Czech. Following that, our directives will be focused on the improvement of the automatic detection of the readability and trust, and prototype improvements of the interface. We are aware of the importance of the real user tests and how crucial they are to the final product which is why we pay a great deal of attention to the results of these tests. Additionally, it is important to keep in mind that an ‘indifferent’ user (user who does not have prior knowledge of HON, Khresmoi or the search engine) is an important factor to minimize bias and introduce new perspectives.

2 Khresmoi Professional

2.1 Research questions

Evaluation of the KPro search system was carried out in two rounds. In the first round of evaluation, carried out in November 2013 detailed user feedback was obtained and used to advance final year prototype development. The final user evaluation conducted from May-July 2014 aimed to evaluate whether implemented user feedback had the expected positive impact on effectiveness, efficiency and usability of the system as well as to obtain insight on usefulness of tools, functionalities as well as the exploitation potential of KPro. In terms of exploitation the main interest revolved around evaluating the « real life usability » and identifying a potential « niche » for Khresmoi Professional in the medical healthcare system. Having learned that subgroups of physicians vary immensely in their resource and information requirements great attention was devoted to evaluating the extent to which each physician group was likely to benefit from using KPro [7]. Table 1 lists the main objectives of KPro evaluation.

	Research question	Sample/comparisons	Questionnaire items
1	Search preferences of physicians. What search preferences do physicians have and what information do they search for?	Search preferences (N=33) Snippet preferences (N=52) Usage of devices (N=33) Search behaviour (N=84) Comparison across gender, age, occupational groups.	Search preferences Usage of devices Snippet preferences Search behaviour (query analysis)
2	KPro user evaluation in year 4 What is the overall reported efficiency, effectiveness and usability of Kpro in Y4?	Year 4 user evaluation (N=84). Comparison across age and occupational groups.	Effectiveness and efficiency Usability Usefulness of Tools, search features and search facets.
3	KPro improvement in year 4? How have efficiency, effectiveness and usability of KPro changed during year 4?	Comparison: User evaluation November 2013 (N=33) vs. User evaluation May 2014-July 2014 (N=51).	Effectiveness Efficiency Usability
4	KPro: the final prototype. Is it useful? Insight on usability, effectiveness and efficiency of the final version of KPro.	User evaluation May-July 2014 (N=51). Comparison across age, occupational groups and different versions of KPro.	Effectiveness Efficiency Usability
5	Exploitation potential of KPro. What is the likelihood of future use and what is useful? Suggestions for improvement?	May-July evaluation (N=51). Comparison across age and occupational groups.	% agreeing to SUS Item 1 „I think that I would like to use this system frequently" Analysis of open feedback

Table 1: KPro user evaluation: Research questions.

2.2 Methodology

2.2.1 Evaluation strategy

In the user tests of the initial search system that were conducted in January 2013 we learned that certain subgroups of physicians are "hard to reach" for user tests [2]. To address this issue we decided to conduct subsequent user tests at medical education conferences. The first round of user tests was conducted during the STAFAM in November 2013, one of the biggest medical events of general practitioners in Austria. The feedback gathered was then implemented in subsequent development of Khresmoi Professional. Final user tests were then carried out from May 2014- July 2014and included an online feedback survey. The methodology of different rounds of user tests was adapted to the evaluation objectives mentioned earlier. The user tests of the initial search system in January 2013 were based on users trying out the system using a number of pre-defined tasks. The initial advancement of the methodology was that in the November 2013 evaluation, users were additionally

asked to conduct a "free browsing task" in which they were asked to search for a question they formulated themselves. Furthermore, the final evaluation of the KPro system, in May-July 2014, tested the system solely on the basis of the « free browsing tasks » encompassing an online field evaluation.

2.2.2 Questionnaire

2.2.2.1 Questionnaire design

A questionnaire was designed by GAW and TUW, in collaboration with UDE, to evaluate the usage of tools, search preferences, efficiency and effectiveness of the Khresmoi Professional search system and made accessible in German and English. System usability was measured using the system usability scale [3] and the German translation proposed by Lohmann and Schäfer [8].

Face-face user tests: Each round of user tests was preceded by a pilot phase where members from the GAW and TUW conducted 2-3 pilot interviews and tests with Austrian physicians at the society of physicians in Vienna. Pilot tests gave insight into validity and reliability of the questions tested the methodological procedure using the testing software and allowed modifications to prepare for the final user tests. Final face-face user tests were carried out by GAW, TUW and UDE and ensured that Khresmoi members from different partners and WP were present to answer questions and address upcoming prototype issues.

Online survey: GAW set up an online survey via survey monkey, promoted to members of the society of physicians, social media platforms and allowing field evaluation of the search system. Prior to online dissemination an « online test period » of 1 week allowed further evaluation of reliability, validity and assessed the adequacy of the length of the questionnaire. Two physicians in Austria filled out and « tested » the questionnaire. In response to physician feedback the questionnaire was shortened and partially modified to ensure high levels of readability.

2.2.2.2 Questionnaire structure

The basic questionnaire structure remained similar across all evaluation events and varied primarily in length. Table 2 illustrates the questionnaire setup in detail. The full questionnaire can be viewed in Appendix 5.1.3, Tables A17-A25.

Evaluation event and total number of questions	Questionnaire			Total number of questionnaire items
	Part 1 Demographics and search preferences	**Part 2** Tasks, efficiency and effectivness	**Part 3** Usability, tools and functionalities	
STAFAM, Austria November 2013	Demographics: 7 items Search preferences: 6 items	Free browsing task: 4 items + Optional: Personal library task : 3 items or Search filter task: 4 items	Tools, facets and search features: 3 items SUS: 10 items Open feedback: 2 items	Free browsing group »: 32 items « Personal library group »: 35 items « Search filter group »: 36 items
Wiesbaden, Germany **May 2014**	Demographics: 4 items	Free browsing task: 4 items	Tools, facets, search features: 4 items SUS: 10 items Open feedback: 2 items	24 items
GAW events, Austria **May-June 2014**	Demographics: 4 items Search preferences: 1 item (= usage of devices)		Tools, facets, search features: 3 items SUS: 10 items Open feedback: 2 items	
Online feedback survey **June-July 2014**				

Table 2: Questionnaire structure across different evaluation events.

2.2.2.2.1 User evaluation November 2013

The final version of the questionnaire used for the November 2013 user tests consisted of 3 parts and a total of 32-36 questions, depending on which evaluation group users were assigned to (i.e. free task, personal library task, or filter task). In part one consent, demographic variables and information on search preferences was obtained. Part 2 measured the efficiency and effectiveness of KPro after the "browsing" period. Part 3 collected feedback on the usefulness of tools/search features /facets and KPro usability. The completion of the test took between 15-30 minutes depending on whether

physicians completed only the "free task" or also the additional "pre-defined tasks" asking them to evaluate specific tools.

2.2.2.2.2 User evaluation May-July 2014

2.2.2.2.3 Face-face user tests

For the second round of user tests the questionnaire from November 2013 was shortened to 25 items. Users only received the "free browsing task" (see procedure for details). Some demographic variables that did not show much variation, ceiling effects or proved redundant during the November 2013 evaluation were excluded. Questions on search preferences in part 1 were also excluded. However, a question on usage of devices was added in the last evaluation event and for the online feedback survey to clarify the importance of device accessibility. To maintain the shorter "25 item" questionnaire the question on "filter preferences by category" was in turn excluded. As a result a shorter test focusing on the evaluation of the system using real scenarios was performed. The survey took about 15 minutes and included a short demonstration on Khresmoi functionalities.

2.2.2.2.4 Online feedback survey

For the online feedback survey the same questions as for the medical events at GAW were used, (25 items) only the formulation of questions was adapted slightly (See Appendix, Table A19). In addition, the first page of the online survey included an item that allowed users to give feedback, in case they tried to access KPro but had accessibility problems. An item asking users to indicate which version of KPro (Java, web browser or mobile version) they tested was also included (Appendix 5.2). The pilot phase revealed that the completion of the survey took about 10 minutes.

2.2.3 Setup and promotion

2.2.3.1 Face-face user tests

Laptop setup: The user tests were carried out on two laptops with the Java desktop version of Khresmoi Professional. Questionnaires were configured with Morae Testing Software [9] (see Khresmoi Deliverable D10.1 [2] for a detailed description). For these user tests, only the screen with mouse movements was recorded. Audio and video recordings were not made (or only in the case of explicit consent) to avoid scaring away physicians with confidentiality issues.

Khresmoi Booth Setup: In order to reach "hard to access " general practitioners a booth was set up at the STAFAM in Graz, Austria in November 2013, the Praxis update in Wiesbaden, Germany in May 2014 and at the society of physicians in Vienna during May and June 2014. User tests were largely carried out during the coffee and lunch breaks of the conferences. At the STAFAM, physicians were attracted with "free coffee" 50 Euro Amazon vouchers, Khresmoi USB sticks and organic wine, which was handed as a thank you gift for participation. For the final user tests in Wiesbaden and the medical society of physicians in Vienna, thank you gifts included 15 Euro Amazon vouchers and Khresmoi USB Sticks. Posters informing about KPro were used to decorate the booth and flyers were handed out to motivate physicians to take part in user tests. In addition, a laptop was set up with a running video demonstration of Khresmoi Professional to attract the attention of passing physicians. Images illustrating the Khresmoi Booth set up can be found in Appendix 5.2.

2.2.3.2 Online feedback survey

The target population of the online feedback survey was of physicians of all specialties all over Europe. To promote the survey the society of physicians in Vienna placed a Khresmoi banner on their homepage www.billrothhaus.at leading to the link of the study and created a promotion text that was sent to 205 physicians encouraging them to participate. Furthermore, the GAW newsletter asked all members to participate. The link to the online survey was disseminated via the homepage of the society of physicians (www.billrothhaus.at) and via the Facebook Khresmoi expert community (see

Appendix). Furthermore, E-mails were sent to 205 physicians by GAW, explicitly asking for survey participation. Further details and screenshots illustrating the online survey and dissemination can be retrieved from the Appendix 5.2.

2.2.4 Timeline

Overall, the first round of user tests took place from 28/11/2013 to 30/11/2014. The second round of user tests took place from 16/05/2014 to 12/07/2014. The medical events at GAW composed of three events in May and June each lasting between 1 and 3 days. The event in Wiesbaden lasted two full days from 16.5-17.5.2014. The online feedback survey was accessible from 03.06.2014-12.07.2014. A pilot phase was conducted one-two weeks before each evaluation period. Details on time line and setting of all places of user evaluation can be found in Table A26 in the Appendix (Section 5.2).

2.2.5 Procedure

2.2.5.1 Face-face user tests

The basic procedure was the same for both rounds of evaluation. The only difference was that that for Part 2 all users in the May-July 2014 evaluation were assigned exclusively to the « free browsing task » group while in the November 2013 evaluation two thirds of the participants additionally completed a « personal library task » or « filter task » (See Table 2).

The following steps were carried out during each user test:

Part 1: Demographic variables:

Users were given information about the study, provided informed consent and filled out questions on demographic characteristics and search preferences.

Part 2: Free browsing task

Participants were asked to think of a typical medical question and use KHRESMOI to find the answer without detailed instructions.

Khresmoi demonstration

Users were then shown the following functionalities by the experimenter and given the chance to ask questions about the system while trying out the functionalities.

- Personal library: storing and tagging information
- Exporting a link
- Filtering, search help and double click restriction
- Translation of the summary of an article.
- Sorting by date/relevance

Experimental group allocation: User feedback on KPro effectiveness and efficiency

Users were then allocated to one of three groups, which determined the next step of the procedure

- **Personal library task:** Participants were then additionally asked to complete a task which required them to use KHRESMOI to find information on LADA Diabetes, use the personal library to store relevant websites, the tag function to attribute labels to the links and the export function to export the link and give detailed feedback on the tool.
- **Filter task:** Participants were additionally asked to either complete a task, which required them to find the answer to a task about the side effects of estradiol or find information on

diabetes education and in both cases use the filter restrictions, sorting option, and translation option.

- **Only "free browsing task":** Users allocated to this group did not receive a pre-defined additional task. Feedback was given based on the experience with the free task only. Time-constrained users were allocated to this group.

Part 3: Final feedback: Usability, usefulness of tools, search facets, search features and open feedback

All participants were then asked to fill out the Standard Usability Scale and asked further questions on the usefulness of tools, search facets and features of KPro, as well given the chance to provide open feedback/make suggestions for improvement.

2.2.5.2 Online feedback survey

After a pilot phase (27.5-02.06.2014) the questionnaire was made available in English and German on the www.surveymonkey.com where it was disseminated electronically to physicians around Europe. The online surveys were accessed using the following links:

English version: http://www.surveymonkey.com/s/khresmoi_en

German version: http://www.surveymonkey.com/s/khresmoi_de

Details on dissemination of the survey can be found in the Appendix.

2.3 Statistical analysis

2.3.1 Overall

Analysis of user test results and statistical evaluation was conducted by GAW using SPSS 21 for Macintosh. All graphs and tables were constructed using Microsoft Office Excel for Mac 2010. Descriptive statistics were used to calculate frequencies for categorical variable and mean values for continuous variables. Open-format answers were categorised and all included in their original form in the Appendix. Frequency tables and graphs were constructed to describe and illustrate demographic characteristics and survey responses. Differences between independent variables were tested using a Mann-Whitney U Test for categorical variables. To determine statistical significance an overall alpha level of $p>0.05$ was used. Percentages are based on the valid sample within each variable. The size of the valid samples will be mentioned in all graphs throughout the analysis.

2.3.2 Indepedent variables: What was compared?

2.3.2.1.1 Gender

A comparison based on gender was solely performed for search preferences.

- Male
- Female

2.3.2.1.2 Age group

For the purpose of statistical analysis age was collapsed into the following five groups:

- <30 year old
- 31-40 year old
- 41-50 year old

- 51-60 year old
- 60 and older

2.3.2.1.3 Occupational group

Physicians were classified by their occupational group. Extensive user analysis as described in an earlier report [4] allowed us to identify concrete subgroups of physicians that have been shown to vary in information and resource requirements [7]. The following occupational groups were classified and compared throughout the analysis.

- Physician in training
- Self-employed general practitioner: included physicians that were primarily active in their own practice as general practitioners.
- Self-employed specialist: included physicians that were primarily active in their own practice as specialists.
- Hospital physicians: included physicians that worked in non-academic institutions. (e.g. rehab centres)
- Research physicians: included physicians that worked in academic institutions and/or in education.

2.3.2.1.4 Time of evaluation:

For the purpose of statistical analysis and to determine KPro improvement in Y4 two rounds of user tests were compared.

- November 2013 user evaluation (i.e. STAFAM)
- May-July 2014 user evaluation (included Wiesbaden Praxis update event, all medical events at GAW and the online feedback survey)

2.3.2.1.5 Type of prototype:

For the purpose of statistical analysis a comparison in usability and accessibility between the two different types of Khresmoi prototypes was performed.

- Khresmoi web browser
- Khresmoi desktop version

We did not make comparisons based on country or language since most of the users were from Austria or Germany and no significant differences were detected.

2.3.3 Dependent variables: What was measured?

The most important variables measured were search preferences, search behaviour, usability, effectiveness and efficiency, usefulness of tools, facets and functionalities.

2.3.3.1.1 Self-reported search preferences

Internet usage, usage of online medical resources, information needs, ranking preferences, and categorisation preferences were assessed solely during the first round of user tests in November 2013. Thus, comparisons across occupational groups are excluded for these items, as they would lack statistical relevance. In addition, snippet preferences were assessed in November 2013 and May 2014. Usage of devices was only measured in the online feedback survey and GAW medical events.

2.3.3.1.2 Actual search behavior

The complete list of questions asked by physicians in the "free browsing task" can be viewed in the Appendix (Tables A1-A12). All questions asked by physicians were assigned to one of six different information groups. Table 3 illustrates the categories with examples. Overview information and abbreviations typically included single worded queries (e.g. « Makulpopathie », « Hyperhidrosis ») or abbreviations (e.g. CCNU). Research queries and news included all queries that asked for theoretical scientific associations and updates. Drug information included queries that included the name of a drug or a class of drugs. Treatment information included all queries that mentioned treatment or guidelines of some sort. Patient information included queries that specifically asked for information for patients. Table 3 illustrates typical examples for each category.

	Type of query	Examples (translated in English)*
1	Drug information	"*Dosage of Amoxicillin for Kids with Erythema Migrans*"
2	Scientific article and news	"*Oesophageal cancer p53*", "*bronchial Ca T3N2M1*"
3	Treatment	"*Diabetes insipidus treatment*"
4	Information for patients	"*Patient information rheumatism*"
5	Diagnostic information	„*Diagnosis of Chusing syndrome*"
6	Overview info /abbreviations	"*Pseudopseudohypoparathyroidism*", "*Hifu*"
*For further details on analysis and allocation of queries to categories, please contact: marlenekritz@gmail.com		

Table 3: Categorization of physician queries made in the "free browsing task".

2.3.3.1.3 Effectiveness

Effectiveness was determined by whether users succeeded in solving tasks and the level of satisfaction of retrieved information. For this purpose participants were asked after each task « did you find the information you searched for using KPro? ».

2.3.3.1.4 Efficiency

Efficiency was determined by the self-reported time and effort invested to solve tasks. In the STAFAM user tests the question asked was "Do you find Khresmoi more time-consuming than normally?" (Answer: Yes/No). In the user tests carried out from May-July the question was: "How time-consuming do you regard it to search for medical information using Khresmoi Professional? (Likert scale: 1 (Not at all) - 5(very time consuming). Answers 1 and 2 were collapsed to mean "not time-consuming", and answers 4 and 5 were collapsed to mean "time-consuming". Answer "3" was regarded as neutral.

2.3.3.1.5 Usability and calculation of SUS percentile ranks

To measure KPro usability the System Usability Scale (SUS) was used. The SUS is a ten-item scale giving a global view of subjective assessments of usability. Users indicated on a scale from 1-5 to what extent they agreed with 10 statements about the system. Positive and negative SUS statements were first analysed separately and then collapsed into a global usability scale. Level of agreement to each of the 10 individual SUS items was compared. To obtain the global usability scale, raw SUS scores were converted into percentile ranks (Range 0-100). To determine the SUS score, first the score contributions for each item were summed. For items 1, 3, 5, 7 and 9 the score contribution was

determined by the scale position minus 1. For items 2,4,6,8 and 10, the score contribution was determined by calculating 5 minus the scale position. The sum of the scores was then multiplied by 2.5 to obtain an overall value. SUS scores had a range of 0 to 100. SUS score interpretations were based on the assumption of Sauro [10]. A SUS Score of 68 was considered as average [10] [11]. He proposes for example any score above 70 is above average and a score above 80 would be classified as an "A" system that is recommended by a friend. A system scoring around 68 is average and below 51 would be a "C" of a system that is unlikely to be used in the future. Furthermore, SUS items number 4 and 10 are used to provide insight into the "learnability dimension" while the remaining 8 Items provide information on the usability dimension [10].

2.3.3.1.6 Usefulness of tools, search features and facets

Usefulness of tools and search facets was determined by designated questions asking users to give feedback at the end of the questionnaire. In addition, questions associated to the "filter task" and the "personal library task" gave further, qualitative, insight on tools.

2.4 Results

2.4.1 Selection criteria and participants

Of the initial ninety-four participants taking part in the study, ten participants were eliminated from the analysis. This resulted in a final sample of eighty-four physicians being included in the analysis. Questionnaires were excluded if less than 70% (7/10 items) of the SUS and/or one of the central feedback questions effectiveness and/or efficiency were not answered. In three cases missing SUS data was replaced using the multiple imputation method [6] (Predictors were other SUS scores, occupational status, gender, age and prototype).

2.4.2 Demographics

2.4.2.1 Location of evaluation

2.4.2.1.1 Location of evaluation: Overall participation in year 4

As illustrated in Figure 1, 40% the overall feedback came from the initial STAFAM event in November 2013, a third came from the online feedback survey from June-July 2014 and the remaining feedback was gathered at medical events held at the society of physicians in Vienna in May and June 2014.

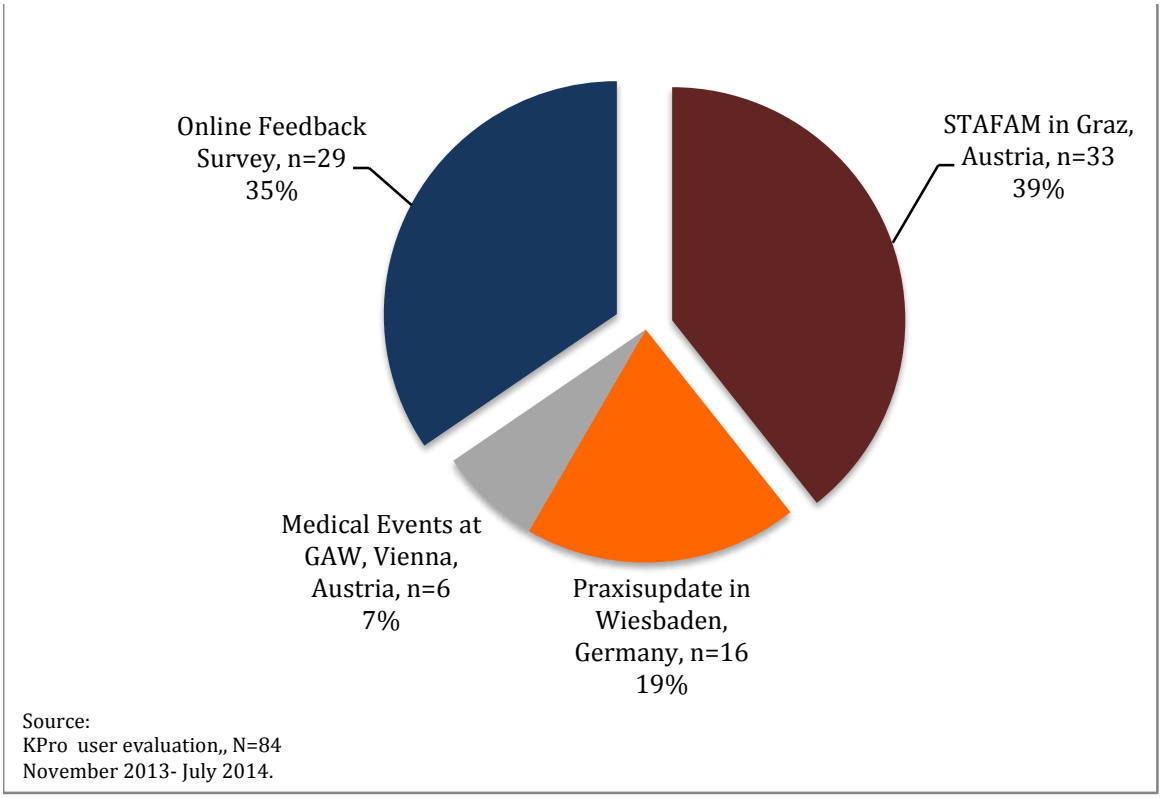


Figure 1 Location of KPro user evaluation conducted in Y4.

2.4.2.1.2 Location of evaluation : Location across different rounds of user evaluation

The first round of user tests was all carried out in a single location while in the final round of user tests several locations for evaluations were chosen (Figure 2). The sample in the first evaluation was biased towards Austrian general practitioners. The General practitioner bias, deliberately created and initiated by carrying out the user tests at a GP conference, aimed to address a prior lack of feedback from « hard to access » self-employed practitioners. The need of self-employed practitioners can be quite different to that of other physician subgroups and their feedback was therefore of crucial importance. In the second round of user tests, shorter feedback from a broader spectrum of physicians, from different locations, including an online field evaluation, was acquired. Furthermore accessing a larger number of physicians allowed insight into KPro exploitation potential.

2.4.2.1.3 Location of evaluation: May-July 2014 user evaluation

51 (22 female, 29 male) participants took part in the evaluation of the final prototype from May-July 2014. Just over half of the feedback was obtained from the online feedback survey and the remaining feedback was obtained from face-face user tests held at medical events in Austria and Germany.

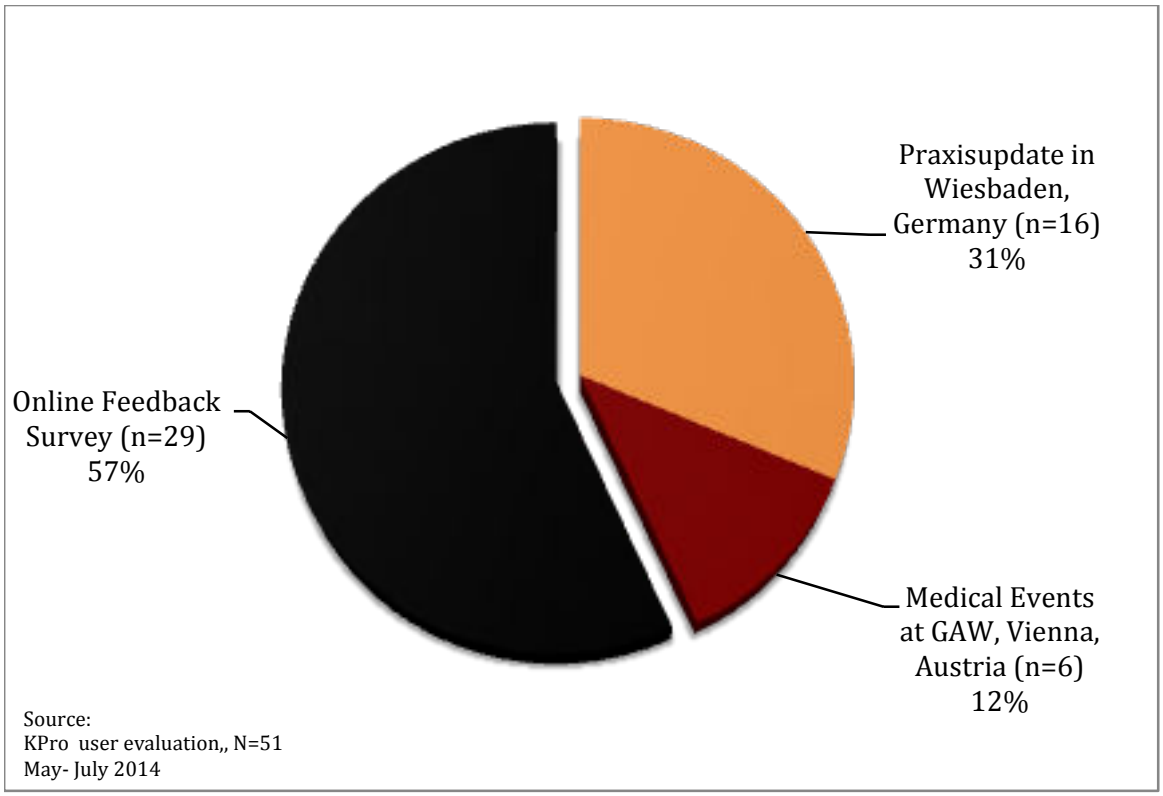
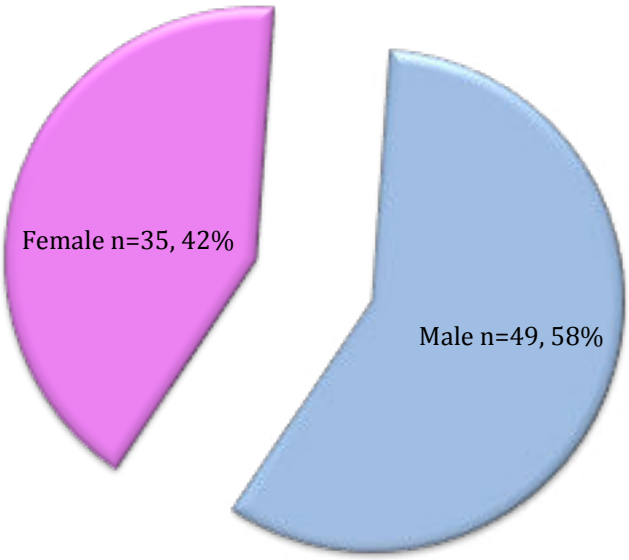


Figure 2 Location of the final KPro user evaluation.

2.4.2.2 Gender

Overall, 42% of the participants were female and 58% male (Figure 3). The highest proportion of males took part in the STAFAM event (61% males). The highest proportion of female participants took part in the Wiesbaden events (50% female). However, despite slight divergences, the gender distribution did not differ significantly across events.

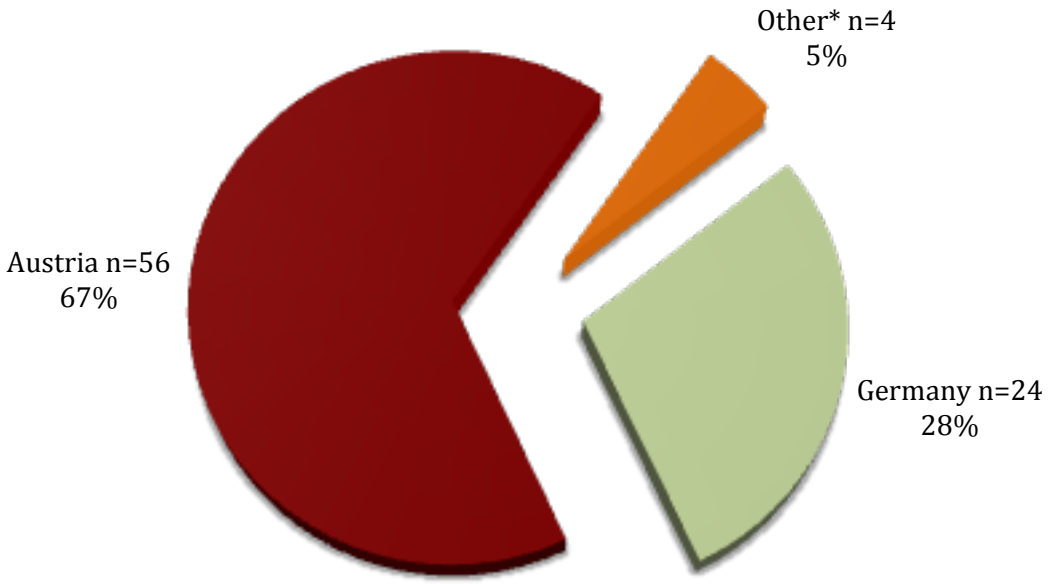


Source:
KPro user evaluation,, N=84
November 2013- July 2014.

Figure 3 Gender distribution in KPro user evaluation conducted in Y4.

2.4.2.3 Country

Most physicians came from Austria or Germany. Only 4 participants from "other countries" took part, where 3 of them were accessed via the online feedback survey (Figure 4).



Source:
KPro user evaluation,, N=84,
November 2013- July 2014.
* The category "Other" includes the
following countries: Spain, United Kingdom, Italy

Figure 4: Country distribution in KPro user evaluation conducted in Y4.

2.4.2.4 Age

2.4.2.4.1 Age: Overall participation in year 4

All illustrated in Figure 5 all age groups were represented in year 4-user evaluation with about half of the participants being between 41-60 years old. The age of physicians participating in year 4 user evaluation events ranged from 25-73 years (Mean: 48 years).

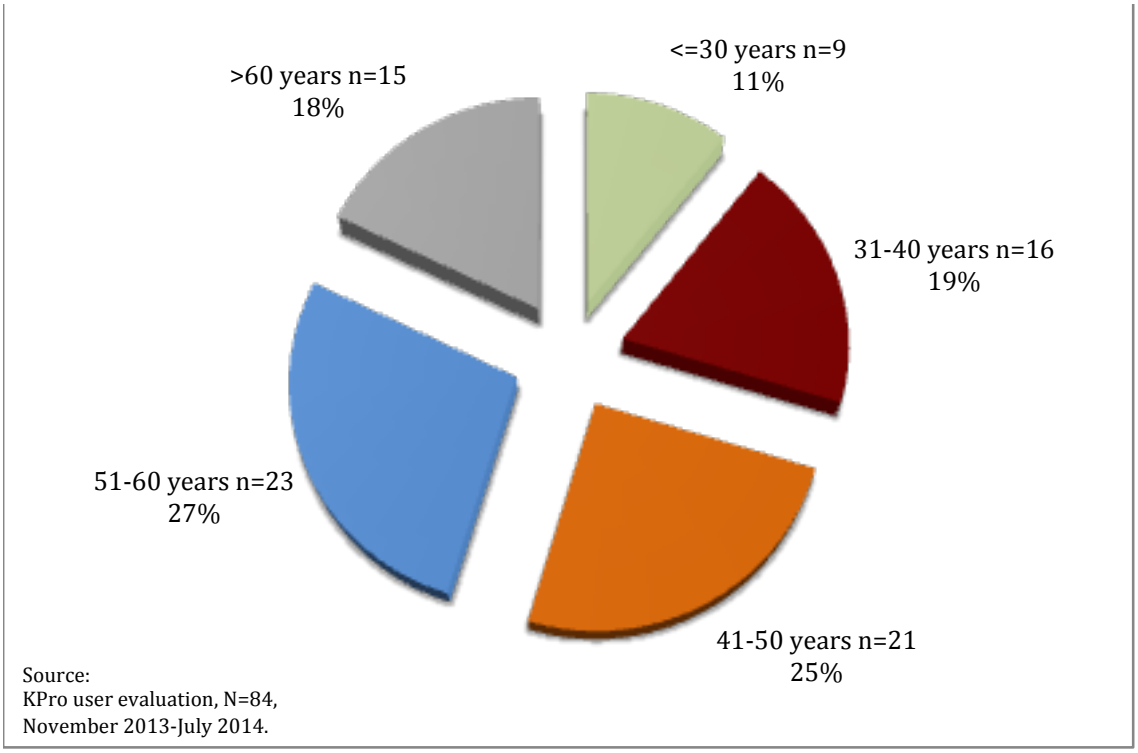
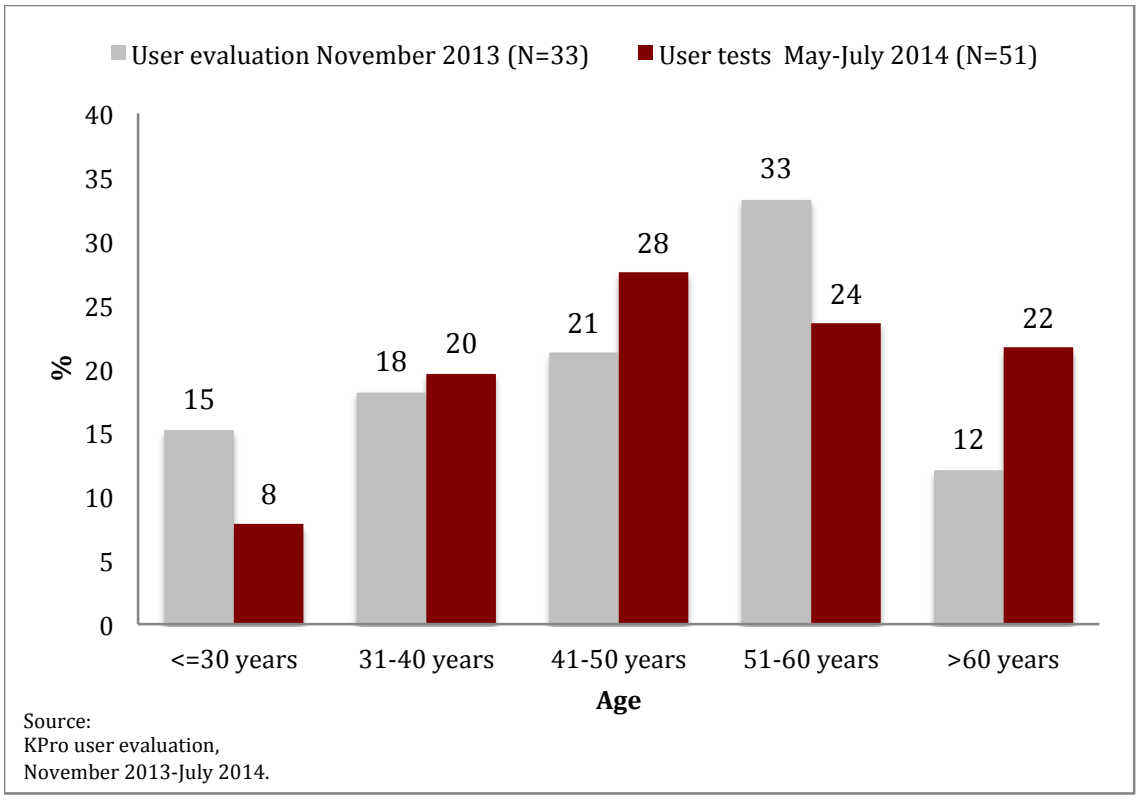


Figure 5 Age distribution in KPro user evaluation conducted in Y4.

2.4.2.4.2 Age: A comparison across different rounds of user evaluation

Physicians were on average oldest at the medical events of the GAW (Mean age: 59 years) and youngest in the online feedback survey (Mean age: 45 years). Across different rounds of evaluation events age groups were represented to a similar extent.

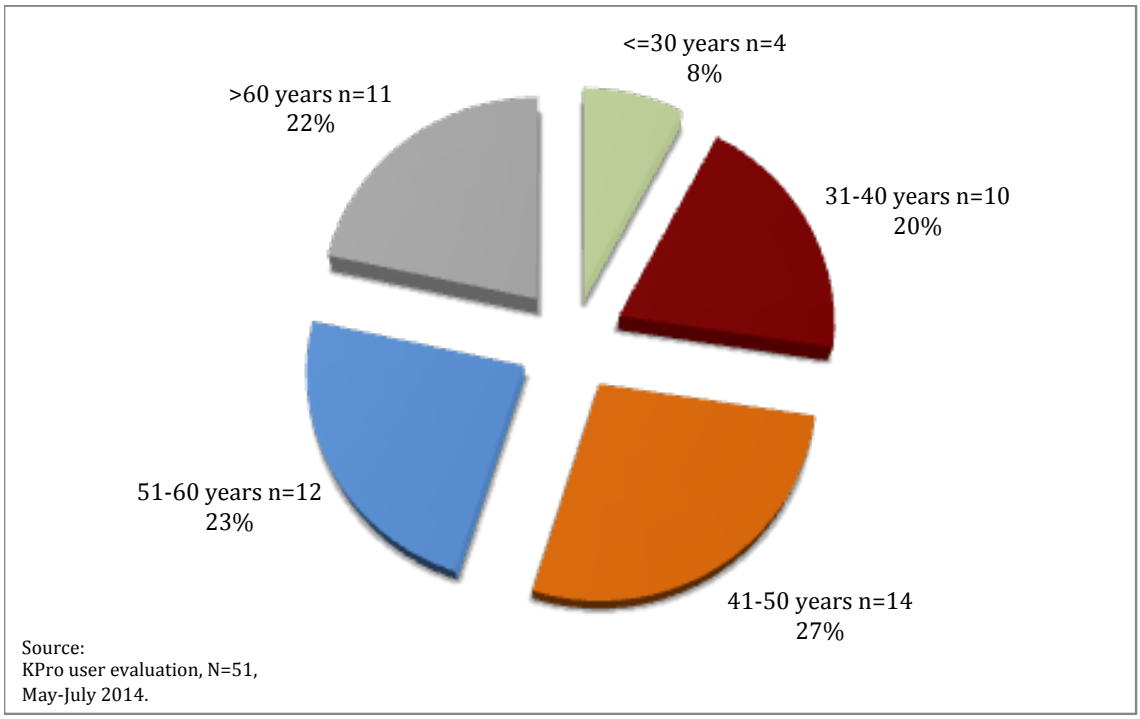


Source:
KPro user evaluation,
November 2013-July 2014.

Figure 6 Age distribution across different rounds of user evaluation.

2.4.2.4.3 Age: May-July 2014 final user evaluation

In the final round of user evaluation half of the participants were 41-60 years old, just over a quarter was below 40 and every fifth participant was more than 60 years old (Figure 7). The overall mean age was 49 years. As expected physicians in training made up the youngest group of physicians (Mean age: 30 years) and research physicians (Mean age: 57 years) were the oldest group.



Source:
KPro user evaluation, N=51,
May-July 2014.

Figure 7 Age distribution in the final KPro user evaluation.

2.4.2.5 Occupational group

2.4.2.5.1 Occupational group: Overall participation in year 4

Overall, all occupational groups of physicians were represented in the Y4 user evaluation, with most working as self-employed practitioners (Figure 8).

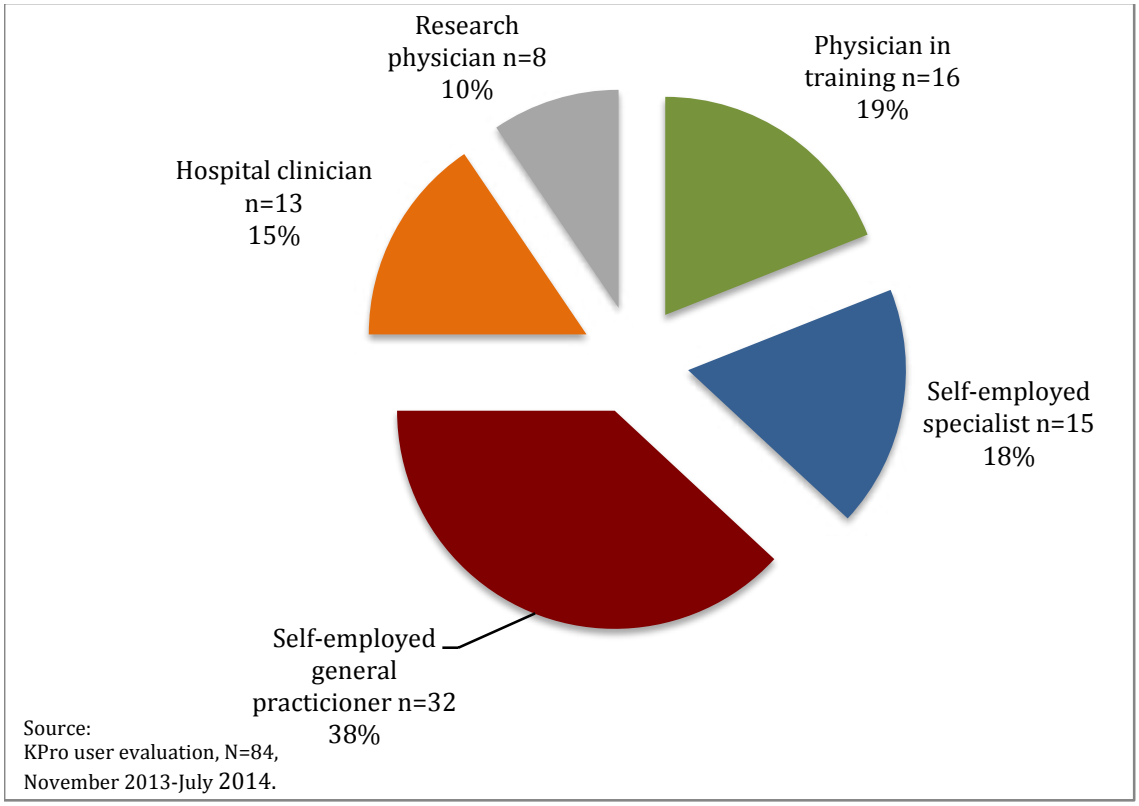


Figure 8 Distribution of occupational groups in KPro user evaluation conducted in Y4.

2.4.2.5.2 Occupational group: A comparison across different rounds of user evaluation

During the user tests in November 2013 participants consisted of general practitioners (N=19) and physicians in training (N=9), a few self-employed specialists. (N=4) and only one research physician. In the final round of user tests all occupational groups were represented (Figure 9).

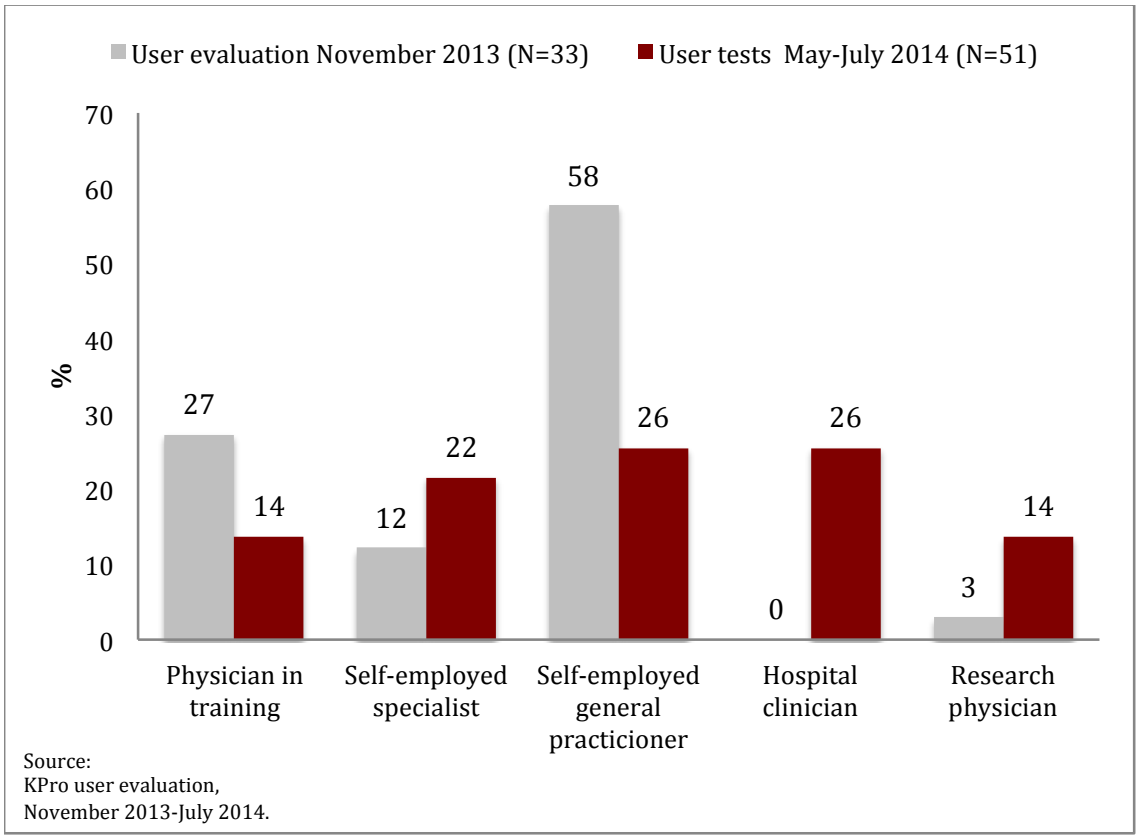


Figure 9 Distribution of occupational groups across different rounds of user evaluation.

2.4.2.5.3 Occupational group: May-July 2014 final user evaluation

In the final user evaluation hospital clinicians, self-employed general practitioners, self-employed specialists, represented in equal amounts about three quarters of the sample. Physicians in training and research physicians were each represented 14% of the sample (Figure 10).

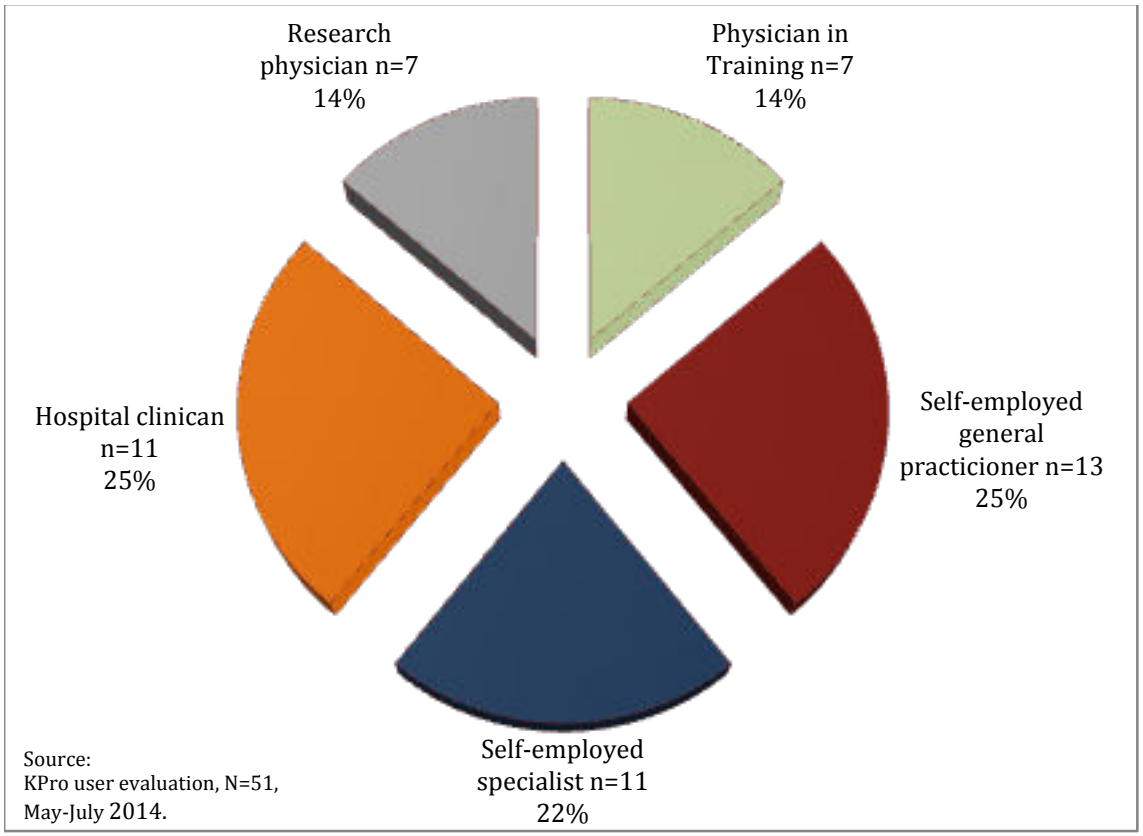
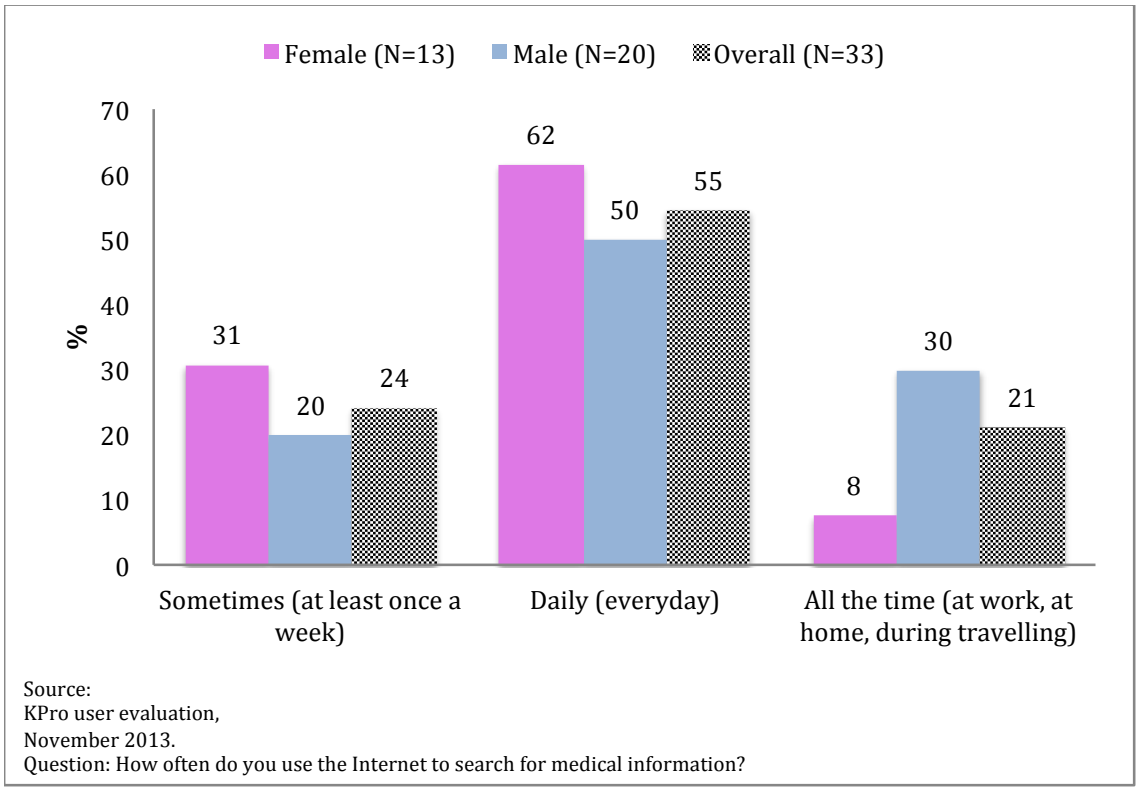


Figure 10 Distribution of occupational groups in the final KPro user evaluation.

2.4.3 Search behaviour and preferences

2.4.3.1 Internet usage

More than half (18/33) of the participants reported using the Internet on a daily basis to search for medical information. An additional 21% reported to use the Internet "all the time" to search for medical information. Males showed a tendency to report higher Internet usage than females. However, the difference did not reach statistical significance.



Source:
KPro user evaluation,
November 2013.
Question: How often do you use the Internet to search for medical information?

Figure 11 Professional Internet use by gender

2.4.3.2 Devices used to access online medical information

As illustrated in Figure 12, the majority of physicians accessed medical information using a Laptop of a PC. 57% accessed via a smartphone of which more than half used an iPhone. Every second physician reported using a tablet of which two thirds used an iPad. The dominance of IOS devices among physicians highlights the importance of medical search system being compatible with IOS systems.

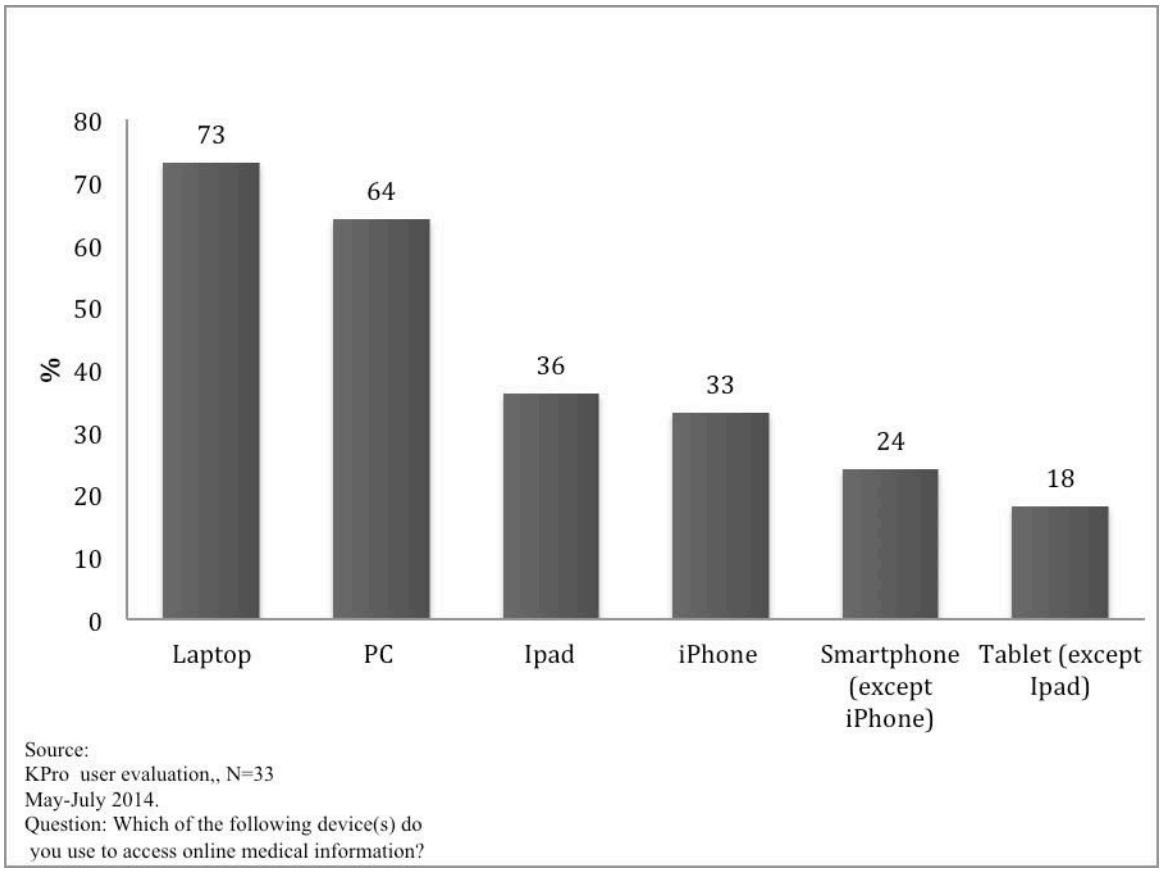
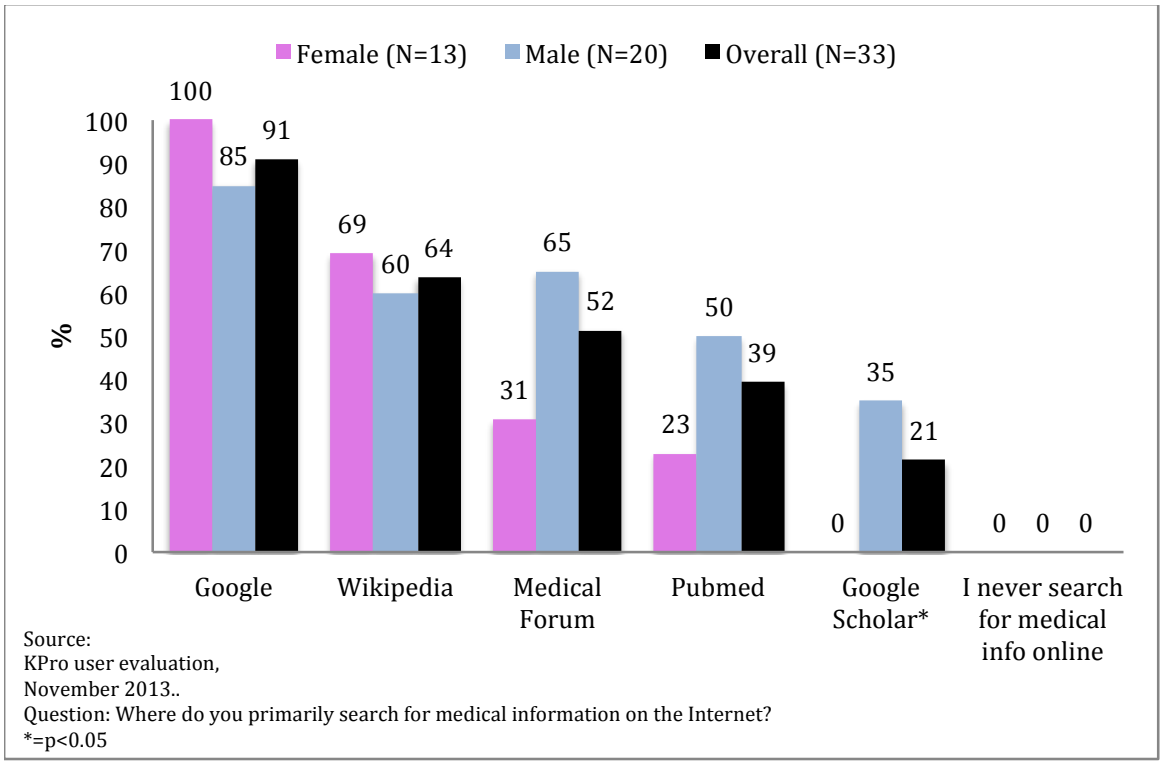


Figure 12 Devices used to access online medical information.

2.4.3.3 Usage of online resources

As shown in Figure 13 almost all physicians reported to use Google (30/33, 91%) to search for medical information. Almost two thirds used Wikipedia followed by every second physician using medical forums. Scientific resources such as PubMed and Google scholar were cited as the least popular resources. However, the sample, which was solely based on the November 2013 evaluation, may have been biased towards the needs of general practitioners who are known to express less interest in scientific articles [7]. Male physicians were more likely to report using Google scholar than female physicians ($p>0.05$).



Source:
KPro user evaluation,
November 2013..
Question: Where do you primarily search for medical information on the Internet?
*=p<0.05

Figure 13 Usage of medical online resources by gender.

2.4.3.4 What information do physicians search for?

2.4.3.4.1 Self-reported information needs

As illustrated in Figure 14 information about drugs, scientific articles, abbreviations and treatment guidelines were cited as the most important online information needs. Furthermore, male physicians were more likely than females to search for scientific articles ($p>0.01$). On the other hand female physicians were more likely than males to use the Internet to search for abbreviations (($p>0.05$) or for patient information ($p>0.01$)).

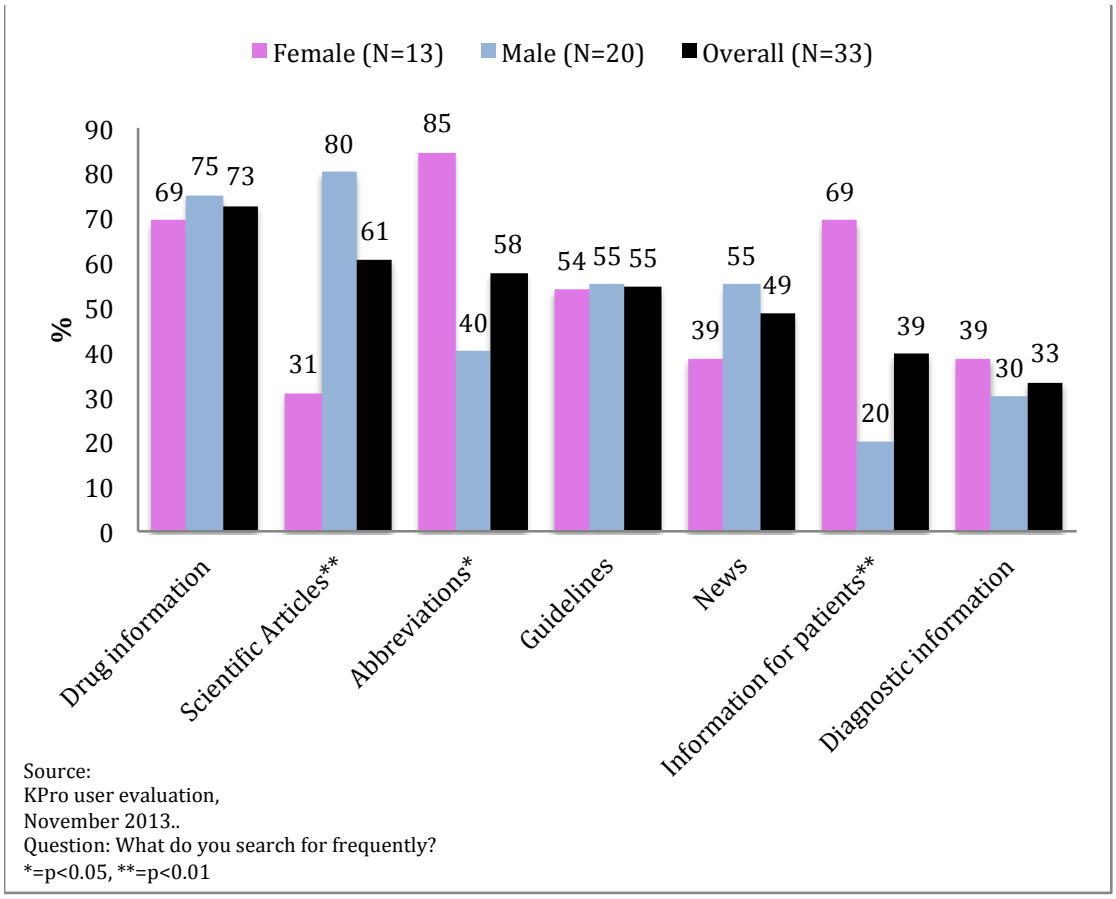


Figure 14 Reported medical information needs by gender.

2.4.3.4.2 Real search behaviour: What information did users in the free task search for?

Figure 15 illustrates the type of queries users searched for in the free browsing tasks. It was found that most users searched for treatment information, diagnostic information, overview information and drug information. Only 10% searched for scientific information. In comparison to reported information needs (Figure 14) it appears that physicians somewhat over-reported the need for diagnostic information and while reporting higher need of scientific articles than is reflected in real search behaviour.

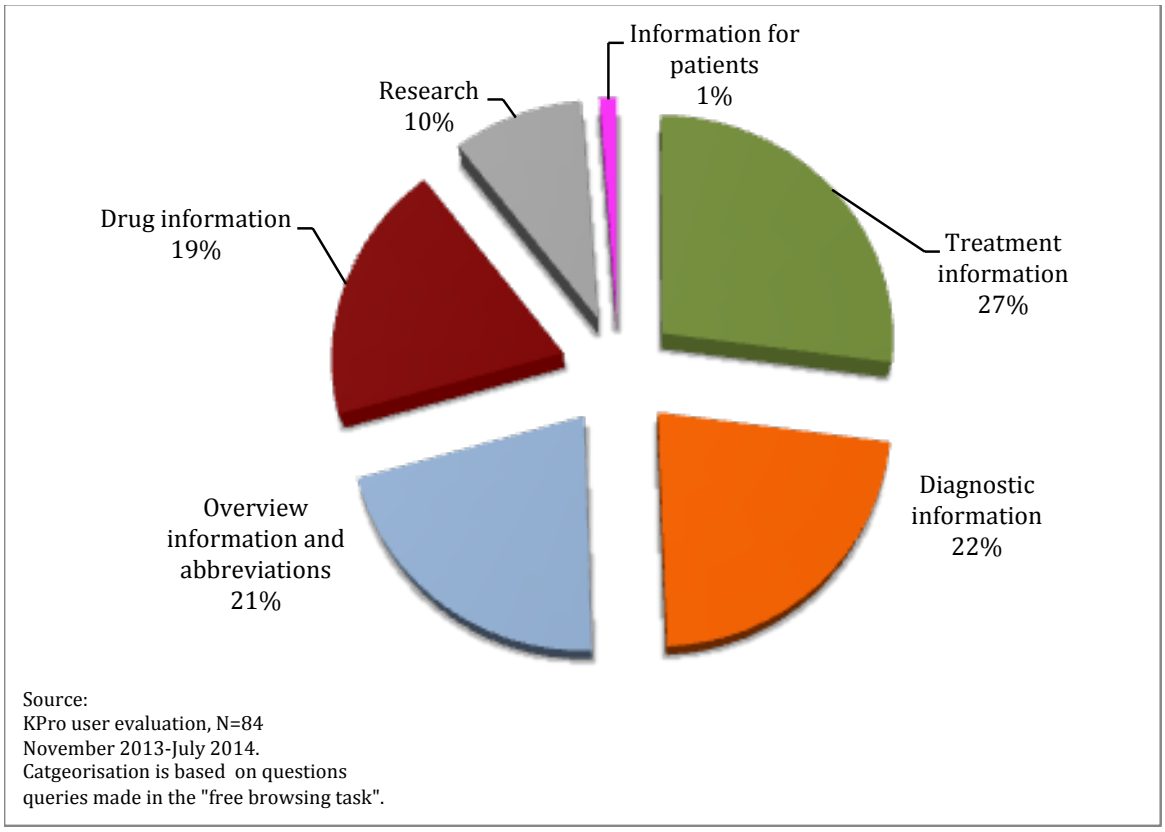
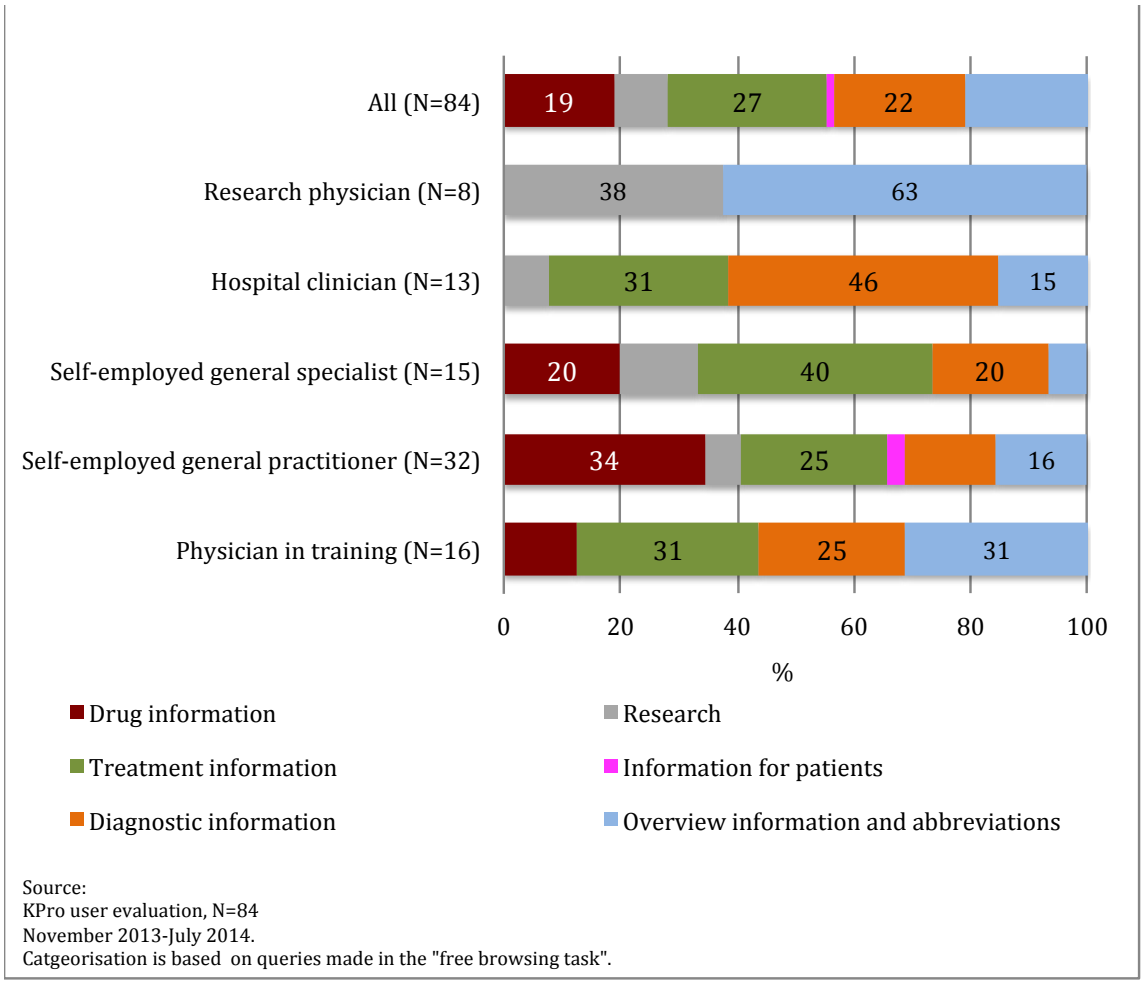


Figure 15 Categories of questions asked in the « free browsing task ».

2.4.3.4.3 Real search behaviour: How did search behaviour differ across different occupational groups of physicians?

As shown in Figure 16, the majority of self-employed specialists searched for treatment information while most hospital physicians searched for diagnostic information. Research physicians searched for scientific articles and overview information while general practitioners searched on all domains. An interesting finding was that research physicians were solely interested in scientific and overview information (63 %!). The finding is in line with previous research where research physicians reported an exceptionally high usage of overview resources such as Wikipedia [7]. However, since only 7 research physicians took part in the evaluation further studies are required to confirm the trend.



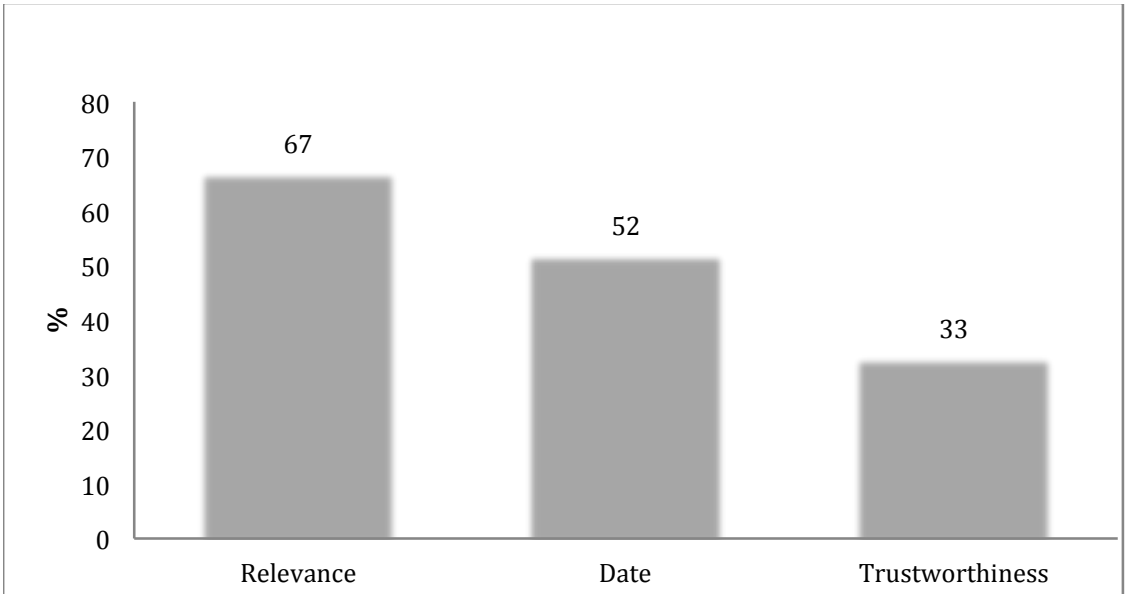
Source:
KPro user evaluation, N=84
November 2013-July 2014.
Catgeorisation is based on queries made in the "free browsing task".

Figure 16 Categories of questions asked in the « free browsing task » by occupational group.

2.4.3.5 Ranking preferences

2.4.3.5.1 Ranking preferences: All groups

As shown in Figure 17, two thirds of the participants (22/33) preferred ranking by relevance, every second participant by date, and a third by trustworthiness.

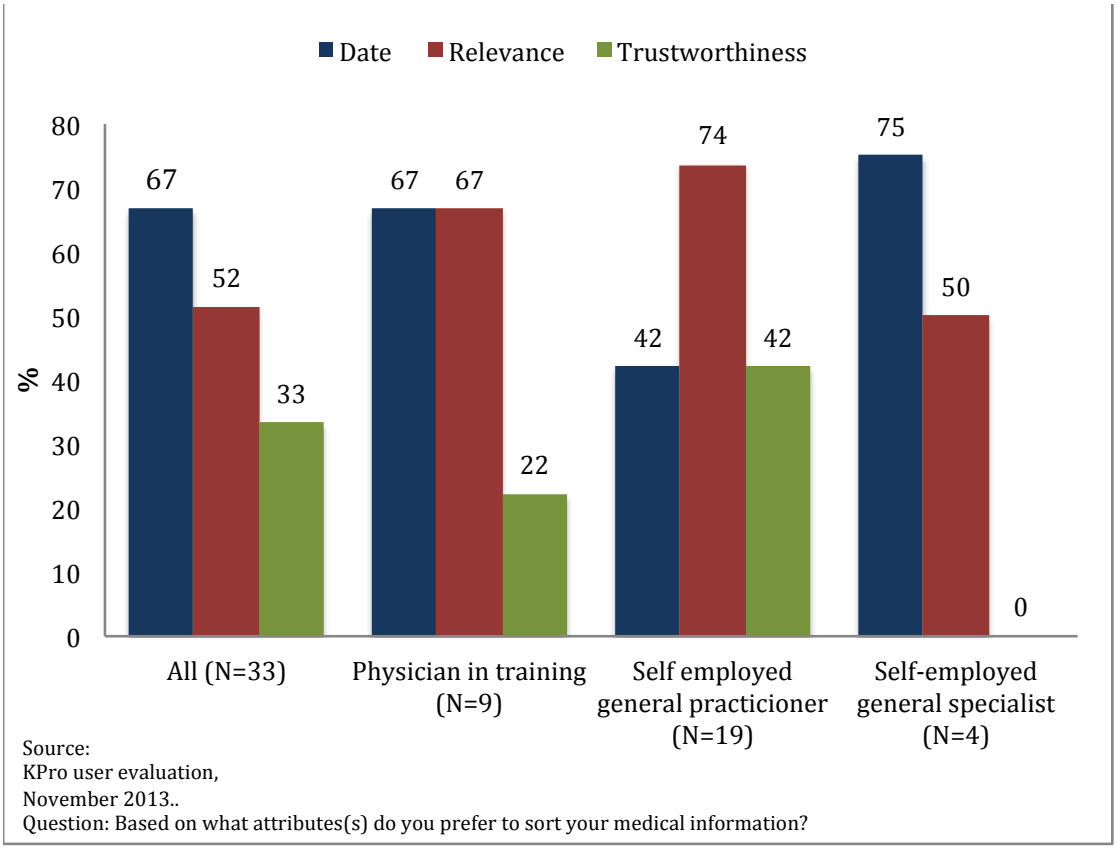


Source:
KPro user evaluation, N=33,
November 2013..
Question: Based onwhat attribute(s) do you prefer to sort your information?

Figure 17 Ranking preferences

2.4.3.5.2 Ranking preferences: A comparison across different occupational groups

As shown in Figure 18, specialists preferred ranking by date of publication, most general practitioners preferred ranking by relevance and physicians in training liked both. Another interesting finding was that none of the asked specialists reported to sort their information by trustworthiness. On the other hand, as much as 42% of general practitioners preferred to sort their information by trustworthiness. However, due to the low number of specialists participating (N=4) in the STAFAM, a bias is possible.



Source:
KPro user evaluation,
November 2013..
Question: Based on what attributes(s) do you prefer to sort your medical information?

Figure 18 Ranking preferences by occupational group.

2.4.3.6 Link snippet

The most important attributes that physicians preferred to see in the link snippet were the title of the document, highlighted query words, the link, the source and the date. Of less importance was the indication of target audience (Figure 19).

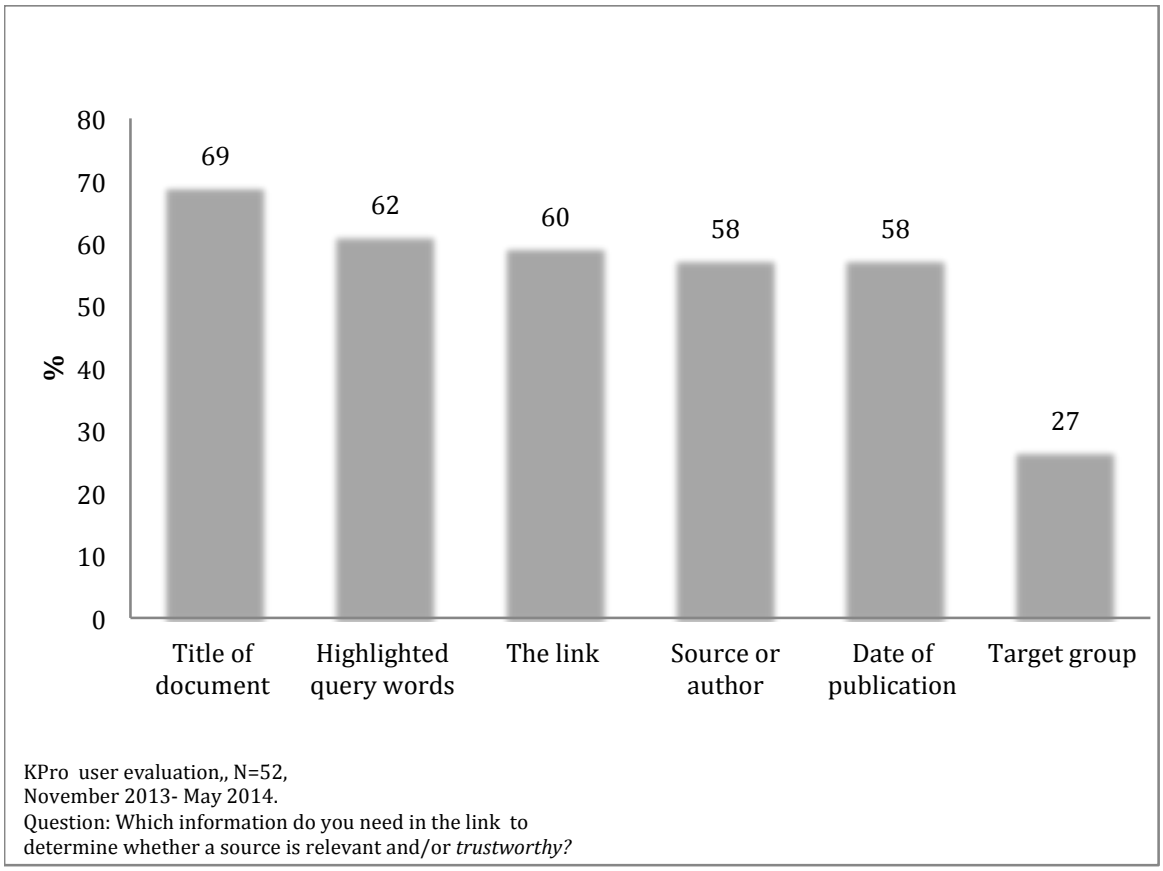


Figure 19 Snippet preferences.

2.4.3.7 Content categorization

With regard to categorization, most physicians reported to prefer categorizing the information by content, publisher and type of source. Almost every second participant preferred to categorize by language and country. Least popular was categorize by medical specialization (Figure 20).

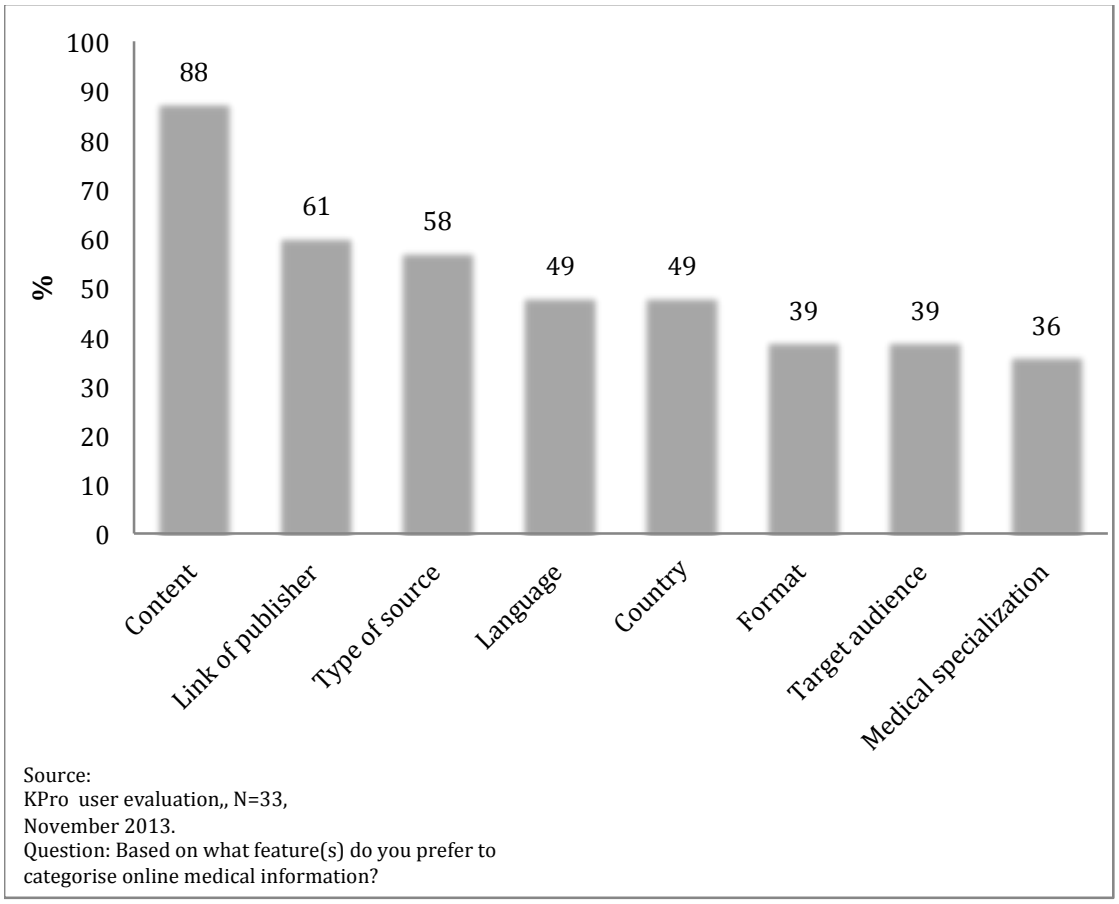


Figure 20 Information categorization preferences.

2.4.4 Lessons learned and improvements made

2.4.4.1 Lessons learned from the November 2013 user evaluation

Table 4 illustrates the most important feedback gathered in response to the first round of user tests. A detailed summary of November 2013 results can be found in Appendix 5.1.

November 2013 user evaluation	
Users liked:	
1.	Tools and search facets: Search facets, personal library, export function.
2.	Multilingualism: Interface, query translation and translation of summary.
3.	Organized and detailed structure of interface.
Users didn´t like:	
1.	Effectiveness: Irrelevant search results for complex queries, too specific search result for general queries, too general search result for simple/concrete queries.
2.	Usability: Interface too complex for older/less IT proficient and time-constrained.
3.	Efficiency: Slow loading time of search results and translation, irrelevant auto-suggestion and spelling correction.
Users suggested:	
1.	Effectiveness: Improve relevance of search results.
2.	System efficiency: Implementation of „quick search“, quicker loading time of results, advance auto suggestion/spelling correction, include “did you mean” function.
3.	Usability: Improve navigation (e.g. start search button, bigger search field etc.), de-clutter interface from unused features (i.e. common words, excerpt).
4.	Development of tools and facets: Expand export formats, expand facet content, add feed, develop collaborative tools, improve usability of personal library and quality of translation in summary.

Table 4: Most important user feedback obtained from November 2013 user evaluation.

2.4.4.2 User feedback implemented into prototype development in year 4

The table below illustrates issues that were solved in response to the first round of evaluation.

November 2013 user evaluation	
Issues raised	**Solution**
Effectiveness: • Relevance of search results: „*Irrelevant search results*“ „*Search query was not considered as a whole*“ • Resources: *„ Quick readable articles were missed*“, “*More German overview articles*”	Ranking improved substantially and missing content was addressed by crawling more resources.
Efficiency: • Quicker loading time: „ *very slow loading time*“ , „ *Abbreviations should appear faster*“ • Improvement of auto-suggestion and spelling correction: “*Lack of search input help for “ Maligne Wirbelsäulen tu*“ *“ Failure of KHRESMOI to do simple spelling correction*”	Substantial improvements.
Usability: • Interface: “*Simplification of interface*” *“ Middle interface is too narrow*” • Navigation: *“ Bigger search bar would be good and start search button*“ *„ I didn´t like the Umlaute were not displayed correctly which makes the preview useless/not readable.*“	Issue solved by excluding unused features and hiding some features by dropdown (e.g. excerpt, common words) Navigation issues solved.
Tools, search features and facets: • Autosuggestion, spelling correction and summary translation: *“ Lack of search input help for “ Maligne Wirbelsäulen tu*“ *“ Failure of KHRESMOI to do a simple spelling correction*”, *„ Umlaute not displayed in translation*“ • Search facets: *“ Restrict date category to 3 facets*” „*Leitlinien.de should not be classified as laypeople info*“	Issues were solved.

Table 5: Issues addressed that were raised in the November 2013 user evaluation.

2.4.4.3 User centred search system? An effective approach?

Feedback from the November 2013 user evaluation was addressed by other WPs and implemented into further prototype development. Important issues raised were addressed in the context of subsequent

prototype development. System efficiency and effectiveness was mainly addressed by including more resources, improving ranking and reducing the loading time. The interface was improved in organisation, bugs were fixed and navigation was simplified. Tools that were perceived as "less useful" in the first set of user tests were placed in the background for final user tests. Further improvements that were noticed related to autosuggestion, spelling correction and the implementation of a narrow date filter (Table 5).

The aim of sections 2.45-2.47 is to address the question of whether the listed implementations have been effective in improving system performance from a user perspective. Insight on the overall effectiveness, efficiency and usability of KPro in year 4 is provided, how it changed in response to user feedback implementation and the status of the final prototype. Each section will address the following questions:

KPro overall feedback in year 4: The overall feedback on efficiency, effectiveness and usability, as well as tool preferences, search filters and search features in year 4. What differences are there across different audiences of physicians (.i.e. age groups, occupational groups)?

KPro improvement in year 4: Comparison of effectiveness, efficiency and usability of KPro across different rounds of user evaluation.

KPro: the final prototype: Where are we now in terms of usability, effectiveness and efficiency? What open-format user feedback has been made in year 4? Differences across occupational and age groups and which KPro prototype scores better in terms of usability, usefulness of tools and access? A detailed analysis of the last round of user tests in May-July 2014.

2.4.5 Effectiveness of Khresmoi Professional

2.4.5.1 Overall KPro effectiveness in year 4

2.4.5.1.1 Overall KPro effectiveness in year 4: All groups

Over the course of Y4 user evaluation, 38 out of 84 of the physicians testing the system were able to find the answer to a medical question they formulated themselves. A third of the participants complained about retrieving irrelevant results. Irrelevant results usually meant that the resources offered were about a different topic than what was searched for. An additional 22% remarked that retrieved results were either too general or too specific to what was expected. A "too general" result typically meant that the physician expected more concrete or higher quality/professional content. A result was typically classified as "too specific" if the physician expected an overview articles and received scientific articles focusing on specific audience (Figure 21).

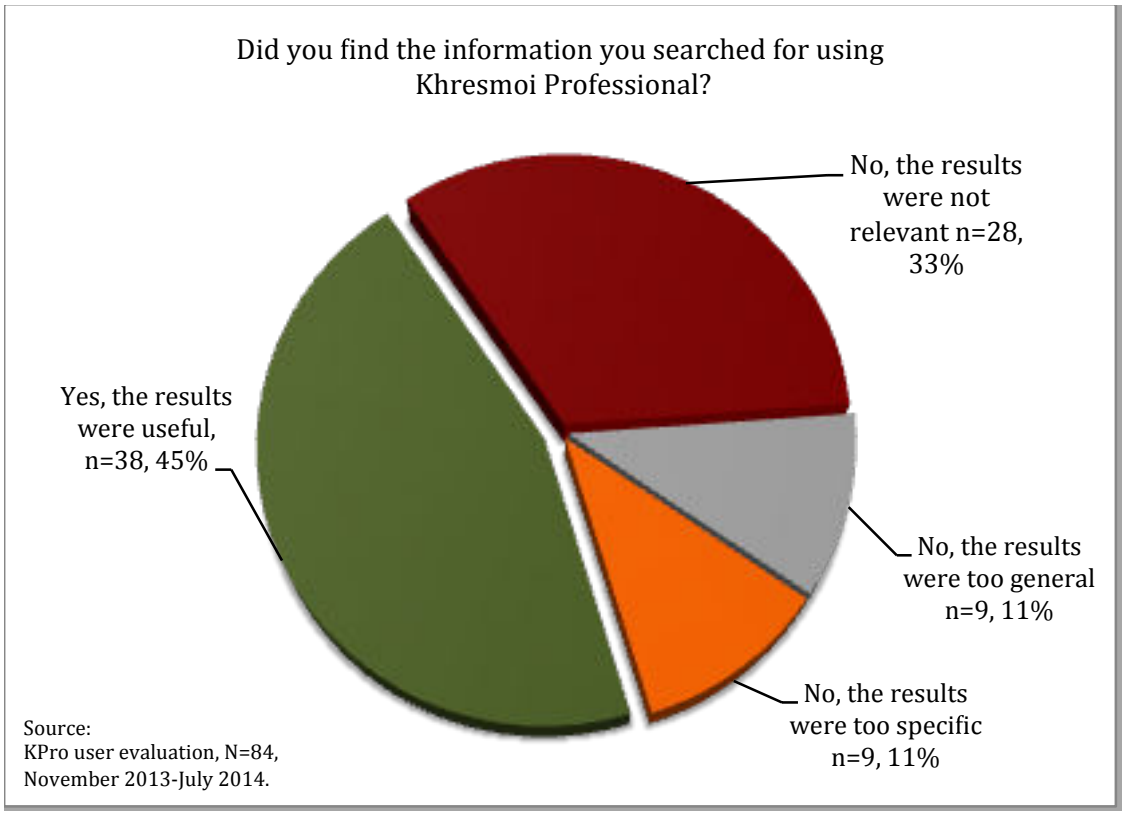


Figure 21 KPro effectiveness in Y4.

2.4.5.1.2 Overall KPro effectiveness in year 4: A comparison across different occupational groups

Research physicians and hospital clinicians were most likely and self-employed general practitioners least likely to find the information they looked for using the KPro search system. Self-employed practitioners (specialist and GP) were more likely than other groups to find irrelevant information. Every fifth general practitioner found the search results too specific (Figure 22).

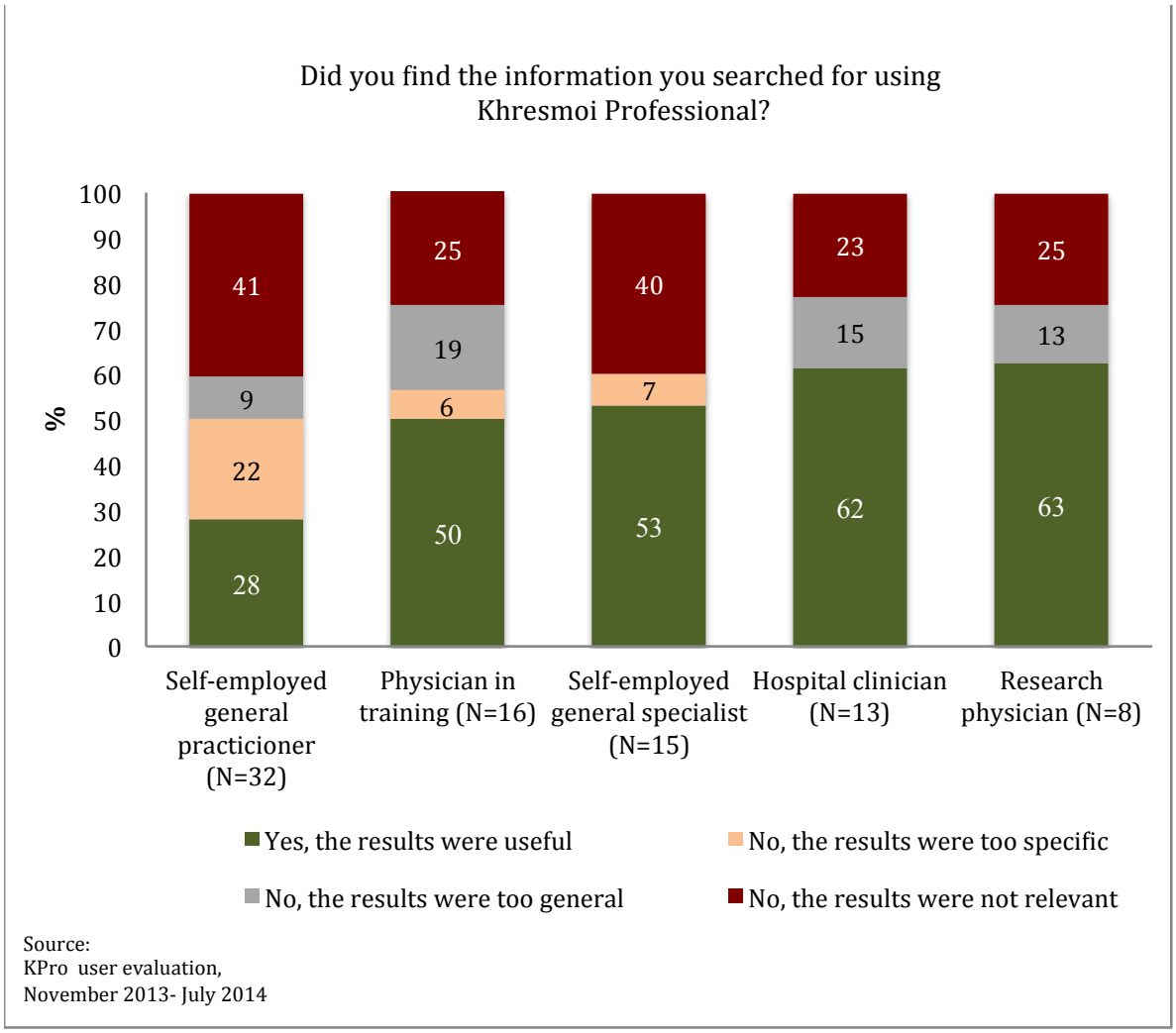


Figure 22 KPro effectiveness in Y4 by occupational group.

2.4.5.1.3 Overall KPro effectiveness in year 4: A comparison across different age groups

More than half of the physicians below 40 years were able to find the requested information using KPro. Physicians between 51-60 years were least likely to be successful in their search. Almost half of the physicians older than 60 years could find useful information. Overall, occupational group appeared to be a higher determinant of effectiveness than age. However, it appears that older, more experienced physicians were more likely to regard found information as irrelevant and somewhat less likely to regard it as "too general". However, the effect does not have significant foundation and requires further exploration (Figure 23).

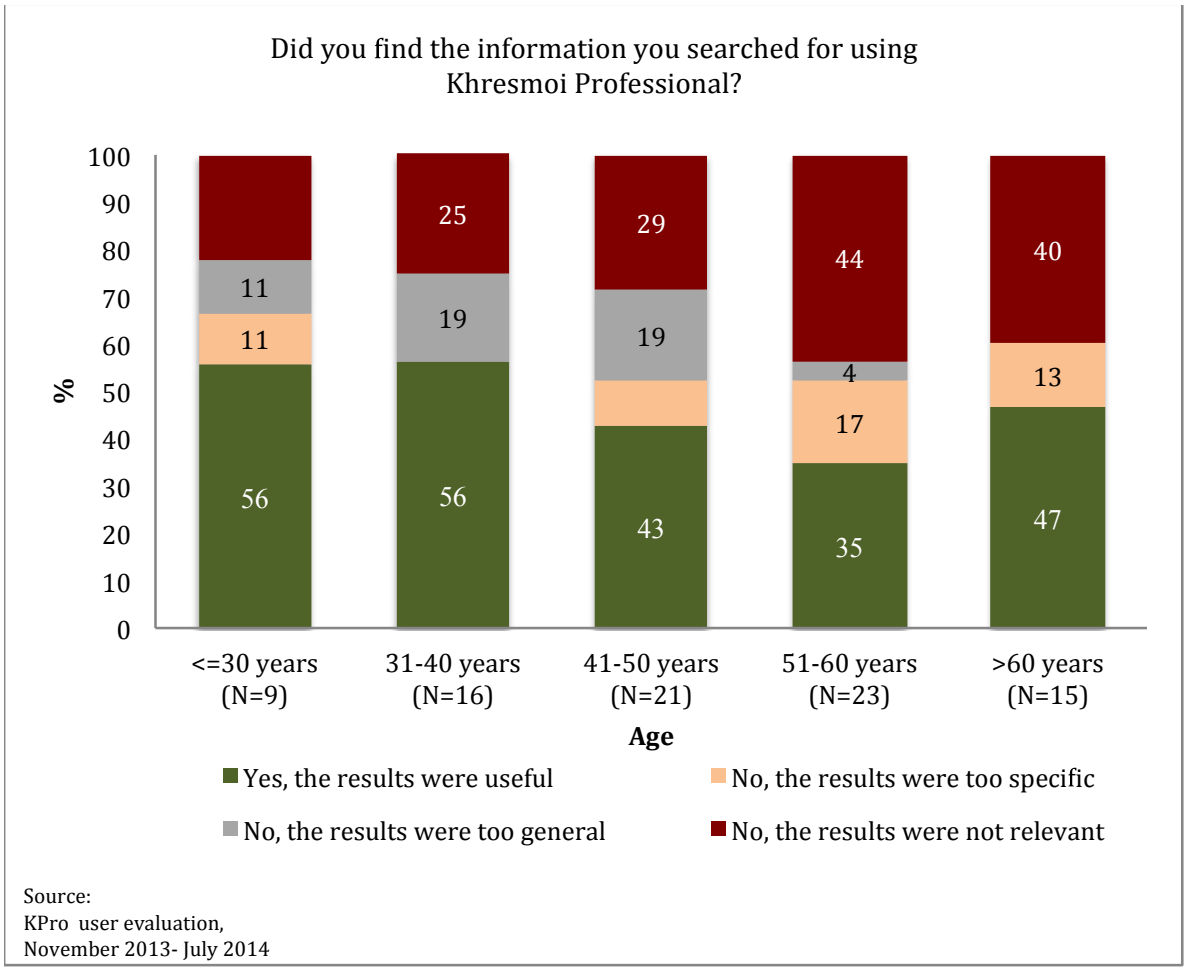


Figure 23 KPro effectiveness in Y4 by age

2.4.5.2 KPro Effectiveness improvement in year 4: A comparison across different rounds of user evaluation.

2.4.5.2.1 KPro effectiveness improvement in year 4: All groups

In the first round of user tests the most common reason why physicians did not find answers was that the offered content was not relevant to their query or too specific/scientific. The prototype was weak in providing relevant content for complex queries. Queries were often not recognized as one entity leading to irrelevant results. This was illustrated by the results obtained for the pre-defined tasks. While most of the participants (6/11) were able to solve the information-gathering personal library task, only 3/12 users were able to find the relevant answer to a task encompassing a concrete question (e.g. side effects of estradiol). Furthermore, specific/scientific articles were ranked first in cases where physicians wanted general/overview information.

As illustrated in Figure 24, substantial improvements in relevance of search results were achieved during the final year of Khresmoi Professional development. Twice as many physicians were able to find relevant answers in the second user evaluation as compared to the first round of tests. Improvements in ranking addressed the issue of information being too specific. The biggest improvements were in relevance of displayed search results and results were less likely to be classified as "too specific" than in the first set of user tests.

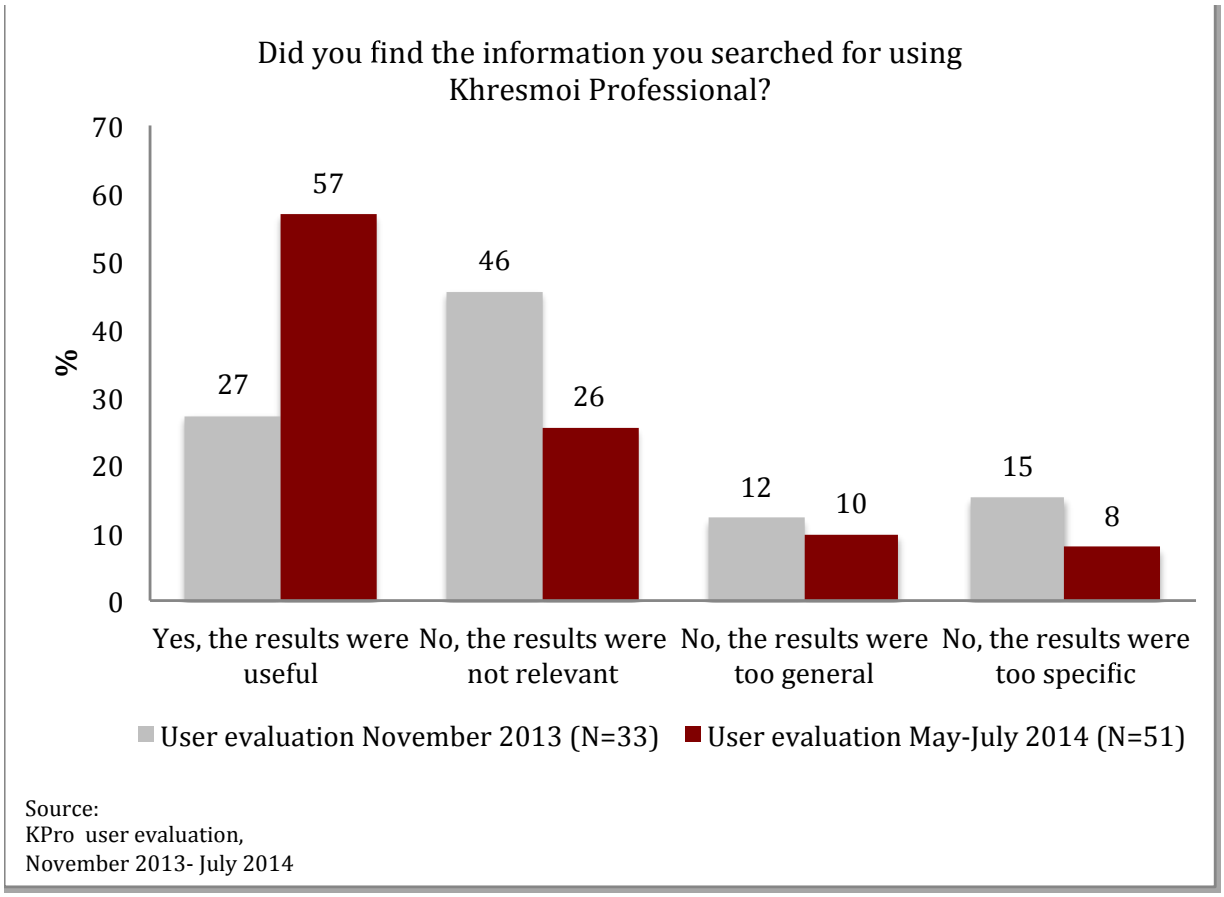


Figure 24 KPro effectiveness improvement in Y4.

2.4.5.2.2 KPro effectiveness improvement in year 4: A comparison across different occupational groups

Increase in KPro effectiveness across different occupational groups

Effectiveness increased across all occupational groups that were evaluated in both rounds of user tests. In the first set of user tests 21% and in the final tests 38% of general practitioners were able to find what they searched for. The most dramatic increase was observed for the physicians in training. (+38% regarded Khresmoi as effective). This may have been due to the improvements in accessibility and inclusion of more overview resources such as Wikipedia (Figure 25).

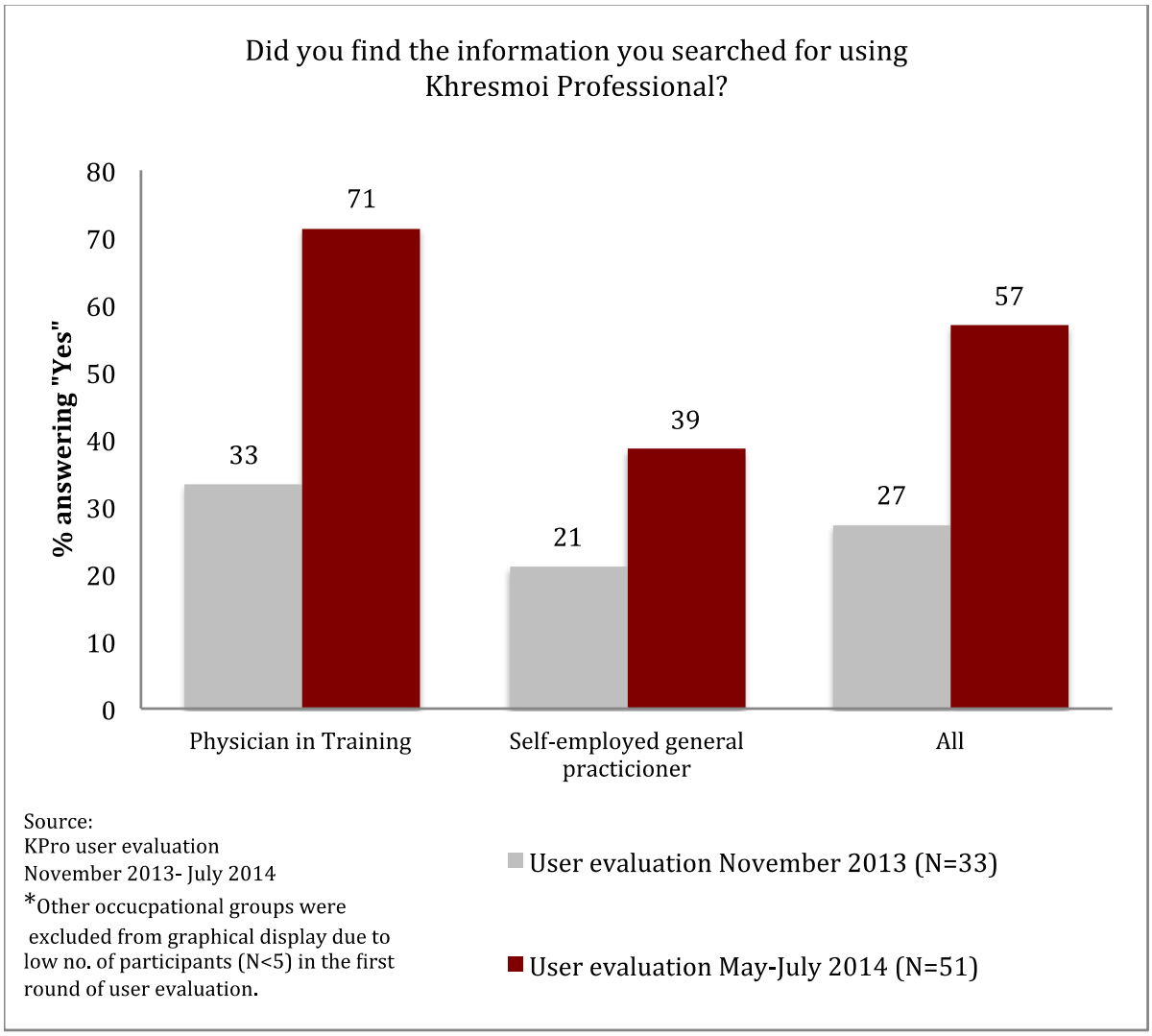


Figure 25 KPro effectiveness improvement in year 4 by occupational group.

2.4.5.3 KPro: the final prototype. May-July 2014 Khresmoi Professional user evaluation

2.4.5.3.1 KPro: Effectiveness of the final prototype: All groups

2.4.5.3.2 Quantitative feedback

As illustrated in Figure 26 the majority of physicians evaluating the final search were able to find what they searched for using Khresmoi Professional. A quarter had problems finding information relevant to their query; about one in five physicians regarded the information as too specific or too general. In the first round of user tests "lack of relevance" was typically experienced due to the system failing to recognise a complex term as one entity. The problems before the onset of the final user tests and relevance issues that remained were largely an issue of lack of, or the wrong type of resources being displayed.

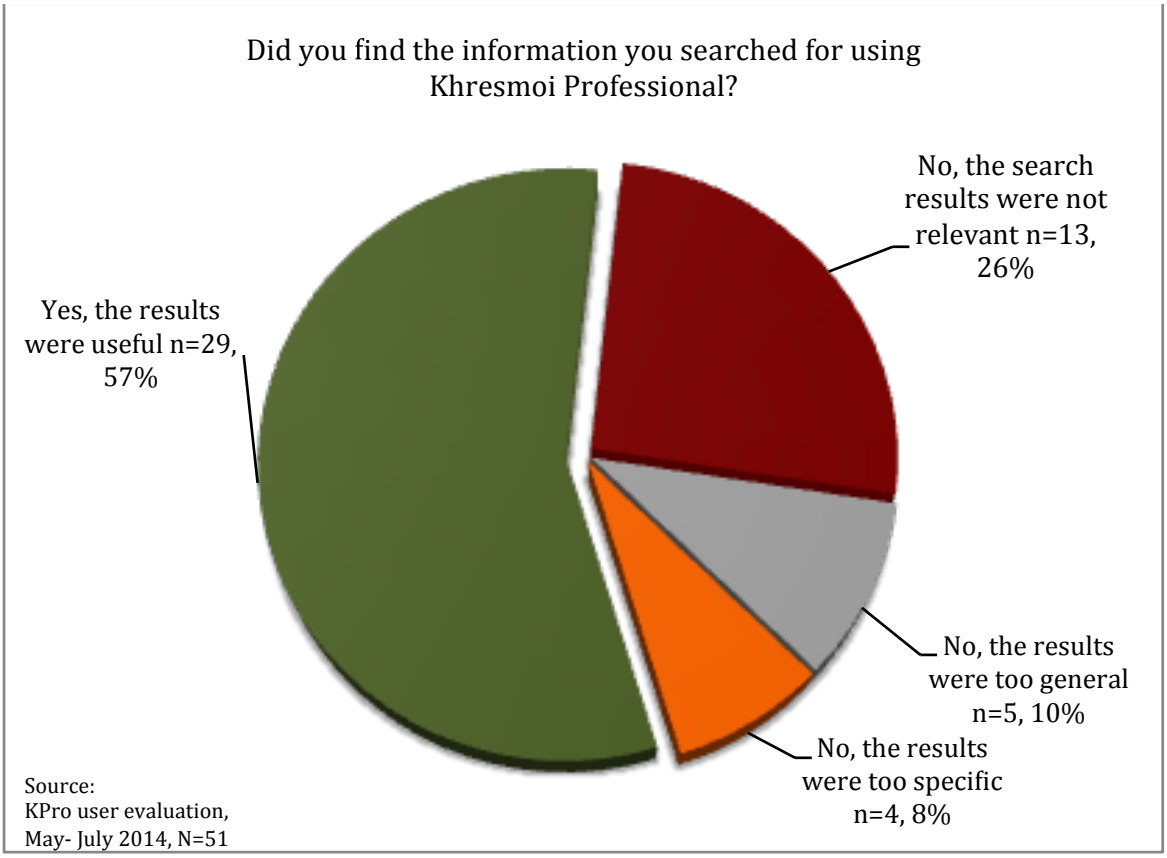


Figure 26 KPro effectiveness of the final prototype.

2.4.5.3.3 Qualitative feedback

Table 6 illustrates some original comments on effectiveness made by users. Further original feedback can be viewed in Appendix 5.1, Table A14. A few users explicitly commented on the good quality of results they obtained. One user even remarked that search results were better than Google because of allowing him to navigate amongst physician resources. The most common problem mentioned was a lack of search results for the query made. Nine users missed concrete type of content such as reviews or guidelines on the topic they searched. An additional eight physicians complained about a lack of German content and five users missed good images for their query. Information was typically regarded as too specific in cases where users expected summary information instead of primary literature on the topic. On the other hand a "too general" complaint usually reflected results that lacked professional content for specific queries.

May-July 2014 user evaluation	
KPro Effectivness	**Examples of comments (no. of users)**
The results were useful and relevant	**Good results (2)** *„Very good results. Compared it to Google which wouldn't recognize the abbreviation so quick since it confuses it with non-medical information. A little bit too specific but useful"*
Results not relevant	**Irrelevant results (5)** *„ Irrelevant results"*, „ Not one relevant website" **Overall lack of content (3)** *„ Resources are somewhat limited"*, *„ The selection is limited and „I am missing reviews and short articles which have clinical relevance."* **Lack of German content (8)** *„ More German resources would be great. The only German source on the topic was for lay people."* *„ Primarily English articles appeared"*, *„ Lack of German articles"* *„ not a lot of search results for the German query septische Arthritis"* **Lack of specific type of content: (9)** *„Lack of guidelines"*, *„ Relevant drug information (Austria Codex) was missed"* , *„ I am missing reviews and short articles which have clinical relevance."* **Lack of images (5)** *"Missing images to dermatological queries"*
Results too general	**Results too general (1)** *„ No relevant information to specific questions"*
Results too specific	**Results too specific (4)** *„Results a bit too specific for a general query. Would have liked more overview articles."* , *„Results are a bit too scientific."*

Table 6: Comments made on KPro effectiveness in the final user evaluation in May-July 2014.

2.4.5.3.4 KPro: Effectiveness of the final prototype: A comparison across different age groups,

Physicians younger than 40 years were most likely to find the information they search for using Khresmoi Professional. Physicians between 51 and 60 were least likely to find relevant results. An interesting finding was none of the physicians aged between 31-50 years perceived information as too specific (Figure 27).

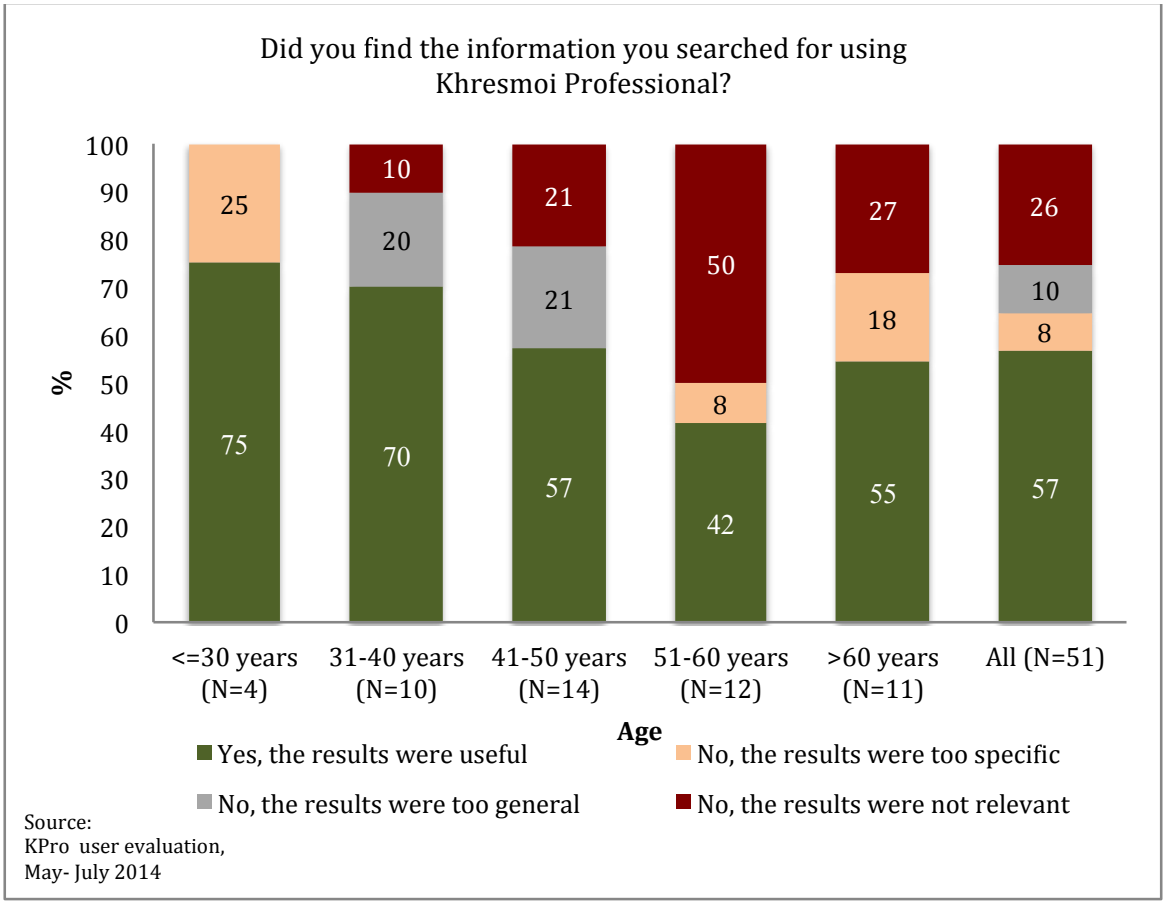


Figure 27 KPro effectiveness of the final prototype by age.

2.4.5.3.5 KPro: Effectiveness of the final prototype: A comparison across different occupational groups

Physicians in training and hospital clinicians were most likely to find the answers to their questions using KPro. Self-employed specialists and general practitioners were most likely to fail at finding relevant information. General practitioners and self-employed physicians were the only group perceiving results as too specific (Figure 28).

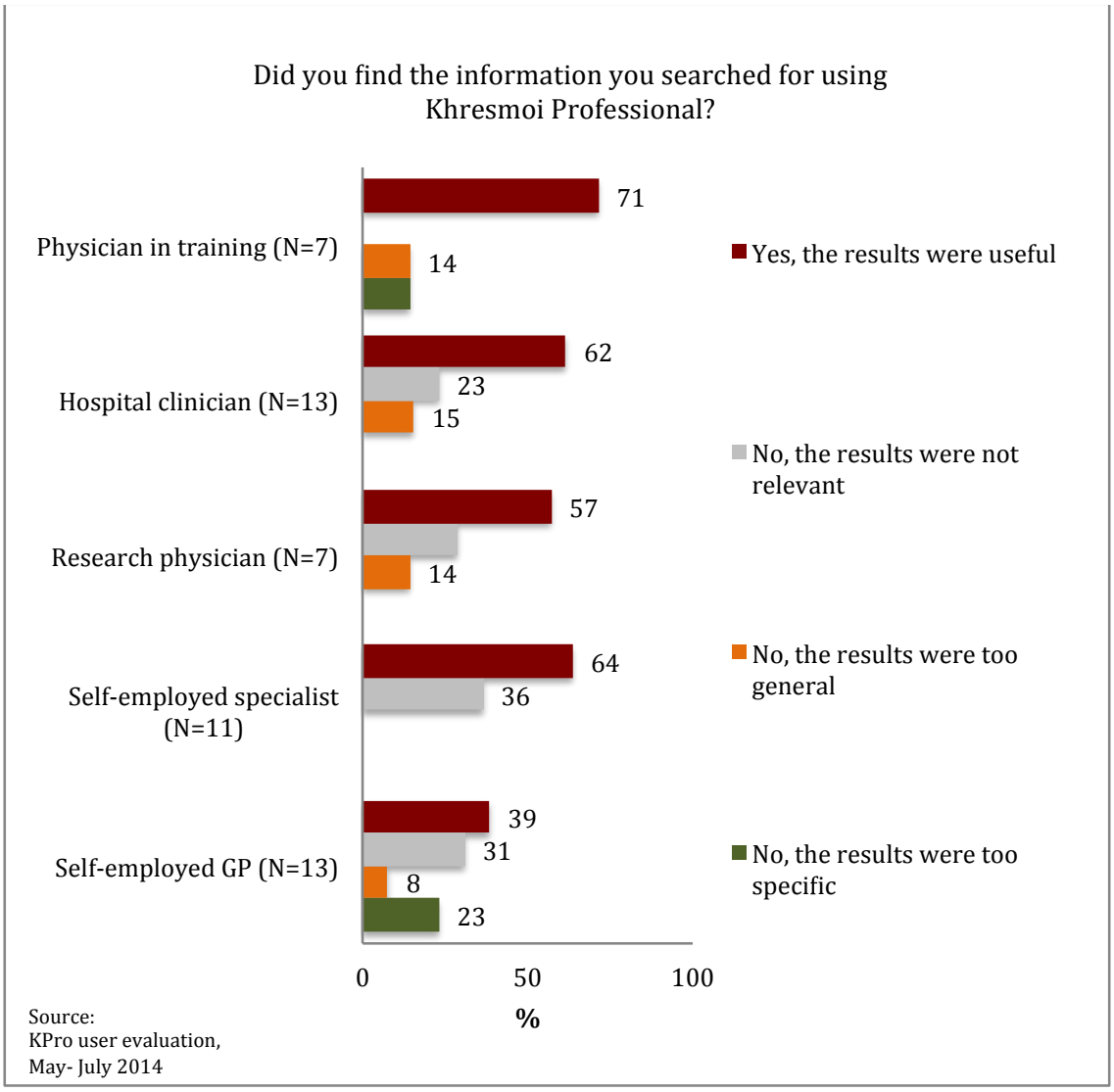


Figure 28 KPro effectiveness of the final prototype by occupational group.

2.4.6 Efficiency of Khresmoi Professional

2.4.6.1 Overall KPro efficiency in year 4

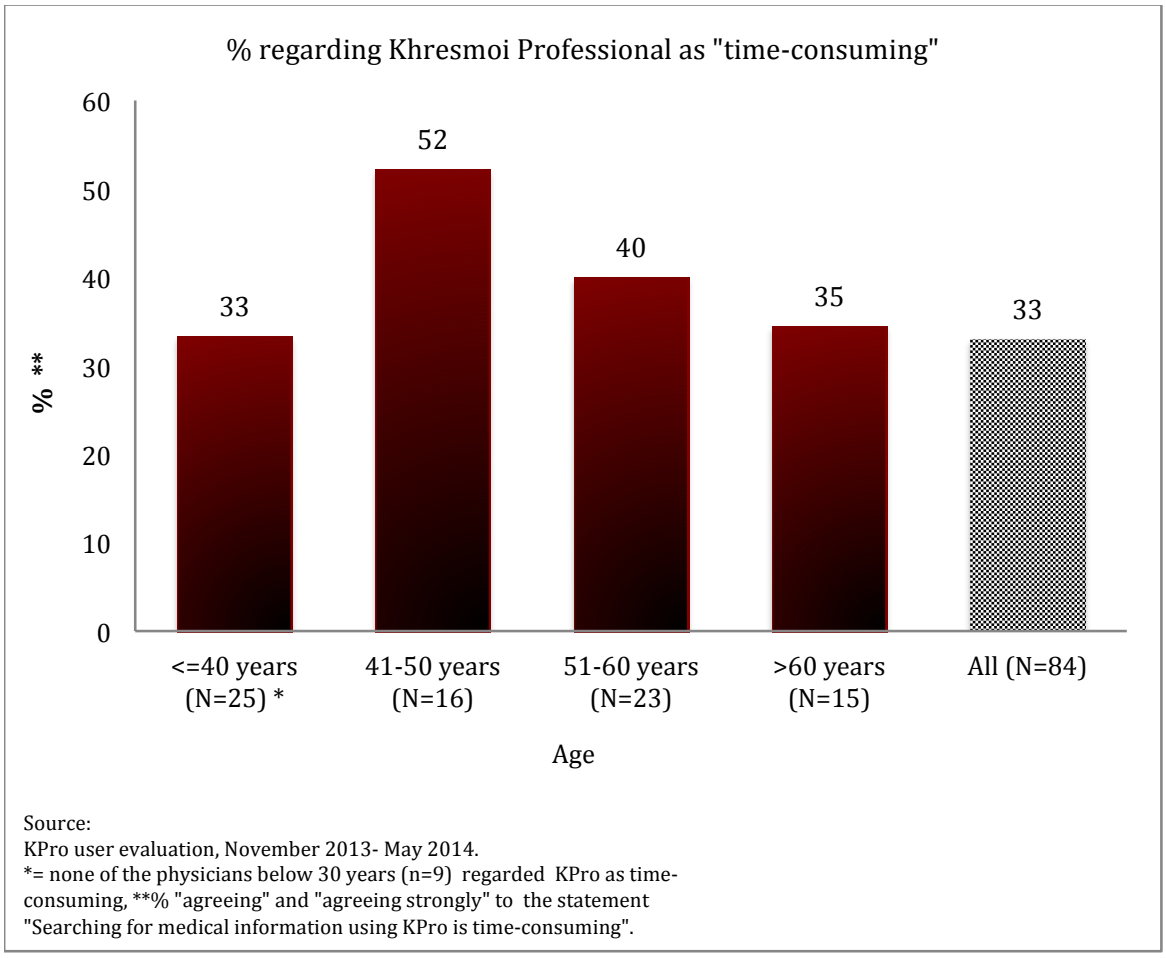
2.4.6.1.1 Overall KPro Efficiency in year 4: All groups

KPro efficiency was measured by the extent to which physicians agreed to the statement that using the system is "time-consuming". Overall, two thirds regarded Khresmoi Professional as efficient (i.e. not time-consuming) and one third as « time consuming ».

2.4.6.1.2 Overall KPro efficiency in year 4: A comparison across different age groups

An interesting finding was that none of the physicians younger than 30 years, perceived Khresmoi Professional as time-consuming. A possible explanation could be that very young physicians "who grew up with IT" get used to a new system more quickly than older users. Every second physician aged 41-50 regarded the system as time- consuming. Presumably, 41-50 year old physicians are the

most time-constrained group and therefore have more sensitive thresholds for efficiency. Only a third of the physicians over 60 years regarded Khresmoi as time-consuming (Figure 29).

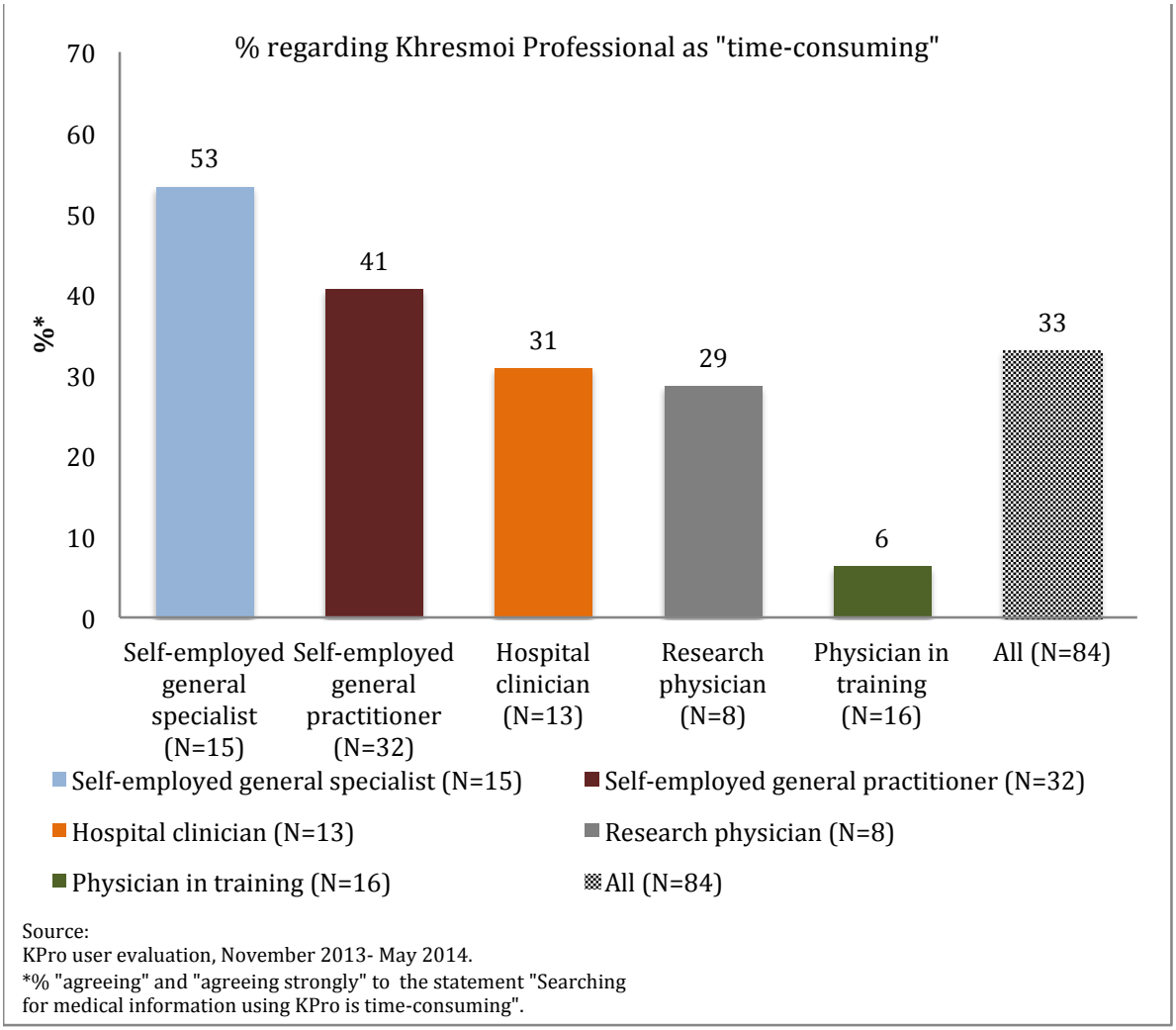


Source:
KPro user evaluation, November 2013- May 2014.
*= none of the physicians below 30 years (n=9) regarded KPro as time-consuming, **% "agreeing" and "agreeing strongly" to the statement "Searching for medical information using KPro is time-consuming".

Figure 29 KPro efficiency in Y4 by age.

2.4.6.1.3 Overall KPro efficiency in year 4: A comparison across different occupational groups

Self-employed practitioners were most likely to perceive the system as time consuming. One reason self-employed practitioners were most likely to be confronted with irrelevant search results. In addition self-employed practitioners tend to be the most time-constrained group, which may lead to a more sensitive efficiency threshold [7]. In contrast only 6% of the physicians in training perceived Khresmoi as time consuming. For the younger physicians age may have confounding impact. However, for physicians over 40 occupational statuses appears to determine whether a physician regards the system as time-consuming (Figure 30).



Source:
KPro user evaluation, November 2013- May 2014.
*% "agreeing" and "agreeing strongly" to the statement "Searching for medical information using KPro is time-consuming".

Figure 30 KPro efficiency in Y4 by occupational group.

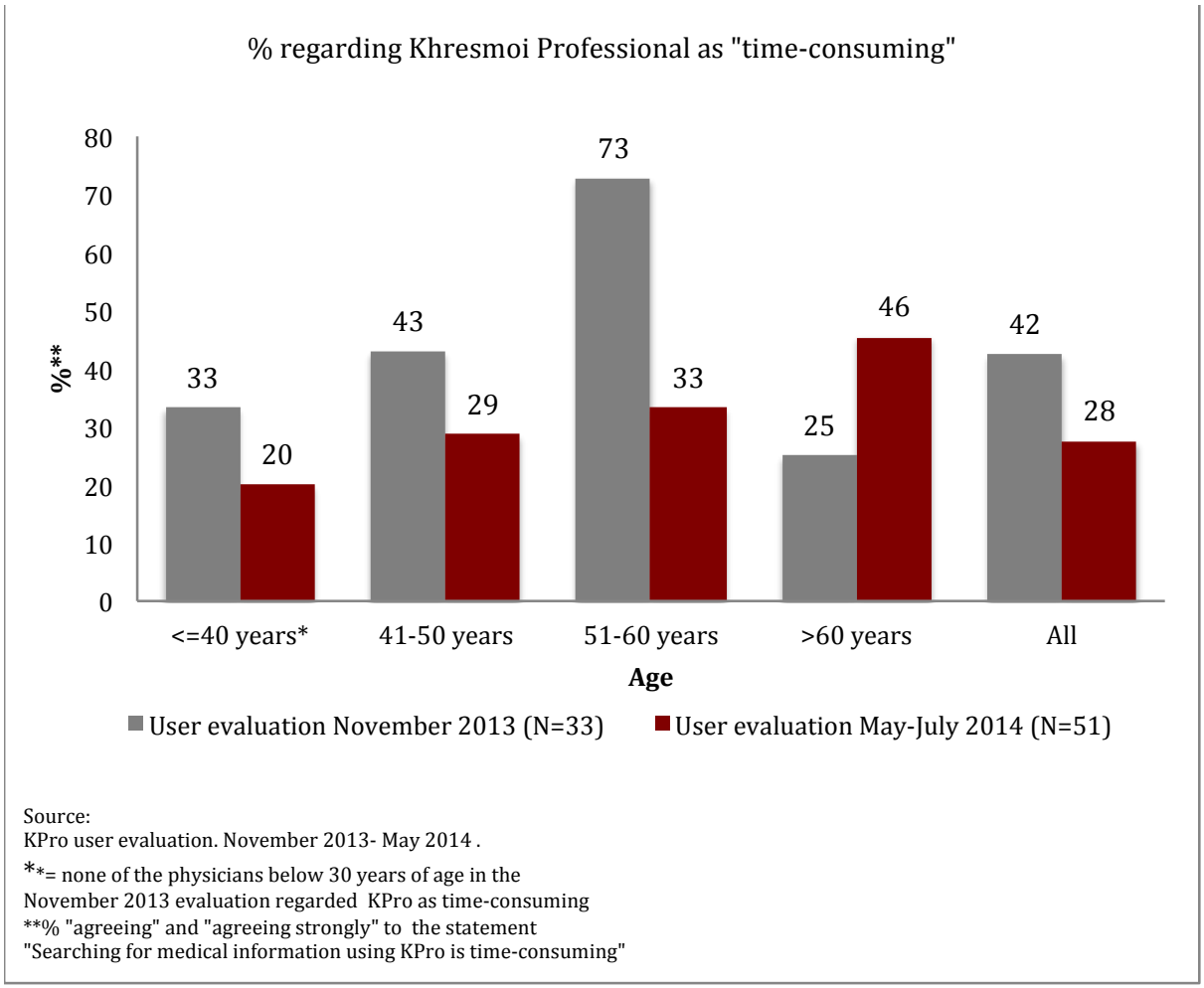
2.4.6.2 KPro efficiency improvement in year 4: A comparison across different rounds of user evaluation.

2.4.6.2.1 KPro efficiency improvement in year 4: A comparison across different age groups

In comparison to the November 2013 evaluation KPro efficiency improved substantially. In the first round of user tests 42% and in the final round only 28% of the users regarded the system as time-consuming. Many efficiency issues experienced in the November 2013 such as extended system loading time and inefficient ranking of results were solved or improved before the start of the final user tests. Detailed original user feedback on efficiency can be viewed in Appendix 5.1.

2.4.6.2.2 KPro efficiency improvement in year 4: A comparison across different age groups

Across different age groups the system was perceived as more efficient across all age groups except the over 60 year olds. For the age group 51-60 years the biggest improvement in system efficiency was found.



Source:
KPro user evaluation. November 2013- May 2014 .

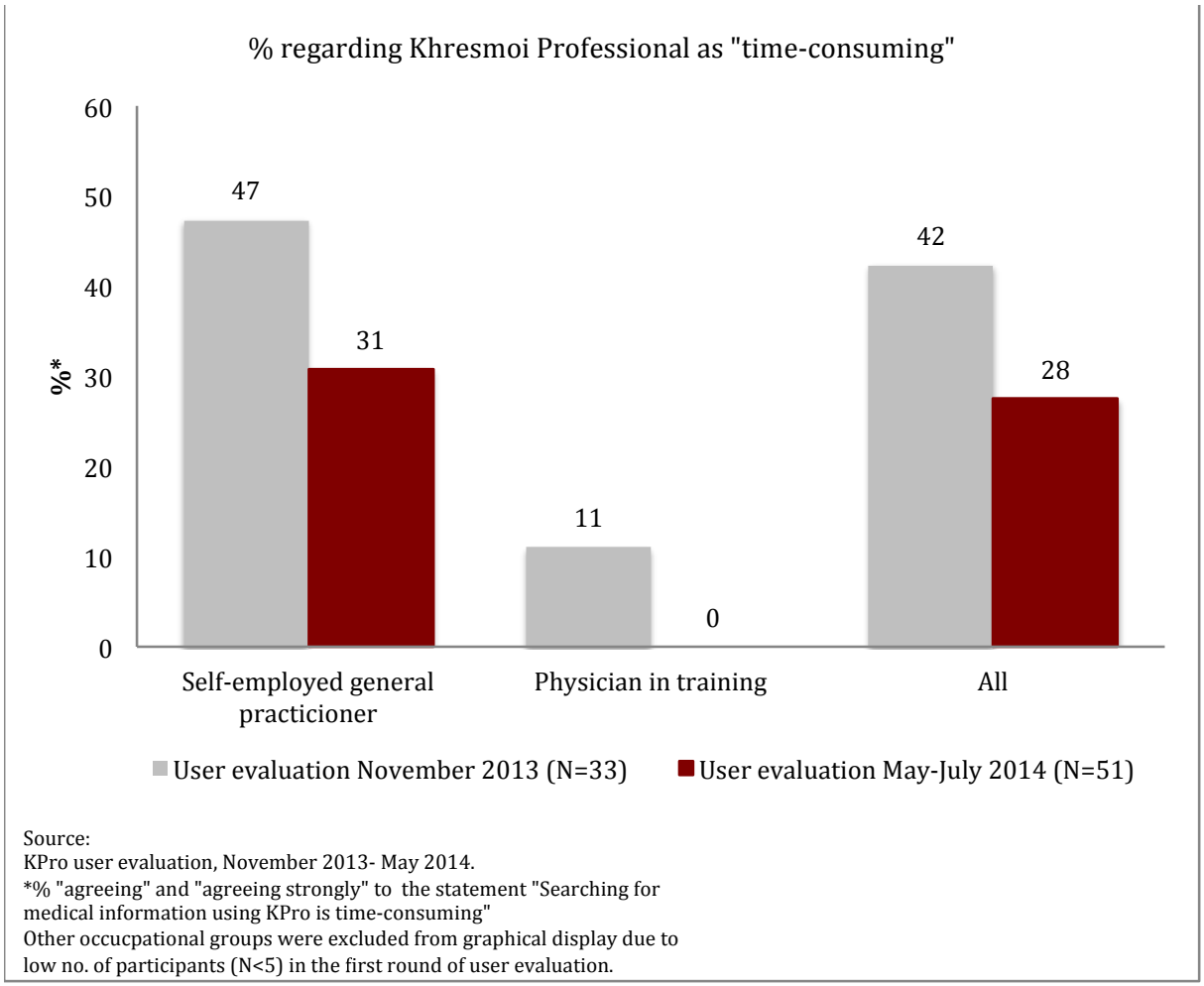
*= none of the physicians below 30 years of age in the November 2013 evaluation regarded KPro as time-consuming

**% "agreeing" and "agreeing strongly" to the statement "Searching for medical information using KPro is time-consuming"

Figure 31 KPro efficiency improvement in Y4 by age

2.4.6.2.3 KPro efficiency improvement in year 4: A comparison across different occupational groups

For subgroups, general practitioners as well as physicians in training, the issue of efficiency was smaller in the final user tests than in the November 2013 evaluation.



Source:
KPro user evaluation, November 2013- May 2014.
*% "agreeing" and "agreeing strongly" to the statement "Searching for medical information using KPro is time-consuming"
Other occucpational groups were excluded from graphical display due to low no. of participants (N<5) in the first round of user evaluation.

Figure 32 KPro efficiency improvement in Y4 by occupational group.

2.4.6.2.4 Overall KPro efficiency of the final prototype

In the final user evaluation, 72% of the physicians evaluating KPro system regarded the system as efficient and 28% as « time consuming ». As illustrated in Table 7 comments of efficiency in the May-July 2014 evaluation, usually related to users not finding relevant articles or being overwhelmed by the complexity of the system.

May-July 2014 user evaluation	
Issue raised	**Examples of comments made in relation to KPro efficiency.**
Search is time-consuming	"*Search is a bit time-consuming because it is hard to find the result*" „*For the quick search at point-of-care a little bit too time consuming. (Due to lack of spelling correction)*" „ *Relevant articles came at the end*"

Table 7: Comments made on KPro efficiency in the final user evaluation in May-July 2014.

2.4.6.2.5 KPro efficiency of the final prototype: A comparison across different age groups

In the final user evaluation, older age predicted a higher likelihood that KPro was perceived as inefficient. The efficiency issues, related to ranking and loading time, raised by 51-60 years appeared to have been solved. However, some usability issues overwhelming older physicians remained.

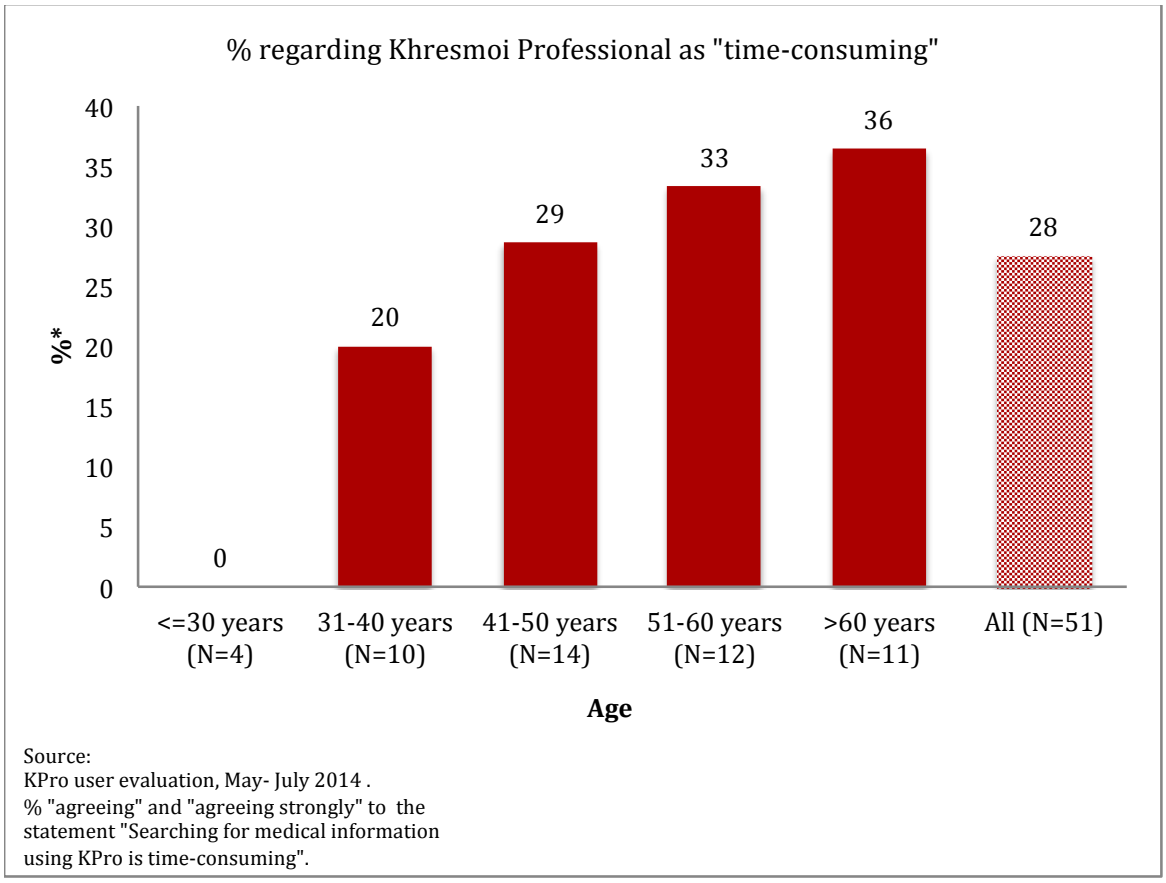


Figure 33 KPro efficiency of the final prototype by age.

2.4.6.2.6 KPro efficiency of the final prototype: A comparison across different occupational groups

In line with the overall trend, self-employed practitioners were most likely to perceive the system as time consuming. However the difference across different groups was not significant (except for physicians in training). In terms of efficiency, age appears to have had a higher impact in the final user evaluation than occupational status (Figure 34).

2.4.6.2.7 KPro the final prototype: May-July 2014 user evaluation

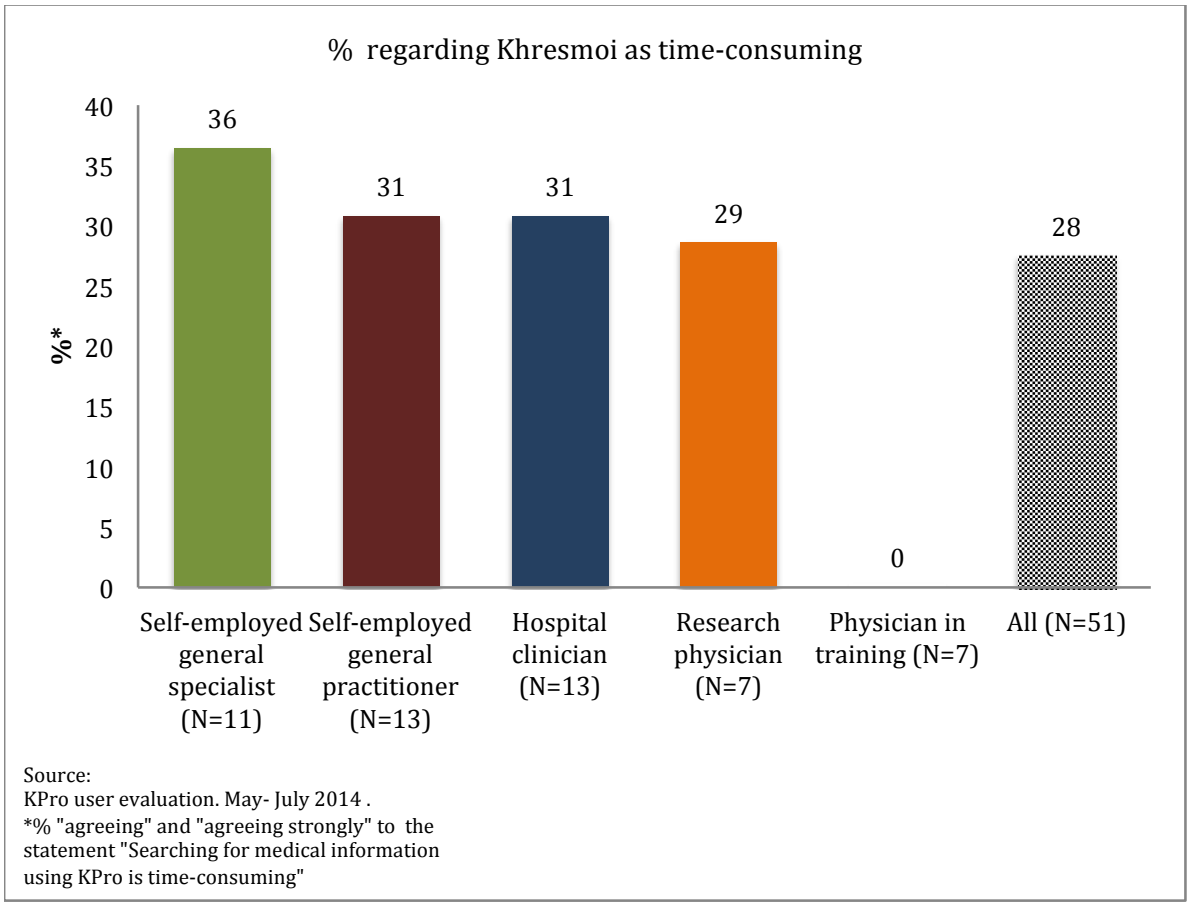


Figure 34 KPro efficiency of the final prototype by occupational group.

2.4.7 Usability of Khresmoi Professional

2.4.7.1.1 Overall KPro usability in Y4: All groups

Usability was determined using the 10-item Standard Usability Scale (SUS). A global usability score was calculated to determine overall usability. Further details on calculation procedure are described in 2.3.3.1.5. For the individual response analysis the level of agreement to the 5 positive and the 5 negative statements about the system was assessed. For graphical display of individual items, answers to the 5-point Likert scale were collapsed into disagree (scale items 1 and 2), neutral (scale item 3) and agree (scale item 4 and 5).

2.4.7.1.2 Global usability score

Across all user evaluations performed from November 2013- July 2014 the mean overall global usability score was 67 (N=84), which is just under what is considered as average usability of a search system [4]. However scores ranged from 38-93, indicating the substantial differences in global

usability that were reported across different age groups, occupational groups and across different points of user evaluation.

2.4.7.1.3 Individual SUS item responses

Figure 35 illustrates the level of agreement to positive statements about KPro among all physicians taking part in the year 4 user evaluation. The system scored best in terms of it being intuitive in its handling. More than half of the users reported feeling confident in using the system, perceiving it as simple to use and imagining that people get used to it quickly.

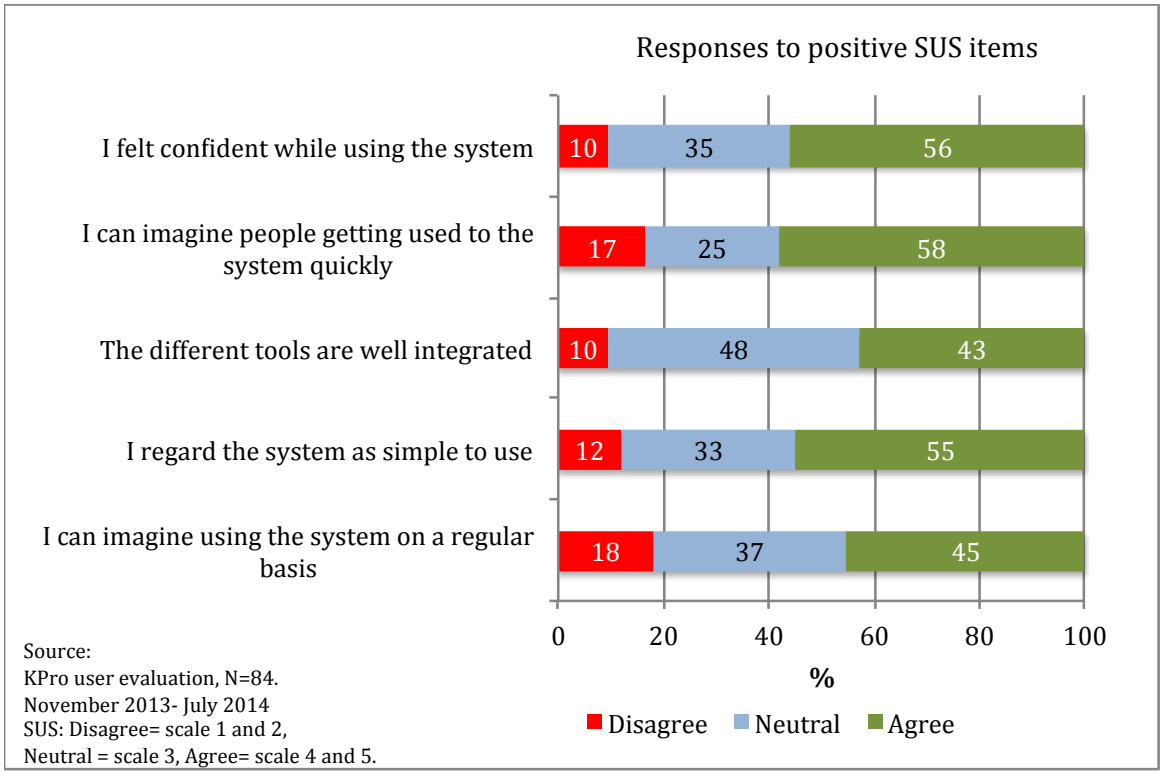


Figure 35 KPro usability in Y4: Positive SUS items.

With regard to negative statements about the system, inconsistencies in the system posed the biggest problem. (13% agreeing). That the system was intuitive to use was confirmed with the finding that 79% of users disagreed with the statement that they had to learn a lot before working with the system (Figure 36).

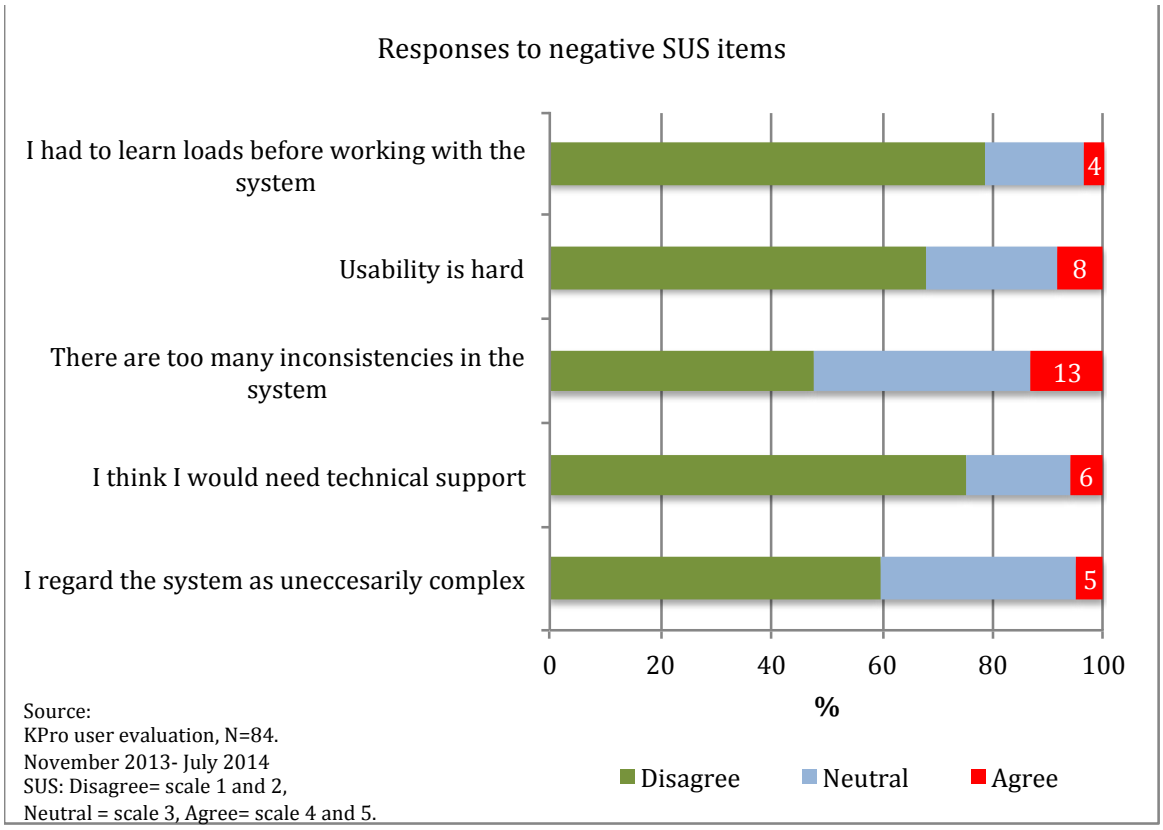


Figure 36 KPro usability in Y4: Negative SUS items.

2.4.7.1.4 Overall KPro usability: A comparison across different age groups

2.4.7.1.4.1 Global usability scores

There was a strong association between age and global usability. As shown in Figure 37, global usability was inversely associated in magnitude to the age of physicians. Younger physicians showed, across all user tests, higher levels of usability than older physicians. When age was controlled, physicians working in hospital or academic settings showed higher levels of usability. Younger physicians got used to the system more quickly and adapted better to new features.

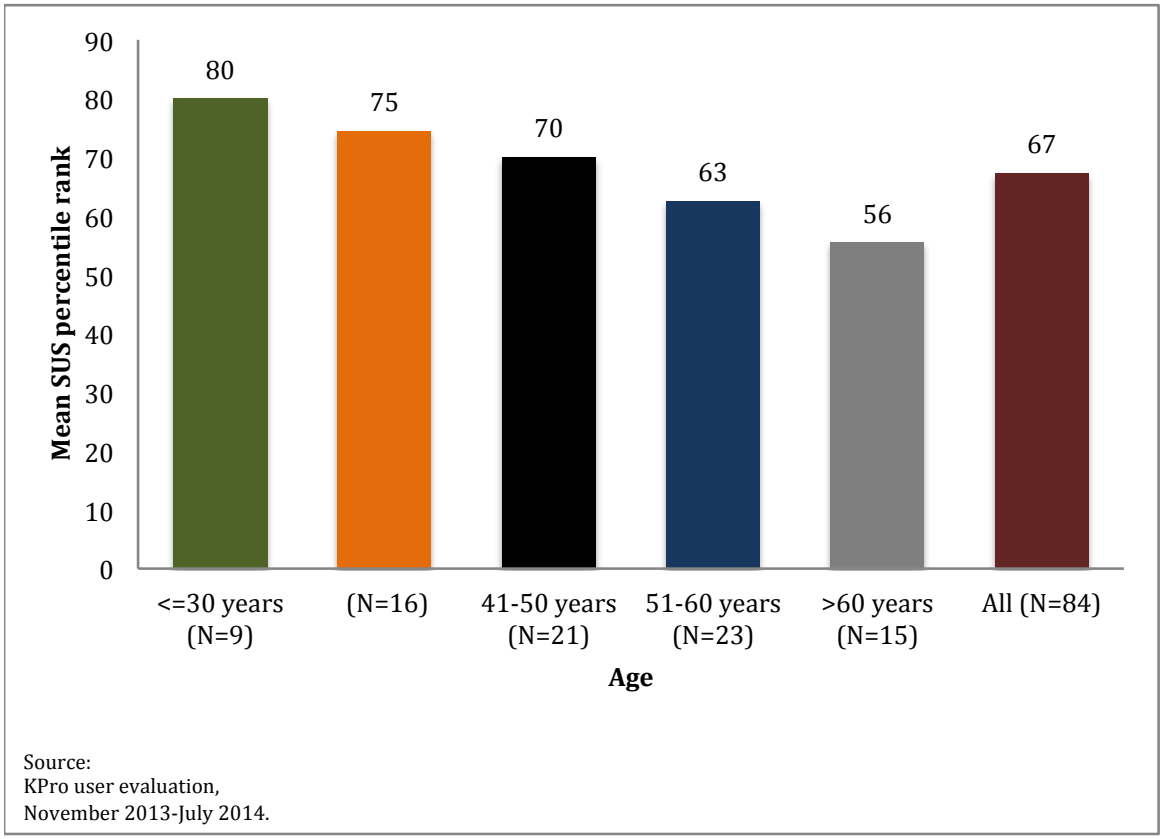
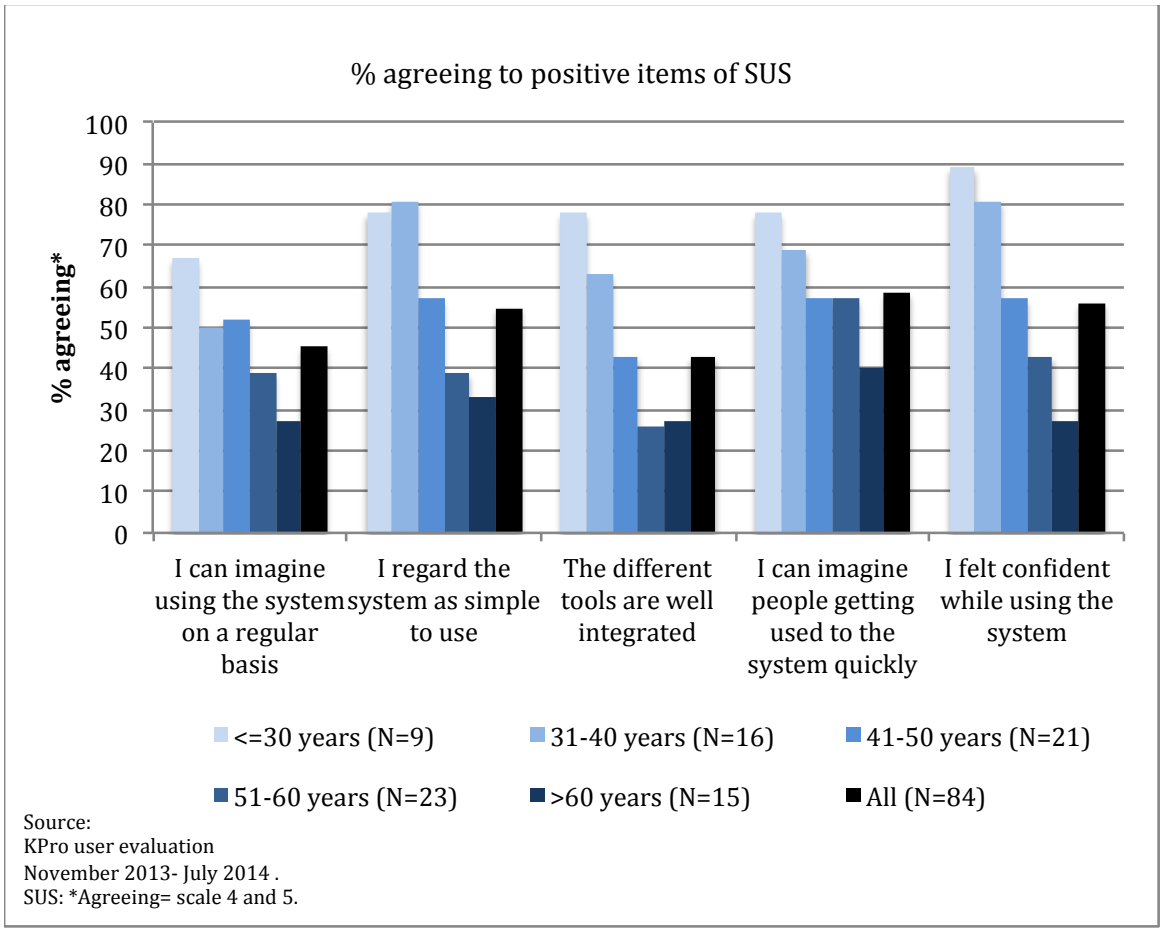


Figure 37 KPro global usability in Y4 by age.

2.4.7.1.4.2 Individual SUS item responses

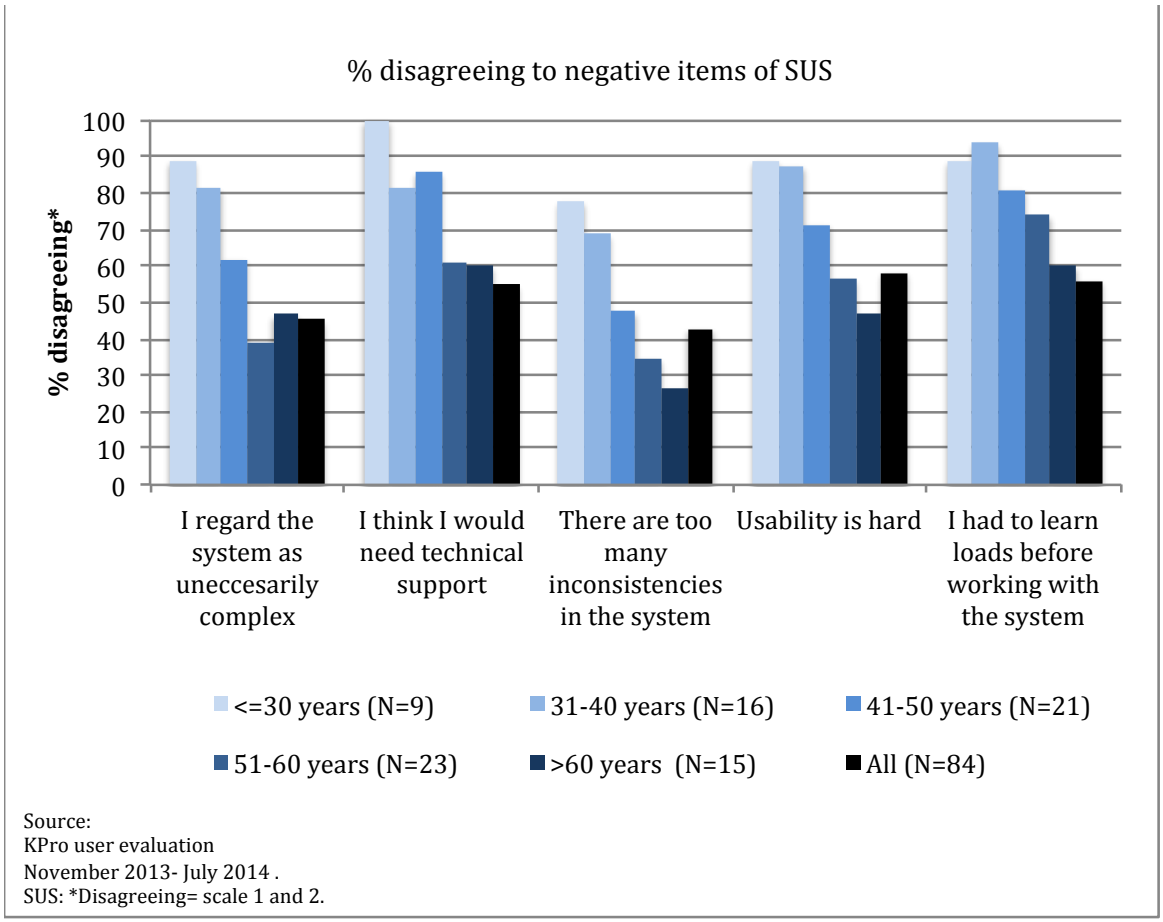
Age appeared to have a strong impact on overall usability across all SUS items. The steepest decrease in usability with age was found for the item "I felt confident in using the system". 89% of the physicians younger than 30 felt years vs. only 27 % of physicians older than 60 years, agreed with feeling confident with using the system. The flattest curve was observed for items "I can imagine using the system on a regular basis" and "I can imagine getting used to the system quickly". Thus, it appears that although younger physicians are more adept in learning the new system, the interest in the system and willingness to learn is more likely to appear across all age groups (Figure 38).



Source:
KPro user evaluation
November 2013- July 2014 .
SUS: *Agreeing= scale 4 and 5.

Figure 38 KPro usability in Y4 by age: Positive SUS items.

With regard to negative statements on the system, a similar trend was observed. Physicians younger than 40 were less likely to regard the system as complex, or be overwhelmed by inconsistencies or to need technical support (Figure 39).



Source:
KPro user evaluation
November 2013- July 2014 .
SUS: *Disagreeing= scale 1 and 2.

Figure 39 KPro usability in Y4 by age: Negative SUS items.

2.4.7.1.5 Overall KPro usability: A comparison across different occupational groups

2.4.7.1.5.1 Global usability scores

Across different occupational groups, mean global usability scores were highest for physicians in training and hospital clinicians. For both groups average global usability ranks were above average. Mean global usability scores were lowest for general practitioners (Figure 40).

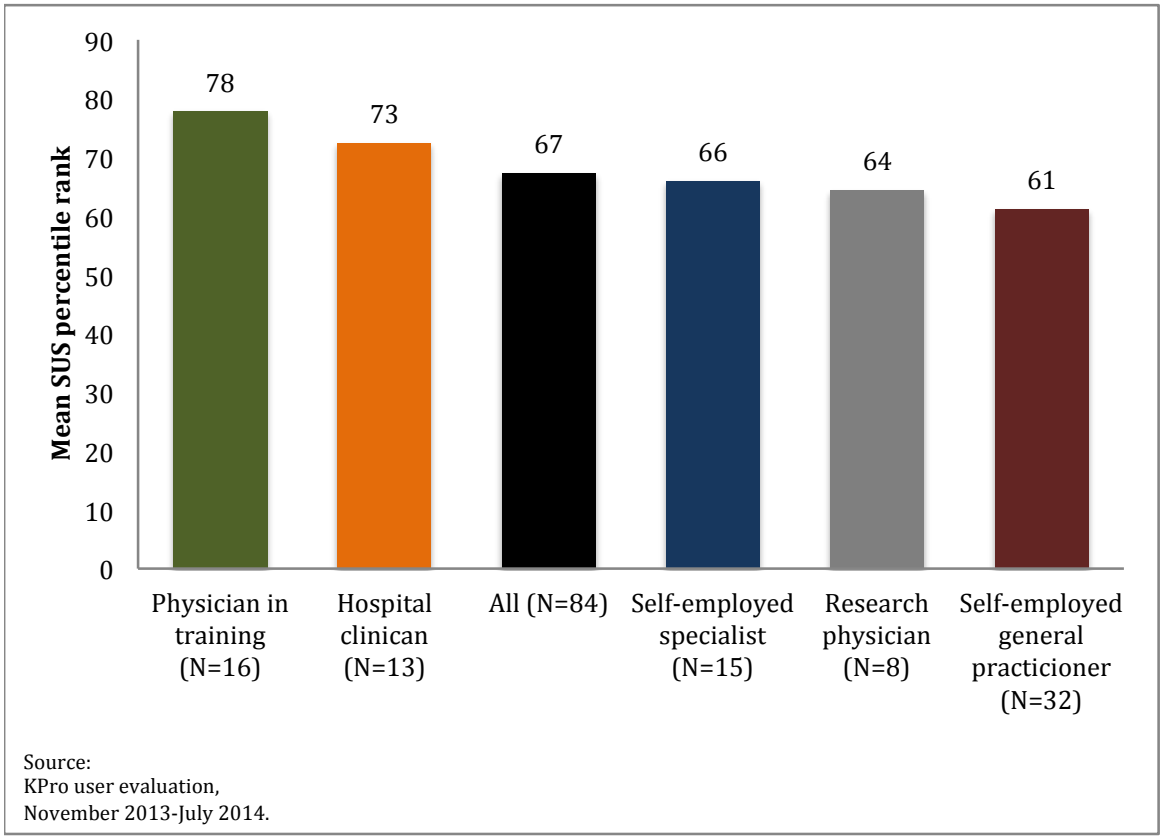


Figure 40 KPro global usability in Y4 by occupational group.

2.4.7.1.5.2 Individual SUS item responses

Physicians in training scored were the most likely group to agree to positive statements about the system. Almost two thirds of hospital physicians reported that they can imagine using the system on a regular basis. However, since all data from hospital physicians was collected in the last round of user evaluation, a sample bias is possible. While most general practitioners appeared overwhelmed by the way tools were integrated, every second GP was confident that people would get used to the system quickly (Figure 41).

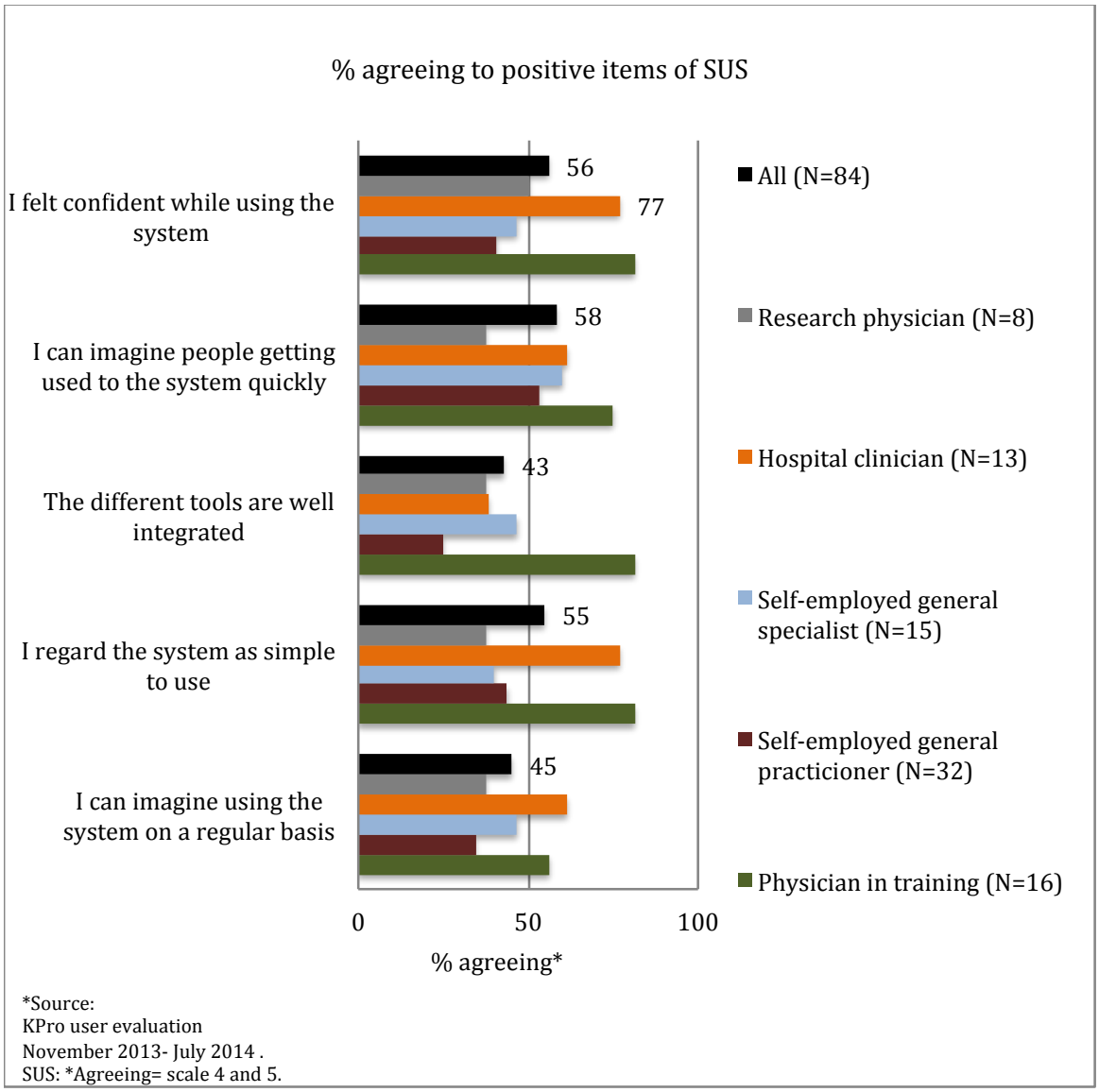


Figure 41 KPro usability in Y4 by occupational group: Positive SUS items.

A similar pattern was observed when analysing the % of users disagreeing to negative items. Hospital clinicians scored highest and general practitioners lowest in terms of usability (Figure 42).

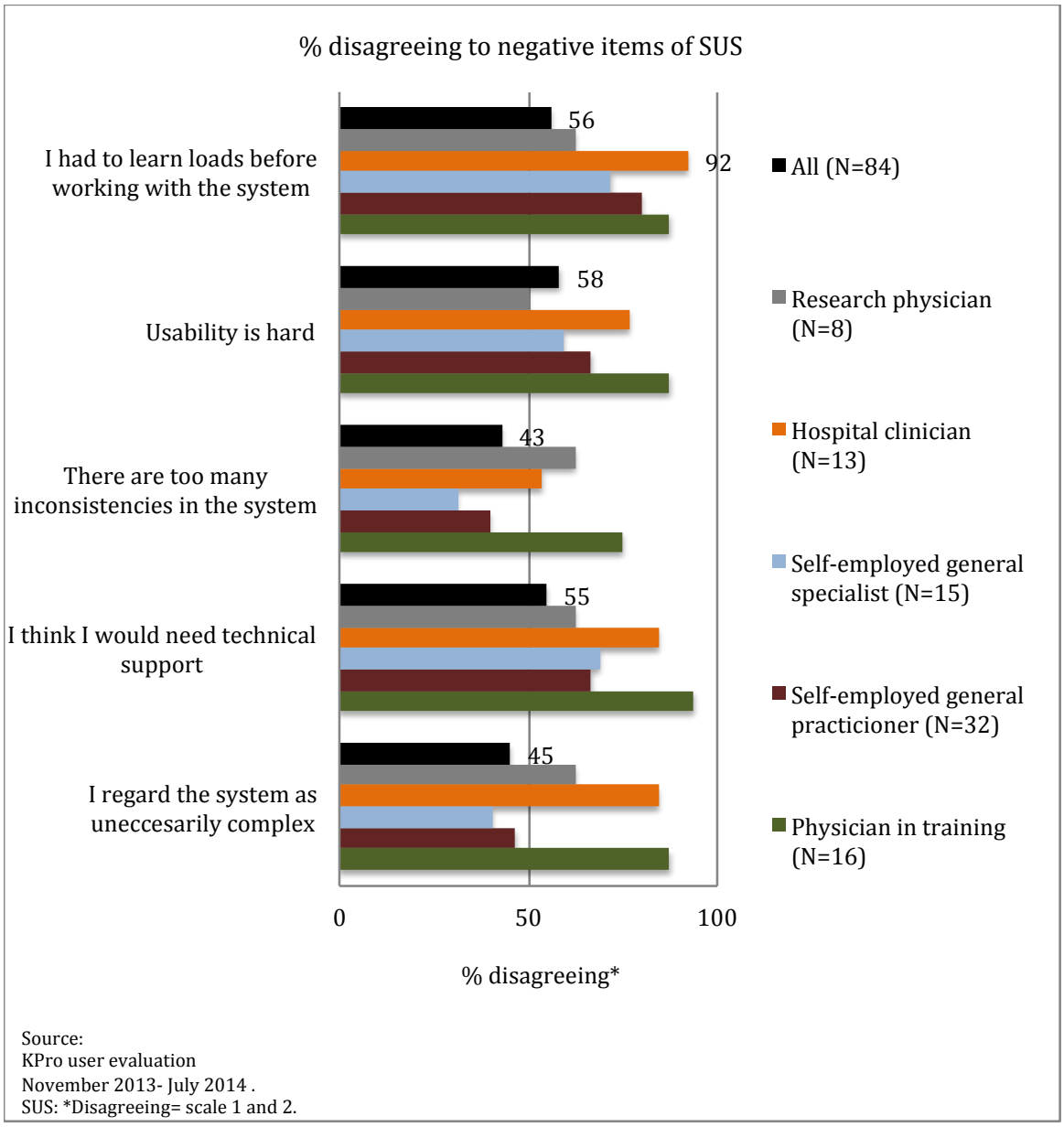


Figure 42 KPro usability in Y4 by occupational group: Negative SUS items.

2.4.7.2 KPro usability improvement in year 4

2.4.7.2.1 KPro usability improvement in year 4: All groups

2.4.7.2.1.1 Global SUS usability scores

Overall global usability scores increased by 5 percentile ranks across the different rounds of user evaluation. (November 2013: Mean=64, May-July 2014: Mean 69, Figure 45).

2.4.7.2.1.2 Individual SUS item responses

It was found that the biggest improvement within usability was that users perceived the system as less complex, felt more confident in using the system, and were more likely to imagine using the system on a regular basis in the second round of user tests (Figure 43).

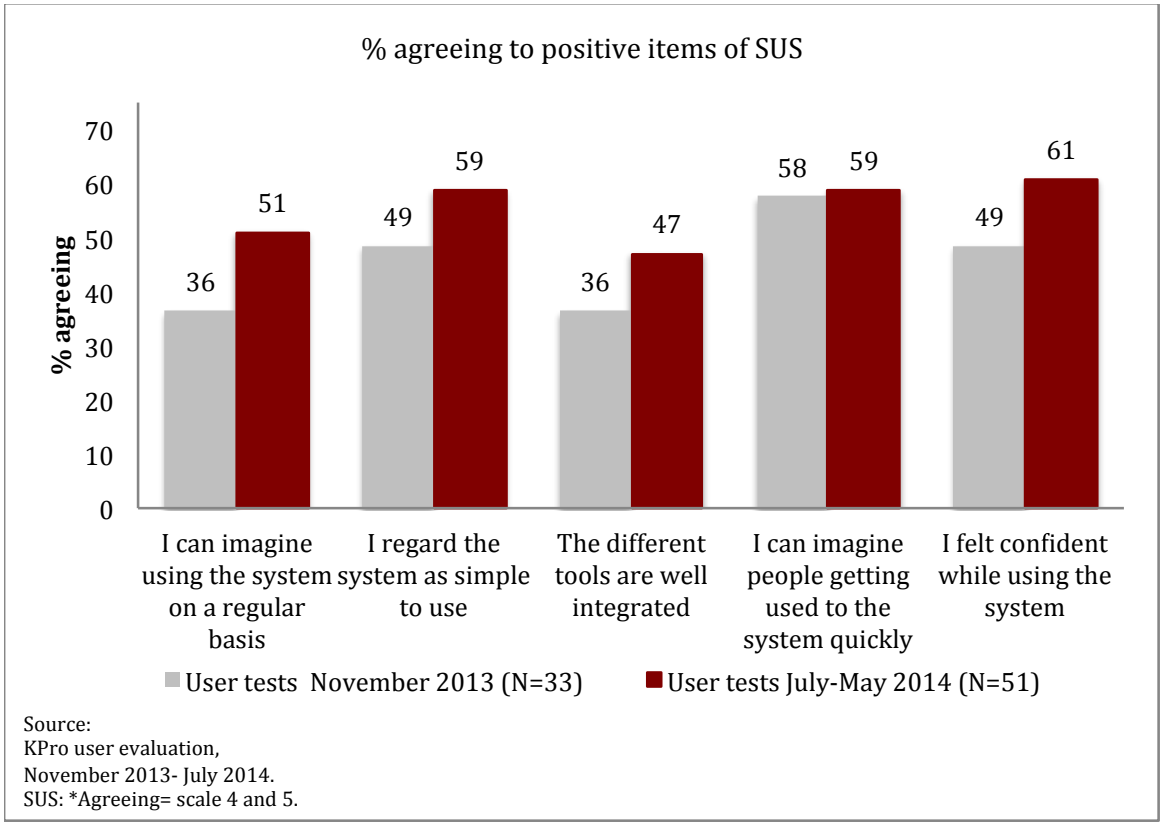


Figure 43 KPro usability improvement in Y4: Positive SUS items.

In the final user tests physicians were much less likely than in November 2013, to perceive that there are too many inconsistencies within the system or that usability is poor (Figure 44).

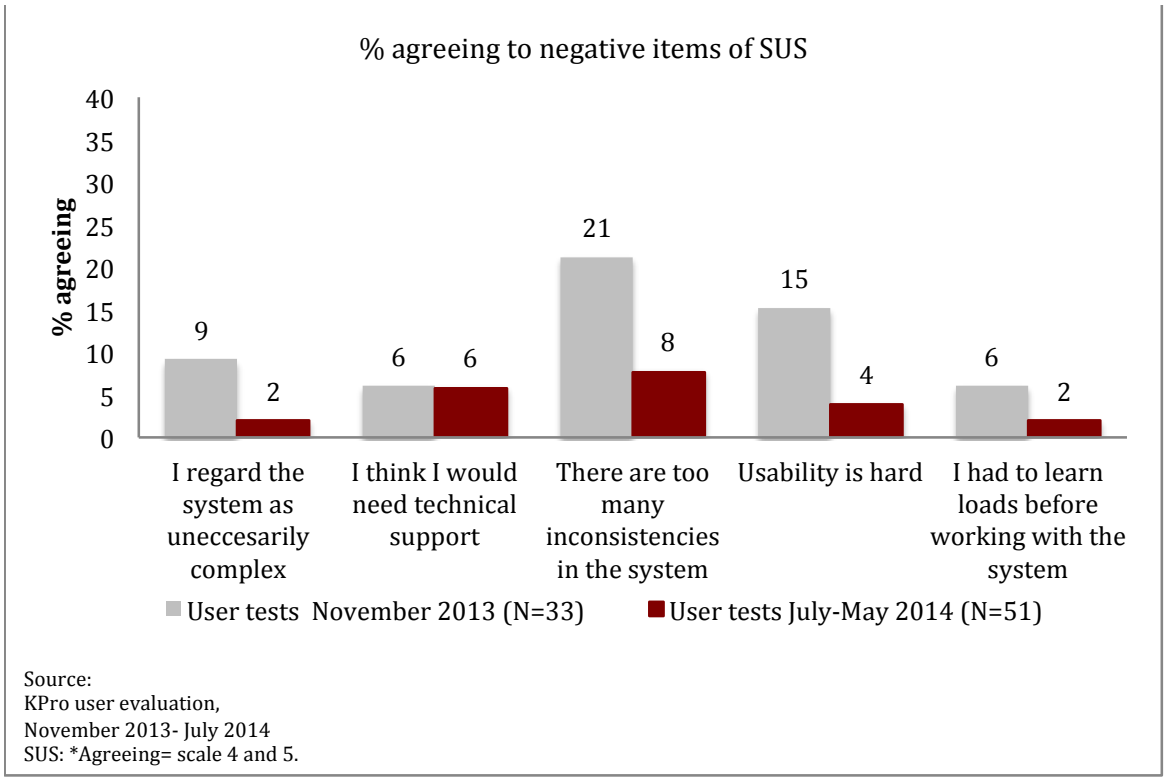
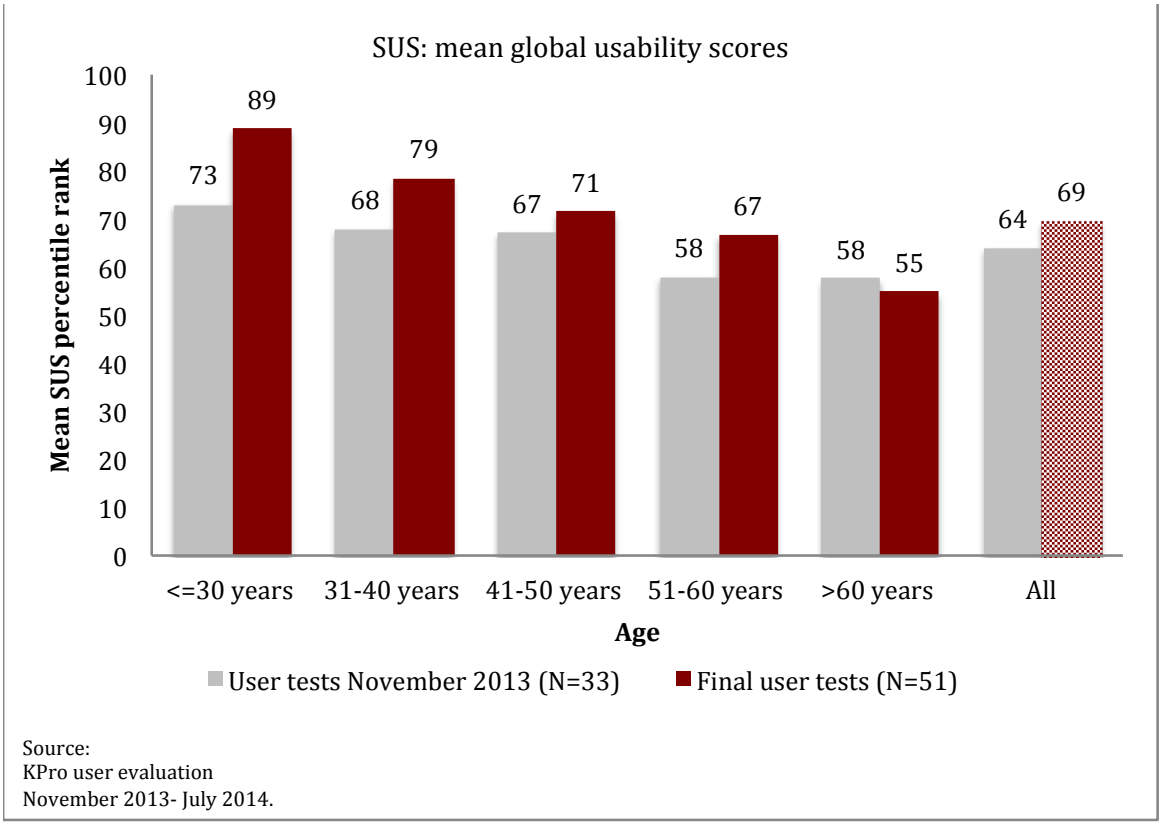


Figure 44 KPro usability improvement in Y4: Negative SUS items.

2.4.7.2.2 KPro usability improvement in year 4: A comparison of SUS scores across different age groups

In comparison with the first set of user tests, usability improved slightly for all age groups, except for the over 60 year olds where it remained stable (Figure 45).



Source:
KPro user evaluation
November 2013- July 2014.

Figure 45 KPro global usability improvement in Y4 by age.

2.4.7.2.3 KPro usability improvement in year 4: A comparison of SUS scores across different occupational groups

Across two rounds of user evaluation overall usability was most likely to increase for physicians in training. For general practitioners it remained somewhat stable (Figure 46).

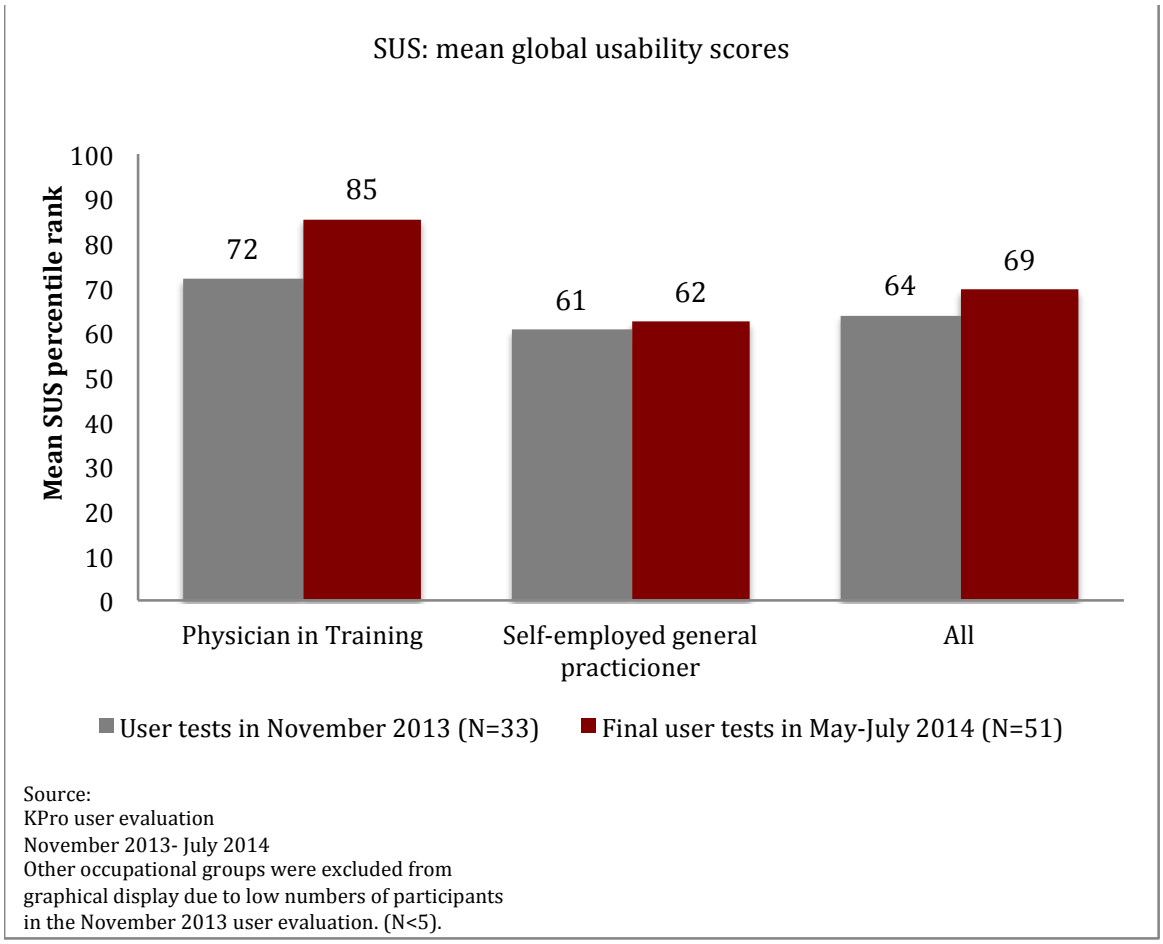


Figure 46 KPro global usability improvement in Y4 by occupational group.

2.4.7.3 KPro usability of the final prototype

2.4.7.3.1 KPro usability of the final prototype: All groups

2.4.7.3.1.1 Qualitative feedback on usability

During the final user evaluation carried out from May-July 2014 several users made positive comments on the interface for being clearly structured, easy to use and organised. However, some, especially older, users regarded the diverse functionalities offered as overwhelming and complex while others suggested larger icons, the exclusion of borders and adapting the layout to known systems. In Table 8 some original, translated user comments on usability are listed.

May-July 2014 user evaluation	
Comments on KPro usability	**Examples of comments**
Positive comments on KPro usability	„*I liked the simple usability*“ „ *The interface is cleary structured*“ „ *Easy to use*“ „ *Good possibility of content organisation*“
Suggestions for improvement	„*Search platform was too complex.*“ „ *I suggest more colored distinction of important content.*“ „ *The interface is a bit old-fashioned with the use of borders.*“ „ *The platform and icons are a bit small.*“

Table 8: Comments made on KPro usability in the final user evaluation in May-July 2014.

2.4.7.3.1.2 Global usability score

The overall global usability score achieved in the final user evaluation (May-July 2014) was 69, which is just above what is considered as average (i.e. 68) for a search system [2]. Scores ranged from 40 and 93 percentile ranks. Substantial difference in usability depended on the age and occupational group users belonged to, as will be illustrated in section 2.4.7.3.2 and 2.4.7.3.3.

2.4.7.3.1.3 Individual SUS item responses

As illustrated in Figure 47, almost two thirds of the physicians evaluating the final prototype felt confident in using the system. Every second physician reported that they can imagine using KPro in the future. The lowest scores were achieved in terms of tool integration and inconsistencies in the system. This is possibly explained by some bugs and system break downs that occurred but were quickly solved (upon efficient communication and response by other partners) during user tests.

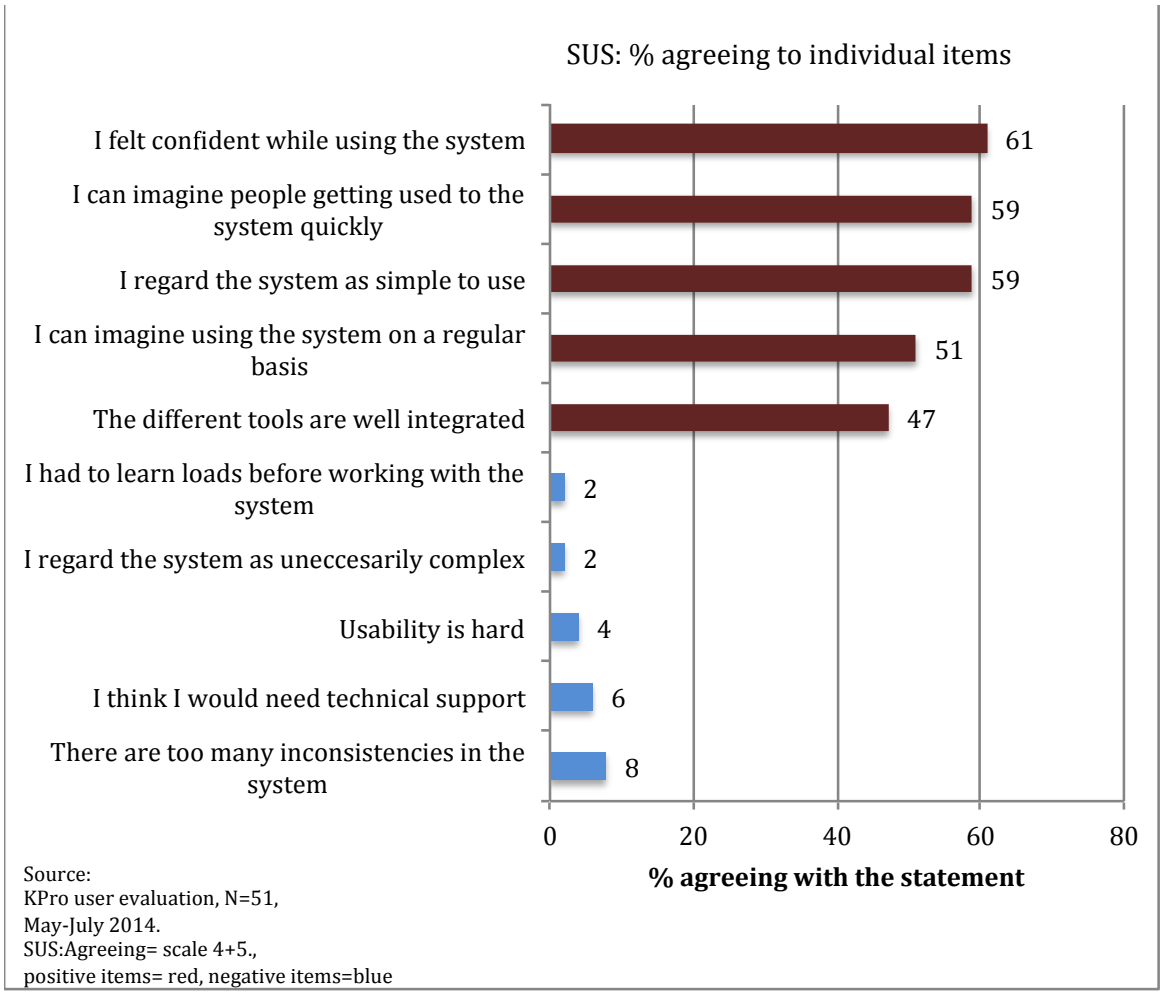


Figure 47 KPro usability of the final prototype: All SUS item responses.

2.4.7.3.2 KPro usability of the final prototype: A comparison of SUS percentiles across different age groups

Overall, a trend of younger physicians reporting higher usability persisted in the final user evaluation. System usability was classified as "very high" for physicians younger than 40 years, as "average" for physicians between 41-60 years and below average for physicians older than 60 years (Figure 48).

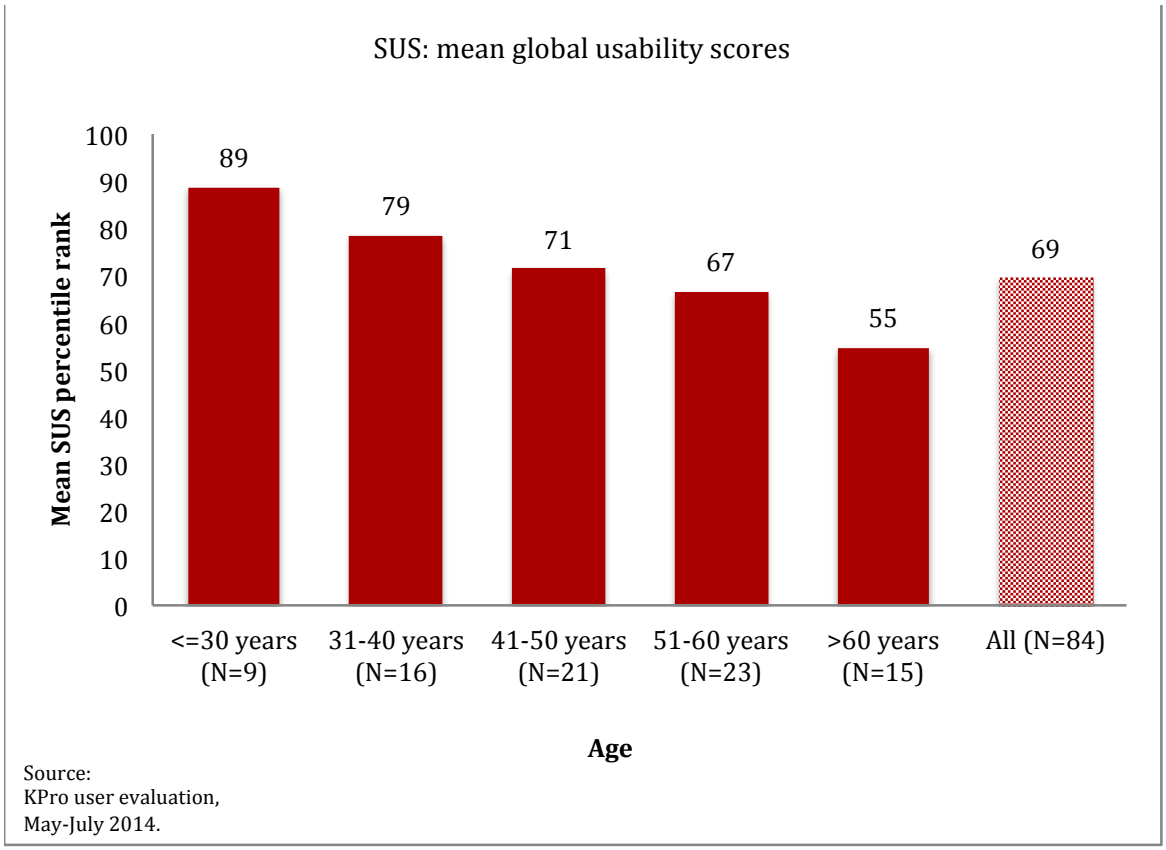


Figure 48 Global usability of the final KPro prototype by age.

2.4.7.3.3 KPro usability of the final prototype: A comparison of SUS percentiles across different occupational groups

Occupational group was less influential than age but showed slight impact between self-employed physicians and hospital physicians. Hospital physicians were, if controlled for age, more likely to score higher in terms of usability than self-employed physicians (Figure 49).

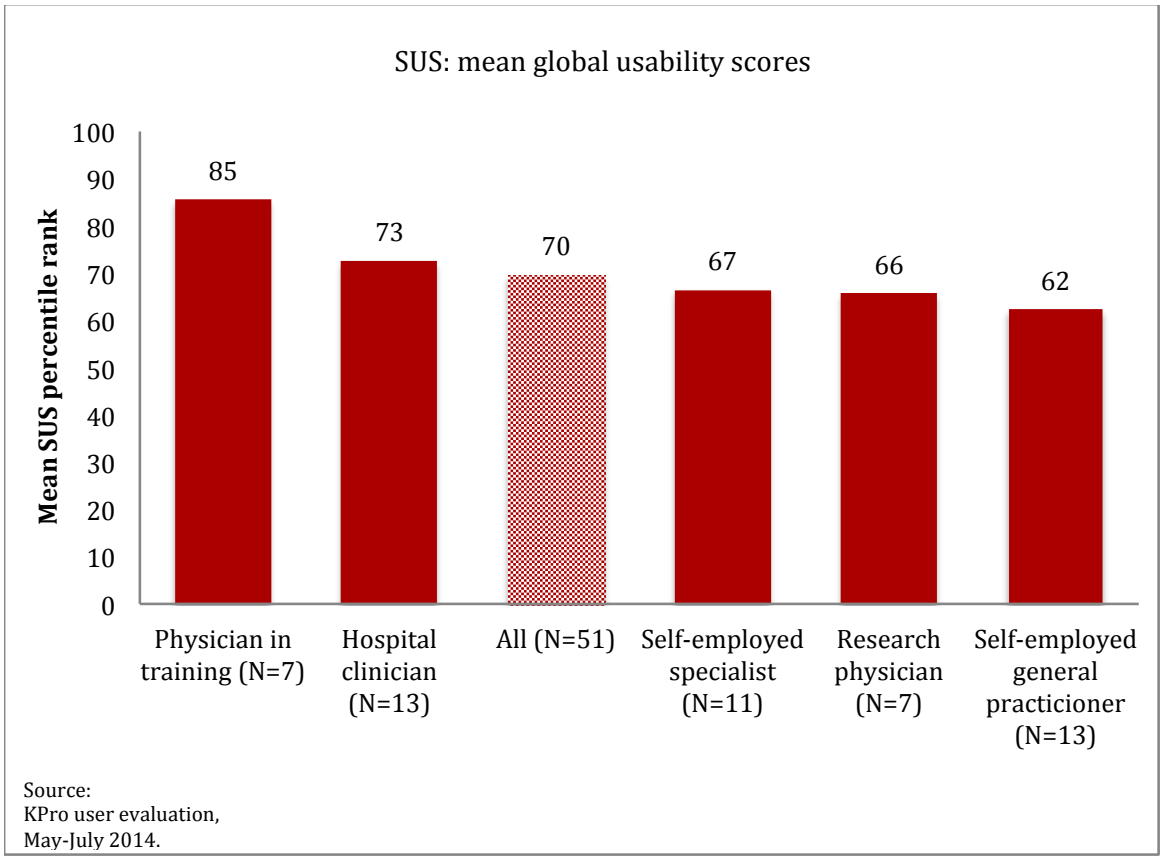


Figure 49 Global usability of the final KPro prototype by occupational group.

2.4.8 Usage of tools, search facets and search features

2.4.8.1 KPro search features: Which are useful?

2.4.8.1.1 KPro search features: All groups

The search facets were rated as the most, and the « common words » the least, useful search functionalities. Furthermore, every second physician rated « viewing similar queries » and the « preview of the article » as useful (Figure 50).

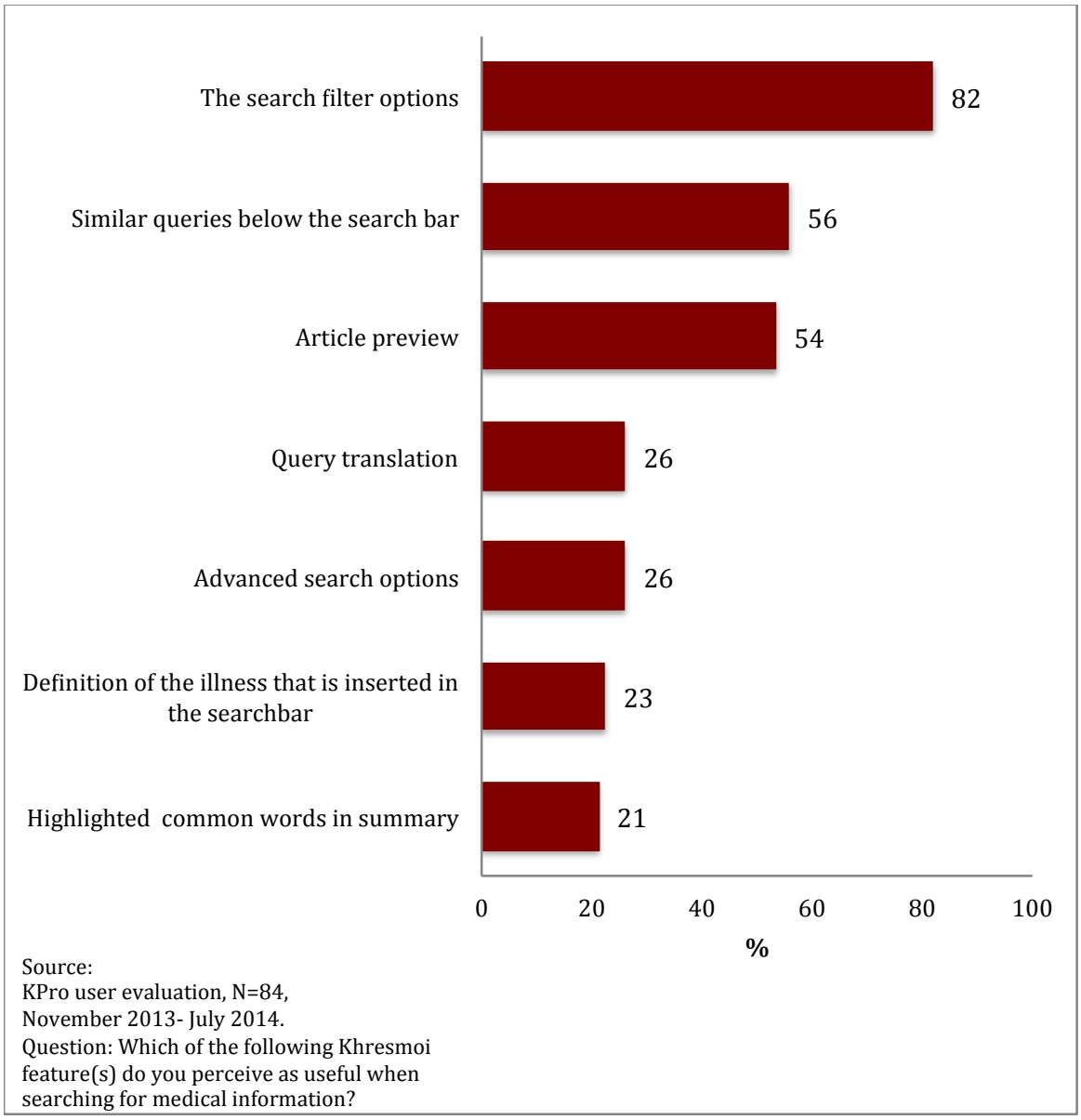


Figure 50 Usefulness of Khresmoi features.

2.4.8.1.2 KPro search features: A comparison across different occupational groups.

As illustrated in Figure 51 physicians in training and general practitioners placed the biggest importance on search filters. Physicians in training were the most likely group to regard query definition and query suggestion as useful. Three quarters of the research physicians regarded article preview and advanced search as most useful.

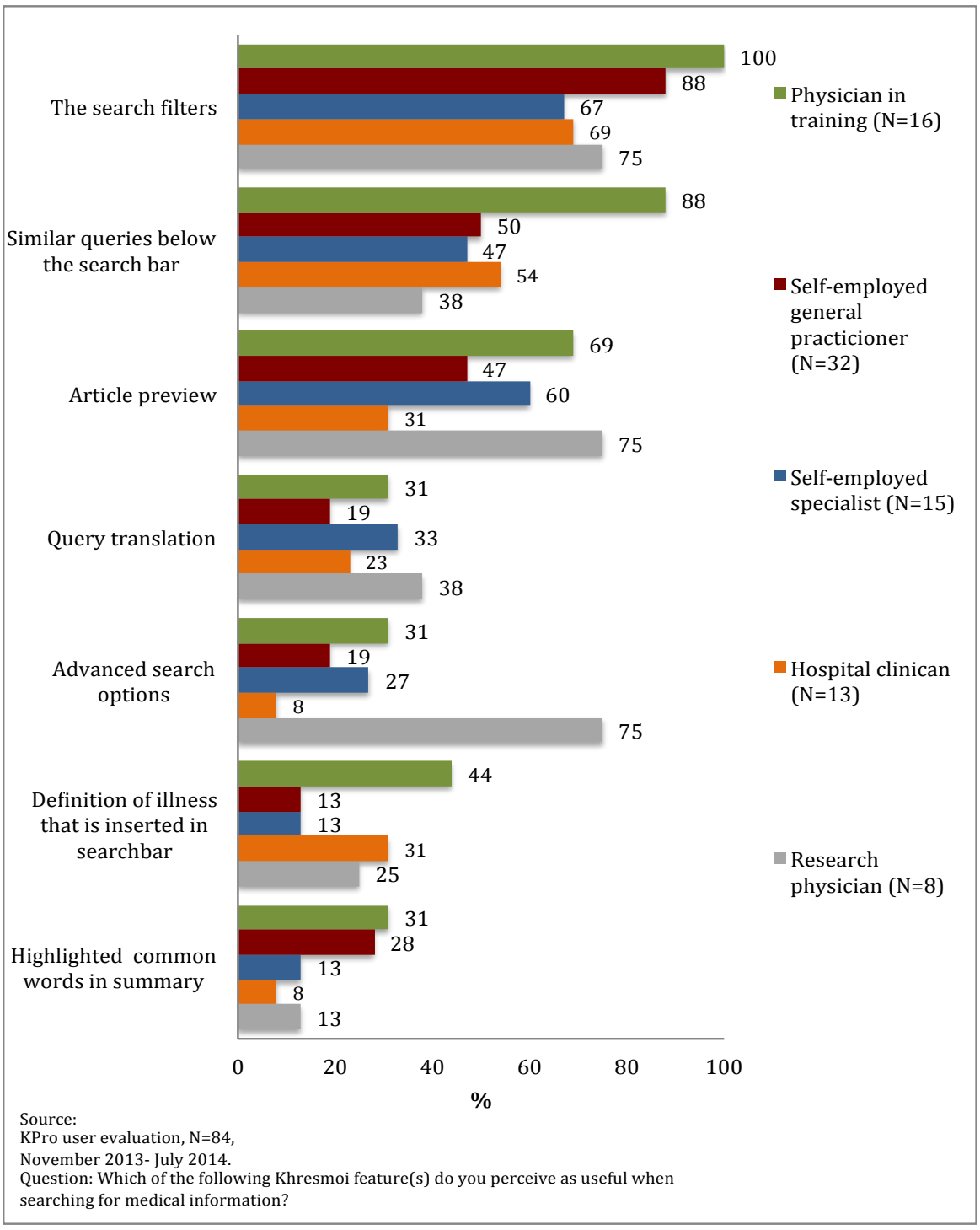


Figure 51 Usefulness of Khresmoi features by occupational group.

2.4.8.1.3 KPro search features: A comparison across different versions of KPro.

As illustrated in Figure 52 users evaluating the web version were more likely to regard query suggestion and article preview as important. One the other hand users, who evaluated the Java desktop version, were more likely to regard "highlighted common words" as useful than online users.

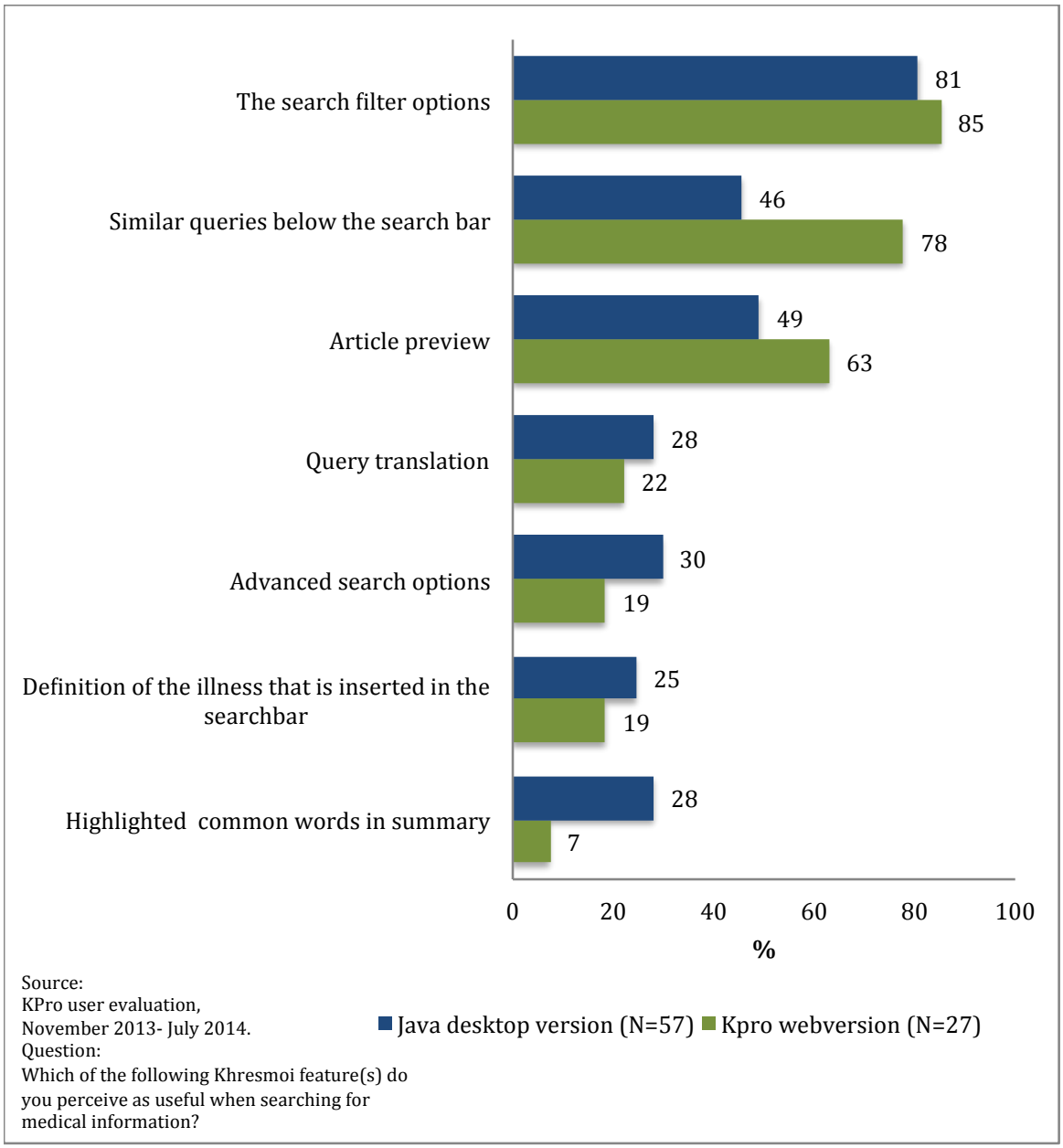


Figure 52 Usefulness of Khresmoi features by KPro version.

2.4.8.2 Pro search facets: Which are useful?

2.4.8.2.1 KPro search facets: All groups

Due to the popularity of search facets in the user tests in November 2013, a question asking about which facets are useful was added in the subsequent user evaluation.

The most popular search facets were “by category”, “by date” and “by language”. Most physicians liked the category restriction, every second physician regarded date restriction and just over a third language restriction as useful. Only one in five physicians regarded restriction by target group, publisher and media as useful. Country and image modality were classified as the least popular search filters (Figure 53).

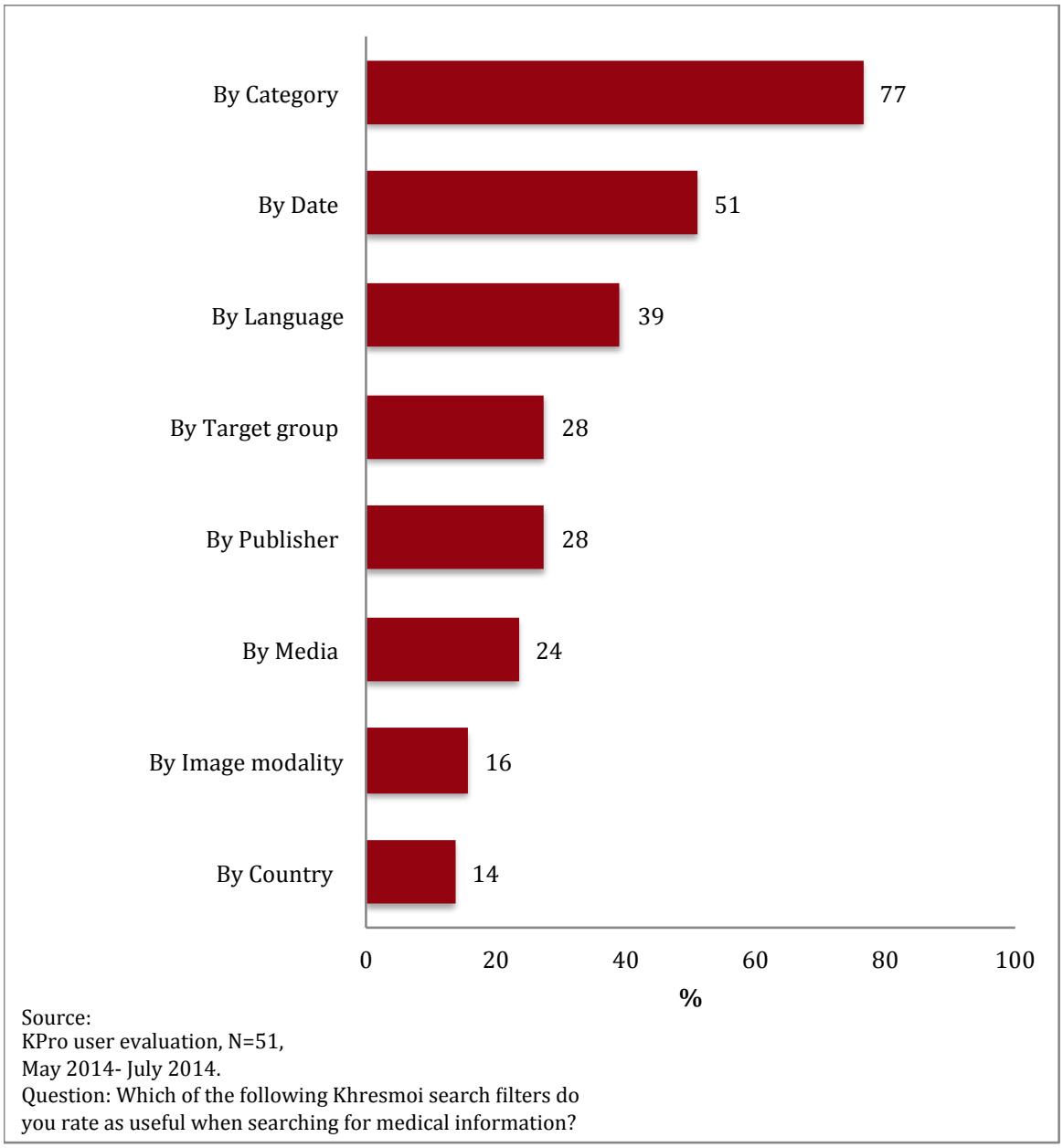


Figure 53 Usefulness of Khresmoi facets.

2.4.8.2.2 KPro search facets: A comparison across different occupational groups

Across all groups, the most popular search facets were “by category”, “by date” and “by language”. Research physicians expressed an additional interest in the filter « by image modality ». General practitioners perceived the filter « by country « as useful. Physicians in training regarded filtering « by publisher » as useful. While general practitioners appeared less interested in the date filter, they showed higher interest in restricting “by target group” and “by media” than other groups (Figure 54).

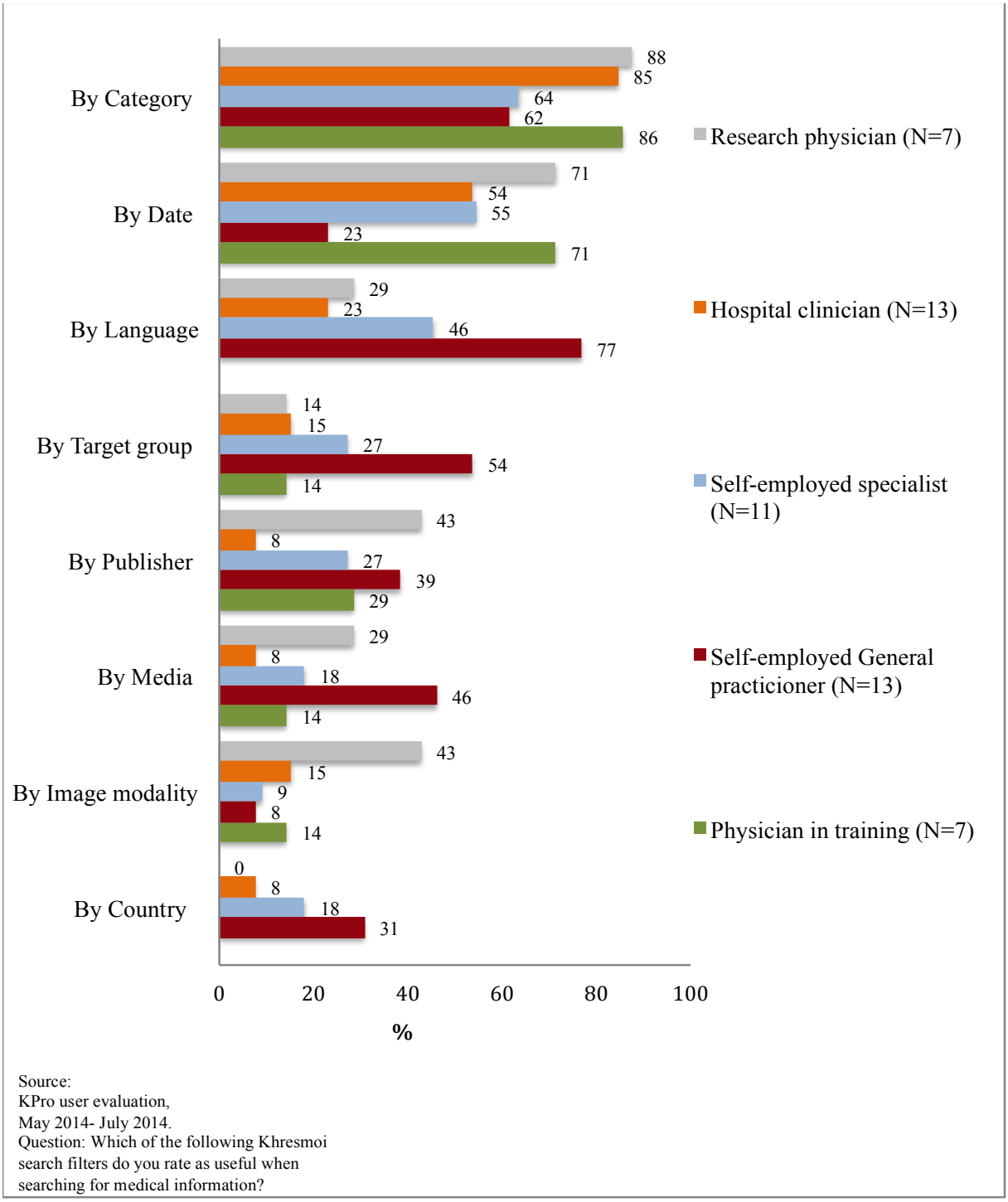


Figure 54 Usefulness of Khresmoi facets by occupational group.

2.4.8.2.3 KPro search facets- the facet "by category": All groups

Search filter "by category"- which restriction?

After pilot tests in October 2013, physician filters "by category" were defined, in response to user feedback and as part of the user case analysis performed for WP8 [4]. These encompassed the feature that users could manually differentiate resources that fitted in one of the following categories: Definition, Medical Education (Sub filters: Online Education, Events), Clinical practice (Sub filter: guidelines, diagnostic information, drug information) and organisational. Within the search filter "by

category" two thirds of the participants rated restriction by "guidelines" and "drug information" as useful. Furthermore every second physician rated the restriction « by scientific articles » and a third the filter by « online education » as useful. The least popular feature was "organizational" only rated by two physicians as useful (Figure 55).

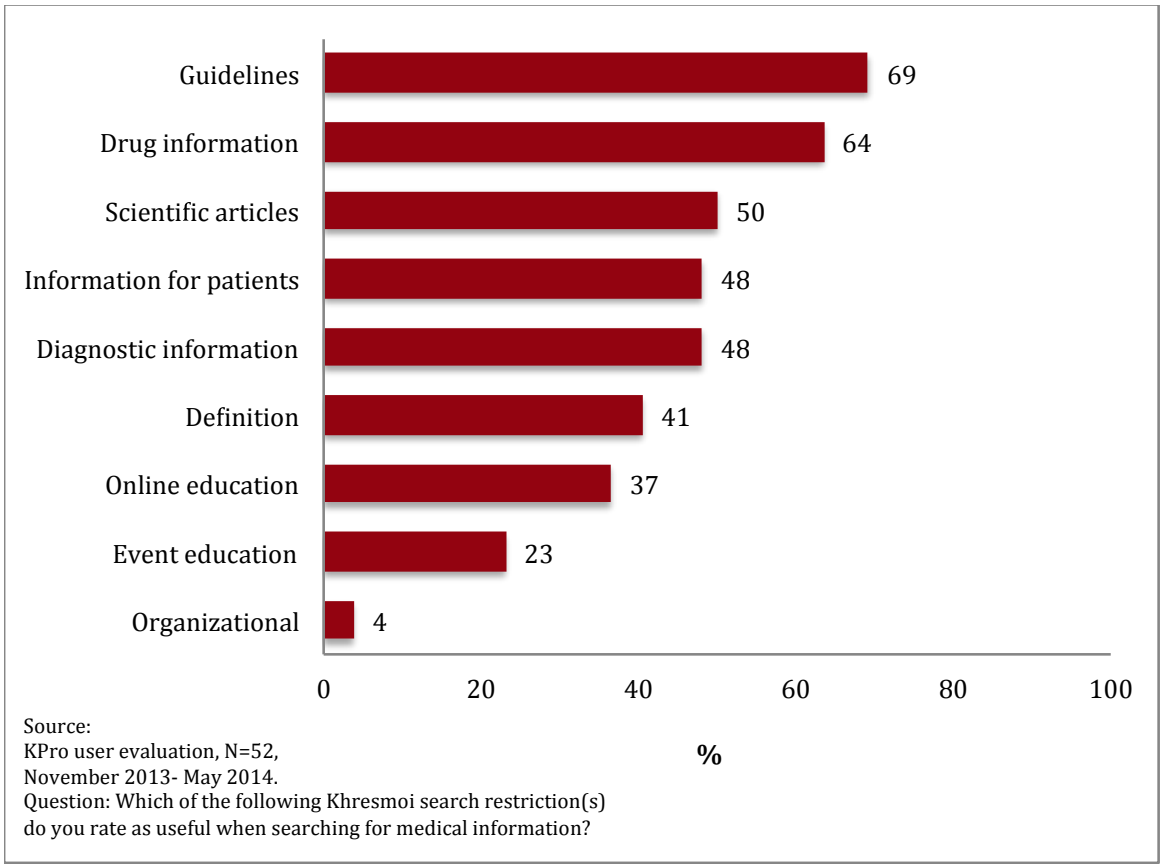
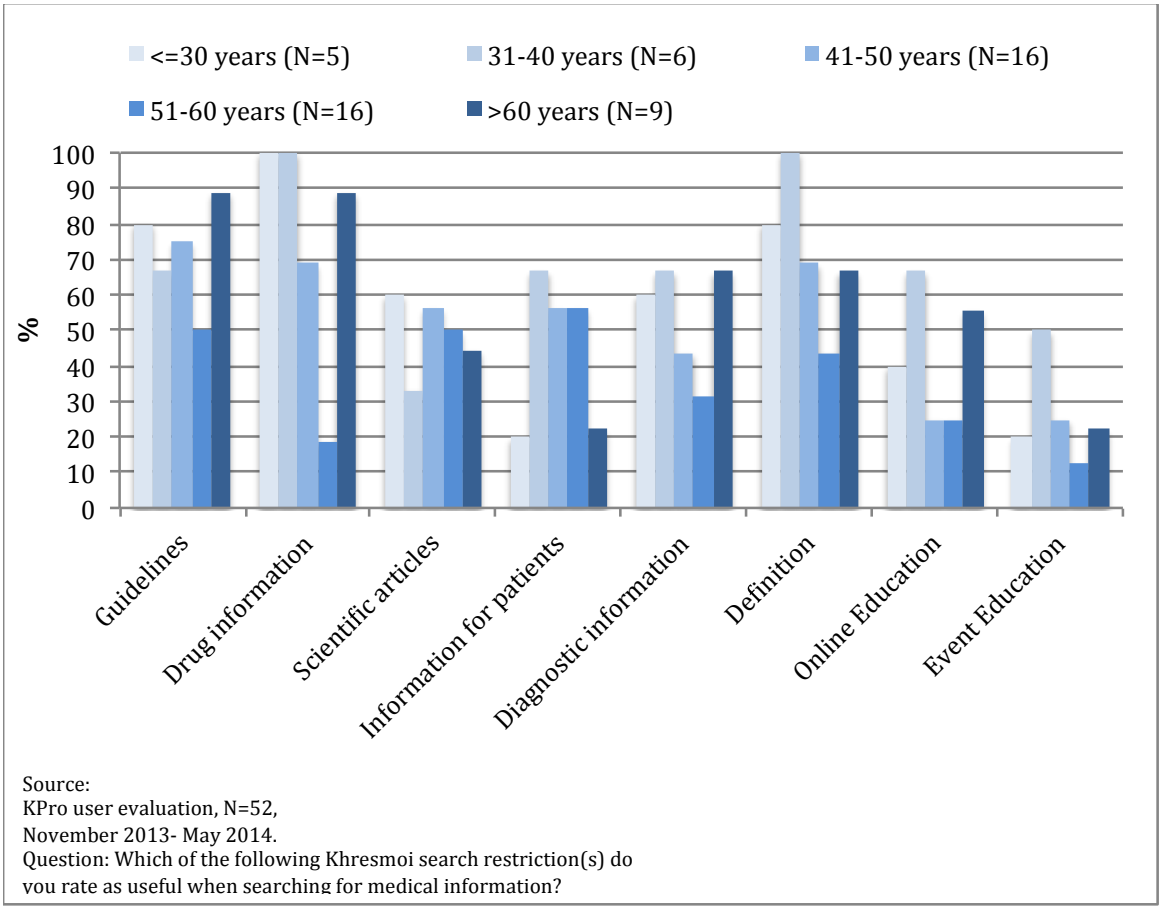


Figure 55 Usefulness of search restrictions within the facet « by category ».

2.4.8.2.4 KPro search facets- the facet "by category": A comparison across different age groups

Across different age groups, the only pattern that could be identified was for the "definition" facet, which appeared to be more frequent among younger physicians. For other facets, age appeared to be of less influence (Figure 56).



Source:
KPro user evaluation, N=52,
November 2013- May 2014.
Question: Which of the following Khresmoi search restriction(s) do you rate as useful when searching for medical information?

Figure 56 Usefulness of search restrictions within the facet « by category » by age.

2.4.8.2.5 KPro search facets- the facet "by category": A comparison across different occupational groups

Amongst self-employed physicians, restriction by «drug information », « guidelines » and « information for patients » prevailed. Research physicians and hospital clinicians expressed strong interest for the « scientific articles » filter and physicians in training rated the « definition » and « online education » restriction as useful. An interesting observation was that physicians in training and research physicians were more likely than other groups to rate the filter by "drug information" ($p<0.05$) and "online education" as useful. A possible explanation may be that self-employed and hospital clinicians usually have access to drug information software in their workspace (Figure 57).

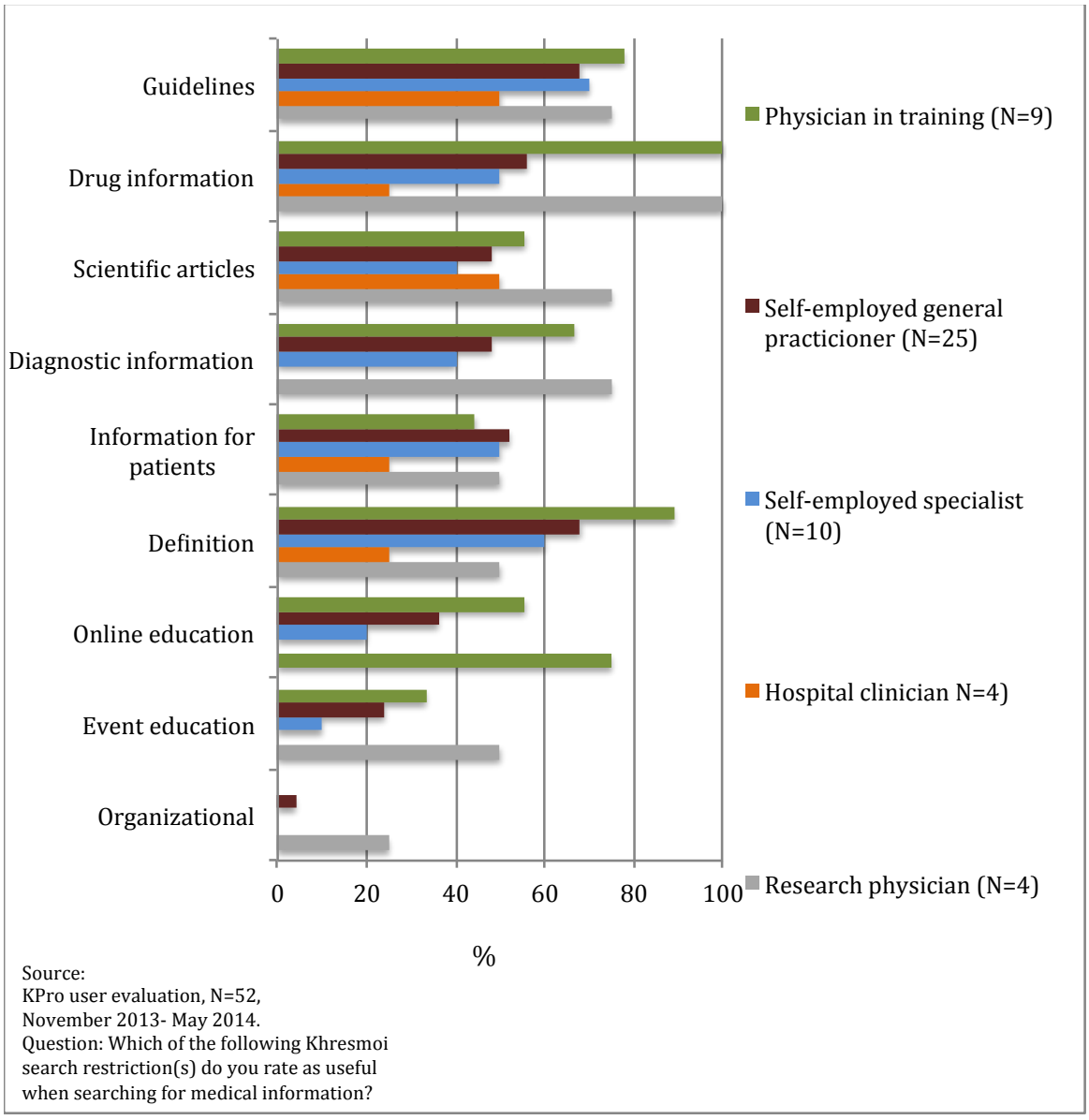


Figure 57 Usefulness of search restrictions within the facet « by category » by occ. group.

2.4.8.3 KPro tools: Which are useful?

2.4.8.3.1 Khresmoi Tools: All groups

The most popular tool was the personal library and associated features such as the tagging and export function. More than half of the participants reported that they can imagine using the personal library, export function and tagging function on a regular basis. Just over a third reported to imagine using the summary translation feature in the future. The personal library and summary translation was most popular amongst research physicians and self-employed physicians. Physicians in training were least likely to report that they could imagine using the summary translation in the future. Collaborative tools such as sharing articles and communicating with colleagues were given less attention (Figure 58).

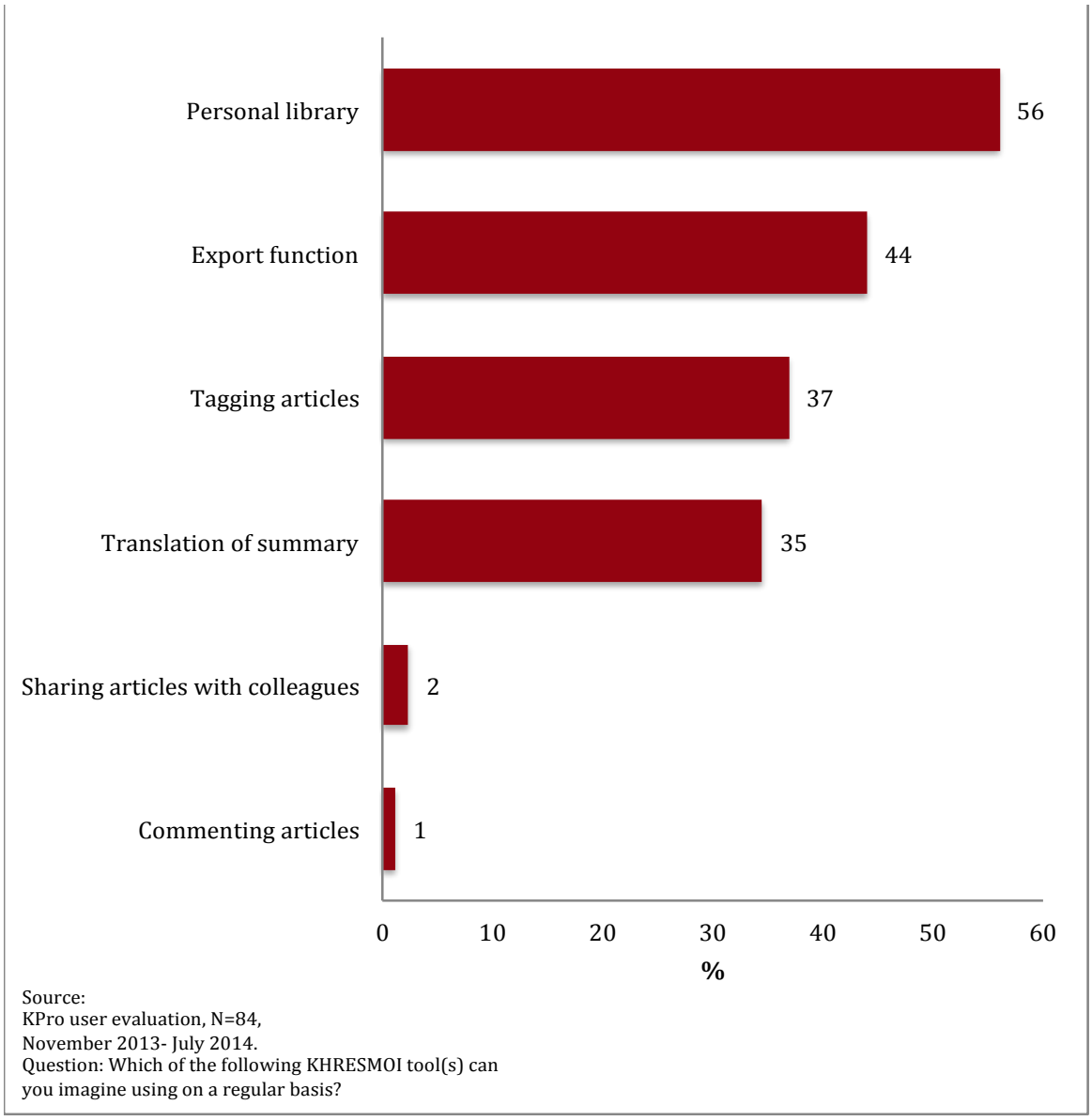


Figure 58 Usefulness of tools.

2.4.8.3.2 Khresmoi Tools: A comparison across different occupational groups.

Three out of four research physicians reported that they can imagine using the personal library, tagging and export function in the future. Over a third of self-employed practitioners perceived the translation of summary feature as useful (Figure 59).

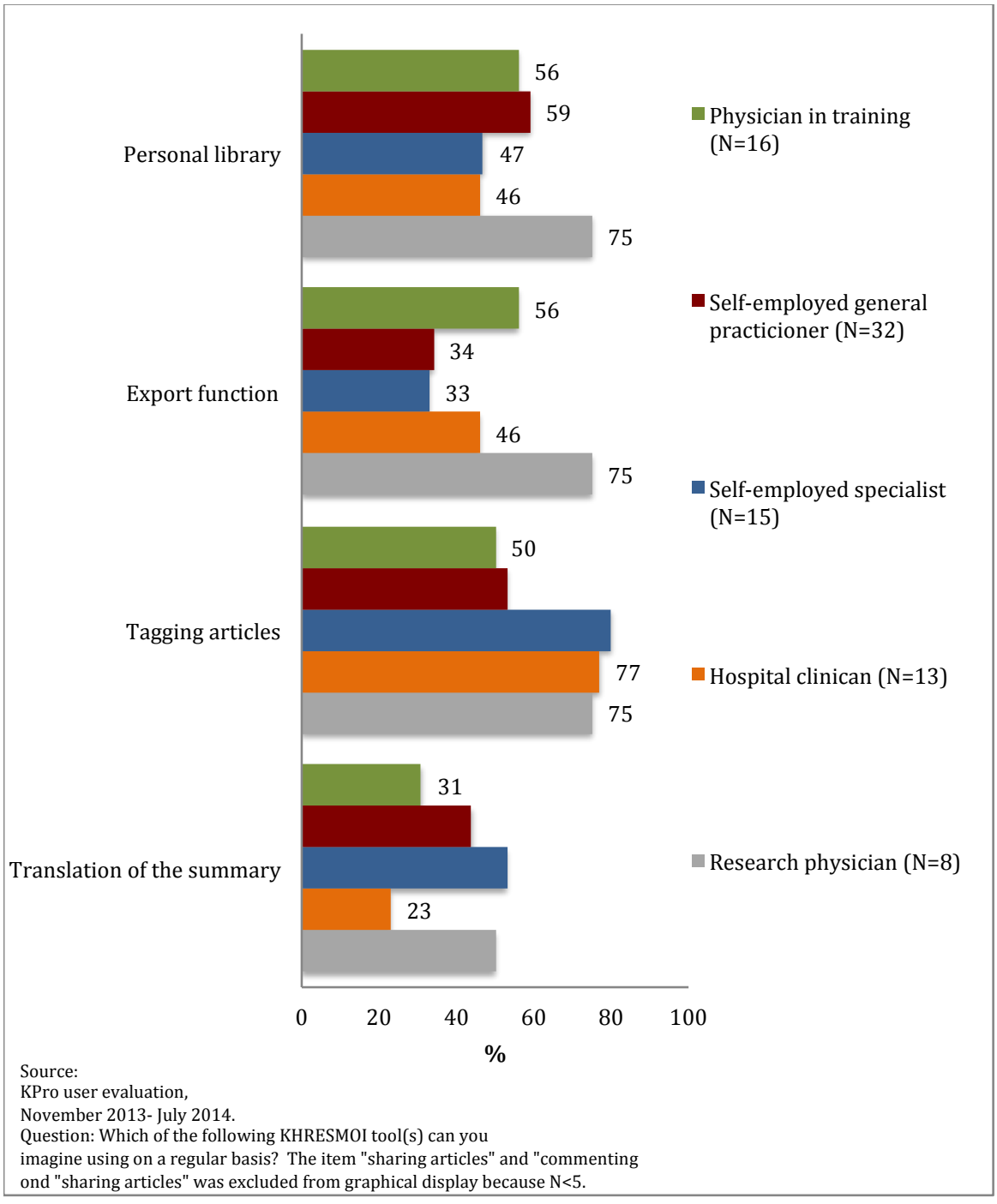


Figure 59 Usefulness of tools by occupational group.

2.4.8.3.3 Khresmoi Tools: A comparison across different versions of Khresmoi Professional.

Figure 60 compares the answers of user testing of different versions of KPro, in their reported likelihood of future use of Khresmoi Tools. The biggest difference was found for the personal library. Two thirds of the users trying the Java desktop version, compared to only one third of the users trying the web version could image using the personal library in the future. The likely explanation is that the personal library did not work in the KPro web version. However, despite having had no chance to "try it out" a third expressed interest. For the tools that worked in both versions, the "expressed" interest

was similar for both versions of KPro. The results suggest that an integration of the personal library feature in the web version would be useful.

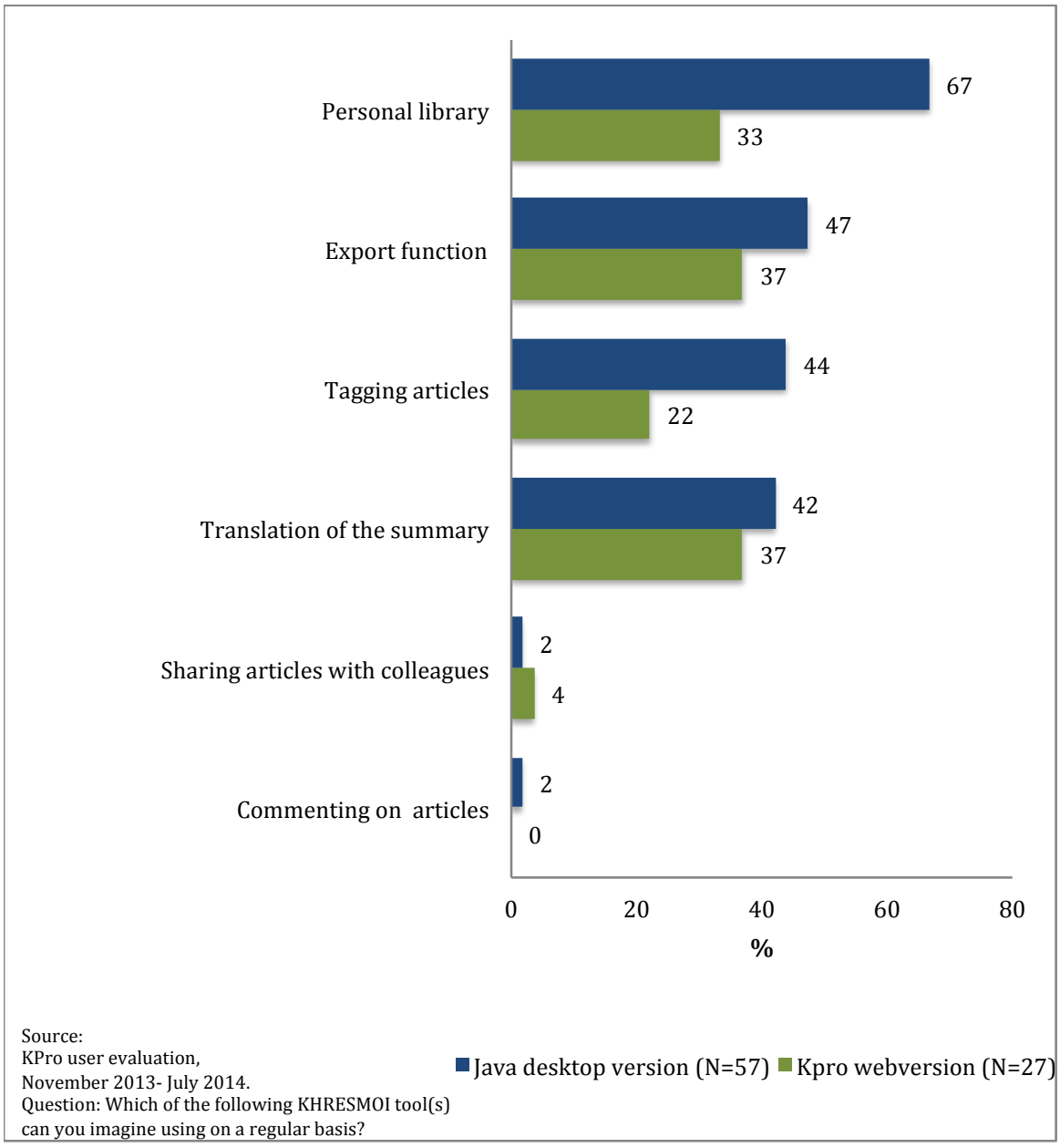


Figure 60 Usefulness of tools by KPro version.

2.4.8.3.4 Khresmoi Tools: Additional feedback from the pre-defined tools tasks in November 2013 evaluation.

In the November 2013 evaluation 22 physicians completed one of two pre-defined tasks that focused on the personal library, the search filters, sorting by date and the double-click restriction. All users trying out the personal library perceived the tool as useful and could imagine using it in the future. Ten out of eleven users perceived the search filters as useful or very useful with only one user stating that they would never use this feature. Both sorting by date as well as the double click search restriction feature, was perceived as useful by nine out of eleven users (82%),

2.4.9 KPro Exploitation

To understand the exploitation potential of KPro, the likelihood of future access, strengths and reasons to return to KPro and user suggestions for improvements were analysed.

2.4.9.1 KPro Exploitation - likelihood of future use of KPro

2.4.9.1.1 Likelihood of future use of KPro: All groups

As illustrated in the previous paragraphs, the usefulness of Khresmoi Professional has immensely improved over the last months. With regard to the estimated "future use" the level of agreement to the SUS item "I can imagine using the system on a regular basis" was analysed in further detail. More than two thirds of the physicians younger than 40 years can imagine using KPro on a regular basis. Physicians above 50 years are significantly less likely to report that they can use the system on regular basis (Figure 61).

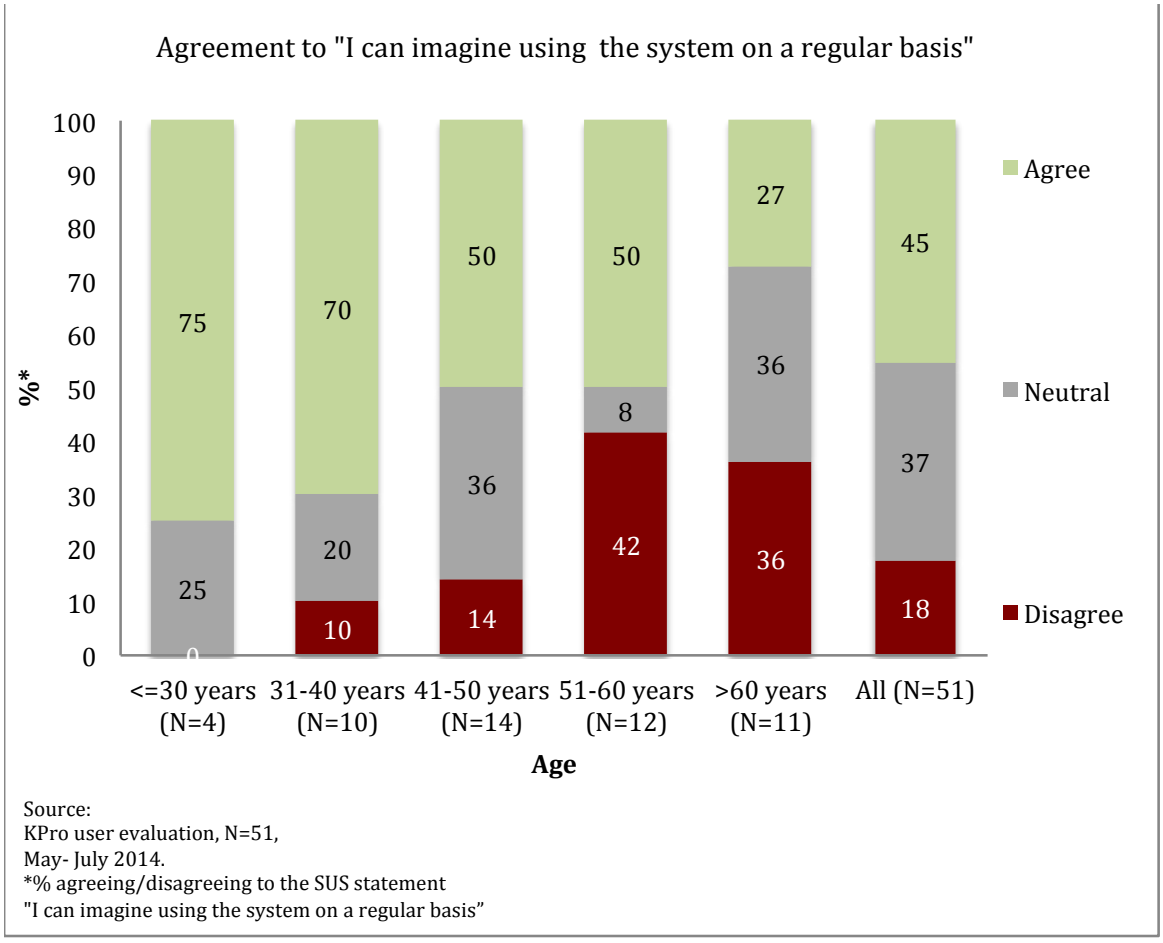


Figure 61 KPro Exploitation: Likelihood of future use of KPro.

2.4.9.1.2 KPro Exploitation-likelihood of future use of KPro: A comparison across occupational groups.

Across different occupational groups, self-employed practitioners were the least likely group to report future use, while physicians in training and hospital clinicians were the most likely group.

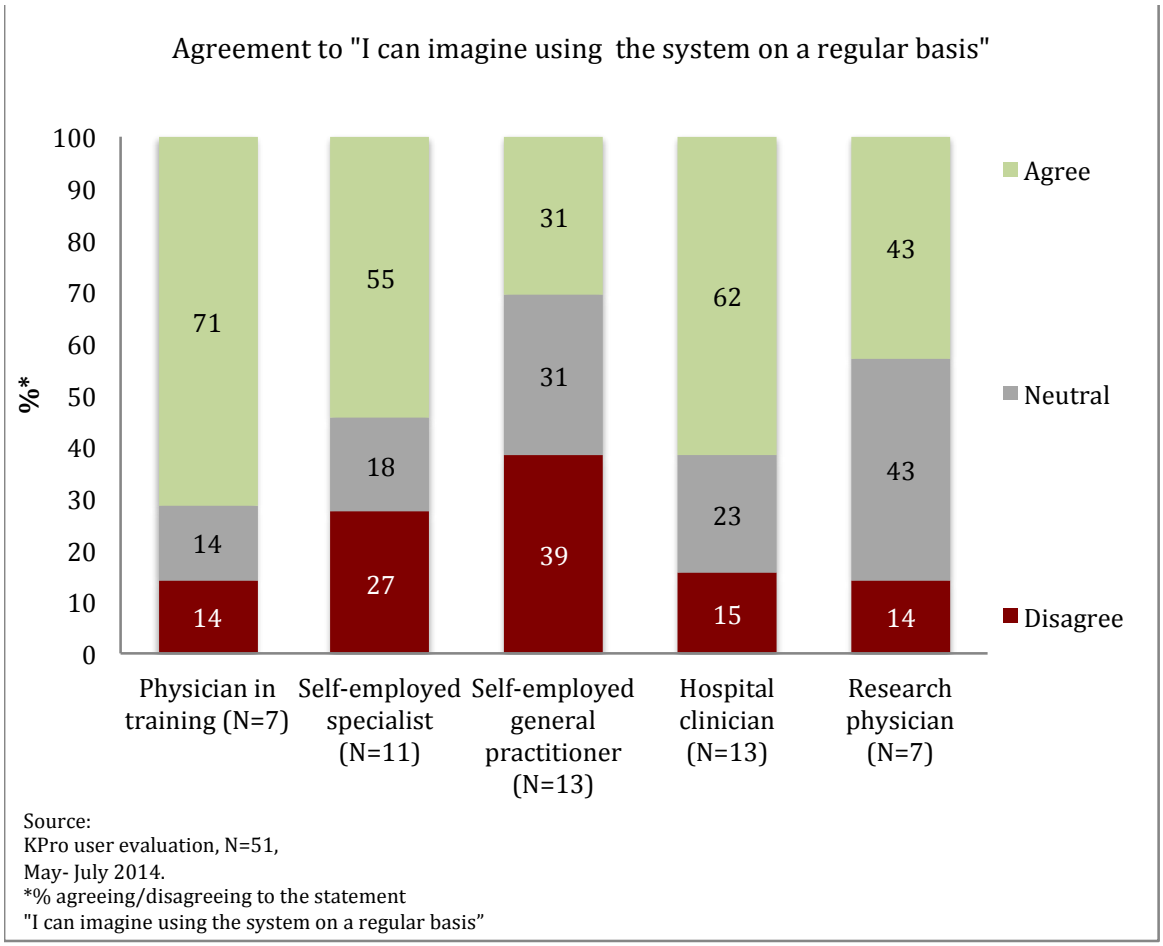


Figure 62 KPro Exploitation: Likelihood of future use of KPro by occupational group.

2.4.9.1.3 KPro Exploitation qualitative insight: Accessibility

When given the choice, most physicians preferred to access KPro via the web browser to downloading the Java desktop version. It was found that 93% (27/29) of the users taking part in the online evaluation of Khresmoi Professional chose, when given the choice, to try out the web version of Khresmoi Professional. This suggests low willingness to download an "unknown" search system and that future exploitation should focus on the advancing the KPro web version and integrate popular tools such as the personal library.

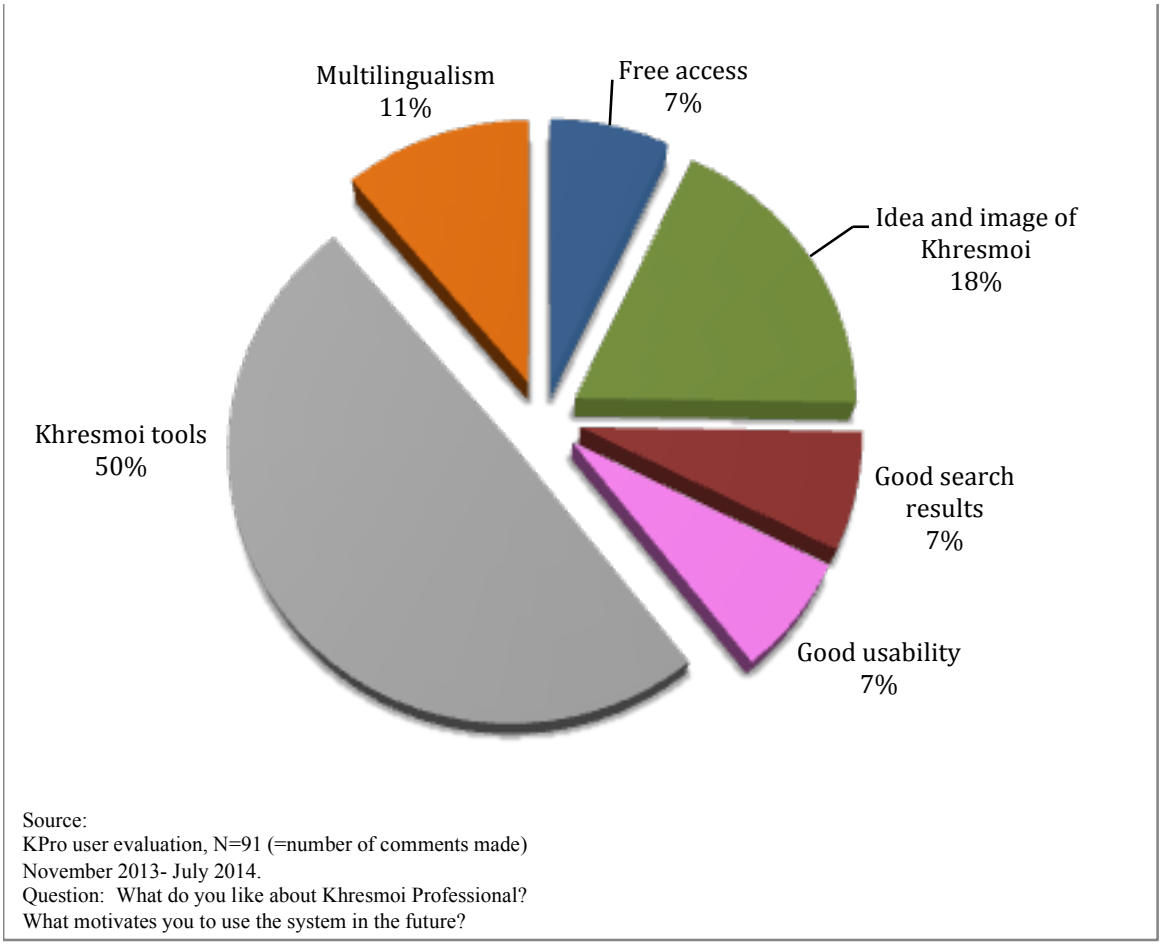
Examples of original comments in relation to system accessibility can be viewed in table 9. Further original user feedback on accessibility can be viewed in Appendix 5.1. Device accessibility posed an important barrier to accessing the KPro system. As illustrated earlier, in Figure 12, most physicians reported to access medical information via mobile devices. Two third of physicians using mobile devices used IOS devices. In the face-face user tests the mobile version was demonstrated to some physicians. Initial interest towards a mobile application was high. However, users commented that they missed the implementation of tools and criticised the lack of IOS accessibility. In the online feedback survey, six physicians gave explicit feedback that they could not evaluate the system due to IOS incompatibility. One user reported having problems downloading the JAVA version and one user complained that the system is not easy to access on the iPad. This suggests that it would be of benefit to develop a more advanced mobile application and make the KPro system accessible for IOS systems.

May-July 2014 user evaluation	
KPro Accessibility	**Examples of comments made**
KPro is hard to access	*„lack of Ipad compatibility“* *„ Java Version with the extra functions didn´t work for on my device.“* *“ Not accessible due to device incompatibility” (6* online IOS users reported that they couldn´t test the system for this reason)

Table 9: Comments made on KPro accessibility in the final user evaluation in May-July 2014.

2.4.9.1.4 KPro qualitative insight: KPro strengths and reasons for future use.

On a qualitative level open feedback was analysed to questions of what users liked/disliked about KPro and what aspects of KPro would make them return. Six users wrote spontaneously that they would use the system as it is. An example of a comment was “I would already use it now” or “Good approach, it is for me (in its current form) interesting parallel to other search systems”. In the second round of user tests, users either liked or disliked the system, with a lower proportion in the “neutral” domain. Open feedback was categorised, counted and assigned to different categories. As illustrated in Figure 63 the most popular reasons to use the system in the future were the tools and facets, the idea and unbiased foundation of KPro and the multilingual aspects of KPro. In Table 9 examples of comments on each dimension are provided.



Source:
KPro user evaluation, N=91 (=number of comments made)
November 2013- July 2014.
Question: What do you like about Khresmoi Professional?
What motivates you to use the system in the future?

Figure 63 KPro Exploitation: Reasons to return to KPro in the future.

2.4.9.1.5 KPro Exploitation qualitative insight: KPro suggestions for improvement

Overall year 4 user evaluation	
	Issues raised (n) and examples of comments made
KPro Accessibility	Free Access *(7) „ Freely accessible“*
Idea and image of KPro	The idea of KPro (7) *„ The idea of an independent, freely accessible search engine for physicians.“* *„Search for medical professionals in the German-speaking domain“* *Unbiased, EU-supported, independent, no adverts (10)* *“ EU-supported”, “ Independent”, “ no pharma”*
Multilingualism	Translation (2) “*Direct translation possible”* Multilingual interface (2) *“ German interface”* Multilingual resources (1) *„ Multilingual articles“* Multilingual filters (5) *„The idea to restrict by language is good.“*
KPro effectiveness KPro efficiency	Good results (5) „Good results and professional information.“ Quick access to good information (2) „Quick retrieval of current guidelines.“
KPro usability	Overview of of the interface (4) *„The interface has very good overview. „* Other comments (3) *„ I like the preview.“ „ I like. Display of the publication year in the link.“*
KPro tools, facets and search features	Good filters (16) *„ Search filters are interesting“* The search filter „by category“ (6) *„I like the filter option by guidelines“ (3)* *„ Restriction by category is very interesting.“* Search filters country, date and image (3) *„ Restriction by date“*, Personal library (13) *„the personal library is very good.“* Other tools (5) *„I like. Sorting by date..“* *„ Search restriction with double click is interesting.“* *„ Sorting by date with recent resources first“*

Table 10: KPro Exploitation potential: What users liked about KPro that would make them return in the future.

2.4.9.1.6 KPro Exploitation-qualitative insight: KPro suggestions for improvement

Users were asked to give suggestions on what would need to be improved for KPro to be a bigger interest to them. Unsolved issues from the November 2013 evaluation and comments from the final user evaluation in May-July 2014 were analysed, counted and categorised. Full tables on all feedback can be found in Appendix 5.1. As illustrated in Figure 64 most suggestions revolved around

improving effectiveness, usability of tools and KPro accessibility. Table 10 and Table 11 list examples of original user comments for each category.

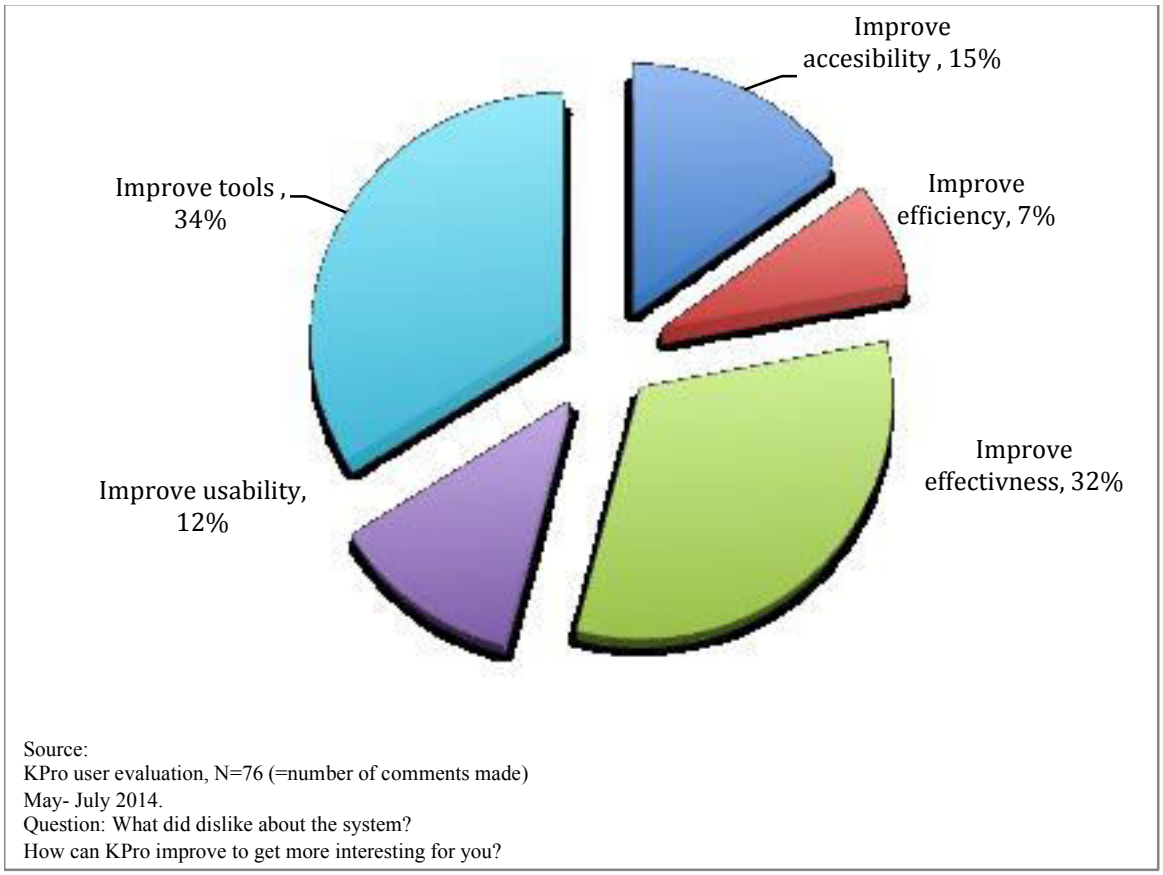


Figure 64 KPro Exploitation: Suggestions for improvement of the final KPro search system.

May-July 2014 user evaluation	
Topic	**Suggestions for improvement and examples of comments (n)**
KPro accessibility	Provide exclusive, access for physicians (1) **"** *Access exclusively for physicians/medical professionals"* IOS Device compatibility (8) **"***Access for Ipad and Safari"*
KPro efficiency	Ensure quicker access/search (6) *„Quicker access to good information than in Google"* *„ Quick accessibility to trustworthy and high quality medical resources"* *„ I can imagine using it if I make good experience with quick search"*
KPro effectiveness	Improvement of ranking (5) *„ Results need to more relevant, not on the whole topic if I ask a specific question."* „ Find everything I am searching for" Improvement of resources (24) Include more resources, images, online CME, German articles, overview articles, alternative medicine articles, patient information, professional information, good scientific resources. *„For the exclusive search it would require more search results."* *„ More articles in German please", „ Suggest the inclusion of more good, scientific resources" „ suggest inclusion of images" " presentations and power points"*

Table 11: KPro exploitation potential: Suggestions for improvement on effectiveness, efficiency and accessibility.

May-July 2014 user evaluation	
Topic	**Suggestions for improvement and examples of comments (n)**
KPro usability	**Simplification and modernization of interface (7)** „User-friendlier, “ „Simplification of search platform. „ Should be more intuitive „ , „less cluttered“ „Modernization of search platform“ ***Improve navigation and fix bugs (3)*** *“ Reset search function (double click) is not clear/intuitive.”* „ I like the preview but unfortunately many buttons don’t work. That makes it confusing.“ „Requires better navigation“
KPro Tools	**Improve usability and consistency of personal library (4)** „The personal library that would be useful if its usability is improved“ *„Personal library sounds interesting but didn´t work form me.„* **Expand export function- more formats (4)** „Export should be possible in pdf“ ***Expand filter content of the search filters (8)*** „The filter approach is very good and is interesting if the content is expanded.“ „The idea to restrict by language is good if more content would be available„ **Improve translation quality (3):** e.g. *„ Low quality of translation“* ***Add PowerPoint filter (3)*** e.g. „Filter by ppt“ **Additional suggestions (9)** *“ Filter „ therapy“ would be useful.”* *„ Indication of publisher sorting would be good (by relevance)“* „For scientific articles- implement a humans/animals filter“ „Option to show reviews only“ „Implementation of a free/Not free filter (sort free first)“ „ Filter by... video would be good“ „Search restriction by Pdf would be good.“ „Better search suggestions and implementation of „did you mean“ Would be great to have the possibility to click away websites that are not interesting.“

Table 12: KPro exploitation potential: Suggestions for improvement on tools and usability.

2.5 Discussion

2.5.1.1 Insight on effectiveness, efficiency, usability and KPro functionalities.

The user evaluation reported in this deliverable was performed to evaluate the efficiency, effectiveness, usability, usefulness and exploitation potential of the final Khresmoi Professional search system. Our findings illustrate that Khresmoi Professional has potential to have positive impact in the clinical routine of physicians. The usefulness of Khresmoi Professional has immensely improved in all dimensions over the last year of the project. The final user tests revealed that most physicians find the information they search for using KPro, regard the system as efficient and appreciate the offered facets and tools. With regard to the KPro functionalities the search filters, the personal library and summary translation received the best feedback. « By category », "by date" and "by language" were the most useful facets. In particular self-employed practitioners profited from translation features and multilingual faceting options.

2.5.1.2 Who is likely to use KPro in the future?

The findings of the Y4 user evaluation confirm and advance earlier research that suggested that physicians vary in resource requirements and information needs [7]. Overall, KPro has proven, in its current form, to be most useful for research physicians, physicians in training and hospital clinicians. Physicians in training, hospital clinicians and specialists reported the highest likelihood of future use. Features that helped restrict the results by country, language and target group were regarded as useful by general practitioners. Perceived usability was primarily predicted by the age of physicians. Physicians younger than 40 years have attributed excellent usability to the system. Older physicians were more likely to struggle with the complexity offered and affected in perceived search efficiency, than younger physicians.

2.5.1.3 Exploitation potential and most important ideas for KPro advancement.

With the development of KPro a strong foundation of a good medical search system has been achieved. Further attention needs to be devoted to system accessibility, ranking and efficiency of search results and the interface. To be of better use for self-employed practitioners, access to more resources with clinical relevance is needed. The implementation of a simplified "quick search" allowing simpler navigation was recommended for less IT competent, older and time-constrained physicians. System accessibility need expansion since most physicians access medical information via mobile, in particular IOS devices.

The following list the ten most important improvements, which are likely to advance and broaden future exploitation potential of the system.

- **Tools**: Advance integration and navigation of the personal library, provide access to pdf guidelines, improve quality of summary translation and integrate export in pdf.
- **Search features:** Integrate a "did you mean" function suggesting similar queries.
- **Search facets**: Expand the content in the search facet "by category", especially access to guidelines and local CME content should be easier, expand image, PPT and video search.
- **Develop the web version of KPro:** Integrate access to the personal library into the web version of Khresmoi.
- **Accessibility**: Allow accessibility from all IOS browsers and all mobile devices.
- **Mobile application:** Advance mobile application to include search facets and the personal library and develop mobile application for IOS devices.
- **Interface**: Simplify the interface to the important tools and functionalities, implement quick search for older and time-constrained physicians. Modernize and adapt the layout of interface the interface to "known system"
- **Ranking:** Improve access to relevant sources by further advances in ranking

- **Resources**: Expand access to resources available in pdf format to include access to more documents with clinical relevance.

2.6 Limitations

We devoted great effort to keeping all methodological procedures standardised to ensure high validity and reliability of the evaluation conducted. Furthermore, the same experimenters carried out face-face user tests across different rounds of user evaluations (Veronika Stefanov (TUW) and Marlene Kritz (GAW)). Limitations of the evaluation include a potential reporting bias due to users being observed and recorded in face-face user evaluation, small sample size for some subgroups of physicians (e.g. research physicians, physicians in training) and bias towards general practitioners in the first round of user evaluation. Furthermore, previous research suggested an inverse relationship between age and global usability scales [1], which was replicated in our study. However, reasons for this relationship have not been clarified and may reflect the fact that the younger generation growing up with computers might find it easier to adapt to a new system.

2.7 Conclusion

Overall, the user evaluation of Khresmoi Professional in Y4 documented substantial system improvements on all dimensions. Unbiased, EU-supported, multilingual access and the availability of Khresmoi functionalities such as the personal library, search facets and summary translation received excellent feedback. Awareness on how different age and occupational physician subgroups react differently to Khresmoi Professional has proven useful. The best response has been obtained from physicians in training, hospital clinicians and research physicians. Self-employed practitioners would profit from quicker accessibility to clinically relevant resources. Furthermore, simplification of usability to cater for older, less IT-adept physicians is suggested. Accessibility issues need attention since most physicians access medical information via mobile, in particular IOS devices. KPro is likely to have exploitation potential if it is made accessible on mobile and IOS devices, ranking is improved, access to PDF and clinically relevant resources is extended, usability is simplified, tools are advanced in integration, popular functionalities are implemented in the web browser, and facets expanded in content.

3 Khresmoi for Everyone

The deliverable D10.1 [2] presented the results of the initial search system, the first prototype produced within the project. This deliverable is a follow up of the previous work, conducted this time on a larger population. These extensive tests conducted on the final search system follow the same protocol and methodology as for the previous tests conducted during Year 3 of the project. Thus, the research questions remain mainly the same. The test sessions in Year 4 were carried out to understand if the updated prototype upon the Year 3 evaluation outcomes fits the user needs.

3.1 Research questions

1. Does the new version of the Khresmoi for Everyone (K4E) prototype, version enhanced in Year 3-4, better meet the General public's expectations?
2. Do layman users get better results and have better user experience (in terms of relevance of search results, speed and more comprehensible results) using K4E compared to their previous experience of online health searches? Are they satisfied with K4E results compared with general search engine results?
3. Can K4E be used by different types of users within the general public population?

   In the previous test in 2013, the research question was "Does the outcome of the search correlate with user profiles (age, gender, mother tongue and grasp of English, Internet/web search experience and health knowledge and experience in a given topic)?"

   However, this research question could be perceived as "Is K4E personalized according to the person profile?" During the identification of K4E requirements the "profiling" feature was not seen as a demand, so it was not developed.
4. Usability of the search system:
   a. Which aspects of the system are already "good enough" and are being utilized by users?
   b. Which aspects of the system need to be changed? Which tools and functionalities are not "good enough"?
   c. What is missing?
5. Did the users feel that having a translation service for K4E is important?

   In the previous test in 2013, this research question was presented as follows: "Study the use of translation services (should be based on user's native language and English language abilities in addition to all other demographic and experience factors)."

   This research question was more dedicated to the evaluation conducted in Prague as most of the participants tried the translation service (due to the limited number of results in Czech). However the participants in Paris and in Geneva had no particular translation need during the tests conducted in 2013, especially because no precise task required the usage of it. We decided to slightly change the research question this year, as no specific task required using this feature year either.

In order to answer these questions, we plan to evaluate the following aspects:

- **Effectiveness**: success with task completion (search results relevant to the question asked or not, user able to find information about a given topic or not). Success in solving tasks is presented in section 3.4.2.2. We can also take into account users' feedback after the evaluation

session. Participants were asked if they found the results relevant (SUS questionnaire, question 15, detailed in section 3.4.2.3).

- **Efficiency**: 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). This is presented in section 3.4.2.2. In addition, users gave their impression on how quickly they could find answers when using the K4E search system (question 11 in the SUS questionnaire, detailed in section 3.4.2.3).
- **Usability of the search system**: whether participants use the available tools in K4E 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. Answers from the SUS questionnaire conducted after the evaluation are analysed in section 3.4.2.3.
- **Overall user satisfaction** (like/do not like, what requires change), detailed in section 3.4.2.4.
- **Overall preference to Khresmoi search compared to usual search engine**. Results from the blind comparison between Khresmoi and Google are presented in section 3.4.1. In addition, answers to the question related to how users perceive the quality of K4E results compared to other search engines (question 16 in SUS questionnaire, detailed in section 3.4.2.3) will help in evaluating this aspect.

3.2 The evaluated prototype: Khresmoi for Everyone (K4E)

The deliverable D8.5.2 [5] describes most of the changes conducted in Year 3 and 4 to follow recommendations given after Year 3 user tests (described in D10.1 [2]). A summary of these recommendations, along with improvements made, is presented in Table 13.

Functionality	Recommendations	Improvements made during Y3-4
Classification and filtering in/out	Presentation in the interface has to be clearer in terms of functionality.	One title “Filter by” with 3 types of different filters.
Filters	- Have a checkbox, allowing checking or unchecking various options simultaneously. - All the documents in all languages should be (correctly) classified	- Done, with the possibility to have multiple choices - Automatic classification improved
Disease section filter	Needs to be added for French and Czech websites, and then displayed in the appropriate language.	Done : French and Czech websites are now classified
Keywords cloud	Remove it or at least hide it.	The keywords cloud now offered as a third filter option. Number of keywords reduced to 5.
Disease definitions under the search bar	Add definitions in French and in Czech.	Done: definition in the interface language when available via the machine translation system.

Query assistance	- Taking into account problems with French accents. - Lack of content in Czech for the query suggestions derived.	- Query assistance in French improved. - Czech content coverage improved.
Images	Relevance of images should be increased.	Ranking of images improved.
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. - Add more local healthcare services information in the index as directories of hospitals, pharmacies, physicians, paramedical professionals, etc.	- 109 major Czech resources added. - The index was cleaned when re-crawling all the resources. - Filters are now disabled when no results are obtained. - Directories of hospitals and physicians in Switzerland, France, Czech Republic, Germany and Spain have been added.
Automatic translation	- Improve quality - Simplify presentation - 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 full page where the user is redirected	- Major effort to improve quality - Presentation of the translation feature redesigned - Snippets of original language are displayed while the snippets are translated in the language of the interface - Google translation is proposed to translate the full webpage (since full page translation is out of the scope of the Khresmoi project)
The interface for non-English users	Should be improved, to make clearer what the "Translation" feature actually does, and to make 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.	Presentation of the translation feature redesigned. The current display has been opted after many discussions. No ideal layout option exists to present intuitively this translation option which is not offered by other search engines.
System speed	Should be increased.	Speed increased.
New functionalities		- Vertical red/green bar as an indication of readability.

		- Horizontal red/green bars as an indication of trustability. - Search Pro (semantic search) interface.

Table 13 Summary of 2013 recommendations along with 2014 improvements

In addition to the modifications made and described in D8.5.2 [5], the interface of both Search Lite (basic textual search) and Search Pro (semantic search) have been fine tuned for features such as the display of filters. Additionally, the translation tool presentation has been improved by the HON team.

Figure 65 to Figure ***68*** show an overview of the prototype tested by the participants. Additional details on the available features in K4E are available in the tutorial presented in Annex 6.2.3.

Figure 65 Simple search interface of K4E during the user tests in 2014

The major update is the translation service display and integration. The query is automatically translated into other languages available and displayed on the top of the result list. Figure 65 shows the query has been translated into English. When clicking on the button with the British flag, the user obtained a list of English results with snippets translated into French (language of the interface). Figure 66 shows the original snippet that has been translated (from English to French in this case). It appears when moving the mouse over the translated snippet.

In addition to having access to the original webpage in English (by clicking on the title), the user has now the possibility to display the full webpage translated (by clicking on the French flag near the URL). For full page translation we use Google translate.

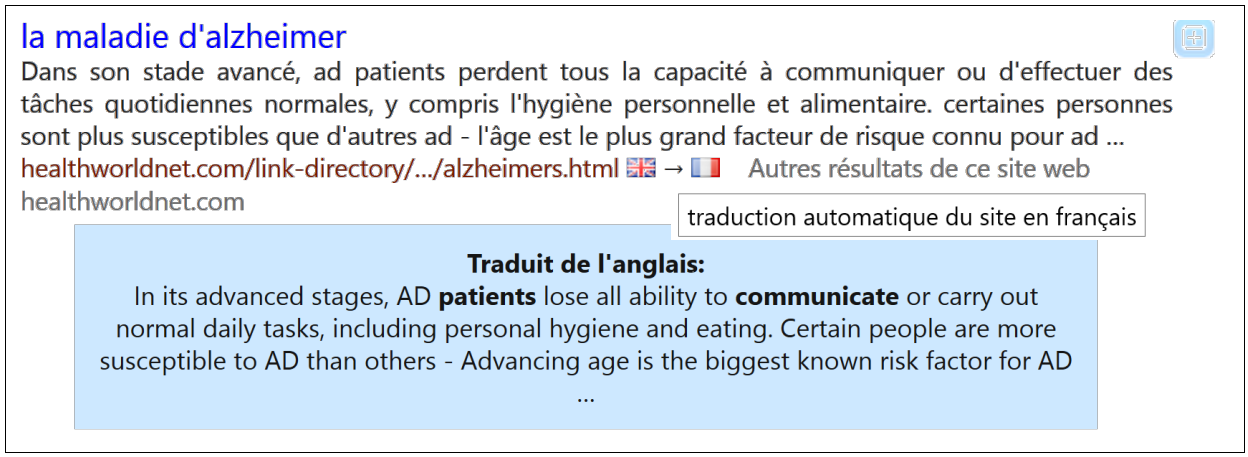


Figure 66 Results of the translation service: snippet translated and display of original content

When available, the definition of the query is presented in a frame on the top of the list of results (Figure 67). The definition is translated into languages provided by CUNI translation system.

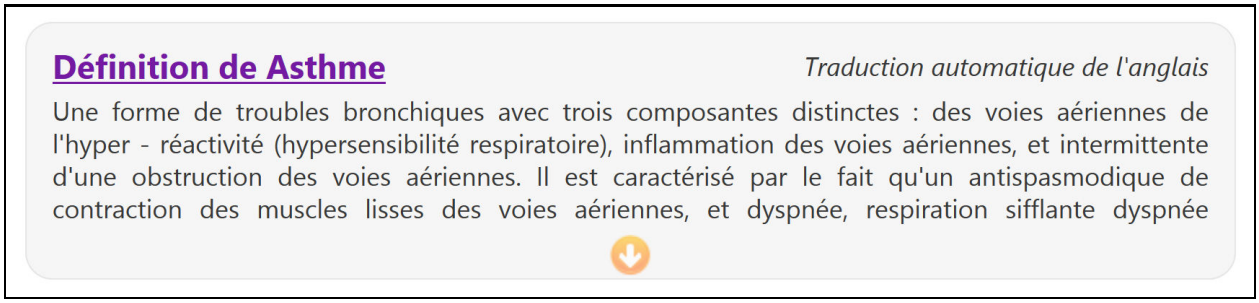


Figure 67 Definition service

In K4E, there are 3 types of filters: a) definition, diagnosis, causes, treatment b) gender and different types of sites, c) cloud of words with a dynamic classification of the results. Filter a) and c) are automatic classification done on statistical process or based on a glossary of terms, which means the term used for the classification appear in the web page. The Filter b) is a manual classification done by an expert, so the term of the class does not necessarily appear in the page content.

Presentation of filters has been enhanced (Figure 68) with a check box allowing cumulating of checked filters and the cloud of terms has been reduced to five terms.

Figure 68 New presentation of the filters and cloud of terms

The semantic search tool, Search Pro, has also been enhanced (cf. Figure 69). Some of the partners (ONTOTEXT, University of Sheffield, HON and CUNI) worked together to improve the system and provide this feature in Czech in addition to English. For now, this feature is not available in French.

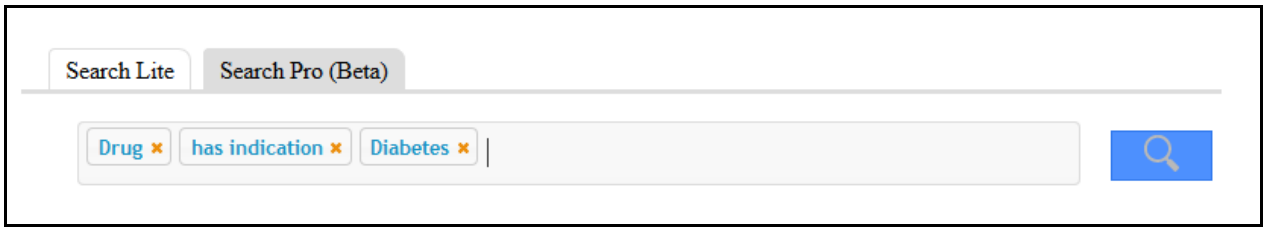


Figure 69 Presentation of the semantic search (Search Pro)

The production version of the prototype is available at: http://everyone.khresmoi.eu/hon-search/.

3.3 Experiment Design

The evaluation tests followed the same procedure as the one conducted in 2013 (detailed in D10.1 [2]). Two stages of evaluation have been proposed:
1) online comparison of search results, and users' preference towards Google or Khresmoi search results. It has been conducted by HON between December 2013 and February 2014.
2) "full user tests" which covered all evaluation aspects. Several tasks and questionnaires have been designed in order to evaluate the Khresmoi search engine. The "full user tests" were conducted by HON, CUNI and ELDA between May and June 2014 in Prague, Geneva and Paris.

The design of experiments is described below and results are presented later in this deliverable (chapter 3.4).

3.3.1 Blind comparison

3.3.1.1 Background

This study is a follow up of blind comparison of search results described in D10.1 [2], section 4.7. The initial study aimed at comparing the top ten search results from two search engines (Google.ch in French and K4E in French) for eight given health tasks/situations. Students of the Faculty of Medicine at the University of Geneva, Switzerland were asked to perform three randomly assigned tasks out of eight available. The evaluation was blind, i.e. participants did not know which search engine the lists came from. It was also conducted in a controlled environment - in one of the rooms of the Library of the Faculty of Medicine during few days over a period of two weeks. The main result of this study is that, despite pronounced quality concerns, participants preferred Google.ch in the majority of cases, most likely because of its wider coverage. In a follow up evaluation we wanted to see whether the fact of being aware of HONcode certification of the search results coming from K4E would influence participants' preferences. Our hypothesis is that once participants/users become familiar with HONcode and trust the web sites holding HONcode certificates, they would give a preference to a search engine providing such results, hence would not compromise quality for quantity.

3.3.1.2 Methodology

For the second stage of the evaluation we changed the method. Main highlights are as follows:

1) Participants were self-recruited via online and email promotion i.e. evaluation was completely online-based with no supervision, it remained completely in French.

2) Each participant was asked to do only one task as opposed to three in the first study.

3) There was no remuneration for participation (in the first study we offered sandwiches and soft drinks to thank participants).

4) The last questionnaire evaluating search results provided by two search engines was significantly simplified. We only asked which list of results they preferred (and give reasons) and specify a maximum of three results from both lists which were the most useful to the participants.

5) Participants were randomly assigned blind or non-blind versions, i.e. some of them were aware which list came from the resources being manually checked against ethical criteria and transparency.

An overview of the task is presented in Annex 6.2.1.

3.3.2 Full user test

3.3.2.1 Background

This study is a follow up of the evaluation described in D10.1 [2], section 4.8. This time, the study was performed on a larger population, 63 participants compared to the 27 during the test on the initial search system. The main characteristics of the search process defined in the D10.1 remain identical. The evaluation goals were completed to understand if the updated prototype fits the users' needs better than the previous prototype (see Section 3.1). Additionally, the same software has been used for the testing: Morae from TechSmith [9]. Only the tasks have been updated.

3.3.2.2 Methodology

Full user tests were conducted with the Morae recording software – similar to physician and radiologists evaluation. We followed the same methodology as in 2013 user tests. We describe the minors changes and improvement made to the procedure below.

3.3.2.2.1 Tasks

Three new tasks have been designed since the evaluation conducted in 2013. Topics for these tasks – passive smoking, Alzheimer's disease and diabetes - are based on the most frequent searches extracted from the logs analysis of the HON search engines. The chosen tasks aimed at reflecting real situations in order to minimize a possible bias in the evaluation. We maintained the starting task, which consists in giving the participants the possibility to perform a free search using the K4E search engine.

An additional task for people speaking English and Czech was added in order to test the semantic search, Search Pro. Some French-speaking participants were able to complete this task, according to their level of English. The topic of this task has been selected with the same process as the other ones. It is related to "hypertension" and "medication".

All the existing functionalities (filters, translation service, visual features) were presented to the user before the evaluation started, to prevent any influence on the usage of functionalities by the task itself.

3.3.2.2.2 Questionnaires

The demographic questionnaire, SUS and success of solving tasks questionnaires were similar to the one presented in 2013. In addition to the 10 standard and 15 specific questions in the SUS questionnaire, two questions linked to the Search Pro tool were added: "I found Search Pro query suggestion useful" and "I found Search Pro results useful".

3.3.2.2.3 Procedure

Each session started with a welcome message, 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 the prototype on their own, performing a typical search they would have usually done,
3. Carry out the four tasks described above (as opposed to three in 2013); after each task they had to respond to a couple of questions (success of solving tasks questions),
4. Fill in the expanded SUS questionnaire which included 10 standard questions and 17 questions specifically designed for Khresmoi.

During the recording, the participant was using a laptop on which the "Recorder" part of Morae was installed and the researcher was using the "Observer" part of Morae on his laptop, both computers being connected through IP address. In addition, the researcher (also called "observer") was adding markers to the recordings (writing observations of all the actions and comments of the participants). Like in 2013, the researcher conducting the evaluation (also called "facilitator") was in charge of presenting the evaluation to the participant and helping in case of technical bugs.

After the recording we asked the participants to sum up one more time their main impressions and feedbacks: what was positive and negative, helpful and distracting etc. Then, the participants were thanked for their time and participation.

The recording was further analysed as presented in Section 3.3.2.3.5 (Post-evaluation process). The Figure 70 gives an overview of a test conducted in Geneva, at the HEGP Hospital.

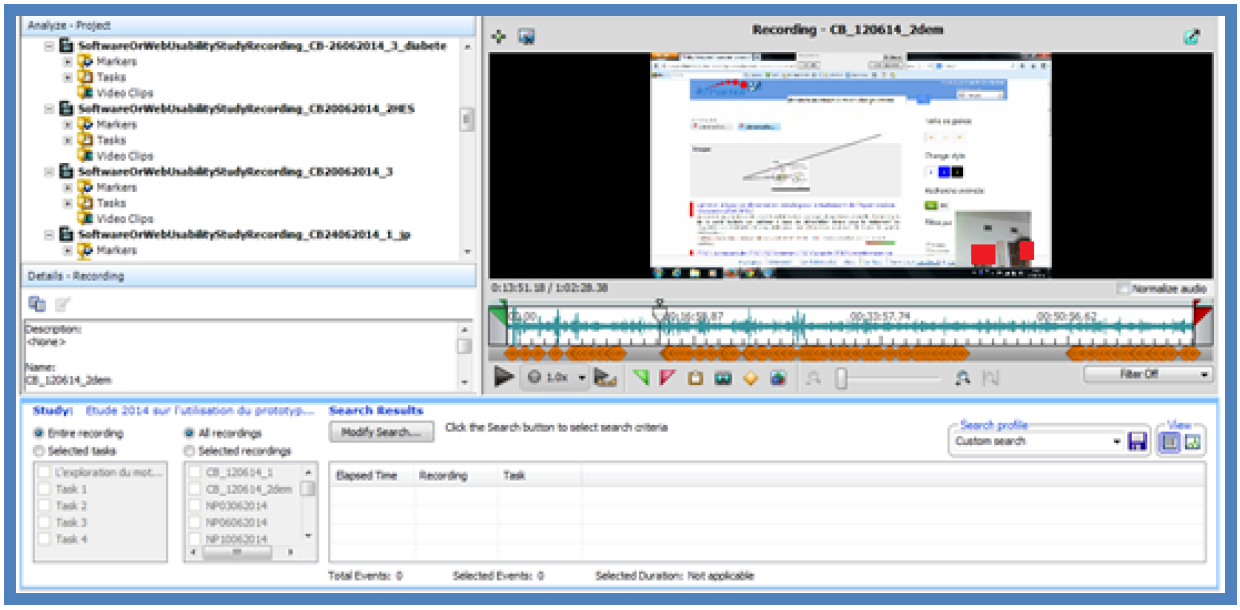


Figure 70 Morae recording of a test conducted in Geneva

3.3.2.3 Setup of the evaluation sessions

3.3.2.3.1 Preparatory tasks

Prior to the evaluation sessions, the following tasks have been performed:

(a) Check the installation of the Morae software.

(b) Localization (translation) of the Morae test configuration file, as provided by the Evaluation group in English, into French and Czech; this included the demographic questionnaire, the SUS questionnaire, and the four task descriptions and their short post-task questionnaires. The files from 2013 were partly reused, but the tasks were new, and there were a few new questions in the demographics as well as in the SUS.

(c) Recording a test session to check the configuration.

(d) Revision of the translation of the information sheet, to be distributed prior to the test, and of the informed consent form, to be distributed and signed by participants prior to the test.

In Geneva and Paris, the following tasks were also performed:

(a) Consent form in French was improved between Year 3 and Year 4 in order to disclose how we will treat the recordings after the end of project and to explain under which law the agreement is concluded. According to guidelines from the CNIL (French Data Protection Authority)[1], a sentence was added indicating that we can give access to their personal recordings upon request. We also informed recordings will be destroyed after the end of the project.

(b) Check the installation of the Morae software on the previously installed computers and install new Observer and Recorder software on two new computers (one in Geneva and one in Paris).

(c) Check the Internet connection: at the Geneva University Hospital we had to connect a hub in order to facilitate the evaluation and guarantee the Internet connection. At the George Pompidou Hospital where the evaluation was conducted at the patient bed side, we had three wireless 3G connection (including one standby spare). The connection was checked the day before starting the test. We noticed that many times, without any warning and no reason, there

[1] http://www.cnil.fr/english/news-and-events/news/article/security-of-personal-data-a-guide-for-action/

were multiple disconnections when the Internet connection was not stable.

In Prague, the following tasks also needed to be performed:

(a) Localization of the 2014 version of the K4E Search interface to Czech (partially done by HON, the rest by CUNI)

(b) Check the installation of the Morae software on the three notebooks used in 2013, update installation (Windows, MS Office, SSH, IE, Chrome, Mozilla, Morae): one for the observer (Zdenka Uresova), one 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 and Jaroslava Hlavacova).

3.3.2.3.2 Recruitment and planification of evaluation sessions

In Geneva, a total of 20 people were recruited and conducted tests. First, participants of last year's evaluation were contacted: 3 persons out of 4 accepted to participate again. We also contacted various different clinics to request their participations, but none of them responded despite our follow up emails.

We used different channels to reach different profiles in order to have the broadest representation of population as possible. Thus, the decision was taken to focus on multiple channels and not restrict ourselves to just a singular one. This strategy was a success as the ***Figure 72*** in the section 3.4.2.1 demonstrates.

Thanks to Professor Ruch and Arnaud Gaudinat, Associate Professor, we were able to diffuse a promotional document at the HES-Geneva intended for students, assistants and researchers of the *Information documentaire HES* curriculum. This curriculum trains librarians and other jobs related to information technologies, and HES-SO is one of the KHRESMOI partners. A few of the tested participants included students of this course who also work part-time. In total, we were able to recruit 3 persons at HES. The small number of participant is due to the fact that end of June is the starting of summer period, so not many students were still there.

To give an idea of the wide range of people contacted, we reached out persons working at the *Télévision Suisse Romande* (a Swiss TV channel); 3 persons participated. Another 3 persons working at the Geneva University Hospital in Health informatics accepted to participate. In addition, we were able to recruit two teenagers and elderly persons using the Internet for health purposes were also contacted. Neighbours were contacted and agreed to participate. All persons accepting to participate were requested to recruit as many acquaintances as possible.

The tests were conducted in various locations: at the premises of the Geneva University Hospital, at HES-Geneva, at the *Télévision Suisse Romande* premises, at home of a few participants when requested and at the Health On the Net premises. We had to adapt to the participants' agenda and so we decided to be flexible in order to be able to recruit as many participants as possible. Providing this additional flexibility considerably improves the acceptance rate from the participants.

All the tests were conducted between June 3 and 25, 2014. Overall, each evaluation session took about 45 minutes to an hour, except when problems with the Internet connection occurred (in connection secured area).

In Paris, a total of 22 people participated in the evaluation. Similar to 2013, we were able to conduct users test at the *Hôpital Européen Georges Pompidou* (HEGP), in the same departments as last year:

- the *Service de Diabétologie, Nutrition et Endocrinologie* (department of Diabetology, Nutrition and Endocrinology), managed by Professor Altman,
- the *Service de Médecine Vasculaire et Hypertension Artérielle* (department of Vascular Medicine and Arterial Hypertension), with the active participation of Dr. Postel-Vinay and Mrs Céline Itié, *cadre de santé* (Health manager), in charge of the supervision of consultations for Nephrology, Arterial

Hypertension (AHT), Vascular Medicine and for the Center of Cardiovascular Preventive Medicine, both at the *Hôpital de jour*[2] (outpatient section) and at the *Hôpital de semaine*[3].

Without the previous contribution of these professionals, none of the evaluation at the Hospital would have been possible. Well in advance we managed to book four days at the Hypertension department and one day at the Diabetes department.

The recruitment of patients was conducted as follows:

a) The Doctor and the Health manager checked the list of patients they had in the outpatient service and in the *Hôpital de semaine* in order to identify potential participants according to their health status and the kind of laboratory tests required. Then a second screening was conducted based on the age of patients, testing the likelihood they used the Internet or not.

b) The physician introduced the Khresmoi user test and asked if the patient wanted to perform the testing of a health search engine. During this demand, Khresmoi researchers were not present in order to avoid influencing the patient. Only a few patients decided not to participate in the study.

c) Once the patients accepted to participate, the observers from HON and ELDA started the user test evaluation, in the room of the patient.

A total of 20 people were recorded in the George Pompidou hospital: 3 of them were patients from the Diabetes department, 16 were patients from the Hypertension department and one was a junior doctor. All the recordings were conducted in June 12, 13, 26 and 27, 2014.

Two other patients who took part in last year's tests accepted to participate again this year. They are members of the AMRO association (French Association Maladie de RENDU-OSLER-WEBER). These evaluations were conducted at the ELDA premises, in Paris, on June 25, 2014.

Typically, each evaluation session took about 45 minutes from patient recruitment to the test itself. Depending on the patients and the medical care some sessions lasted for 35 minutes.

In Prague, there were a total of 21 subjects recruited. Similarly to 2013, we used staff and researchers of several age groups and genders from the Faculty of Mathematics and Physics, Charles University in Prague. The youngest participant was 24 and the oldest 78 years old; education varied from vocational/technical school or college level to Ph.D. All of them, except for one long-term resident, were Czech native speakers (the one remaining was a Russian native speaker, with over 12 years of residency and very good Czech, even though an accent was displayed when speaking), with several being highly fluent in English. 9 subjects out of the 21 took part in the 2013 tests, and therefore had some memory and experience, albeit weak, since they have not used any version of K4E since the past year.

All the recordings were made at Charles University, between May 17 and 25, 2014. Typically, an hour and half was necessary for the whole process including going through the forms, explaining the initial setup, doing the recording, and organizing the paperwork and the machines (making file backups etc.) after the test.

3.3.2.3.3 Technical setup

In Geneva and Paris

The setup was identical to the 2013 K4E user tests in Paris. The observer was alone with the participant. We decided that the facilitator was not needed and that the observer could very easily represent the facilitator when needed. The facilitator was only present for a few recordings.

[2] The *Hôpital de jour* includes patients who come in the morning and leave the hospital around 5pm.

[3] The *Hôpital de semaine* is for patients who come few days within a week and leave on Friday around 5pm.

In Geneva, four laptops were used for the tests. One had Morae Recorder, Observer and Manager installed, while two others had Observer software and one had the Recorder. The two laptops with the Recorder were used by the participants along with the mouse and external microphone. Network varied depending on the location, it could either be hospital public WiFi network or the University encrypted network.

In Paris, we had to install two additional computers from ELDA in order to be able to leave two laptops in Geneva for parallel testing. One of the ELDA's computers requested to install Windows to be able to install Morae Observer. We tried to give the participants laptops with the AZERTY keyboard (in Switzerland the keyboard is a QWERTZ). Few participants had to use the QWERTZ and were somehow lost while trying to type the queries. Headphones with external microphones and a mouse were provided to participants. At the George Pompidou Hospital, we used the 3G wireless keys provided by ELDA for accessing K4E on the web and for connecting together the observer and participant's laptops.

During the test the observer was marking up the recording and making observation notes. The observer was seated during the Geneva test and often standing up during the George Pompidou evaluation session as it was not easy to seat in the patient room. The observer was able to communicate directly with the participant.

We had some issues with the Internet connection during the sessions at the George Pompidou Hospital, and we had one case where the headphone speaker did not work so we were not able to record the voice. One time we had to change computers as we were not able to obtain the connexion between the Observer and the Recorder laptops.

In Prague

A special "clean" notebook was used for the subjects themselves, with a fully upgraded version of Windows 7 Professional SP1. The notebook was running only a virus/firewall software, MS Office and a few utilities for transferring files, plus the Morae Recorder Software. We provided the users with a mouse to make pointing more convenient than the notebook's integrated touchpad. This setup was identical to the 2013 K4E user tests.

The observer used her own notebook with the Morae Observer software installed, connected to the subject notebook over the 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 was sitting at her desk, not visible to the subject who was sitting in the middle of the room at a large table, facing the west window. The observer did have the subject in view, but could have overheard if there was communication between the facilitator and the subject.

The facilitator (mostly Jaroslava Hlavacova, substituted for several recordings by Jan Hajic) was sitting next to the subject, and provided assistance by explaining the project, the setup and the interface at the beginning, as well as answering questions during the test. The observer was in charge of distributing the info sheet and the consent form, and in charge of collecting and organizing the signed consent forms. The observer was also checking and preparing the subject's notebook between recordings, transferring the observed files and making backups, as well as solving some technical issues. The facilitator used his own notebook and plain paper for comments, but it was not running the Morae software while conducting the evaluations and it was not connected to the two other notebooks. This setup has been similar to the 2013 setup, except a different room was used (which was used in 2013 for only the first two recordings). The noise and disturbance level have been low this time, so moving the setup to SU1 (as done in 2013) was not necessary.

This setup worked well except the Morae Observer "stalled" often at the beginning of the recording, but resumed after a minute or so, which was not critical and the recording was not been affected (this has been the same issue as in 2013). After several recordings were made, we attempted to correct a typo in one of the Morae questionnaires, which resulted in mismatch when re-importing the files to the Morae Manager. Thus the import became more complicated and since Morae does not have good full

export function, it is highly recommended NOT to change anything in the Morae questionnaires and configuration after the user tests start with the first subject and its recording is completed.

At the Post-evaluation stage, when re-analysing the recordings in June 2014, no bug revealed itself like in 2013, and the analysis and editing of the markers went smoothly.

3.3.2.3.4 Organizational setup and staff

In Geneva, one people at a time was running the tests: four tests were conducted by Natalia Pletneva (HON), three tests were conducted by Vincent Baujard (HON) and the others were conducted by Celia Boyer (HON). The technical part was managed by Vincent Baujard who installed the software and backup all the recordings.

In Paris, the tests were run by one person at a time: Priscille Schneller (ELDA) and Celia Boyer (HON) conducted the tests at the George Pompidou Hospital. Jérémy Leixa (ELDA) conducted the evaluation of one member of the AMRO association.

In Prague, three people were running the Prague tests: Jan Hajic (CUNI) as the technical person, organizer and facilitator, Jaroslava Hlavacova as the facilitator for the most of the tests and Zdenka Uresova (CUNI) as translator, localizer, recruiter, observer, and post-test evaluator. Milan Fucik consulted the check of the notebooks re-used from 2013 and Katerina Stuparicova has handled all necessary administrative tasks.

3.3.2.3.5 Post-evaluation process

In Geneva, all participants were given "thank you" gift cards of CHF 50 (approximately EUR 40) for a local supermarket chain.

In Paris, the only gift given was the KHRESMOI pen. The medical team (nurses and physicians) received Swiss chocolates.

In Prague, all the 21 subjects were rewarded with vouchers to buy books in one of the large bookstore chains in the Czech Republic (NeoLuxor), valued at EUR 11 each.

For all the tests, the observer had to save the recordings and merge the recorder file with the markers file, then to import this file into the Morae Manager software. Next, all the recordings were re-played and markers and comments edited by the observer (first pass). Short summaries were written for each participant recording (while re-playing the recordings when needed).

All the data from the questionnaires were exported from the Morae software to several .csv files for further analysis. The analysis of the data was conducted following the same protocol as last year (see Annex 6.2.2). Each participant has been given an anonymized ID and no mention of their name is displayed in the analysis.

The signed Consent forms have been stored and scanned.

3.4 Results

3.4.1 Blind comparison

Evaluation took place between December 2013 and February 2014 with some email and social media promotion. 124 users accessed the study, but only 22 largely completed the study.

16 participants were from France and 6 others from neighbouring countries (Switzerland, Italy and Germany). Correspondingly, for 20 of them the mother tongue was French. The study covers 15 women and 7 men. Age varied significantly with an average of 51.4 years old, where 15 participants were aged over 50. They represented users with various education levels: 7 had completed technical

schools and 11 in total had university degrees including Bachelors, Masters and PhD. We also asked participants about their occupation, 10 out of 22 turned out to work in health-related areas (nurse, physician etc.). Only 7 out of 22 participants judged their English level as good or very good. Nevertheless, half of the participants indicated they searched in English on a regular basis (at least once a week).

All of the participants were using PC/laptop to connect to the Internet, 8 of them also used mobiles/smartphones and 5 – tablets in addition to PC/laptop (overall 3 participants used all three types of devices). A vast majority of the participants (77%) declared to connect to the Internet a few times a day. When connected, 21 out 22 participants are using search engine at least a few times a week with a vast majority using it on a daily basis. Also, 21 out of 22 participants judged themselves as experienced or having a good level of online search ability.

Regarding online health information and whether participants tend to trust it and be satisfied with its quality, the majority (19 out of 22) did not express a definitive opinion and reported that quality should be improved and online health information varies on case-by-case basis. In an open answer format, participants expressed their opinions on the main criteria to trust online health information. The most popular answer (7 participants shared it) was “reputation” of a website and organization behind it, i.e. hospital, university, governmental agency etc. An equal number of participants (6) mentioned the following as criteria of trust:

1) a website being certified or accredited by an officially recognized state organization

2) when the same information is found across several websites and seems to be objective.

Less popular criteria of trust (and their frequency):

- presence of references (3)
- health professional involved in content editing (do not trust layman forums) (3)
- a website is recommended or further double-checked with a physician or another medical/paramedical professional (2)
- information is validated against original articles in PubMed (1)
- information found only on websites restricted to physicians (1)
- relevance of terms (1)
- number of comments (1)

After filling in questionnaires, participants were presented with one out of eight health situations and given a query and corresponding ten top results retrieved with Google.ch (in French) and K4E (in French). Participants could receive a blind or non-blind version. Eventually, 12 participants completed a blind version, and 10, the non-blind one. Overall, 4 out of the 12 respondents receiving blind scenarios chose Khresmoi, and 6 participants out of the 10 having the non-blind task chose Khresmoi. An overview can be seen in Table 14.

	Blind	Non-blind
K4E	4	6
Google.ch	8	4

Table 14 Repartition of search engine’s choice between blind and non-blind scenarios

As for the scenarios, we can see in Table 15 their repartition amongst the participants.

Scenario	EN translation	Number of occurrences
(1) Maladie ou problème médical spécifique (Goutte)	(1) Disease or specific medical problem (gout)	6
(3) Comment perdre du poids ou contrôler votre poids	(3) How to lose weight or how to control your weight	3
(4) Sécurité alimentaire	(4) Food safety	2
(5) Sécurité des médicaments ou médicaments dont vous avez vu la publicité	(5) Drug safety or drugs you saw advertised	4
(7) Résultats d'examens médicaux	(7) Medical test results	3
(8) Prendre soin d'un proche ou parent âgé	(8) Caring for an ageing relative or friend	4

Table 15 List of scenarios and their occurrence

Due to the fact that there are few participants and many variables such as blind-non/blind and types of scenario, we cannot draw statistically valid conclusions from the data given. Nevertheless, we have analysed the results on a per scenario basis.

The first task was regarding finding an adequate diet for gout. Six participants completed this scenarios, four blind and two non-blind versions. Three of the participants chose Khresmoi (both of those who did non-blind test and one of the four who did the blind test). The only participant previously familiar with the topic chose Khresmoi. The main reason for choosing Google for this scenario was the more specific and appealing search results as some of the Khresmoi results seemed of high quality, but were not accessible and one participant also commented that the Khresmoi list (in a blind test) contained more commercial results. As for specific results, we asked participants to choose the three best results according to them from both lists. Not all of them marked three results which would satisfy their information needs. For example, for this scenario, 6 participants overall indicated 13 choices/"votes" corresponding to 7 search results: 2 in K4E and 5 in Google. The most common chosen results for this scenario were:

1. Second results of Khresmoi from a web site docteurclick.com (4 "votes")
2. Third result of Google from a website sante-medicine-commentcamarche.net (3 votes)

The third task was completed by three participants on the topic of rapid weight loss. All of the participants made a blind test and two of them preferred Google over K4E. One of the participants who choose Google commented that its results were more exact and credible. Three participants attributed 6 votes to 5 search results.

Task number four was dedicated to a topic of food safety and asked participants to find a piece of information about the risks of drinking raw milk. Two participants completed this task, both non-blind versions. The first one chose Google as it seemed to him/her to provide more reliable and scientific results. The second one chose K4E as he/she did not trust Wikipedia as a top Google result. Two participants gave 5 votes to 5 websites, the first votes were given to:

1. K4E first result from e-sante.fr
2. Google seventh result from health.gov.on.ca

Task number five was dedicated to security of medications, and more specifically to a risk posed by excessive painkiller intake. Four participants completed this task, two blind and two non-blind versions. One of the two in each group chose Google, and another one K4E. One of those who chose Google in a blind search commented that the titles were more precise. Four participants gave only 5 votes attributed to 3 search results: two from Google and one from K4E list. The top were:

1. Second K4E result from santelog.com (one first and one second choice)
2. Third Google result from utile.fr (one second and one third choice)

The seventh task was dedicated to a medical test and specifically where to find help for possible AIDS infection in Geneva. Three participants completed this task: one who did non-blind test chose K4E, and two who did blind test chose Google. 6 votes were attributed to 5 results. The only one result which received two coinciding votes was the first one from Google from a website (check-your-lovelife.ch/).

The last, eighth task was dedicated to caring for an elderly person and specifically searching for a nurse who could take care of an elderly family member. This task has been completed by four participants: one who did a blind test chose K4E, as well as one out of three who did a non-blind test. 8 votes were given to 6 results across two lists. The coinciding results were the following:

1. Third results from Google from a web site alamaison.ch (two first choice voices)
2. Second result from K4E from a web site infirmiers.com (one second and one third choice voice).

3.4.2 Full user tests

Between May and June 2014, 63 users participated in the evaluation: 20 in Geneva, 22 in Paris and 21 in Prague. Demographic information on participants is described below and the answers they gave to the success-in-solving-tasks questions and SUS questionnaires are analysed in the following sections.

3.4.2.1 Demographic questionnaire analysis

A total of 63 recordings were taken into account for the analysis. One of these recordings will be excluded from further analysis (presented in section 3.4.2.2 and 3.4.2.3) because the patient did not meet the main inclusion criteria, i.e. searching for online health information at least occasionally: by personal choice, his use of Internet is limited to emails and social networking. We chose to include him in the demographic analysis, because this category of people must be taken into account. Hence in some cases the sample is 63 or 62 persons (indicated by N=63 or N=62).

The participants mostly originated from France[4] (26), from the Czech Republic (19) and Switzerland (12). The remaining 6 people originated from other countries:

- 2 originally from Belgium, but living in Switzerland,
- 1 from Germany but living in France,
- 1 from ex-Yugoslavia but living in France,
- 1 from Slovakia but living in the Czech Republic,
- 1 from Russia but living in the Czech Republic.

4 Geneva being closed to France's border, this explains why some of the participants in Geneva are originally from France.

The sample covers 35 females and 28 males of various age groups: the most prevalent were aged 60-69 (23%) and 20-29 (21%), followed by the groups of 30-39 (19%), 40-49 (17%) and 50-59 (13%). Compared to the 2013 evaluation, the different age groups are more equally represented. This year, we have younger people (below 20) whereas the youngest participants were older than 20 in 2013. We also have one participant older than 80 years old, whereas we did not have any participants of such advanced age in 2013. The largest represented age group in 2013 was 30-39 (37%), and this category is more equally represented this year (19%). Figure 71 shows age repartition in our sample.

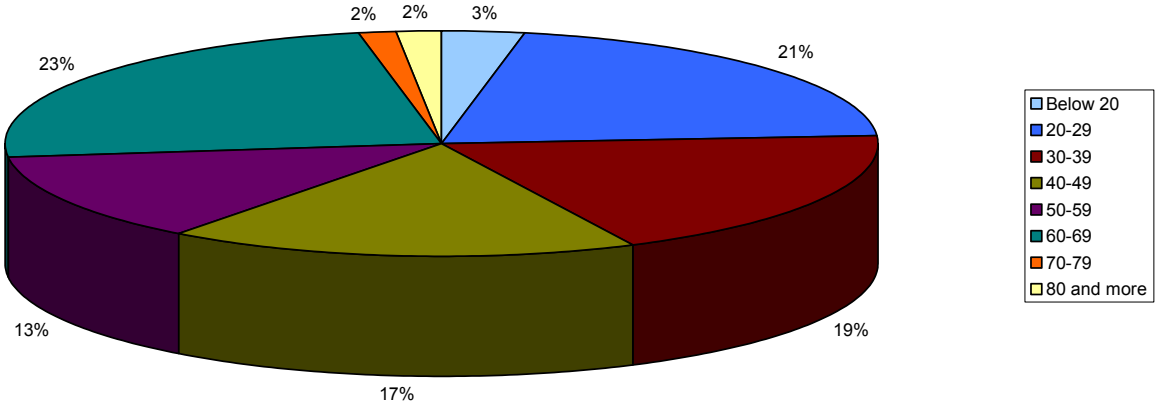


Figure 71 Participants' age groups

If we compare the sample evaluated in Prague and the one evaluated in Paris and Geneva, we notice a different age repartition (see Figure 72). Most of the people aged from 30 to 39 were from Prague, whereas most of the people between 40 and 80 years old were from Paris and Geneva, in addition to the two youngest people (13 and 14 years old).

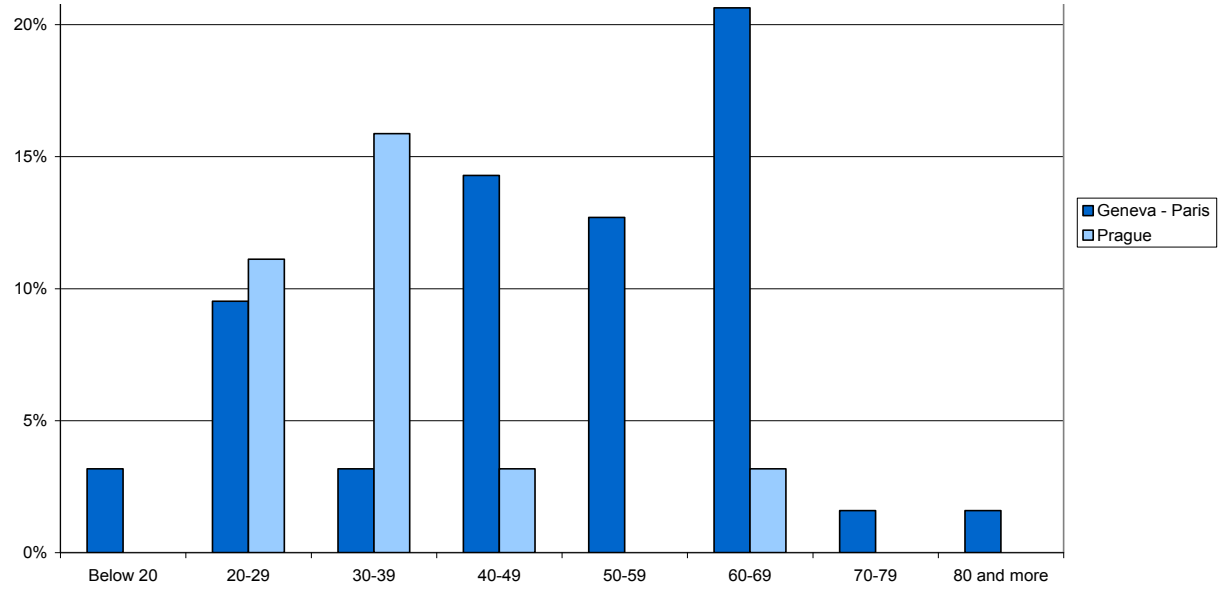


Figure 72 Participants' age groups by location of evaluation

Regarding their level of education, the majority of participants hold a Masters degree (43%), while all education levels were present (see Table 16). Most of the participants from Prague hold a Master degree or a PhD whereas participants from Geneva and Paris are more equally represented (see Figure 73).

High school	11%
Vocational/technical school	16%
University graduate (Bachelor)	14%
Master degree (MSc, MA, MBA etc)	43%
PhD	16%

Table 16 Participants' level of education

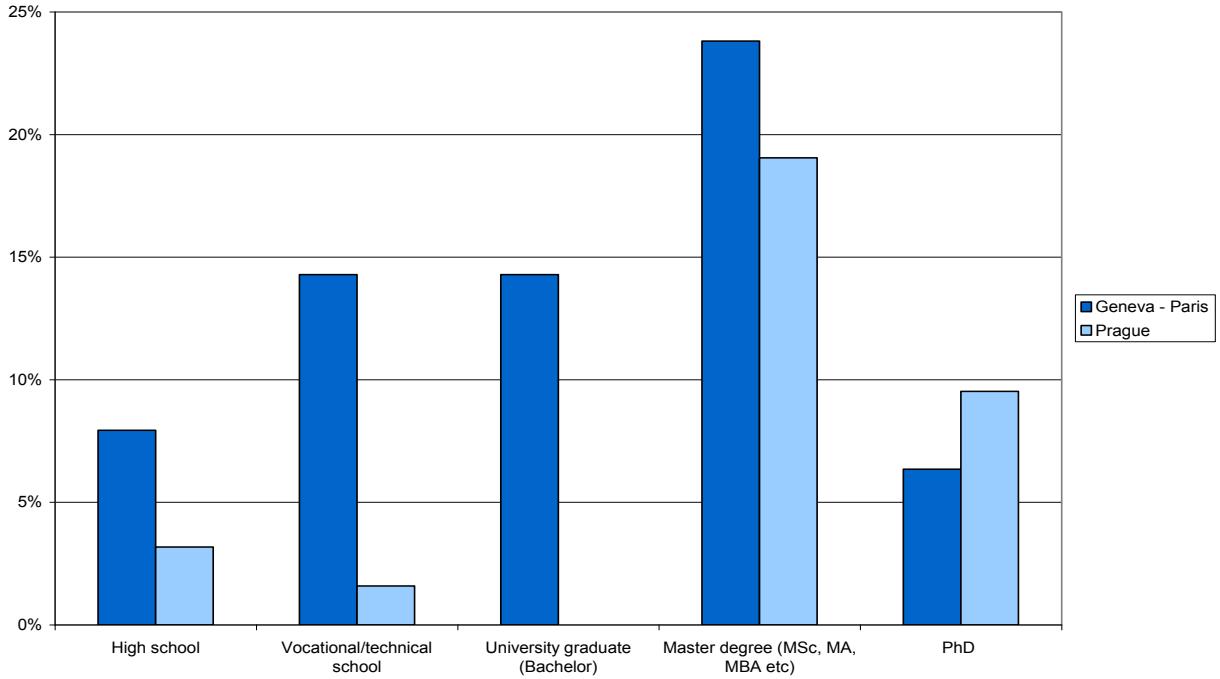


Figure 73 Participants' level of education by location of evaluation

Most of the evaluated participants are working in research (21%) and IT areas (17%)[5]. People working in the Health domain, Business/Trade and Secretariat/public administration domains are equally represented (11%). In the same way, participants working in the Media area have the same representation as teachers or people involved in educational programmes (6%). A few participants are students (3%) and others are working in the field of information technologies/libraries (3%). The 'other' category includes participants working in various domains such as banking, insurance, architecture, translation and craft. The total number of participants includes five retired people, who have been represented in the area in which they used to work. Figure 74 illustrates the participants' distribution per professional domain.

5 The represented segmentation was produced by categorizing answers from participants (open-ended question).

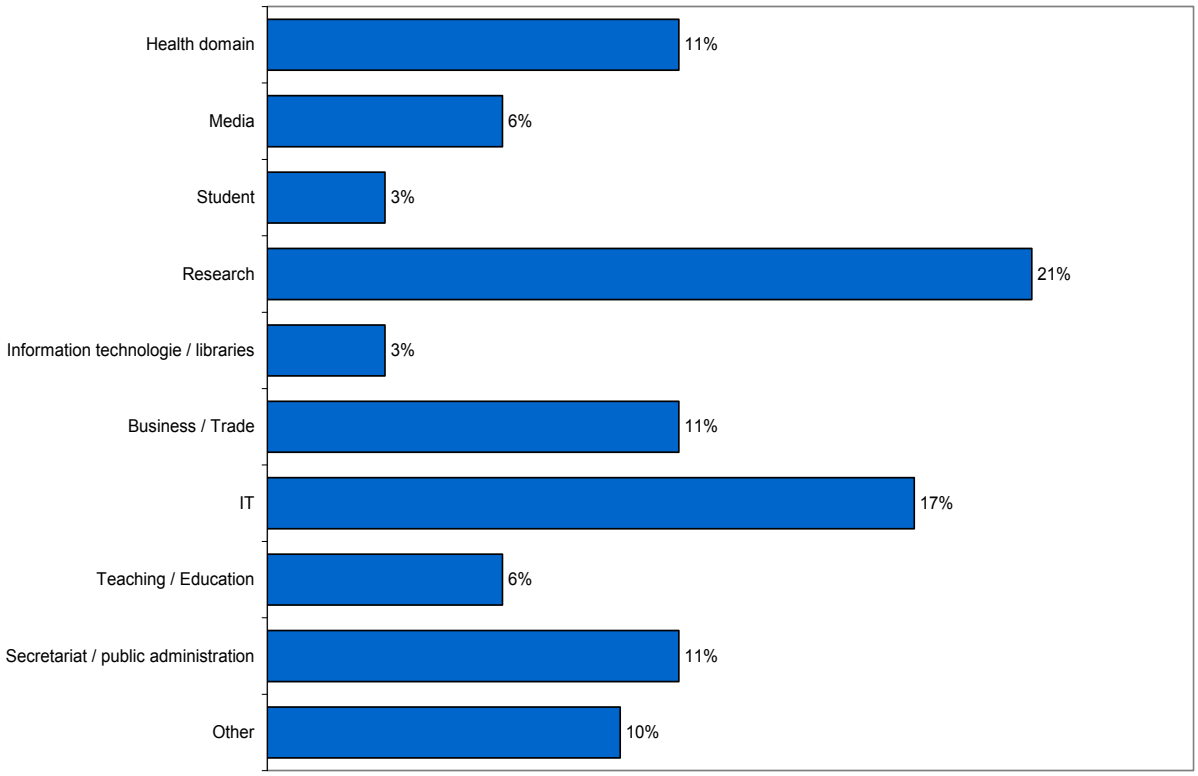


Figure 74 Categorized answers to the question "In which area do you work?"

If we look at each group specifically (Figure 75) we can see:

- that a large part of people evaluated in Geneva are working in IT (20%), Health (15%), Media (15%) and in the Secretariat / public administration domain (15%), the sample also including students (10%) and teachers (10%) and people involved in Information Technologies (5%),
- patients evaluated at the Pompidou Hospital in France are working in various domains: Business / Trade (23%), Health (18%), Secretariat / public administration (14%), IT (9%), Media (5%) and teaching (4,5%),
- people evaluated in Prague are mostly researchers (62%) and IT specialists (24%). The sample also includes librarians, teachers and secretaries.

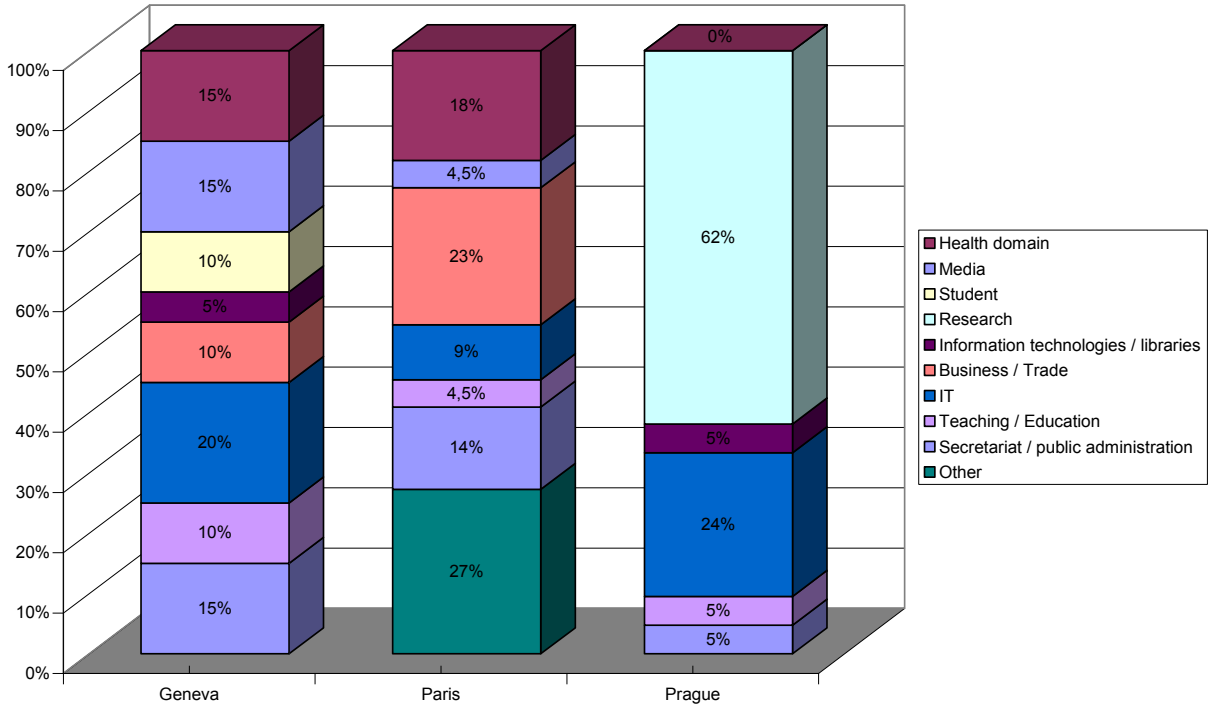


Figure 75 Categorized answers to the question "In which area do you work?" by location of evaluation

The demographic questionnaire also includes questions related to the use of Internet, usage of English in online search and frequence of searches related to the health domain. These are detailed below.

3.4.2.1.1 Internet use

All participants (N=63) were active online: 57 (90%) indicated they used Internet on a daily basis, and 3 (5%) did so several times a week. Two of the participants said they connected to the Internet several times a month only and one of them said he did not use it directly but through other people (friends, family).

To go on the Internet, 98% of the participants use a PC or laptop whereas 2% are using their smartphone only. This question was presented as a multiple choice question: 30% of the participants indicated using a tablet in addition to their PC, and 30% using also a smartphone.

Regarding usage of Internet at work, 89% of them indicated they used the Internet for work or in their studies, while 6% indicated they did not work at the moment and used it for personal inquiries. A group of 5% responded they did not use the Internet for work.

Almost all the participants said they conducted online searches: 53 people every day (representing 84%), 6 others several times a week (10%) and two (3%) several times a month. Two of them (remaining 3%) indicated they did not search for information on the Web: one is not searching directly for online information but asking other people to do so for him, and the other one is not using Internet for searching information (only for emails and social networking).

Everyone (among people looking for online information) indicated Google as the main search engine being used. Three of the Swiss participants indicated using also Bing[6], one using Exalead[7], and another

[6] http://www.bing.com/

[7] http://www.exalead.com/search/

one mentioned searching through bibliographic databases. One Czech participant mentioned also using Yahoo[8], another one using Seznam.cz[9] exceptionally. A French participant added DuckDuckGo[10], and another one mentionned Pubmed[11].

Almost half of the participants (49%) said they were very confident with web search and considered themselves as expert users. 43% were slightly less confident, they reported having problems in finding information from time to time. Only one person (2%) judged her skills as average and reported often having problems with online search. Four people (6%) declared they were not confident at all, being new-comers to online searches.

3.4.2.1.2 Mother tongues and usage of English for online searches

Among the participants, 62% stated their mother tongue as French, 30% as Czech. The other participants reported the following mother tongues: German, Serbo-Croatian, German and Greek (bilingual), Slovak and Russian. The first three communicated in French and the last two in Czech.

When asking for their level of English, 9% declared having no knowledge at all, and 30% being fluent. The intermediary levels are presented in Table 17:

No knowledge	9%
Basic	13%
Average	21%
Good	27%
Fluent	30%

Table 17 Participants' level of English

Almost half of the participants (44%) reported searching for or reading information in English on the Internet on a daily basis. On the contrary, 18% of them said they never looked for information in English on the Web. The intermediary levels are presented in Table 18:

Everyday	44%
Several times a week	19%
Once a week	3%
Several times a month	8%
Once a month	3%
Rarely	5%
Never	18%

Table 18 Frequency of searches in English

Frequency of searches in English is not equal depending on the location where people were evaluated: most of the Czech participants being from the University, they have a high percentage of people searching online information in English everyday (24% of participants looking for information in English everyday were evaluated in Prague). Users recorded in Geneva tend to be quite used to searches in English, as 14% of the participants who look for information in English are from Switzerland. On the contrary, 16% of the participants who never look for English information on the Web are patients from the Pompidou Hospital in Paris. As shown in Figure 75 presented above, the sample of users recorded in the Czech Republic is made of people with a completely different background from the ones in Paris. Figure 76 shows an overview of answers classified by location where participants were evaluated.

[8] https://www.yahoo.com

[9] http://www.seznam.cz/

[10] https://duckduckgo.com/

[11] http://www.ncbi.nlm.nih.gov/pubmed/

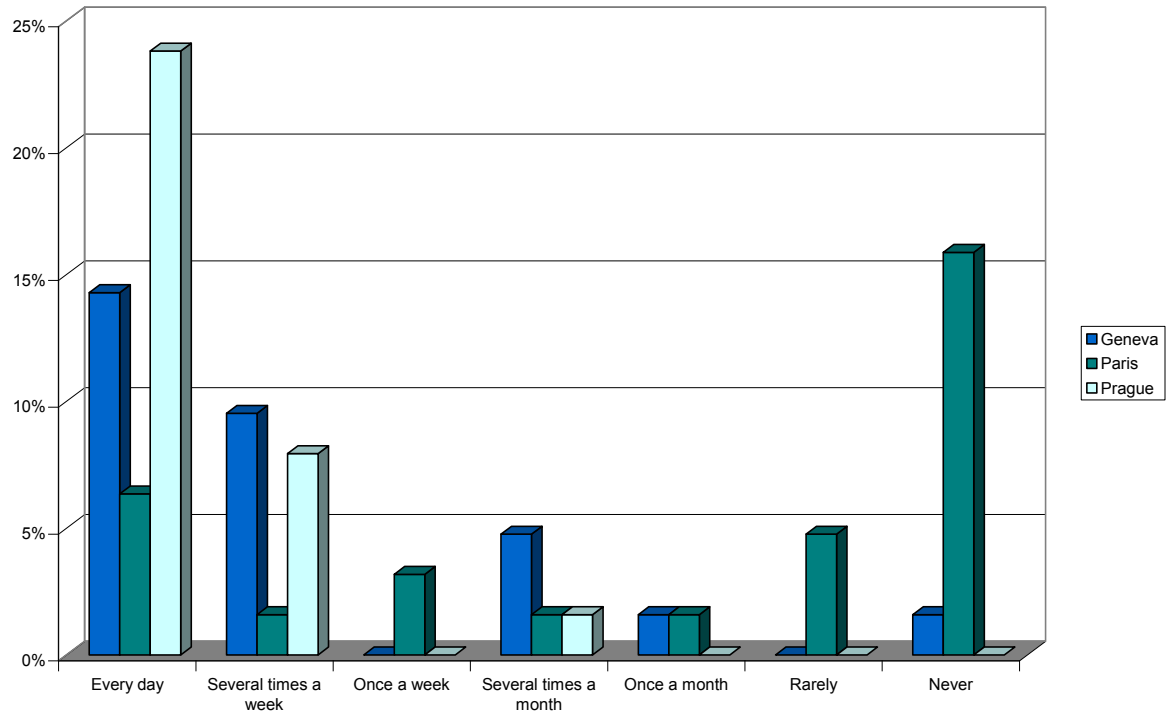


Figure 76 Frequency of searches in English by location of evaluation

3.4.2.1.3 Online health search

Participants were asked how often they search for health information related to them or their family/friends health, and most of them (34%) answered they did it occasionnally (less than once a month). 19% of them reported they looked for medical information at least once a month, 15% several times a month and a cumulative percentage of 29% answered they did it between once a week and every day (details in Table 19). A small percentage of 2% answered they never looked for medical information.

Every Day	6%
Several times a week	10%
Once a week	13%
Several times a month	15%
Once a month	19%
Less than once a month	34%
Never	2%

Table 19 Frequency on which participants search for online health information

All participants gave preference to Google when looking for health content, except for one using exclusively Pubmed[11] and another one who did not know which one he was usually using. This person using Pubmed is the same as the one mentionned in section 3.4.2.1.1. This is one of the French patients, working as medical scientist. The participants had the possibility to give several answers: two Swiss participants also mentioned using Bing[6], two other participants using Pubmed (the junior doctor evaluated at the Pompidou Hospital in Paris and a Swiss participants), one Czech participant using exceptionally Seznam.cz[9] and one person also using HON code Hunt.[12]

[12] http://www.hon.ch/HONsearch/Patients/hunt.html

We then asked participants which types of online health information they were looking for (multiple choice question). As seen from Figure 77 most of them sought for information about a specific disease or medical problem (89%). The second category of retrieved information is related to medical treatments and procedures (54%) and the third one to food safety and recalls (37%).

In the "Other" category, participants mentionned information related to medical centers (opening hours, description of health services), medical indications before travelling (vaccination, etc), alternative medecines, description of drugs and specific searches for work (prevention in collective health for instance).

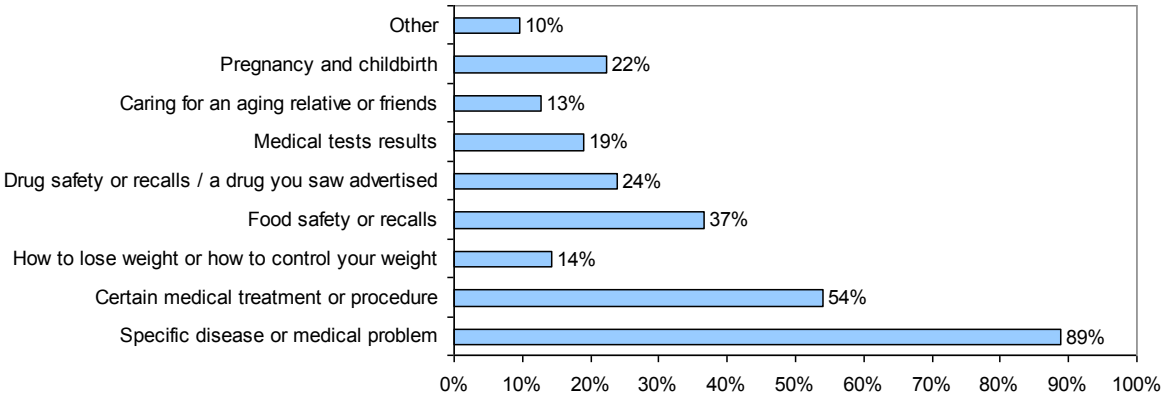


Figure 77 Types of online health information participants are looking for

When asked if they had been diagnosed with chronic or/and life-threatening diseases, 22 respondents out of 63 (35%) answered yes, mostly for more than three years (15 out of 22). As can be foreseen, most of the participants diagnosed with such diseases are the users evaluated in Paris (15 people out of 22). 4 of the 20 participants in Geneva declared having such diseases. This was also the case for 3 out of 21 participants in Prague.

Among people with chronic or/and life-threatening diseases, 68% of them (15 out of 22) reported that their Internet use to search for online health information did not change after the diagnosis, while 7 of them (32%) increased their Internet use in their search for medical information.

We also asked whether the participants had encountered the situation of a family member or a friend being diagnosed with any chronic or life-threatening disease: 32 reported having such experience, while 28 did not. A few participants did not answer this question. Regarding those participants having reported such experience, we do not see much difference between location of user evaluation: 11 out of the 20 participants from Geneva, 12 out of the 22 from Paris and 9 out of the 21 in Prague.

3.4.2.2 Success in solving tasks

Participants had four tasks to solve using the Khresmoi for Everyone prototype. This section describes the tasks and gives an overview of rapidity with task completion and system performance. This is one of the aspects taken into account to measure the system efficiency. Here we have a sample of 62 users (N=62).

The average time spent for each task (summarized in Table 20) is closed to what was expected in the protocol of evaluations, except for the "free search" task, which actually lasted less than expected (9.43 minutes instead of the 15 foreseen). For all the other tasks, the person running the evaluation suggested to stop and go to the next task when users could not find results after 5 minutes. Thus success in solving tasks, which has been measured by questions to participants after each task, can be compared effectively. This also confirmed that the protocol established during the 2013 evaluation was correct in terms of methodology.

	Average time per task (in minutes)	Foreseen time per task (in minutes)
Free task	9.12	15
Task 1	5.19	5
Task 2	5.27	5
Task 3	4.24	5
Task 4	4.23	5

Table 20 Summary of time spent per task

Before starting, participants were asked to perform a typical health search they would normally do, in order to get accustomed to the search prototype. Most of the users did personal searches and some of them tried to use the functionalities showed by the researcher running the evaluation session.

The first task was dedicated to finding information on whether inhaling tobacco smoke potentially damages health of passive smokers or not. After this task we asked participants if they had any experience on searching information about health risks on the Internet and 44% answered yes. Then regarding the success in resolving the task, 94% reported they were able to find the information about the risks of inhaling tobacco smoke using K4E. Some participants had to reformulate their request several times to find relevant results, others rapidly found the more appropriate keywords to find relevant results. For instance a French-speaking participant typed successively *'nocivité tabac'* (tobacco noxiousness), *'dangers tabac'* (tobacco dangers), *'fumée passive'* (passive smoke), then a long request *'que faire si mon compagnon fume à la maison'* (what do to if my partner smokes at home), and again a shorter request *'fumée passive dangereuse?'* (passive smoke dangerous?). He finally found relevant results with this query. Some other people directly typed 'passive tabagism' and found very quickly relevant results.

The second task was concerned with finding how to communicate with patients suffering from Alzheimer's disease. Only 15% of the participants had earlier experience in searching this type of information. At the end, 77% of them were able to find information about communication with elderly patients with Alzheimer's disease using K4E. Again, most of the time, the reason why participants could not find the requested information was because of the keywords they used. Some of them typed general requests (like "Alzheimer communicate") and then narrowed down the search by adding keywords, such as "how to communicate with Alzheimer patients?"

The third task aimed at finding information about glucometers for giving advice to someone who had been diagnosed with diabetes. Participants were asked to find existing devices and how to use them. After the task, they were asked whether they had previous experience in searching information about medical devices on the Internet and 89% answered they did not. A total of 79% considered they were able to find information about glucometers using K4E. We noticed the fact that two sub-questions were included in this task (find types of glucometers and find how to use them) caused confusion. Most of the users were able to find how to use glucometers, however people looking for comparative studies between glucometers could not find relevant information, neither the ones who searched for a list of glucometers available on the market.

The last task was intended to evaluate the semantic search (Search Pro interface). This functionality is available in English and Czech languages only, so 11 French speakers who did not understand either

English or Czech skipped the task. Participants were asked to find all possible diseases where "Lasix" can be prescribed. Among the 51 participants who were able to do this task, 75% of them had a previous experience of searching drug information on the Internet. A total of 61% considered they were able to find information about the use of the drug Lasix for various diseases. However they said they would not have been able to do so without explanations from the person running the evaluation (who explained how to type the request, i.e. using specific vocabulary proposed by the system). In addition, some participants declared not having evaluated the relevance of results (would need advice from a doctor) but only the success in finding results. A few of them did not reply to this question because they said they could not judge the results.

The correlation between previous experience in searching the requested information and success in resolving the task is presented in Figure 78. We see that almost half of the participants had already searched for information on passive tabagism (Task 1). This may have influenced the success in finding relevant information (94% of them declared they succeeded in this task) but we can also imagine this is due to the quantity of results available in K4E on this specific subject. On the contrary, only 15% and 22% of users had previously searched for information requested in Task 2 and 3 respectiveley. At the end, more than three quarters of them succeeded in these tasks, which means the tested prototype allowed to answer the given tasks correctly. Regarding the last task (semantic search), success in solving the task was less important than for tasks on the simple search, although 75% of the users had already searched for drug information on the Internet. This is obviously due to the fact that this is a completely different way of looking for information: keywords are autosuggested and cannot be written directly on the search bar (see Figure 69). There is no display of the keywords from the query in summaries of results (on purpose because results are containing keywords semantically related to the keywords in the query), and the layman users cannot really judge if the results are relevant because medical background is needed.

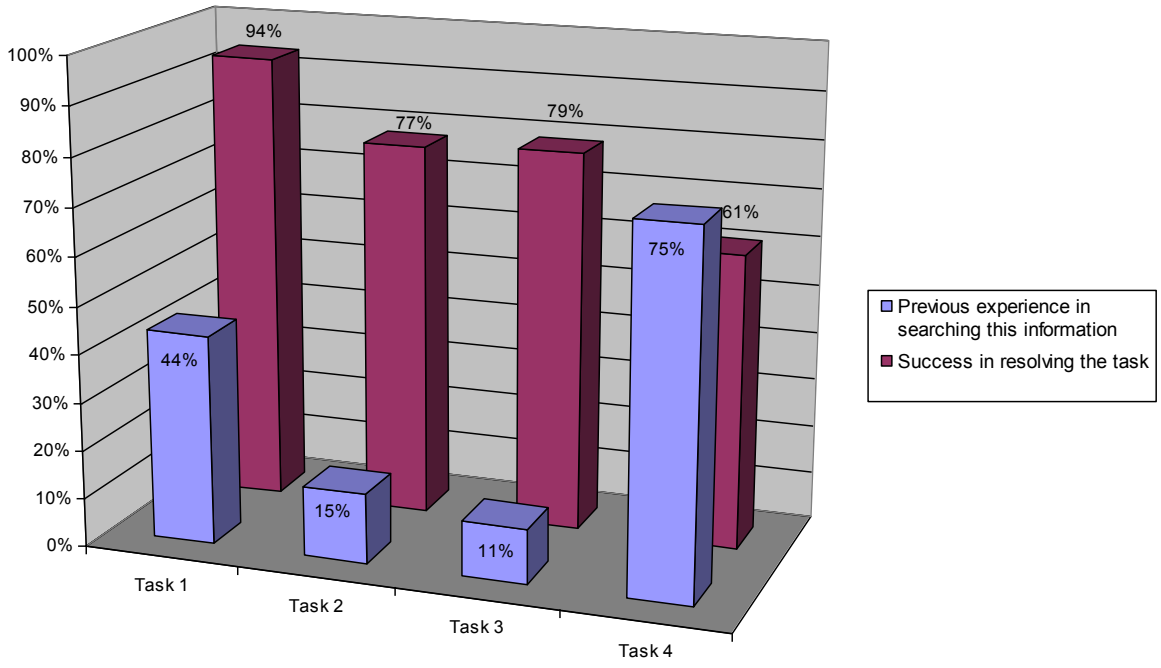


Figure 78 Correlation between having previous experience in searching the requested information and success in solving the task

Finally, most of the users were able to find information about a given topic in a given time. This contributes to demonstrating effectiveness and efficiency of the search system.

3.4.2.3 User satisfaction

A specific questionnaire was developed to measure users' satisfaction with the prototype. It is inspired from the System Usability Scale (SUS) commonly used for measuring perceptions of usability [3] . We used ten standard statements of the SUS questionnaire and 15 specific ones related to Khresmoi (previously added in the last round of evaluation in 2013). In 2014, two additional statements were dedicated to the last task evaluating the semantic search (Search Pro).

Participants had to grade each statement using the Likert scale from one to five, from strongly disagree (1) to strongly agree (5) accordingly.

For the analysis we split all statements into positive (19) and negative items (8). Positive statements reflected positive user feedback, for such statements the higher the grade was, the more satisfied users were with the prototype and their experience. The negative statements reflect disappointment and dissatisfaction with the system when the users are choosing high scores (strongly agree). On the contrary, it shows satisfaction when they scale the statement with a low score.

Here we have a sample of 62 users (N=62).

3.4.2.3.1 Overview of the SUS questionnaire answers

We first present the results collapsed by 'disagree' (scale items 1 and 2), 'neutral' (scale item 3) and 'agree' (scale items 4 and 5). Due to the large number of positive statements we present in Figure 79 the statements related to the system itself, in Figure 80 the statements related to the functionalities of the system, and in Figure 81 the statements more specifically related to the semantic search. Figure 82 shows results for negative statements.

Globally the participants found K4E easy to use (51 out of 62), they thought results they found were understandable (50 out of 62), and agreed with most of the statements presented in Figure 79. The statement which shows less agreement is related to how they perceived the quality of results compared with other search engines. Only 28 out of 58 (4 people did not reply here) perceived a better quality in results from K4E than those of other search engines. 13 users answered neutrally and 17 said they did not perceive such a difference.

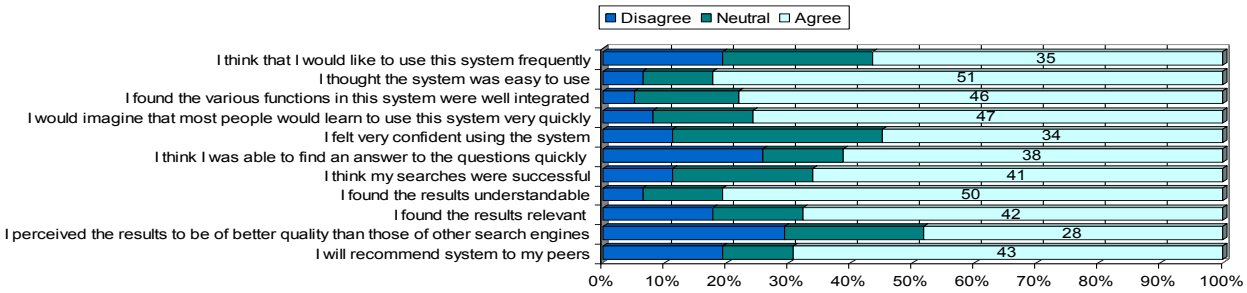


Figure 79 Answers to positive items of SUS (statements related to the system itself)

A majority of participants had positive feedback on the functionalities presented in Figure 80. The most popular feature is automatic translation (appreciated by 43 users out of 55 respondents). Then, participants found the following functionalities useful: assistance in query formulation (38 out of 61), filters (37/51), images (36/61), readability and trustability red/green bars (35/53) and disease definition under the search bar (29/53).

Some users did not reply to these questions because they considered they could not evaluate features they did not use. We added them (in grey) in the representation in Figure 80. Functionalities which seem to have been used less than others are the filters and the disease definition. This can be explained by the task themselves, which did not require the participant to use them: it was not specified to filter results to a certain category, there was no search for specific diseases so the disease definition bar was not displayed, except during the 'free search' task. There may be more people who did not use these

functionalities because the 'Neutral' category includes both people who did not use the features and users who had no categorical opinion on them. However, users who did reply gave their opinion on the usefulness of such functionalities even if they did not use them during the tasks of the evaluation session.

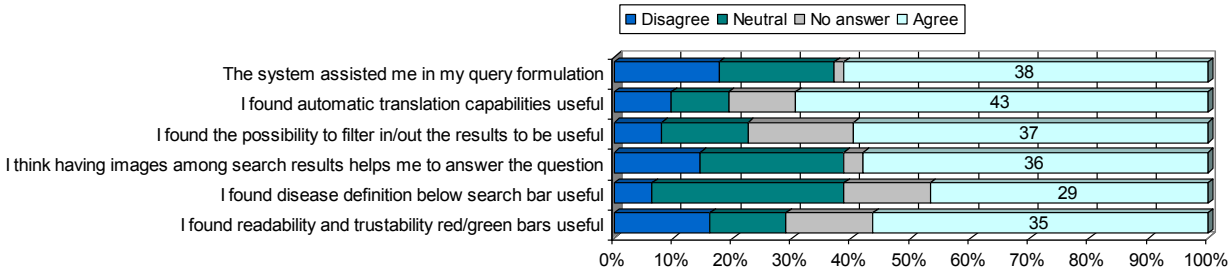


Figure 80 Answers to positive items of SUS (statements related to the system's functionalities)

Regarding the semantic search (Figure 81), we see that most of the participants declared they found Search Pro query suggestion useful (31 out of 48 who did answer) and almost the same proportion found the results useful (29 out of 47 users). However, they insisted on the fact that they would have had difficulties in using it if the observer/facilitator had not explained how to do it before the task. They highly recommend providing a tutorial to end users. Some of the users did not find this feature useful and said this could be useful to medical doctors.

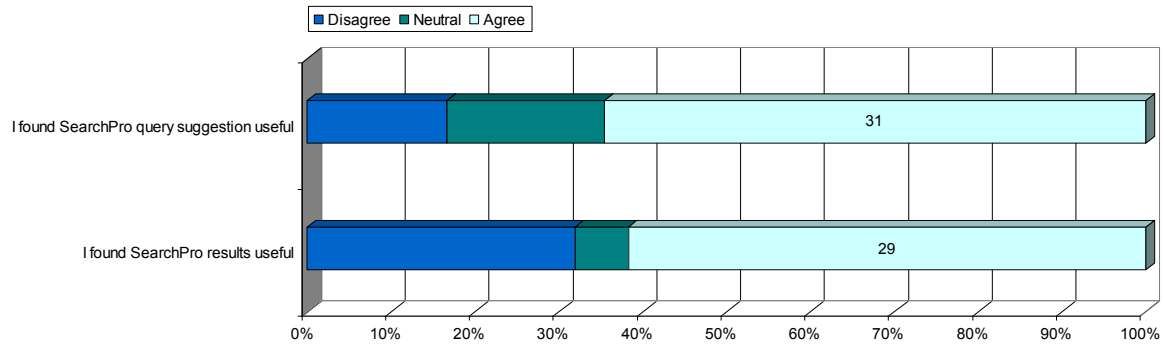


Figure 81 Answers to positive items of SUS (statements related to the semantic search only)

When looking at negative statements presented in Figure 82, we notice that most of the participants disagree with them, which means they were mainly satisfied with the system and their experience. Two statements with higher rate of agreement must be taken into account though, as they show unsatisfactory comments: 'I did not have enough results' and 'I thought there was too much inconsistency in this system'. The analysis presented in the following sections will try to answer why.

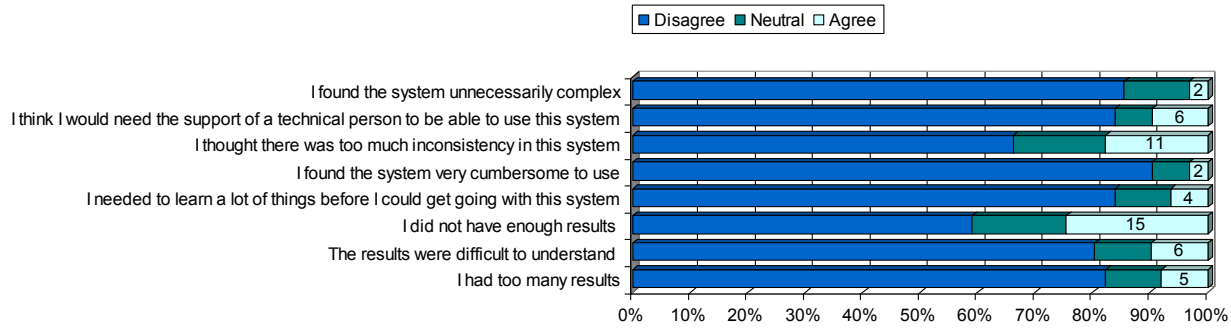


Figure 82 Answers to negative items of SUS (statements related to the system itself)

3.4.2.3.2 Comparison between Czech and French-speaking answers

We split the responses of Czech and Francophone participants and analyzed whether there was a difference between weighted means of the scores for each statement. Table 21 shows in red when there is a significant difference of 20% or more between the Czech and the French population. Positive numbers mean higher grades given by Francophone participants, and the negative correspond to higher grades given by Czech participants.

We can see for instance that the high grade for "I did not have enough results" (statement 13) is mainly due to unsatisfaction of the Czech participants (even if the difference is less than 20% with the French population) as they pointed out the lack of results in their own language.

The main differences between the French-speaking and the Czech population are subject to their final perception of the search engine. Many more French-speaking participants answered they would use the current system and they would recommend it to their peers. Regarding their personal experience with the system, Czech participants were less satisfied than the French and Swiss ones: they retrieved fewer relevant results and the procedure was slower. They declared not to be very confident with the system and not many of them found that K4E results were of better quality than those of other search engines. We will try to answer why there is such a difference between these two categories in section 3.4.2.4.1, based on the feedback from observers/facilitators during the Czech evaluation.

	Statement	Mean (total)	FR	CZ	Difference
1	**I think that I would like to use this system frequently**	**3,63**	**4,15**	**2,62**	**1,53**
2	I found the system unnecessarily complex	1,68	1,54	1,95	-0,42
3	I thought the system was easy to use	4,19	4,34	3,90	0,44
4	I think I would need the support of a technical person to be able to use this system	1,66	1,76	1,48	0,28
5	I found the various functions in this system were well integrated	4,10	4,24	3,86	0,38
6	I thought there was too much inconsistency in this system	2,13	2,02	2,33	-0,31
7	I would imagine that most people would learn to use this system very quickly	4,06	4,15	3,90	0,24
8	I found the system very cumbersome to use	1,53	1,29	2,00	-0,71
9	**I felt very confident using the system**	**3,66**	**4,07**	**2,86**	**1,22**
10	I needed to learn a lot of things before I could get going with this system	1,68	1,66	1,71	-0,06
11	**I think I was able to find an answer to the questions quickly**	**3,48**	**4,12**	**2,24**	**1,88**
12	**I think my searches were successful**	**3,79**	**4,20**	**3,00**	**1,20**
13	I did not have enough results	2,36	1,98	3,10	-1,12
14	I found the results understandable	4,18	4,24	4,05	0,20
15	**I found the results relevant**	**3,74**	**4,17**	**2,90**	**1,27**
16	**I perceived the results to be of better quality than those of other search engines**	**3,26**	**3,70**	**2,48**	**1,23**
17	The results were difficult to understand	1,77	1,65	2,00	-0,35
18	The system assisted me in my query formulation	3,62	4,00	2,90	1,10
19	I had too many results	1,73	1,73	1,71	0,02
20	I found automatic translation capabilities useful	4,07	4,06	4,10	-0,04
21	I found the possibility to filter in/out the results to be useful	4,06	4,07	4,05	0,02
22	I think having images among search results helps me to answer the question	3,78	4,10	3,19	0,91
23	**I found disease definition below search bar useful**	**3,81**	**4,34**	**3,00**	**1,34**
24	I found readability and trustability red/green bars useful	3,91	4,34	3,24	1,11
25	**I will recommend system to my peers**	**3,85**	**4,41**	**2,76**	**1,65**

26	I found Search Pro query suggestion useful	3,73	4,11	3,24	0,87
27	**I found Search Pro results useful**	**3,43**	**4,23**	**2,43**	**1,80**

Table 21 Comparative analysis between the Francophone and the Czech population

3.4.2.3.3 Comparison between 2013 and 2014 answers

Some of the statements scores have increased when compared to those obtained last year. We can see, for instance, the "Query formulation" question, which was evaluated with a score of 2.81 last year while it has obtained 3.62 now.

Table 22 presents positive statements sorted by higher grade given by participants in 2013, compared to the grades given to similar statements in 2014. The most appreciated characteristics in 2014 are compared to 2013 and indicated in bold. Statements which received more than 4 points out of 5 (i.e. tending to "strongly agree") were most appreciated.

We notice more people have found the system easy to use in 2014 (mean of 4.19 in 2014 compared to 4.04 in 2013). A larger number of people declared that the retrieved results were understandable (mean of 4.18 compared to 3.78 in 2013) and that the average score for success in searches slightly improved (+0.01 point). In 2014, more people were satisfied with the integration of functionalities than in 2013, which shows an improvement in the K4E interface. In addition, the translation capabilities and disease definition displayed below the search bar were more appreciated this year.

Positive statement	Corresponding statement number in 2014 SUS	Mean in 2013	Mean in 2014	Mean for FR users in 2014	Mean for CZ users in 2014
Results classification (useful to filter in/out the results)	21	4.15	4.06	4.07	4.05
Most people would learn to use the system quickly	7	4.07	4.06	4.15	3.90
Feeling of confidence when using the system	9	4.04	3.66	4.07	2.86
The system was easy to use	3	4.04	**4.19**	4.34	3.90
Would recommend system to my peers	25	4.00	3.85	4.41	2.76
Found relevant results	15	3.93	3.74	4.17	2.90
Results understandable	14	3.78	**4.18**	4.24	4.05
Found disease definition below search bar useful	23	3.78	**3.81**	4.24	3.00
Successful searches	12	3.78	**3.79**	4.20	3.00
Translation capabilities useful	20	3.67	**4.07**	4.06	4.10
Functionalities well integrated	5	3.48	**4.10**	4.24	3.86
The system assisted me in my query formulation	18	2.83	**3.62**	4.00	2.90

Table 22 Comparative analysis of the positive statements between 2013 and 2014

The same analysis was conducted for negative statements of the SUS questionnaire (see Table 23). Statements which received lower grades in 2014 than in 2013 were considered as more appreciated by participants[13].

Globally, fewer participants found the system unnecessarily complex and cumbersome to use in 2014 than in 2013. They found fewer inconsistencies in K4E this year. The fact that results are considered as understandable (see positive statements above) is confirmed by the lower scale given to the statement 'The results were difficult to understand'. The contradiction between the statements 'I did not have enough results' and 'I had too many results' comes from the fact that this is a subjective point of view. It shows people have various expectations from a search engine.

[13] When the mean of answers decreases, this means more people disagree to the negative statement, hence more people are satisfied. For instance, if someone says he strongly disagrees with the negative statement 'I did not have enough results', it means he is satisfied as he thinks he had enough results.

Negative statements	Corresponding statement number in 2014 SUS	Mean in 2013	Mean in 2014	Mean for FR users in 2014	Mean for CZ users in 2014
I did not have enough results	13	2.59	**2.36**	1.98	3.10
Inconsistency in the system	6	2.41	**2.13**	2.02	2.33
Unnecessarily complex	2	1.93	**1.68**	1,54	1,95
Results difficult to understand	17	1.85	**1.77**	1,65	2,00
I had too many results	19	1.81	**1.73**	1,73	1,71
The system is cumbersome to user	8	1.81	**1.53**	1,29	2,00
Needed to learn a lot of things before using the search	10	1.37	1.68	1,66	1,71
Technical support needed to be able to use the system	4	1.22	1.66	1,76	1,48

Table 23 Comparative analysis of the negative statements between 2013 and 2014

3.4.2.3.4 Global usability score

We also calculated the global usability scale, which gives a global usability measure. It is computed from raw SUS scores converted into percentile ranks (Range 0-100). We used the same method as for the Khresmoi Professional evaluation (see section 2.3.3.1.5). To determine the SUS score, the score contributions for each item were summed (items here are the ten first standard statements presented in Table 21). For items 1, 3, 5, 7 and 9 the score contribution was determined by the scale position minus 1. For items 2, 4, 6, 8 and 10, the score contribution was determined by calculating 5 minus the scale position. The sum of the scores was then multiplied by 2.5 to obtain an overall value. SUS scores had a range of 0 to 100. A SUS Score of 68 is considered as average, a score above 70 is above average and a score above 80 would be classified as an "A" system that is recommended by a friend [10].

We present in Figure 83 a mean of SUS percentile rank for each group by location of evaluation: Geneva, Paris and Prague. This confirms the difference of perception between Francophone and Czech participants: in Geneva and Paris the average score given is respectively 82 and 80, whereas the score is of 69 for Czech users.

However, the global usability score for Czech users is still within the usability satisfaction bound (69, when the average is 68). This shows the K4E system can be considered as 'usable' by the evaluated users.

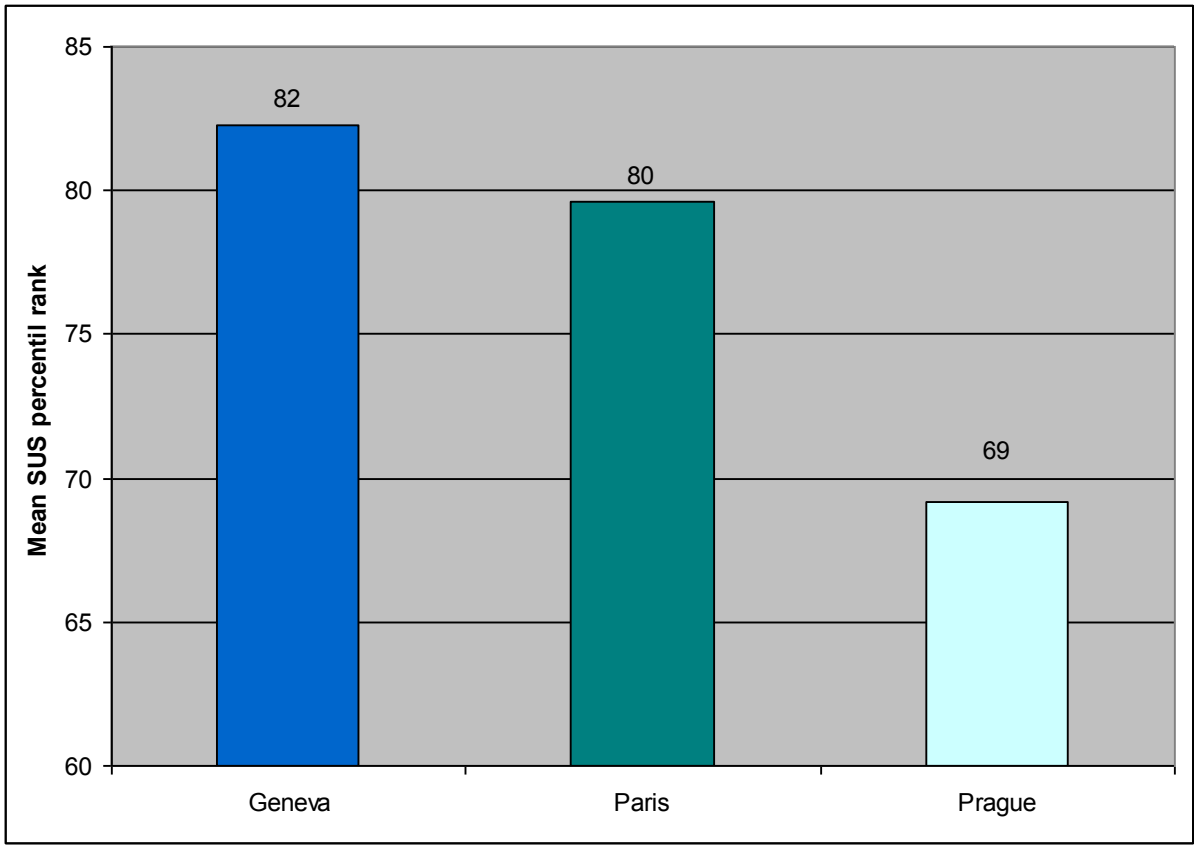


Figure 83 K4E global usability by location of evaluation

Based on these findings, we can conclude that the overall perception was positive, while there is still a need for final tuning of specific features before launching the K4E online and promoting it. Recommendations given by users are summarized later in this Deliverable, in Section 3.6.

3.4.2.4 Overall users feedback

In order to complete the quantitative analysis presented in the section 3.4.2.3, we present qualitative data obtained during the evaluation sessions. This will help analysing users' feedback on the usability of the search system. Due to the language barrier, this part of the results is split in two, depending on the language of the participant and evaluation: Czech-speaking and Francophone.

3.4.2.4.1 Czech evaluations

Overall, the Czech-speaking users have been rather critical to various components of the system, down to small details. In this section, we will provide a summary of the findings based on the detailed summaries for each participant written by the person running the evaluation session, on the results of the Task questionnaires and the final SUS questionnaire, and based on the notes taken by the observer and both facilitators in general.

The feedback will be structured as follows: first, general feedback on content and search results and their properties will be given, then the user interface (UI) feedback will be described including the observed differences in users' perception of the two variants (K4E Lite and K4E Pro) and finally, we will summarize the conclusions, primarily explaining the apparent differences in the SUS between the group of French-language and Czech-language speakers (tests). When referring to the SUS questionnaire, we are using the numbering of SUS questions and the mean scores as presented in Table 21 (p.120).

a. Content, Search Results and Relevance

i. Generally, the Czech subjects complained more often than their French counterparts about the lack of (relevant) contents in the set of search results (despite the fact that the Czech resources have been increased more than 10 times between the 2013 and 2014 user evaluation tests), as can be seen with the mean of only 2.9 for the statement 15 in the SUS questionnaire (*I found the results relevant*).

ii. On the positive side, the Czech users found the system simple ("not complex", statement 2 of the SUS), easy to use (statement 3), usable without the help of a technical person (statement 4 – the Czech mean here is actually better than the French mean score), well integrated (statement 5, even though some people questioned 'what integration means here'). They thought the system was learnable quickly (statement 6), without need to learn many things beforehand (statement 10). They also found the results understandable (statement 14) and were not overwhelmed by them (statement 19). They also found the translation function and result filtering very useful (statements 20 and 21).

iii. Clearly negative was the view of the Czech users on the K4E Search Pro interface (semantic search), as tested by Task 4. While they scored the query suggestion high, many commented verbally and in the SUS comment that this statement is misleading, since the query suggestion method is the only available method to get the results. Also, the difference between the French participants and Czech participants is very large on the Search Pro interface.

iv. Due to the missing definitions of diseases entered in the search window in Czech, the score of the appropriate SUS statement is "neutral" (3.0) due to the formulation of the question; we have learned that it would be more appropriate to allow for "N/A" here. This is apparently the problem of the UMLS dataset, which does not contain definitions in Czech. This problem will be fixed as soon as UMLS official translation to Czech is completed (this is however not part of the Khresmoi project and thus it cannot be influenced from within Khresmoi).

v. The last content-related difference was observed in the "readability" bar, which was not available in documents originally in Czech, since the automatic system for readability assessment was not possible to train on Czech due to lack of manually assessed documents in Czech. This point is external to the project and relatively easy, even if laborious, to fix, since the technology, which is machine-learning based and thus virtually language independent, is available.

vi. Code page identification in the crawled Czech pages was sometimes not working, resulting in wrong diacritics display in the snippets (results page).

b. The User Interface

While the users have been less influenced by the (in)convenience of the User Interface, there was a number of detailed observations which will be listed here for the sake of completeness.

i. The translation, while judged as an important and useful feature, has been implemented in the UI in such a way that every time the page with the results (i.e. with the page snippets) has been shown, the translation has been run anew, even if the page was still the same (i.e. after activating the "Back" button of the browser). This slowed down the experience.

ii. The users appreciated the way the buttons for re-searching using additional pages translated from English has been implemented (after this had been criticized as unintuitive in the 2013 version), with the query translated to English so that it could be immediately seen if the query translation is reasonable (at least for those with some knowledge of English). However, there was a minor problem with this, since the query has been entered on the page where some results are already available, the buttons have not been updated to contain the translation of the current query (the previous query was displayed on them).

iii. UI localization problems: “Governmental organizations” not translated; filter function help missing; “back” button not present on help (FB, Twitter Khresmoi) pages.

iv. Some users suggested they could make use of logical operators in the query, such as “and”, “or” and “not”.

v. Missing help and explanation page for the working and “mechanics” of the K4E Search Pro interface, which has been confusing at least at first, before it was verbally explained to the participants by the facilitator.

To conclude, one thing remains to be explained, or at least an attempt to be made to explain: the difference between the averaged results between the French-speaking user tests participants and the Czech-speaking user tests participants. At first, this phenomenon is confusing, since in the comments made both into the questionnaires and recorded by the observer and facilitator(s), the number of complaints on “bugs” or clear problems of the interface, speed, or complaints about the lack of content in Czech is actually smaller than in 2013. We thus concentrated on other factors: the tasks and the evaluation scenario as a whole, the participants themselves and their demographics, and the environment including the people running the tests (observer, facilitator(s)).

a. The Participants

While some of the participants have been the same in 2013 (8), most of them (13) are new. We have been careful to select people with similar average backgrounds to those in 2013. However, compared with the French participants, most of the Czech participants are well-educated (half with doctorate degree), i.e., their education level is substantially higher than the average of the French participants, inducing differences also in their Internet search experience and level of English knowledge. Other characteristics are comparable (age distribution, gender, experience with diseases, etc.).

b. The Tasks

There were four tasks as opposed to three in 2013, and one of them (the last in the order in which the participants have been asked to perform them) was directed at the use of the K4E Search Pro interface mode. None of the three K4E Search Lite tasks have been as easy as the first task in 2013 (the BMI index query), but overall, they have the same level of difficulty to complete. The K4E Search Pro task has been relatively simple to complete, but as the participants’ comments attest, the behaviour of the system was unexpected to them: most of the comments hinted that even if they understood that the search is semantically-based (and most of those commenting on it even understood the formal “database” and query model behind it, being computer scientists), they would expect a “hybrid” result, where both the diseases cured by the drug Lasix would appear *and* about the drug itself, rather than just getting documents about the diseases themselves without any relation with the drug used to search for them. There was also no way to combine the Lite and the Search Pro search, which could – to a certain extent – circumvent the problem.

c. The Scenario

While the ordering of tasks is assumed irrelevant for the participants’ assessment, in the case of the Lite and Search Pro ordering, it seems that the confusion arising from the limited Search Pro interface has influenced the ratings entered by the participants into the SUS questionnaire, being the last one in the list. This happened despite the fact that the facilitator has explained (at least in most cases) that only two statements in the SUS are specifically related to the Search Pro interface; “the damage” has been done by the participants having to answer the SUS immediately after finishing Task 4, the semantic search one.

d. The Environment

As described in the chapter 3.3.2.3 (Setup of the evaluation sessions), the environment has not changed much, except for the room in which the tests have been made. There was the same notebook on which the tests were run in 2013, the same notebook and software for the observer as well. The

room is a regular office instead of a lab, which we consider to be a more pleasant physical environment which should not have influenced the participants negatively, and certainly more inviting than the environment at a hospital, where many of the French tests were conducted. The observer was the same as in 2013, and since her environment has been the same too, this has most likely no effect on the outcome; also, the observer was not communicating with the subject. A difference was in the facilitator: while in 2013, the facilitator was the same throughout the tests, in 2014 there were two: one, who did a few tests, and a new one assisting in most of the tests.

Summarizing from the points (a) - (d), and also based on verbal comments made by the participants the facilitator(s) and the observer, we have concluded that the following factors contributed substantially to the negative difference between the ratings of the French and Czech participants:

- The major factor was subject's (high) expectation, based on the fact that they were informed both from the Khresmoi project information sheet and again verbally by the facilitator that this is the final evaluation of the Khresmoi project. This has apparently translated to their extremely high expectation on the level of "something substantially better than Google". While the system performed perhaps slightly better, even minor flaws (as listed above, and despite being less substantial compared to 2013) influenced the participants negatively. By comparison, the participants in the French tests have only been told that this is simply a test, and therefore (we believe) their expectation was rather for an experimental system.
- Another substantial factor, according to our opinion, was the average education level of the Czech participants, their prevailing knowledge of English and their occupation (cf. above). Apparently, researchers are more critical than people from other environments, since they are aware of what the system could do, but does not (for whatever reason, not distinguishing that these are often outside of the scope of Khresmoi). The translation quality was apparently not a problem, but their experience with primarily English search with Google seemed to appeal more to them than a specialized Czech system.
- As a third major factor, we have identified the evaluation scenario where the K4E Search Lite interface has been mixed with an additional task with the (much less accomplished, and untested in 2103) K4E Search Pro task. This should have been separated and much more clearly indicated which SUS statements are related to which interface, and perhaps run completely separately.
- Minor contributing factors were also identified: the lack of certain UI and content features for Czech (cf. above – wrong English query translation button label, lack of disease definition and readability score for Czech pages), which became immediately visible when the English pages were invoked; accent problems on the Czech pages; the personality of the new facilitator, who has perhaps been too "frank" in acknowledging system faults to the participants, in some cases reported by the observer as unnecessary.

3.4.2.4.2 Francophone evaluations

Overall, French-speaking users have appreciated the K4E system. As can be seen from Table 21, they declared they would frequently use the system (average score of 4.15/5 at SUS questionnaire) and would recommend K4E to their colleagues and friends (4.41/5). In this section, we will provide a summary of the findings based on the detailed summaries for each participant written by the observer, on the results of the Task questionnaires and the final SUS questionnaire, and based on the notes taken by the observer in general.

a. Content, Search Results and Relevance

French-speaking users appreciated that K4E is a search engine dedicated to medical information only, which avoids pollution in results engendered by non-medical websites. They noticed there is no advertisement and ranking is not influenced by commercial websites.

Most of them found they obtained quicker access to medical information (clustering of webpages from a single website, fewer websites because already filtered).

During the evaluation they estimated having successful searches (average score of 4.20/5). Similarly, they declared having found relevant information (4.17/5) and understandable results (4.24/5).

On the contrary, a few people found some of the selected websites were not serious enough (for instance www.e-sante.fr or www.viesaineetzen.com), some others providing too basic information (vulgarization websites for example).

Some of the participants did not appreciate to have results related to words closed to the keyword typed in the search bar. For example, when using the verb *communiquer* (communicate) in the search, it often mislead results because of a close French word which is *communiqué* (report / release in English). This is due to the fact that keywords are stemmed before retrieving information in order to give as many results as possible. But when the derived words have a different meaning to the original one, this is leading to bad results.

Most of the users said they could not really judge results from the semantic search tool because medical knowledge is needed to check results' relevance. At the *Georges Pompidou* Hospital we had the occasion to have it evaluated by a medical doctor (not included in the evaluated sample because he only tested this functionality). He declared the tool provided relevant results, although partially complete for some specific requests. He recommended to add a disclaimer in order to inform users the retrieved content does not guarantee to cover all the existing answers (for instance results to the request 'drugs + treating + nausea' may not include the new drugs just arriving on the market).

b. The User Interface

The two French patients who participated to the evaluation in 2013 found the overall presentation of the search engine improved a lot in terms of colours (visual aspect showing seriousness), and said organization of information on the page was much clearer.

Some people thought K4E should have a different presentation than Google in order to emphasize the difference. Some others found the system was easy to use because it looks like Google precisely.

Users had various opinions on the existing functionalities and their presentation. We present in Table 24 a summary of what they did appreciate in particular or not, along with their suggestions.

Functionality	Positive feedback	Negative feedback	Suggestions
Accessibility tools	Used by people who needed it (elderly users)		
Filters	• They found this is a useful feature • Someone said the filter 'governmental organizations' will give a wider visibility to such websites.	• People noticed the 'filter' term is not always appearing in the list of results (except for the category of filters related to 'diseases'). • No reset of filters except if the user goes back to the Home page. Not disturbing when the user types a new request on the search bar, but when clicking on the back button of the navigator, there is a problem with the filter which stays clicked although filter is not taken into account.	• Would be interesting to have a filter 'content written by doctors'

Translation system	• Possibility to have the query translated (not available in other search engines). • No comment about the quality of the translation.	• Few users have been blocked after launching the translation while trying to select another feature. • The size of the "translation" button was too narrow and long query translated was nearly impossible to read.	• Have the possibility to correct the translation.
Indication of readability	Many participants found the readability function interesting and important to help them identify adapted results to their level of understanding of the topics.	• Different colours should be better explained with more information in the integrated description (such as 'easy to read' for green bars). • More information should be given in order to explain why the user should trust the result.	• One participant suggested having this functionality by default and not as an advance search tool. • Someone advised to show one star / several stars instead of a vertical bar.
Indication of trustability	Some participants used the trustability indication for choosing results.	• Few find the trust bar with green to red bizarre and not explicit • Explanations displayed when the mouse is over the bar should be translated into French • Someone wondered why we have this tool if all the content is trustable compared to Google.	• A participant suggested it should be taken into account in the ranking.

Table 24 Summary of French-speaking users' feedback on K4E functionalities

c. Feedback on the evaluation itself

Participants appreciated the tasks were related to everyday situations (smoking dangers, Alzheimer disease, diabetes).

None of the participants felt that they were judged on their personal knowledge, as it had been the case in 2013. This is due to a better explanation at the beginning of the evaluation session.

Still a few patients had difficulties in understanding these were fictitious tasks for testing the prototype (saying they already knew the answer to the task and did not need to use the system).

d. Suggestions by users

Many participants recommended adding an explanation of the different functionalities because they felt the features were good and useful, but they would not be able to easily understand the real use by themselves. A participant suggested providing explanations in bubbles when this is the first time someone is using the search system.

In addition, many participants emphasized the need to mention on the KHRESMOI main entry page that the search tool has no advertisement and that they are not influencing its ranking: "Ranking 100% natural", not influenced by additional components other than the query relevance.

Few people said they did not understand the name 'Khresmoi' and thought this should be changed.

A few people suggested adding a functionality for sorting the results (by 'most recent content' for instance, or 'more understandable content', etc.)

3.5 Answering research questions

In the beginning of this chapter we listed main goals of the general public evaluation tests and the main questions to be answered. We attempt to respond to the research questions and judge whether we have achieved the goals initially planned for KHRESMOI in term of search engine adapted to the General Public. It should be noted that the tasks proposed were not prepared to test particular features of K4E. The researcher conducting the evaluation always took around 5 minutes to present all the features of the K4E letting the participants decide what he/she wanted to use during the tasks. The search Pro was not included in the 5 minutes demonstrations.

1. *Does the new version of the Khresmoi For Everyone (K4E) prototype, version enhanced in Year 3-4, better meet the General public's expectations?*

   The participants who conducted both evaluations in Year 3 and 4 were impressed and highly appreciated the evolution of the second K4E prototypes. This is especially true for the evaluations conducted in Paris and Geneva with participants from many different levels of society. In Prague the participants were expecting a better finalized prototype as described in the section 3.4.2.4.1. This can be explained by the intellectual level of this particular population which consists of 62% of researcher and 24% of IT people.

   In general, the system was found easier to use and searches more successful compared to the previous evaluation conducted in 2013 (detailed in section 3.4.2.3.3). Functionalities are considered as better integrated and the interface enhanced in a positive way (looking more serious). Finally, translation capabilities and disease definition displayed below the search bar are more appreciated this year.

   This year, time response was considered as good enough (even with a 3G wireless key), whereas a few people complained about speed last year.

2. *Do layman users get better results and have better user experience (in terms of relevance of search results, speed and more comprehensible results) using K4E compared to their previous experience of online health searches? Are they satisfied with K4E results compared with general search engine results?*

   As demonstrated in the Blind/Non blind study (section 3.4.1) we can say that presenting K4E as a search engine offering trustworthy websites did affect the choice of the participants, i.e. six out of ten participants who made a non-blind test have chosen K4E over Google.ch, while there was only four participants out of twelve who have chosen K4E in a blind test.

   In addition, during the full user test, all the participants found interesting the fact that K4E has only HONcode certified website and selected website (when presenting them the system before to start the evaluation). An average score of 3.85/5 was given to the assumption *I would recommend the system to other people*. We believe this is partly due to the "quality/safe" inclusion of websites in K4E.

   Many participants spontaneously emphasized that with K4E the answer is found without going around too long and without being distracted by other non-medical results or ads. The results

are not manipulated and reflect the query search. So the results are found quicker than for example Google (SUS satisfaction: mean of 3.48/5). As seen in section 3.4.2.2, participants were able in general to find the answers with K4E within a limited time frame.

Moreover, we had few medically knowledgeable users amongst the participants and they appreciated the higher quality of Khresmoi index over a general search engine.

3. *Can K4E be used by different type of users within the general public population?*

   We had participants from many different age ranges (13-85 years old) and they all were able to use K4E and find results adapted to their search during the free task at the beginning of the evaluation.

   Two teenagers were able to evaluate K4E and they found the system to be extremely well-adapted, very intuitive and easy to use. They appreciated the 'safe' feeling they had while using the K4E while with Google, they came across ads not adapted to their age. In addition they also emphasized that they were not distracted by other non-medical results and could find what they were looking for very easily and faster than Google.

   Disease knowledgeable users with chronic disease were using the right medical terms and choosing information with a higher readability level. In some cases, discussion forums added value, in some not. However in both case, participants appreciated that they could filter them out on K4E. We noticed these users tended to use longer queries.

   We could have thought that Internet-experienced users were going to look for more search functionalities. But we noticed this assumption is not totally correct, because it depends more on the level of the education of the person and the type of profession. For example IT people investigate all the functionalities and try to use them. On the contrary, some other people did not try to use specific functionalities (although used to Internet searches). Very often, they tried to use the advanced function when they were not very satisfied with the result on the first page.

4. *Usability of the system*

   a. *Which aspects of the system are already "good enough" and are being utilized by users?*

   Query reformulation with spelling correction was highly utilized and greatly appreciated.

   The auto-completion while typing the query was also very much used and useful.

   The participants used K4E as they use Google and appreciated that it works. For instance the majority of participants used very long queries of more than 5 words, such as a question they would ask a person. This shows the retrieval system is working well.

   The accessibility tools were appreciated and used by aged people.

   Even if filtering was not much used, people appreciated that it exists and that forum results can be isolated.

   The French-speaking participants who typed queries related to specific diseases during the 'free search' task found the disease definitions appearing under the search bar very useful. Few participants wondered why this feature is not provided for all the queries. The Czech participants could not evaluate this feature as it is not yet available in Czech.

   Readability and trustability features were appreciated. Some participants used the trustability indication for choosing results.

   Translations capabilities were appreciated by most of the users, even if few people used it spontaneously. Few asked if it was manual or automatic but there were no comment regarding quality. Most of the users appreciated to have the query translated into their language. They were quite impressed with the quick translation of summaries.

Images were not used by a lot of people, but a majority thought this can be useful (average score of 3.75/5 in SUS questionnaire).

b. *Which aspects of the system need to be changed? Which tools and functionalities are not "good enough"?*

Few users were confused with filtered results because they did not understand why some results were displayed but the keywords (corresponding to the filter) not appearing in the results. For example: if you use the filter "diagnosis" you will see the same term in the results. However when you check the "women" or "senior", "governmental organisation" you will not see these terms in the results as they did not necessarily appear (as different types of filters: some are defined automatically and other are manually annotated, see details in section 3.2).

The readability bar encountered many interests but a few people wonder what level of trust they can have in it because a lot of "difficult to read"[14] documents were retrieved. The different colours should be better explained with more information within the integrated description. Also more information should be given in order to explain them why the user should trust the result.

The trustability bar was found as an interesting feature, but a few people wondered why we have this tool if all the content is trustable compared to Google. This feature could be integrated into the ranking for example.

The semantic search (Search Pro) is not usable as such. In addition to the lack of information on how to type a request with controlled vocabulary, the presentation of results should be enhanced.

c. *What is missing?*

Many users suggested a tutorial should be provided for explaining advanced search functionalities. In addition, the semantic search tool must be previously explained to the users so that they can use it (for writing of the request and explanation of results), and a disclaimer should be added to inform results may not cover all the existing results.

One participant suggested that the system should integrate the choices of webpages visited by other users for ranking the results. This leads to another comment which is taking into account feedback from users. This could be done for the relevance of results (asking users to score relevance of results), but also for the automatic translation system (scoring quality of translation).

5. *Did the users find it important to have a translation service for K4E?*

Even though many participants from Geneva and Paris did not use the translation function, many of them valued this possibility as their level of English was not high or even non-existent, especially for the ones interviewed in the George Pompidou Hospital in Paris.

The ones who used it were impressed by the speed of the translation tool.

In Prague, most participants tried the translation service (i.e., to formulate the query in Czech but get pages originally written in English, translated back to Czech). As it is commented in section 3.4.2.4.1, they considered the quality of the translation sufficient for understanding the snippet, and they considered the quality of the translation of the query (as it always appeared on the UI buttons) to be very good. They also appreciated the ability to access the pages themselves directly and automatically translated to Czech (in this case, by the standard Google translate mechanism), even though this was not considered to be a priority feature due to their relatively high knowledge of English.

[14] The readability vertical bar is red when the document is considered as 'difficult to understand', green when it is considered as 'easy to understand, and orange in-between.

3.6 Conclusions and recommendations for the K4E

In this section we draw overall conclusions from the "full user tests" conducted over the period from May to June 2014 in Prague, Geneva and Paris. It has to be noted that this year the user (general public and real patients) tests have been a success in terms of persons recruited in both Geneva and Paris, a total of 43 participants. To recruit 20 patients within the European George Pompidou Hospital was a real experience for KHRESMOI partners and for the medical team facilitating the user acceptance.

Overall, in comparison with the user tests conducted in 2013 we can see a substantial progress regarding evaluation setup, protocol and prototype. The researchers involved in this year evaluation were able to anticipate technical issue and were more comfortable with the usage of the Morae software. In addition, the K4E prototype in year 4 was well advanced in terms of user experience and functionalities. We were able to anticipate possible technical issue and no re-indexing was proceeded during the period of the evaluation (as it had been the case in 2013).

In our tests, and especially for the French-speaking population the sample of participants was very representative of online health information seekers, i.e. Internet users of both genders, all ages, different education levels and professional backgrounds. They also varied in their web search experience and skills and personal health experience. In this evaluation we had a wide representation of the "normal" population with 8 different types of domain of professional activities (health, media, student, IT, librarian, education, secretariat, architect, bankers) except for the Czech–speaking population with mainly researchers and IT professionals.

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

Having analysed users' feedback and experience we propose the following steps be taken in the near future, before the full online promotion of K4E. Table 25 shows recommendations to improve the basic search interface (Search Lite) and Table 26, steps to be taken so that the semantic search functionality (Search Pro) can be used by the general public.

Functionality	Feedback from user in Y4	New recommendations
Classification and filtering in/out	Sometimes the "filter" term cannot be found in results: for the filter "disease" the chosen terms are found in the results but not for the others filters (targeted people, source of information and keyword cloud)	Explain how the different filters are coming from (manual annotation or automatic classification).
Filters	Filters are staying ticked when you click on the Back button	Technical bug that need to be corrected
Disease definitions	Definitions should be visible for all the queries	Identify how to inform the user that the definition is only available for a disease query and not for a combination of medical terms
Query assistance	- Previous queries should be available as query assistance - The search engine is extremely sensitive to accents and does not retrieve results when missing accents in the query	- Investigate the possibility to use the query of the users in the auto-completion "as Google does". Maybe adding queries from questions on diseases or made in services such as "Ask the doctor" could be proposed to the users. - Allow the search engine to retrieve both results with and without accent and propose a spelling suggestion correcting the missing accent.
Images	Few dead link for images	Find a way to re-rank the image in order to hide the dead images
Index	- Czech results are still limited - Includes dead links	- Czech index to be expanded - Dead links to be hidden by automatic measure or using the selection of results from others users.
Automatic translation	- Users should wait for the translation of all the snippets on the page before to do anything else (change the request, etc.). - Sometimes the translation is bizarre, not possible to propose correction. - Some people are confused as clicking on the title of the snippet (translated) brings to the original webpage. - Some people are confused as clicking on the title of the translated disease definition below the search bar brings to the original English definition. - The size of the "translation" button was too narrow and long query translated was nearly impossible to read.	- The translation should be stopped when the user click on other elements of the interface - The translation option should be editable by adding the translated query in the search box maybe - Move the links providing translation of the full page (flags corresponding to the original and translated languages) near the snippet's title instead of being near the URL on the bottom of the snippet. - A link should be added near the title, for instance "show the English", in order to avoid incoherence. - Enlarge the "translation" buttons according to size of text.
Readability bar	- Different colours should be better explained. - Too many 'difficult to read' documents are identified automatically.	- Improve the "readability level" discrimination between red, green and orange, improve the explanation of the bar (to be integrated). - Check the automatic classification algorithm

Trustability bar	- Not clear for everyone - Details appearing when passing the mouse over are in English only	- Improve the presentation - Translate trustability criteria details into other languages
Interface	Add a clearer link to Home (in order to empty the search bar and reset filters)	Make clearer that Khresmoi logo in upper left corner is a 'Home' link.
Czech interface	UI localization problems: "Governmental organizations" not translated; filter function help missing; "back" button not present on help pages (FB, Twitter Khresmoi)	To be fixed

Table 25 Recommendations for Search Lite (basic search) as outcomes of the user tests in Y4

Functionality	Feedback from user in Y4	New recommendations
SearchPro	Confusion on what can be done with the Search Pro in terms of query and results: - query: Search Pro not able to find free text (only controlled vocabulary) - presentation of results: need to check all the pages of results to have a complete view of answers	- Add an explanation on the specificity of semantic search and a tutorial on how to type requests. - Search Pro query formulation should be enhanced and the type of accepted queries should be clearly shown. - Display a list of results automatically clustered when containing the same keyword, and then link to a list of results for each term.

Table 26 Recommendations for Search Pro (semantic search) as outcomes of the user tests in Y4

4 Acknowledgments

We would like to particularly thank the collaboration of Professor Patrice Degoulet, Director of the Medical Informatics division at the Hôpital Européen Georges-Pompidou (Georges Pompidou European Hospital) in Paris and the Dr. Line Kleinebreil Chair of DESG (Diabetes Education Study Group) and former Diabetologist at the George Pompidou Hospital.

5 References

[1] Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An Empirical Evaluation of the System Usability Scale. International Journal of Human-Computer Interaction, 24:6, 574-594.

[2] Baroz F., Boyer C., Gschwandtner M., Goeuriot,L. Hajic J., Hanbury H., Kritz M. ,Leixa J., Palotti J., Pletneva N. Ruiz de Castañeda R., Sachs A., Samwald M., Schneller P., Stefanov V., Uresova Z. Report on user tests with initial search system. Del. 10.1. Khresmoi. 2013 http://www.khresmoi.eu/assets/Deliverables/WP10/KhresmoiD101updated1113.pdf

[3] Brooke, J. 1996, 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. http://hell.meiert.org/core/pdf/sus.pdf

[4] Boyer C., Gschwandtner M., Hanbury A., Kritz M., Pletneva Natalie, Samwald M., Vargas A. Use case definition including concrete data requirements. Del. 8.2 2013. http://khresmoi.eu/assets/Deliverables/WP8/KhresmoiD82.pdf

[5] D8.5.2 (Dec. 2013), Prototype of a second search system based on feedback, Khresmoi project deliverable, http://www.khresmoi.eu/assets/Deliverables/WP8/KhresmoiD852.pdf

[6] He Y. Missing data analysis using multiple imputation: getting to the heart of the matter. Circ Cardiovasc Qual Outcomes. 2010 Jan;3(1):98-105. http://dx.doi.org/doi: 10.1161/CIRCOUTCOMES.109.875658

[7] Kritz M, Gschwandtner M, Stefanov V, Hanbury A, Samwald M: Utilization and perceived problems of online medical resources and search tools among different groups of European physicians. J Med Internet Res 2013, 15:e122. http://dx.doi.org/doi:10.2196/jmir.2436

[8] Lohmann K. ;Schäffer J. 2013, System Usability Scale (SUS) – An Improved German Translation of the Questionnaire. In Coremedia. http://minds.coremedia.com/2013/09/18/sus-scale-an-improved-german-translation-questionnaire/

[9] Morae usability testing software. 2014. Version 3.3.3 http://www.techsmith.com/morae.html

[10] Sauro J. 2011: Measuring Usability With The System Usability Scale (SUS) http://www.measuringusability.com/sus.php

[11] Sauro, J., & Lewis, J. R 2012. Quantifying the user experience: Practical statistics for user research. Morgan Kaufmann, Waltham MA, USA.

6 Appendix

6.1 Khresmoi Professional

6.1.1 Open responses from November 2013 user evaluation

No.	Profession	Question
1	Self-employed GP	Diagnosis and treatment of Psoriasis?
2	Self-employed GP	Prevention of arteriosclerosis with Amlodipin?
3	Self-employed GP	How long is the incubation period in meningitis?
4	Self-employed Specialist	What is the anatomy of the deltoid ligament?
5	Self-employed GP	Why is there lack of weight gain in breastfed baby?
6	Physician in training	Information on macroangiopathy in Diabetes
7	Physician in training	Obstruction of the bile ducts without increase of ALP?
8	Self-employed GP	New anticoagulants?
9	Self-employed GP	What is CCNU? Effect of Glioblastom as a medication?
10	Other (Research)	Information about usage and effectiveness of Blopress?
11	Self-employed Specialist	Malignant tumor on the spine/Sarkom of spine?
12	Self-employed GP	Information on cholesterol and lipid metabolism.
13	Physician in training	Risk factors of developing depression in old age?
14	Physician in training	Therapy recommendations for chronic back pain?
15	Self-employed GP	Drug interaction between clopidocrel and xarelto?
16	Physician in training	What are the side effects of dapagliflozin?
17	Physician in training	Prevalence of postpartum hemorrhage (PPH) in Austria?
18	Self-employed GP	Association between Vitamin 3 levels and osteoporosis?
19	Self-employed GP	Information on the malignant paraganglion?

Table A1: Questions asked by physicians in the November 2013 evaluation (No.1-19).

No.	Profession	Question
20	Physician in training	Hepatocellular steatosis in HCV infected patients?
21	Self-employed Specialist	Exercise as medication for typ 2 Diabetes?
22	Self-employed GP	Amoxicillin dosage in children with erythema migrans?
23	Self-employed specialist	NAFL, Insulin resistance and expression of P450
24	Physician in training	Best antibiotic for the treatment of Lyme-Borreliosis?
25	Self-employed GP	Is somastatin a suitable treatment for malignant ascites?
26	Self-employed GP	What are treatment options for pavk?
27	Self-employed GP	Looking for patient information on Arthritis?
28	Self-employed GP	What are treatment options for Diabetes insipidus?
29	Self-employed GP	Diagnosis and treatment of anemia?
30	Self-employed GP	Association between Ceramid and Depression? (news)
31	Self-employed GP	Safety of statins safe in patients with myasthenia gravis?
32	Physician in training	Treatment options of Morbus Brechterew?
33	Self-employed GP	How to best diagnose and treat the Sjögren's syndrome?

Table A2: Questions asked by physicians in the November 2013 evaluation (No.20-33).

No.	Question*	Effectivness?
1	Diagnosis and treatment of Psoriasis?	Yes
2	Prevention of arteriosclerosis with Amlodipin?	No, not relevant.
3	How long is the incubation period in meningitis?	No, not relevant.
4	What is the anatomy of the deltoid ligament?	Yes
5	Why is there lack of weight gain in breastfed baby?	No, to specific.
6	Information on macroangiopathy in Diabetes	Yes
7	Obstruction of the bile ducts without increase of ALP?	No, too general.
8	New anticoagulants?	Yes
9	What is CCNU? Effect of Glioblastom as a medication?	Yes
10	Information about usage and effectiveness of Blopress?	Yes
11	Malignant tumor on the spine/Sarkom of spine?	No, too general.
12	Information on cholesterol and lipid metabolism.	Yes
13	Risk factors of developing depression in old age?	Yes
14	Therapy recommendations for chronic back pain?	Yes
15	Drug interaction between clopidocrel and xarelto?	No, not relevant.
16	What are the side effects of dapagliflozin?	No, not relevant.
17	Prevalence of postpartum hemorrhage (PPH) in Austria?	No, not relevant.
18	Association between Vitamin 3 levels and osteoporosis?	No, too general.
19	Information on the malignant paraganglion?	Yes
20	Hepatocellular steatosis in HCV infected patients?	No, not relevant.

Table A3: Individual KPro Effectiveness in the November 2013 evaluation (No.1-20).

No.	Question*	Effectiveness?
21	Exercise as medication for typ 2 Diabetes?	No, not relevant.
22	Amoxicillin dosage in children with erythema migrans?	No, too specific.
23	NAFL, Insulin resistance and expression of P450	No, not relevant.
24	Best antibiotic for the treatment of Lyme-Borreliosis?	No, not relevant.
25	Is somastatin a suitable treatment for malignant ascites?	No, not relevant.
26	What are treatment options for pavk?	No, to specific.
27	Looking for patient information on Arthritis?	Yes
28	What are treatment options for Diabetes insipidus?	No, not relevant.
29	Diagnosis and treatment of anaemia?	No, too specific.
30	Association between Ceramid and Depression? (news)	No, not relevant.
31	Safety of statins safe in patients with myasthenia gravis?	No, not relevant.
32	Treatment options of Morbus Brechterew?	No, too general.
33	How to best diagnose and treat the Sjögren's syndrome?	No, too general.

Table A4: Individual KPro Effectiveness in the November 2013 evaluation (No.21-33).

Responses to "What did you dislike about KPro?"*
The interface: • Interface is too complex for quick search. • Interface in the middle is too narrow. • Interface/display of resources lacks overview. • The window in the middle is too narrow. • Search result platform too small. • Irritated by "unknown" category in the country filter. • Too much information in the preview (abstract enough?). **Relevance of search results:** • Resources either in the wrong language (Spanish) or too specific. • Expected more choice of relevant articles. Lack of news articles, lack local, german resources • Resources too general. More professional content was expected. • Leitlinien.de should not be displayed as information for patients **Navigation:** • Complex navigation. • Navigation requires too much effort. (access should be direct link not only website. • Slow loading time, no possibility of quick content differentiation. • I didn't like the double click when changing tabs, navigation too complex. • Irritated by multiple links from the same main source (compendium.ch). • „Start search" not intuitive, Start search with the mouse is a hassle. • Reset search function: How to delete double click category is not clear/intuitive. • Not intuitive on how to go „back" after doing a focused search via double. • New search but definition of „old" search remains in the preview. • No useful search queries/spelling correction, common search queries). • Automatically translated queries are interesting if positioned less prominently. • Failure of KHRESMOI to do a simple spelling correction (sclechte-→schlechte). • Lack of search input help for "Maligne Wirbelsäulen tu" or spelling correction. • Lack of auto-suggestion/spelling correction for complex terminology. • Spelling correction was useless, as small and only single-worded. • Lack of common queries. • No useful/irrelevant autosuggestions of search are offered. • Suggested queries are not useful, A query suggestion of sopor (two times!) below the word „sport" is not relevant. • No useful query suggestion/ spelling help for complex terminology (myasthenia gravis). **The tools and facets:** • The German "Umlaute" (e.g. ä, ü, ö) are not identified. This makes the German preview/translation unreadable and useless. Not useful translation, since bad quality • Personal library: not user friendly, partially chaotic/lacks overview. • Unusual to separate tags with a instead of a space. • The search filters are not useful if there is no content. • The presented results/filters must be more precise if only 100 search results are offered.
* original responses were translated to English and categorized by Marlene Kritz (GAW)

Table A5: KPro problems reported in November 2013 user evaluation.

Responses to "What suggestions do you have for improving KPro?" *
The interface: • Interface should be simplified and modernized. • The search bar should be more visible. The search field should be bigger. **Resources and ranking:** • In the search ALL words should be considered with AND (as one query/entities). • More relevant results with a bigger choice. • Easier access to overview articles, Make scientific resources more accessible. • Include more local and German articles. Abbreviations should be identified quicker. **Navigation:** • Navigation should be more intuitive. • Automatic setting should be on "show all" not on the more restrictive "all categories. • An option to enlarge the window without making other windows smaller. • A more intuitive "start search" and "stop search" and return should be implemented. **The preview:** • Ideal interface: preview should be less cluttered- without summary and thumbnails. • Common words and text excerpt are useless and should not be shown as it makes the interface unnecessarily complex. Highlighted words in the excerpt should link to the source/query. **The snippet:** • Link description: free/not free (like in „High word Stanford page. • Source in link is essential and should be maintained. Add date integration in link. **The search:** • Translation of difficult words would be useful; translation and similar search/query. correction must be improved. Improve spelling correction especially for rare, complex terms. • Improve relevance of suggested queries. • Integrate similar queries. Add „did you mean...." function like in Google. • Display of similar queries is more important than frequent words. **The tools:** • Usability of Personal library needs to be improved. • Filters should expand in content and precision, narrow date filter. • Implement filter „treatment" (instead of guidelines). • Would suggest direct export into a database. • Export should be possible in pdf, endnote/word plugin. **Additional suggestions:** • Additional categorization based on medical application medical specialization. • Sub-filter: scientific (human/animals). • Option of assigning more weighting to scientific articles. • Free/not free filter, Restricted access/open access filter. Filter to look at reviews only. • A "Feed/RSS" where I can save my search and am mailed for a specified query. • Option of "quick search" to cater for us time-constrained physicians with simple queries. • Button for quick search, encompassing simpler interface. • A KHRESMOI button in drug software/word would be nice/make it more accessible.
* original responses were translated to English and categorized by Marlene Kritz (GAW)

Table A6: Suggestions for KPro improvement in November 2013 evaluation.

Responses to "What motivates you to use KPro in the future?"
The idea • The idea of a free, independent search engine for physicians • EU-supported, restriction to trustworthy resources • Freely accessible search allowing focused search for trustworthy resources • Scientific basis, good tools • Search for medical professionals in the German speaking domain • Independent Information, not sponsored by pharmaceutical industries Multilingual access • Inclusion of German resources • Translation of summary • German search possible The interface and navigation • Good overview • Clear structure of interface • Easy to handle • Good usability Efficiency and resources • Quicker access to good information than in Google • Quick access to trustworthy resources • Integration of PubMed/Google/Web Tools: • Personal library • Translation of summary • Filter options for physicians (guidelines etc.) • Large variety of relevant filters • Restriction via double click is interesting. • Sorting recent resources first
Source: KPro user evaluation, November 2013 . Question: What do you like about KHRESMOI? What motivates you to use KHRESMOI in the future? * Original responses were translated to English and categorized by Marlene Kritz (GAW)

Table A7: Reasons for future use of KPro reported in the November 2013 evaluation.

	Open original feedback translated*
1	Abbreviations should appear fast, quick search
2	For scientific articles- implement a humans/animals filter.
3	Better search results. option to show reviews only, free/Not free filter (sort free first).
4	Not satisfied with result. restrict date category to 3 facets.
5	The restriction of the results must be more specific if only 100 results are displayed.
6	Simpler interface
7	Simpler search platform
8	A bit too complex, irrelevant results
9	I would benefit if there was more focus on scientific articles
10	I did´t like the double click when changing tabs
11	I didn´t like the Umlaute were not displayed correctly which makes the preview useless.
12	Bigger searchbar would be good. Leitlinien.de should not be classified as laypeople.
13	Slow and lack of overview
14	Articles were too specific . I wanted an overview article
15	More german overview articles.
16	Not finished.
17	Interface should be more intuitive. More export format.s
18	Personal library- usability was difficult and lacked overview, results should be more relevant
19	Resources, content. a simpler platform Library is not user-friendly, export in more format.
20	Quick readable articles were missed, simpler platform.
21	Very interesting approach. Wikipedia for physicians- would be great i fit works.
22	Search query was not considered as a whole.
23	Search function for all words, consideration of AND.
24	Search bar should be bigger.
25	Problem with Translation, resources too general and too little content.
26	Low quality of translation. „Quick search for simpler queries should be possible“
27	Tiresome, Too cluttered

Source: KPro user evaluation, November 2013, Question: What did dislike about the system? How can KHRESMOI Professional improve to get more interesting for you? * Original responses were translated to English for this report by Marlene Kritz (GAW).

Table A8: Individual answers on what users disliked about KPro in Nov. 2013

	Open original feedback translated*
1	Personal library, more relevant search results when system is finished.
2	The idea that can be developed further.
3	The idea of an independet, freely accesibly search engine for physicians, filter options.
4	The search filter by category and the personal library is very good.
5	Direct translation possible, organisation in personal library
6	EU-supported, german search. the personal library would be useful if usability is improved,
7	EU- supported/trustworthy
8	EU-supported, independent, no pharma. Good filters.
9	Content restricted search, easy to use
10	A filter „therapy“ would be useful. Easier access to information for physicians.
11	I am waiting for substantial improvements
12	Clear organisation, german
13	Personal library
14	Good tools
15	Quick accesibility to trustworthy resources and high quality medical resources.
16	Quicker access to good information than in google.
17	Has room for development
18	Very detailed (a lot of features)
19	Good overview, user-friendly, quick. Freely accesibly andtrustworthy information.
20	Search for medical professionals in the german-speaking domain
21	Translaiton, sorting by date with recent resources first
22	Independet information, filters relevant for me if the content is expanded. personal library .
23	Indepdendence, free, better scientific functionality
24	I can imagine using it if I make good experience with quick search (e.g. button)
25	If the search is quicker and the filters are better. Results need to more relevant.
26	Indication of publisher sorting would be good (by relevance)

Source: KPro user evaluation, November 2013 . Question: What do you like about KHRESMOI? What motivates you to use KHRESMOI in the future? *Original responses were translated to English for this report by Marlene Kritz (GAW).

Table A9: Individual answers on what users liked about KPro in November 2013.

6.1.2 Open responses from May-July 2014 user evaluation

	Question*	Effectiveness?
1	Chronischer Husten Therapie	Yes
2	Hyperventilation Atem Depression	Yes
3	Bronchial Ca. T3, Bronchial Ca. T3N2M1	Yes
4	ABPA	Yes
5	Akne Rosazea Differentialdiagnosen (Bilder)	No
6	Akromioklavikulargelenkluxation	No
7	Anorexie (retired)	Yes
8	Antiretrovirale Therapiemöglichkeiten bei Kindern	Yes
9	Arteriitis Temporalis Diagnostik	Yes
10	Ass bei sekundär Prophylaxe Apoplexie	No
11	BEMER-Therapie	Yes
12	clostridium diffizile Kolitis ist rezidivierend, Therapie?	Yes
13	Clostridium-diffizile-Infektionen Therapeutische Möglichkeiten	Yes
14	Kolitis ulcerosa Schwangerschaft Therapie	No
15	CPH Therapy	Yes
16	Depressio bei Demenz	Yes
17	Diabetes mellitus adiponectin	No
18	Diabetische Nephropathie	No
19	Diagnostik bei Morbus Cushing	Yes
20	Facies Rubra Chusing	No
Source: KPro User evaluation, May-July 2014, Question: Did you find the information you searched for using KPro?		

Table A10: Individual KPro Effectiveness in the May-July 2014 evaluation (No.1-20).

	Question*	Effectivness?
21	haemochromatose aktuelle Leitlinien	Yes
22	Hifu	Yes
23	HPV Impfung bei Erwachsenen	Yes
24	Hyperhidrosis	No
25	Hypotonie im Alter	No
26	Infantile Hämangiome bei Säuglingen- Therapie und Diagnostik (Leitlinien)	Yes
27	Information on chicken pox	Yes
28	Inkubationszeit Pertussis	Yes
29	Lupus erythematosus neue Therapien	Yes
30	Makulpopathie	No
31	Meds im Ausland verordnet und nicht in Deutschland bekannt sind	Yes
32	Medikamente eingeschränlte nierenfunktion- wie reduzieren?	No
33	Ösophaguskarzinomen p53	No
34	phytopharma testestoron	No
35	Wird die Eisenaufnahme durch PPI Einahme blockiert?	Yes
Source: KPro User evaluation, May-July 2014, Question: Did you find the information you searched for using KPro?		

Table A11: Individual KPro Effectiveness in the May-July 2014 evaluation (No.21-35).

	Question*	Effectivness?
36	Präpulsinhibition Schlafentzug	No
37	produktsuche Arzneimitel dermatologie	No
38	Pseudopseudohypoparathyreoidismus (retired)	Yes
39	Psoriasis Laborwerte	Yes
40	Rektumkarzinom	No
41	rheumatoide arthritis etanercept dosierung	Yes
42	schistosomiasis diagnostik und therapie	No
43	Welches antibiotikum in schwangerschaft?	No
44	septische arthritis leitlinie/Labor	No
45	Suche nach Medikament	Yes
46	Therapie bei CPH	No
47	Therapie der Rechtsherzinsuffizienz	No
48	Therapieresistente Clostridium difficile Infektion	Yes
49	treatment neurodermitis	Yes
50	Verlauf psoriasis zumbusch	Yes
51	WIE IST DIE drG VON mci	No
Source: KPro User evaluation, May-July 2014, Question: Did you find the information you searched for using KPro?		

Table A12: Individual KPro Effectiveness in the May-July 2014 evaluation (No.36-51).

	Open original feedback translated*
1	Resources is somewhat limited
2	In filter "guidelines" there was no content relevant to results and missing scientific focus.
3	Limited resources. English query on the same topic has led to more relevant German articles
4	Requires better navigation, Personal library is hard to find, lack of Ipad compatibility, improve ranking of search results, modernize search platform.
5	More information for patients, image that help convey illness to patients.
6	I couldn´t find the guideline I search for.
7	The idea is good but requires further development
8	The platform and icons are a bit small. In the personal libary the date span till 2019 doesnt make sense. Very good are the search filters and the personal library.
9	I liked the simple usability. I suggest more colored distinction of important content.
10	Easy to use. Good possibility of content organisation. Lack of images.
11	Results a bit too specific for a general query. Would have liked more overview articles. The interface is cleary structured but a bit old-fashioned with the use of borders.
12	Results are a bit too scientific. Missing reviews, short articles which have clinical relevance.
13	Missing are more articles with clinical relevance. (Ärztecodex)
14	Search is a bit time-consuming because hard to find the result. Search suggestions are useless.
15	Missing images to dermatological queries and a good image search which is crucial for me as practicioner. Primarily English articles appeared.
16	freely accesible more websites for online cme
17	Include more,scientific resources, more languages, images, presentations and powerpoints.
18	Didnt correct my typing mistake. For the quick search at point-of-care a little bit too time consuming. Am used from Google that I can also make a spelling mistake.
19	I was trying to use the web-function in english, accessing from Austria, but everything remained german. I noticed a strong bias on german/austrian publications.
20	Irrelevant results

Source: KPro user evaluation, May-July 2014.

Question: What did dislike about the system? How can KHRESMOI Professional improve to get more interesting for you?

* Original responses were translated to English for this report by Marlene Kritz (GAW).

Table A13: Individual answers on what users disliked about KPro in the May-July 2014 evaluation (No.1-20).

	Open original feedback translated*
21	Java Version with the extra functions didn´t work for on my device.
22	No relevant information to specific questions.
23	Was able to find the relevant result only via the focused use of search filters.
24	More physician information. Was hoping for more than 2 lay people articles on the topic. Expected physician resources (diagnostic information, therapy, prognosis), image search?
25	Not one relevant website, repeating links, typing mistakes were not corrected.
26	I like that its neutral
27	Not a lot of search results for the german query). Lack of german articles/guidelines.
28	Relevant drug information (Austria Codex) was missed.
29	Relevant articles came at the end.
30	Very good search results.
31	Useless because too specific and mainly english resources.
32	Very good results. Compared it to google which wouldnt recognize the abbreviation so quick since it confuses it with Non-medical information. A little bit too specific but useful
33	More german resource. Would be be able to click away websites that are not interesting.
34	Little relevant search results on the topic.
35	Filters did not have the expected content. Search platform was too complex.
36	Lacks clinical relevance.
37	Too much specific primary literature, repeating links of the same website. Clinical relevance questionable. Personal library couldnt be opened.
38	Lack of guidelines.
39	Resources is somewhat limited

Source: KPro user evaluation, May-July 2014.

Question: What did dislike about the system? How can KHRESMOI Professional improve to get more interesting for you?

* Original responses were translated to English for this report by Marlene Kritz (GAW).

Table A14: Individual answers on what users disliked about KPro in the May-July 2014 evaluation (No.21-39).

	Open original feedback translated*
1	Find everything I am searching for
2	Idea good but need more relevant search results. Filter by ppt ,video would be good.
3	Access exclusively for physicians/medical professional, guidelines very important, images and presentations.
4	Expansion of the search filters. Search restriction by ppt and pdf would be good.
5	Expansion oft the content and a more simple platform would motivate me to use Khresmoi in the future.
6	If search results are more relevant I can imagine using Khresmoi if it stays free.
7	Better search suggestions. Simplification of search platform.
8	The idea to restrict by language is good if more content would be available.
9	Currently free and in the context of my evaluation a reliable search option.
10	The interface has very good overview. Filters are helpful but lack content.
11	The results were good and I liked the filters.
12	Export should be possible in pdf, the personal library and the filters are interesting.
13	I like the filteroption by guidelines, sorting by date and display of the publicationyear in the link.
14	The filters are good. Restriction by date, language, country and category is interesting but ranking should be advanced.
15	The filters are interesting but more articles in german please and more images.
16	The filter approach is very good and is interesting if the content is expanded. The Personal library.
17	freely accesible more websites for online cme
18	Good articles, quick finding of information, search filters

Source: KPro user evaluation, May-July 2014.

Question: What do you like about KHRESMOI? What motivates you to use KHRESMOI in the future?

*Original responses were translated to English for this report by Marlene Kritz (GAW).

Table A15: Individual answers on what users liked about KPro in the May-July 2014 evaluation (No.1-18).

	Open original feedback translated*
19	Good results and professional information. EU-supported.
20	Good approach. Is interesting for me in parallel to other search systems. Fort he exclusive search it would require more search results.
21	I will use it
22	I would already use it now
23	if it remains free
24	No laypeople information, personal library sounds interesting but didn´t work
25	More german content. The personal library.
26	If it contains more information about alternative medicine.
27	More content to important topics.
28	More content. Restriction by language and image search.
29	Multilingual articles. Filters need to be expanded. Restriction by language and category is very interesting.
30	Its unbiased and free from advertisments.
31	Personal library, search filters, restriction by guidelines.
32	Useful was for me the language restriction.
33	Quick retrieval of current guidelines. If there would be more content to select from, the search suggestions would be better, it would be useful. The personal library seems interesting but didn´t work.

Source: KPro user evaluation, May-July 2014.

Question: What do you like about KHRESMOI? What motivates you to use KHRESMOI in the future?

*Original responses were translated to English for this report by Marlene Kritz (GAW).

Table A16: Individual answers on what users liked about KPro in the May-July 2014 evaluation (No.18-33).

6.1.3 Full Questionnaire

Question German/English	Demographics Choice of responses (German/English)	Evaluation
Geschlecht /Gender	○ Männlich/Male ○ Weiblich/Female	All Y4 tests (N=84).
Land/Country	○ Österreich/Austria ○ Deutschland/Germany ○ Schweiz/Switzerland ○ Sonstiges/Other	All Y4 user tests. (N=84)
Ihre Berufsgruppe/ Occupational status	○ Facharzt-Selbstständig/Self-employed specialist ○ Praktischer Arzt-Selbstständig/Self-employed GP. ○ Spitalsarzt-angestellt in Spital, Klinik, Rehabzentrum etc./ Hospital clinician-employed in non-academic institution (e.g. hospital or health institution). ○ Universität-Forschung und Lehre/Research physician- primarily employed in an academic institution or involved in research and/or teaching.	All Y4 user evaluation. (N=84)
Vorwiegende derzeitige Tätigkeit/ Current work	○ Angestellt/Employed ○ Selbstständig/Self-employed ○ Arbeitslos oder Pension/Unemployed/retired	November 2013 user evaluation only. (N=33)
Berufserfahrung in Jahren /Work experience in years.	○ 0-5 ○ 6-10 ○ 10-15 ○ 15 oder	November 2013 user evaluation only. (N=33)
Höchste abgeschlossene akademische Ausbildung/Highest completed education	○ Dr.med. ○ Dr.med.+ Univ.Prof. ○ Dr. med.+PhD	November 2013 and online feedback survey only.(N=33)
Welche(s) Endgeräte verwenden Sie um nach medizinischer Information zu suchen?/Which of the following devices do you use to access online medical information?	○ PC/Desktop ○ Laptop ○ Iphone ○ Smartphone (NOT iphone) ○ IPad ○ Tablet (NOT iPad) ○ Ich suche nicht nach medizinischer Info im Internet./I do not search on the Internet.	Online feedback survey and medical GAW events. (N=33)

Table A17: Questionnaire Part 1a- Demographics.

	Search preferences		
	Question (German/English)	**Choice of responses (German/English)**	**Evaluation**
1	Wie oft nutzen Sie das Internet zur medizinischen Suche? / How often do you use the internet to search for medical information?	○ Ständig/Always ○ Täglich/Daily ○ Gelegentlich/Sometimes ○ Selten/Rarely ○ Nie/Never	Evaluated in November 2013 user evaluation only. (N=33)
2	Wo suchen Sie derzeit im Internet vorwiegend nach medizinischen Informationen? /Where do you currently search online for medical information?	○ Google ○ Google Scholar ○ Wikipedia ○ Pubmed ○ Medizinische Foren/Medical forums ○ Ich nutze das Internet./I never use the Internet.	
3	Wonach suchen Sie am häufigsten?/What do you search for frequently?	○ Medikamenteninformation/Drug information ○ News/News ○ Wissenschaftliche Artikel/Scientific articles ○ Leitlinien zur Behandlung von Krankheiten/Guidelines for treatment ○ Diagnosehilfestellungen/Diagnostic information ○ Information für Patienten/Information for patients,	
4	Nach welcher Eigenschaft sortieren Sie ihre Information am häufigsten? /Based on what feature(s) do you prefer to your sort information?	○ Datum/Date ○ Relevanz/Relevance ○ Vertrauenswürdigkeit/Trustworthiness	
5	Nach welchen Eigenschaften möchten Sie die dargestellte Information differenzieren können? / Based on what feature(s) do you prefer to categorise information?	○ Inhalt/Content ○ Art der Quelle/Source ○ Herkunft/Author ○ Format/Format	
6	Welche Information benötigen Sie in der Vorschau um zu entscheiden ob die Quelle relevant und vertrauenswürdig ist? /Which information do you need in the link to determine whether a source is relevant and/or trustworthy?	○ Hervorhebung der Suchwörter ○ Den genauen Link/The link ○ Den Titel des Dokuments/The Title of the document ○ Die Zielgruppe/The target group ○ Autor oder Quelle/The author/source ○ Das Datum der Veröffentlichung oder der letzen Modifikation/The date or publication or last modification.	Evaluated in all tests except the online feedback survey. (N=52)

Table A18: Questionaire Part 1b- Search preferences.

	Task description and related questions for free browsing task		Evaluation
	Question: German/English	**Response choices: German/English**	
1	a) For face-face user tests: • Beschreiben Sie eine medizinische Fragestellung die Sie in letzter Zeit hatten mit einigen Worten./Please describe a medical question you recently had in your own words. • Bitte nutzen Sie KHRESMOI um die Antwort auf die soeben formulierte Frage zu finden! (= FREE TASK)/Please use Khresmoi to find the answer to the question you have just formulated. b) For online Version: Nach welcher Information haben Sie mit KHRESMOI Professional gesucht? /What information did you look for using Khresmoi?	Physician formulates question and then tries out KPro	Used and evaluated for all Y4 user evaluation. (N=84)
3	**Effectivness**: Haben Sie die Information gefunden die Sie gesucht haben? /Did you find the information you searched for?	○ Ja, die Resultate waren relevant und nützlich/Yes, the results were relevant and useful. ○ Nein die Resultate waren zu spezifisch/ NO, the search results were too specific. ○ Nein, die Resultate waren zu allgemein/ NO, the search results were too general. ○ Nein die Resultate ware inhaltlich irrelevant/ NO, the search results were not relevant to my query.	Evaluated in all Y4 user evaluation. (N=84)
4	**Efficiency** a) Empfanden Sie das Finden von Informationen zeitaufwendiger als sonst?/Was finding information more time-consuming than normally? b) Das finden von relevanter Information is aufwendig./Finding relevant information is time-consuming.	a) Ja/Nein/ Yes/No a) Skala 1- Stimme gar nicht zu- 5 Stimme voll zu/Scale 1-I disagree- 5 I agree (Answer from a. and scale answers were collapsed into one item. (Yes = scale 4+5,)	Evaluated in all Y4 user evaluation. (N=84)

Table A19: Questionaire Part 2a-Free browsing task.

	Personal library task		Evaluation
	German **Task and Questions (Answer)**	**English** **Task and Questions (Answer)**	
1	Stellen Sie sich vor Sie suchen nach einer Online-Fortbildung über Diabetes. a) Bitte nutzen Sie KHRESMOI um Kurse zu finden, die für Sie relevant sind. b) Bitte nutzen Sie die Persönliche Bibliothek um mindestens zwei relevante Webseiten zu speichern c) Bitte kategorisieren Sie die Webseiten mit der Tags Funktion. d) Bitte exportieren Sie einen link.	Imagine you are searching for online education about Diabetes. a) Please use KHRESMOI to find courses that are interesting for you. b) Please use the personal library to save at least two relevant websites. c) Please use the tagging function, categorise the saved websites with a topic of your choice. d) Please use the export function to export a link.	Used in November 2013 user evaluation only (N=11).
2	Konnten Sie die Fragestellung der Suchaufgabe mit KHRESMOI lösen? (Ja/Nein)	Were you able to find relevant information on the topic using KHRESMOI? ○ (Yes/No)	
3	• Konnten Sie die Links in der Persönlichen Bibliothek speichern und kategorisieren? (Ja/Nein)	• Were you able to save the links and categorise them in the personal library? (Ja/Nein)	
4	• Wie nützlich ist die Persönliche Bibliothek für Ihre alltägliche medizinische Suche? 1 (Ich werde diese Funktion sicher nie verwenden)- 5 (Diese Funktion ist sehr nützlich und ich kann mir vorstellen Sie regelmässig einzusetzen)	• How useful is the personal library for you in your daily medical search for information? (1- This tool is very useful and I can imagine using it on a regular basis-5-I will never use this tool)	

Table A20: Questionnaire Part 2b: Personal library task.

	Task description and related questions for filter library task		Evaluation
	German Task and Questions (+Response choices)	**English Task and Questions (+ Response choices)**	
1	Eine Patientin fragt Sie nach Nebenwirkungen einer hochdosierten Estradiol Therapie. Bitte nutzen Sie KHRESMOI um die Antwort zu finden.	a) A patient inquires about the side effects of a highly dosed estradiol therapy. Please use KHRESMOI to find the answer/information.	Used for November 2013 user evaluation only (N=11)
2	a) Bitte sortieren Sie die Information nach Datum. b) Bitte nutzen Sie die Suchfilter und die Doppelklick function um die Suche einzuschränken und finden Sie einen fremdsprachigen Artikel zu dem Thema. c) Bitte übersetzen Sie den Artikel mittels der automatischen Übersetzung in dies Deutsche Sprache. Dann gehen Sie bitte auf END Task	a) Please sort the information by date. b) Please use the search filters and the double click restriction to restrict your search and find an article in a forgein language. c) Please use the translation feature to translate the summary to your mother tongue. (German). Then go to END Task.	
3	Konnten Sie die Fragestellungen der Suchaufgabe mit KHRESMOI lösen? (Ja/Nein)	Were you able to find relevant information on the topic using KHRESMOI? (Yes/No)	
4	Wie nützlich fanden Sie die Sucheinschränkung mittels der eingezeigten KHRESMOI Filter? (1- Die Filter sind sehr nützlich und ich kann mir vorstellen sie regelmässig zu nutzen- 5- Ich werde diese Filter sicher nie verwenden.)	How useful did you find the Khresmoi search filters? (1- The filters are useful and I can imagine using them on a regular basis- 5-I will never use those filters).	
5	Wie nützlich war für Sie die Funktion “Sortieren nach Datum”? (1- Diese FUnktion ist sehr nützlich und ich kann mir vorstellen Sie regelmässig einzusetzen-5-Ich werde diese Funktion sicher nie verwenden.)	How useful did you perceive the function “ double-click search restriction” (1- This tool is very useful and I can imagine using it on a regular basis-5-I will never use this tool).	
6	Wie nützlich war für Sie die Funktion “Übersetzung der Vorschau”? (1- Diese Funktion ist sehr nützlich und ich kann mir vorstellen Sie regelmässig einzusetzen-5-Ich werde diese Funktion sicher nie verwenden).	How useful did you perceive the tool “ summary translation? (”1- This tool is very useful and I can imagine using it on a regular basis-5-I will never use this tool).	

Table A21: Questionaire Part 2c- Search filter task.

	Search features and tools.		Evaluation
	Question: German/English	**Answer: German/English**	
1	**Khresmoi search features:** Welche der angebotenen KHRESMOI Suchhilfen finden Sie für die medizinische Suche nützlich? /Which of the following Khresmoi features do you perceive as useful when searching for medical information?	○ Suchvorschläge während der Eingabe/Search suggestions while typing in the query ○ Definition der Krankheit die in der Suchleiste eingegeben wird. /Definition of illness that is inserted in search bar. ○ Automatische Übersetzung eingegebener Suchwörter um auch in fremdsprachige Artikel einzusehen/ Automatic translation of query to allow multilingual access. ○ Vorschau des Artikels rechts/Preview of article ○ Hervorhebung von Wörtern die im Artikel häufig vorkommen./Highlighted common words. ○ Erweiterte Suchoptionen/Advanced search options. ○ Gezielte Eingrenzung der Suchergebnisse mit KHresmoi Filter./Search filter options.	Evaluated in all Y4 user evaluation (N=84).
	Khresmoi Tools: Welche der angebotenen KHRESMOI Funktionen können Sie sich vorstellen regelmässig zu verwenden? /Which of the following Khresmoi tools can you imagine using on a regular basis?	○ Die Persönliche Bibliothek/The personal library ○ Das Beschlagworten gespeicherter Artikel/Tagging articles ○ Die Möglichkeit einen Link zu exportieren/ Exporting a link. ○ Übersetzung der Zusammenfassung fremdsprachiger Artikel/Translation of the summary of articles in a forgein language. ○ Das Teilen von Artikeln und die Kommunikation mit anderen Khresmoi Nutzer/Sharing articles and/or communicating with other KPro users. ○ Das Kommentieren und BEwerten einzelner Artikel/Commenting and rating single articles.	

Table A22: Questionaire Part 3a- Search features and tools.

	Search facets.		Evaluation
	Question (German/English)	**Response choices (German/English)**	
1	Welche der angebotenen KHRESMOI Suchfilter finden Sie nützlich für die medizinische Informationsuche?/ Which of the offered KHRESMOI search filters do you rate as useful when searching for medical information?	• Nach Datumsabschnitten/ By date (e.g. last year, last 3 years vs. older than 3 years) • Nach Herausgeber/ By publisher (e.g. Wikipedia vs. Pubmed) • Nach Land /By Country (e.g. Austria vs. Germany) • Nach Zielgruppe/By audience (e.g. for health professionals vs. for laypeople) • By Sprache/By language (e.g German vs. English) • Nach Medium/By media type (e.g. Image vs. Website) • Nach Bildmodalität/By Image modality.	Evaluated in May-July 2014 user evaluation only. (N=51)
2	Welche der Sucheinschränkungen sind für Sie für die medizinische Informationssuche nützlich? /Which of the following Khresmoi search restrictions do you rate as useful when searching for medical information?	• Definitionen/Definition • Medikamenteninformationen/ • Drug information • Leitlinien/Guidelines • Diagnostische Hilfe/Diagnostic • Information • Online Fortbildung/Online CME • Medical Education events • Wissenschaftliche Artikel/Research • Information für Patienten/Information for patients • Organisatorisches/Organisational	Evaluated in November 2013 and Wiesbaden user tests May 2014 only. (N=52)

Table A23: Questionaire Part 3b- Searcb facets.

Item no.	Standard Usability Scale		Evaluation
	German	English	
	In wie weit treffen die folgenden Aussagen über KHRESMOI Professional für Sie zu? (1- trifft gar nicht zu- 5- trifft voll zu)	To what extent do you agree with the following statements? 1- strongly disagree- 5 strongly agree	
1	Ich denke, dass ich dieses System gerne häufig nutzen würde.	I think that I would like to use this system frequently.	Evaluated in all Y4 user evaluations. (N=84)
2	Ich fand das System unnötig komplex.	I found the system unnecessarily complex.	
3	Ich denke, das System war einfach zu benutzen.	I thought the system was easy to use.	
4	Ich denke, ich würde die Hilfe eines Technikers benötigen, um das System benutzen zu können.	I think that I would need the support of a technical person to be able to use this system.	
5	Ich halte die verschiedenen Funktionen des Systems für gut int8egriert.	I find the various functions in this system were well integrated.	
6	Ich halte das System für zu inkonsistent.	I thought there was too much inconsistencyin this system.	
7	Ich kann mir vorstellen, dass die meisten Leute sehr schnell lernen würden, mit dem System umzugehen.	I would imagine that most people would learn to use this system very quickly.	
8	Ich fand das System sehr mühsam zu benutzen.	I found the system very cumbersome to use.	
9	Ich fühlte mich bei der Nutzung des Systems sehr sicher.	I felt veryconfident using the system.	
10	Ich musste viele Dinge lernen, bevor ich das System nutzen konnte	I had to learn loads before working with the system.	

Table A24: Questionaire Part 3c- Standard usability scale.

	Open feedback		Evaluation
	Question (German/English)	**Response choices (German/English)**	
1	Was gefällt Ihnen an Khresmoi Professional? Was würde Sie motivieren das System in der Zukunft zu verwenden?/ What do you like about Khresmoi Professional? What motivates you to use the system in the future?	• Nach Datumsabschnitten/ By date (e.g. last year, last 3 years vs. older than 3 years) • Nach Herausgeber/ By publisher (e.g. Wikipedia vs. Pubmed) • Nach Land /By Country (e.g. Austria vs. Germany) • Nach Zielgruppe/By audience (e.g. for health professionals vs. for laypeople) • By Sprache/By language (e.g German vs. English) • Nach Medium/By media type (e.g. Image vs. Website) • Nach Bildmodalität/By Image modality.	Evaluated in May-July 2014 user evaluation only. (N=51)
2	Was stört Sie an Khresmoi Professinal? Was kann Khresmoi Professional verbessern um interessanter für Sie zu warden? What did dislike about the system? How can KHRESMOI Professional improve to get more interesting for you?	• Definitionen/Definition • Medikamenteninformationen/ • Drug information • Leitlinien/Guidelines • Diagnostische Hilfe/Diagnostic • Information • Online Fortbildung/Online CME • Medical Education events • Wissenschaftliche Artikel/Research • Information für Patienten/Information for patients • Organisatorisches/Organisational	Evaluated in November 2013 and Wiesbaden user tests May 2014 only. (N=52)

Table A25: Questionaire Part 3d- Open feedback.

6.1.4 Setup and images of user tests

6.1.4.1 Overall

Event name	Details	Audience	Where?	When?
STAFAM	Biggest conference for general practitioners in Austria	Mainly general practitioners.	Graz Stadthalle, Austria	28.11-30.11.2013
Praxis update Wiesbaden	Medical CME conference for practitioners	Mainly general practitioners	Wiesbaden Kurhaus, Germany	16.5-17.5.2014
Medical events at GAW	Symposium für den Endokrinen Kreis (Conference for Endocrinologists)	Mainly specialists, hospital clinicians and research physicians.	Society of physicians, Vienna, Austria	23.5-25.5.2014
	Veranstaltung AESCULAP	Mainly research physicians.	Society of physicians, Vienna, Austria	11.6-12.6.2014,
	„Medizin im ersten Weltkrieg“ Fortbildungsveranstaltung der Gesellschaft der Ärzte in Wien (CME Event of GAW)	Mainly specialists, hospital clinicians, research physicians.	Society of physicians, Vienna, Austria	17.6.2014
Online feedback survey	English Version German Version	All	Link dissemination took place via e-mail, social media platforms, and the GAW homepage and during all Khresmoi dissemination events held at the society of physicians in Vienna.	03.06.2014 - 13.07.2014

Table A26: Timeline and setting of Y4 evaluation events.

6.1.4.2 Face-Face user tests

6.1.4.2.1 Khresmoi Booth

Image A1: Setup of Khresmoi Booth at STAFAM conference

6.1.4.2.2 “Thank you” gifts at user evaluation November 2013.

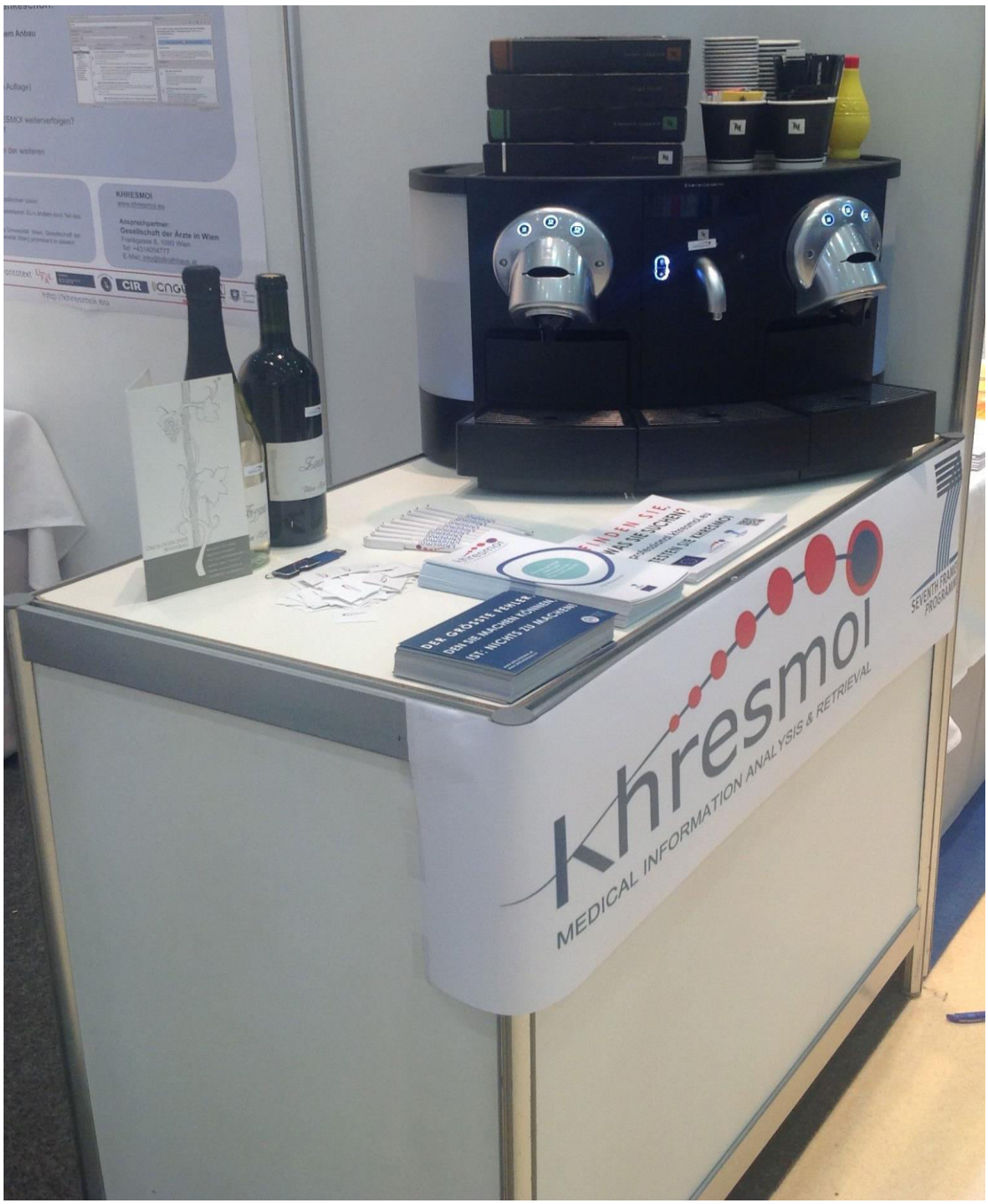


Image A2: “Thank you” gifts and free coffee to attract physicians.

6.1.4.2.3Flyers to motivate users to participate

Image A3: “Thank you” gifts and free coffee to attract physicians.

Image A3: “Thank you” gifts and free coffee to attract physicians.

6.1.4.2.4Online Feedback Survey- Page 1

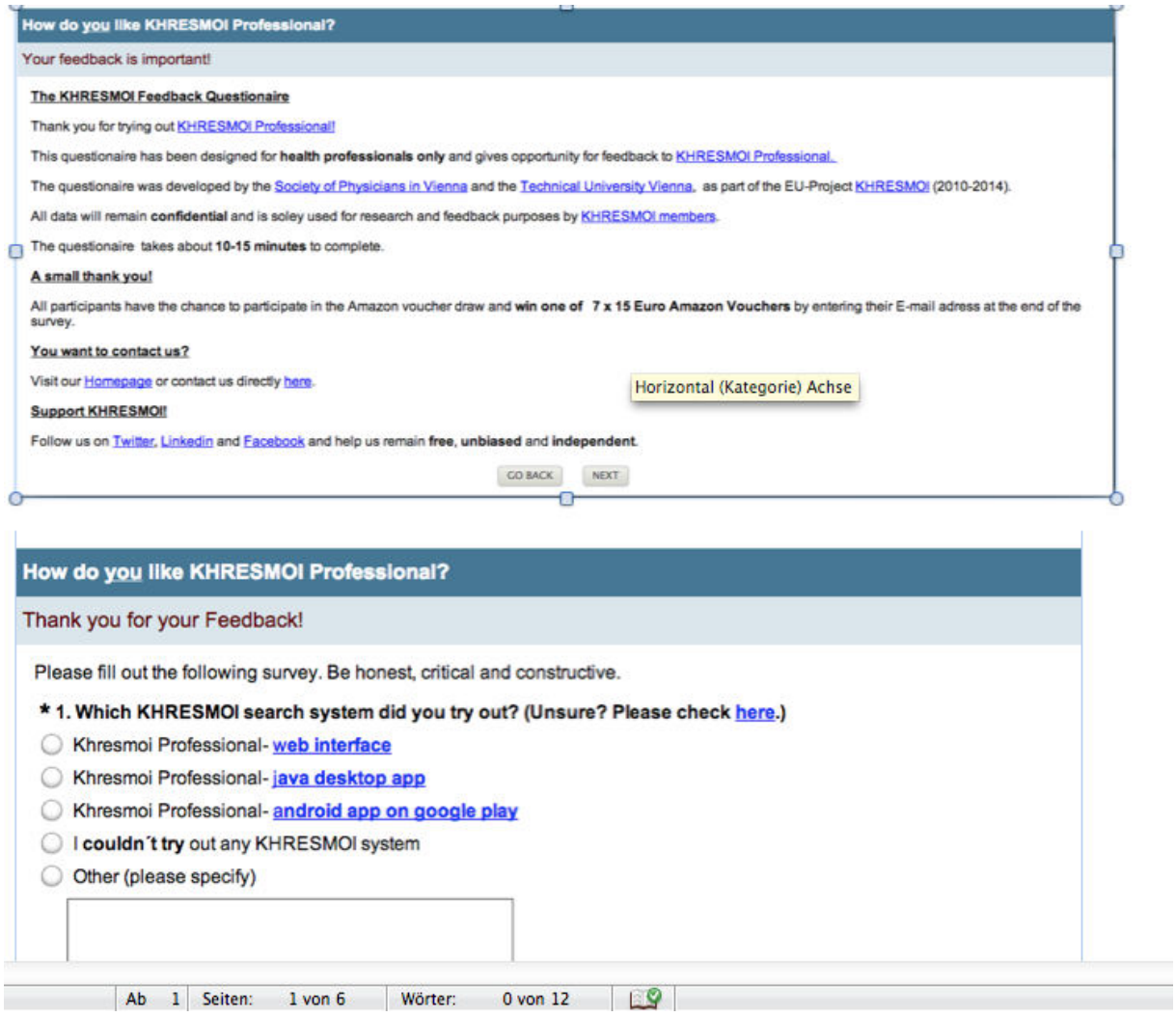


Image A4: The online feedback survey.

6.1.4.2.5 Link dissemination via the Khresmoi Expert community

The Khresmoi Expert Community was set up by GAW after the first round of user tests to allow continuous user feedback and maintain contact to physicians interested in helping KPro development. 10 physicians have joined the German English Community and 12 physicians are members of the English expert community. The screenshot below illustrates link dissemination via the community.

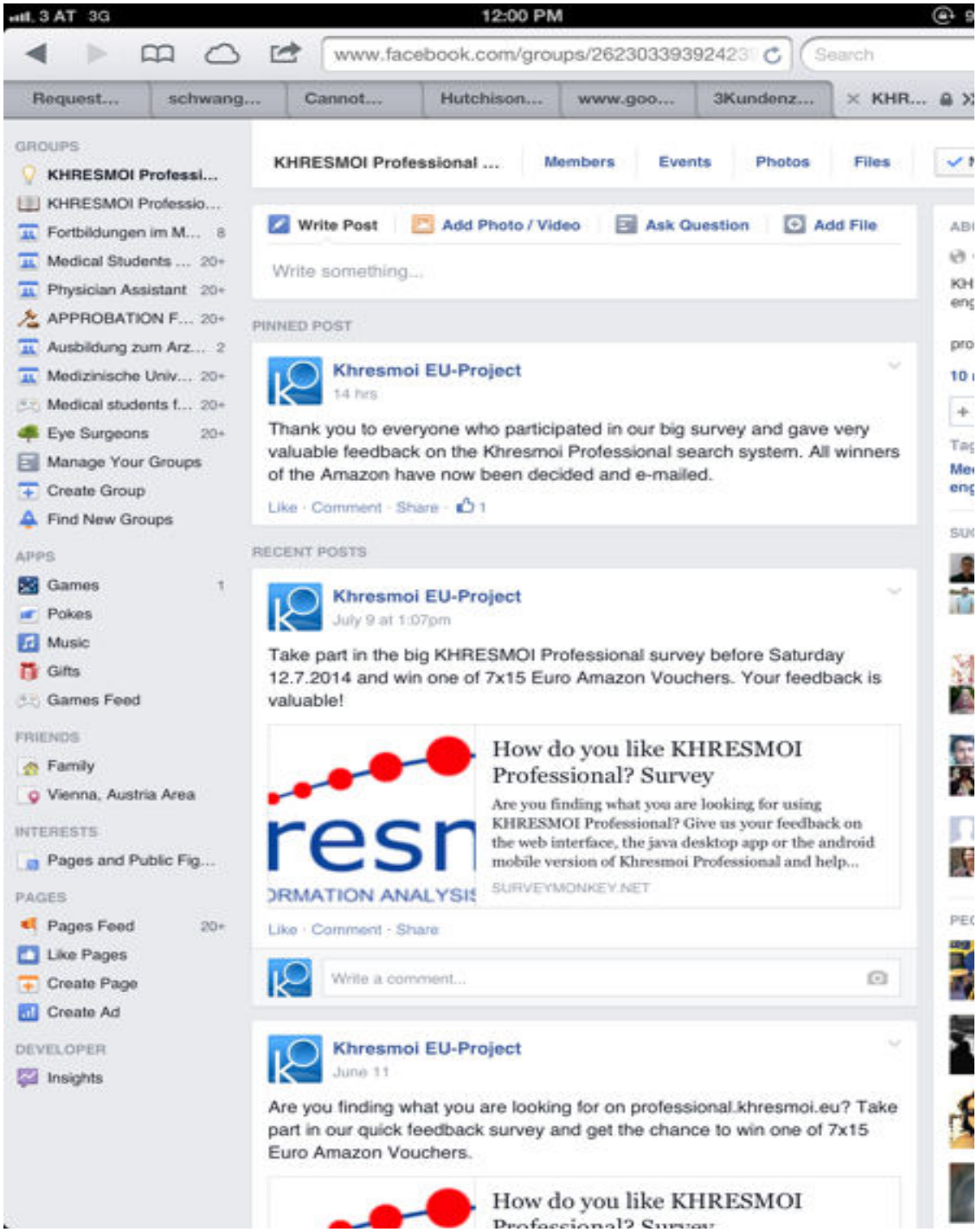


Image A5: Online link dissemination via the KPro expert community.

6.2 Khresmoi for Everyone

6.2.1 Blind vs non-Blind Google vs K4E

Figure 84 Introduction of the new Blind vs non-Blind study

Figure 85 Online dynamic presentation of the platform to the participant

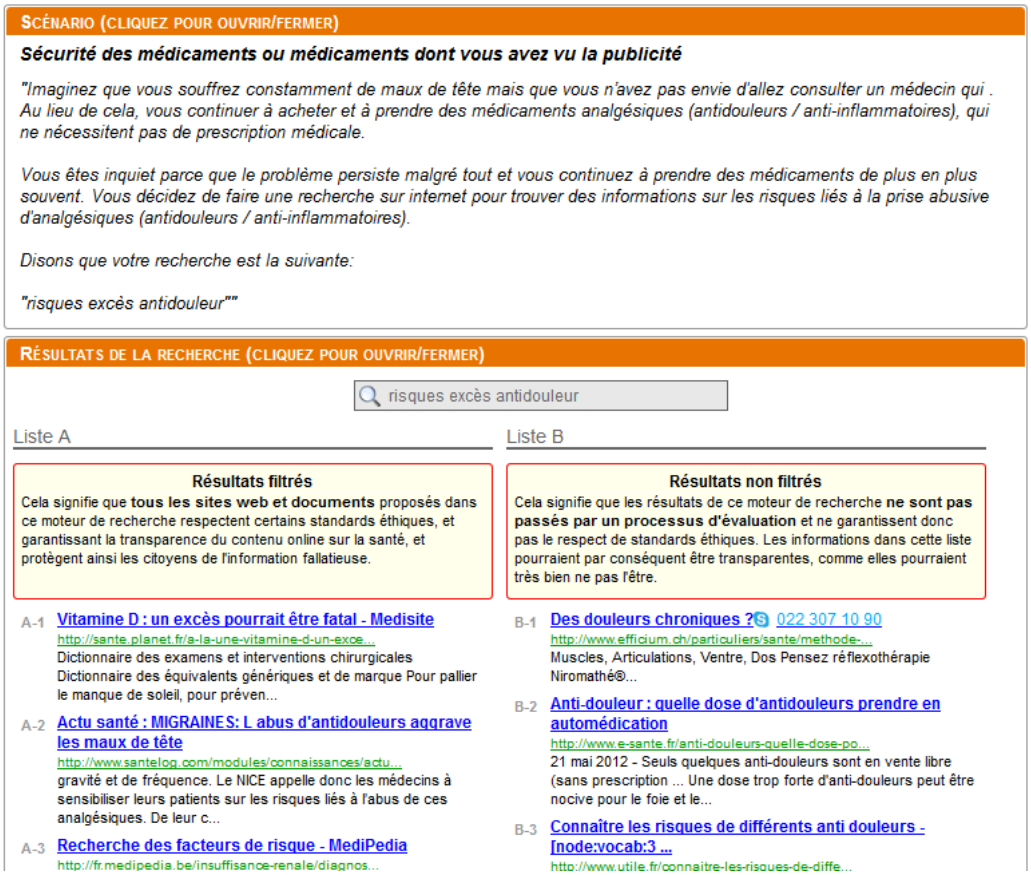


Figure 86 Example of task with the presentation of the results and explanation of the result type

6.2.2 Methodology of the analysis of recording for K4E users tests

Methodology of the analysis of Morae recordings consist in the following steps:

1) data extraction
2) data preparation
3) analysis of questionnaires
4) analysis of markers
5) time of each task completion and its success
6) analysis of summaries written by the observer/facilitator (overall impression from the test and main points, like problems, obstacles etc.).

PART 1. How to extract data from Morae Manager

Once recordings are open in Morae Manager, you should select View menu, then Survey results. In a popped-up window you should choose "Export Survey Results" for each questionnaire (Demographic, 1st task, 2nd task, 3rd task and SUS). The file is saved in a .csv format which can be opened in MS Excel. The data should then be merged in a table to allow further analysis.

PART 2. Data preparation before the analysis

Before starting the analysis, data should be cleaned, i.e. not-completed entries should be excluded:

1) not recorded due to some reasons: if it is possible to add reconstruct the missing data by watching the video, this should be done (for example in Paris in 2013 a part of a recording of one participant was lost, but we could "reproduce" his answers by watching how he responded to the questionnaires).

2) those who did not meet the main inclusion criteria, i.e. searching for online health information at least occasionally. To avoid the latter problem, potential participants should be asked this question before start of the evaluation session, in case they say no, the test should not be performed with them.

Another preparatory task was a reverse translation of questionnaires from French and Czech (languages of evaluation tests) into English (language of document) for the analysis, and ensuring there was no overlap or shift.

PART 3. Analysis of questionnaires (demographic and Internet use, tasks responses and expanded SUS questionnaire):

This data is retrieved from Morae.

All the data from Morae questionnaires is extracted after finalizing the evaluation tests in each location. As the original files in Morae Manager offer questionnaires in two languages (Czech and French), it was important to extract them separately and then merge in an excel sheet (translated into English to facilitate further analysis and presentation of results). Once data is ready (see above Preparation of data), it is analysed by the means of descriptive statistics.

The demographic and Internet use questionnaires are statistically analysed. For SUS analysis we calculated mean of scores for each statement. To see whether there was a difference between Francophone and Czech-speaking participants' satisfaction, the difference of mean between them was calculated. The difference considered significant when it was at least 20%.

PART 4. Analysis of markers:

This data is retrieved from Morae.

All markers were extracted from Morae recordings and merged into an excel spreadsheet, keeping time of event (marker) and ID of participant. All the messages recorded for each participant were read

and reread several times to exclude "noise" and extract "main". "Main" messages stood for meaningful messages containing useful and important information about prototype and interface. They were not quantified, but rather proceeded (if few participants commented on the same problem) and grouped by functionality. Both positive and negative feedback was mentioned, although deliverable was focused on problems participants encountered.

More details: sometimes few main messages had the same theme, for example, "query suggestion was not used", or "does not understand keyword filter", or ""used filter/translation etc.", they were not repeated but were associated with corresponding functionality. Basically, for each functionality, there were several main messages. For example:

Functionality	Corresponding main messages
Forum filter	Not interested in checking forums, does not trust
	Is interested in checking forums depending on topic
Query suggestions	Not used/noticed by many participants
	Suggestions not useful
	Suggestions did not correct spelling errors
	Difficulty with French accents (etc.)

Some messages were beyond the functionalities, for example "losing" a prototype page.

PART 5. Time of each task completion and its success:

This data is retrieved from Morae.

Time of each task was also extracted from Morae and calculated in MS Excel. Task success rate was based on the analysis of questionnaires for each task. Each of such questionnaires followed the task and contained two questions: whether users were familiar with the topic/had an experience and whether they completed the task (Yes-No).

PART 6. Analysis of summaries written by observer/facilitator:

The data is provided by observers from each location of evaluation test.

This qualitative analysis was based on subjective impressions of an observer/facilitator from each evaluation experience. They were detailed to a different extent depending on the person writing, hence could not be analysed quantitatively. Their content analysis was combined with markers analysis, i.e. summaries was used to get a general picture and markers for more specific details. Eventually all "main messages" from both sources entered the same field in the Excel spreadsheet mentioned above. Then, the comments were grouped by topic (overall experience and based on functionality). The choice of topics arises naturally based on the most recurrent messages given/problems encountered by the participants. The list of topics is also contrasted with main functionalities being evaluated (i.e. translation, query support, filters etc.), and based on these findings, are presented in a concise way.

6.2.3 Tutorial of the K4E prototype presented in Year 4

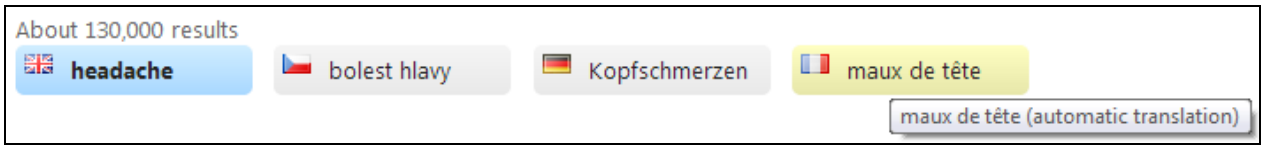
Accessibility

To facilitate access to Khresmoi for Everyone for different categories of users, we have included two accessibility options for a better user experience:
- Font change: you can increase or decrease the font size
- Style change: you can change the background and font colours.

Automatic results translation

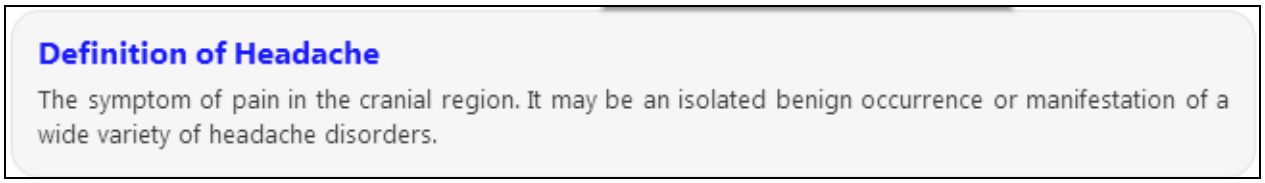
Automatic translation from Czech, German and French to English and vice versa is available. First of all, your query is translated into respective language of choice, and once results are retrieved, they are being translated back into your language (corresponds to the language of the interface). This translation is done automatically by machines and may contain some errors as it has not been validated by a qualified translator.

To get search results translated from other available languages (Czech, German and French) into English, you should click on the corresponding icon with your query.



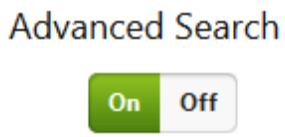
Definition of a search term

When you search for a specific medical condition or symptom, Khresmoi for Everyone offers you a definition which is originally in English and then automatically translated into other languages (depending on the language of the interface).



Advanced search

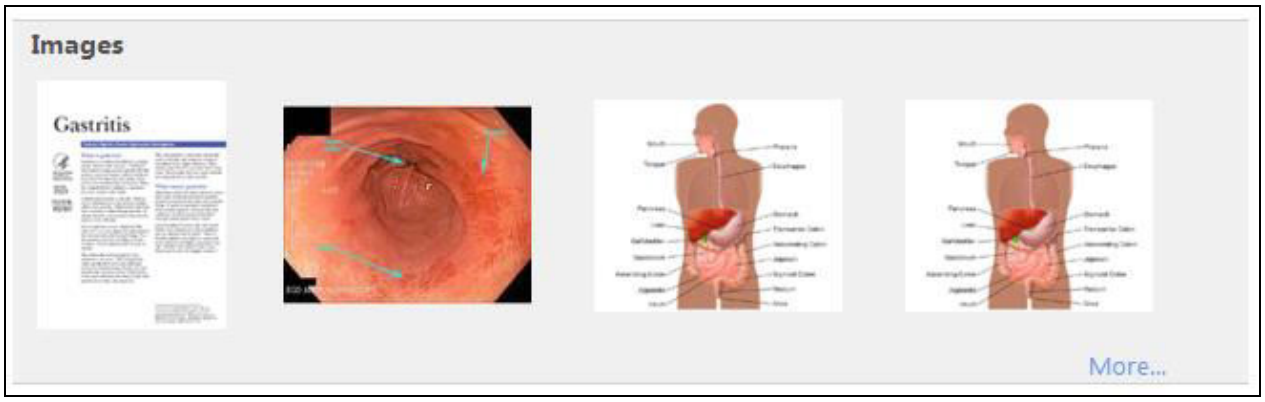
Along with the "Lite" search, Khresmoi for Everyone offers its users several functionalities developed specifically for a health search engine, namely the Advanced Search mode which can be utilized once you have retrieved the results. Advanced Search can be turned on using the following switch:



The available features in Advanced Search Mode are presented below:

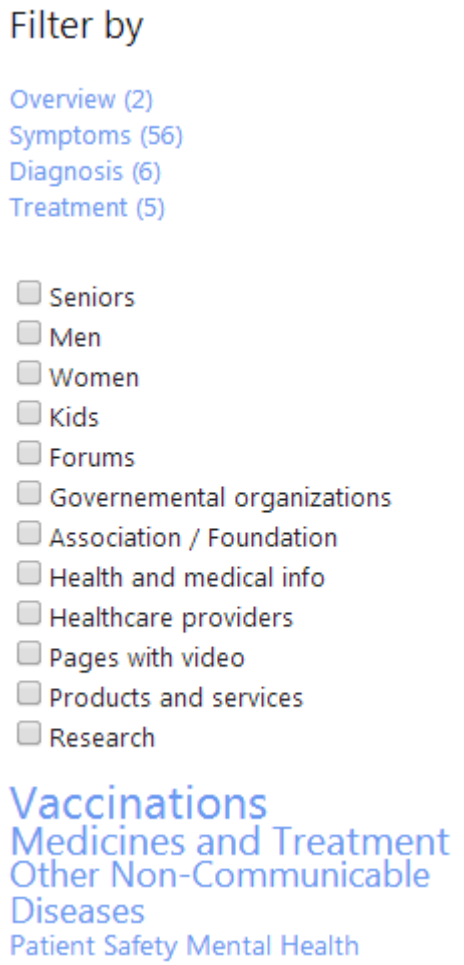
➢ Images

In an advanced search mode you will see images corresponding to your search query.



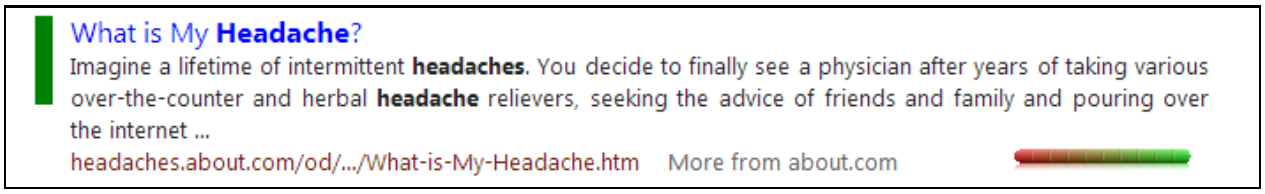
➢ Filters

The advanced functionality offers you a variety of filters. "Filter by" allows you to filter in specific types of online health information. You can filter in specific subsections of any disease description, for example, "symptoms" or "treatment". This possibility exists for English and French-speaking users. You can also filter in/out any information targeted for kids, elderly, men or women. By ticking relevant boxes, you can select to see information from forums, governmental organizations or healthcare providers. For a full list of options, see an image. The last filter gives you four to five main health topics, associated with your query: each result returned by the search engine is attributed to one or more health topics out of 36 topics available based on the number of specific keywords contained in it. A health topic attributed to a larger number of results appears in bigger letters. Hence, it is possible to refine your search by clicking on a health topic, thus keeping only the search results attributed to the chosen topic.



➢ Readability and trustability of search results

In the advanced search mode each search result retrieved has two red-green indicators: vertical bars relate to readability while horizontal bars relate to trustability. Both of these indicators are estimated automatically.



Readability is defined as the ease of reading and understanding the text. This is especially important within a health/medical domain which contains many complex terms often used by physicians, but not familiar to the general public. Khresmoi for Everyone automatically analyses all search results returned to your query and assigns them into one out of two categories: **easy-to-read** or **difficult-to-read**.

Tip: When you start discovering a new subject we recommend you to start with easy-to-read results. Once you feel more knowledgeable about the topic, you may wish to proceed to difficult-to-read search results.

The readability level is estimated for the documents written in either English or French languages. A lexical method is used for English, while for French a machine learning approach is used. The method used has been shown to be more than 90% reliable.

Trustability in Khresmoi for Everyone is defined as automatic detection of web page compliance to the HONcode principles. The system identifies HONcode principles being respected on any given page while the score is calculated on the website basis. It represents the percentage of the HONcode principles detected in all crawled pages from the given website. As the process is automatic, sometimes there may be a difference between the estimation by the system level of trust and the real one. It is an important indicator as not all the web sites contained in Khresmoi for everyone are HONcode certified.

The reliability of the HONcode principle estimation for each individual principle is given in the table below.

HONcode principle	Criteria	Precision
Principle 1 - Information must be authoritative	Authors' names are provided	70%
Principle 2 - Purpose of the website	Statement that the information is designed to support, not replace, the relationship between a patient and his/her physician is provided	90%
Principle 3 - Confidentiality	Privacy policy is present	96%
Principle 4 - Information must be documented: Referenced and dated	References are provided	62%
Principle 4 - Information must be documented: Referenced and dated	A date of last review or page modification is provided	95%
Principle 5 - Justification of claims	Information is presented in a balanced way.	51%
Principle 6 - Website contact details	Contact details are provided	94%
Principle 7 - Disclosure of funding sources	Web site funding is explained	79%
Principle 8 - Advertising policy	Advertising policy is provided	74%

For more details please refer to an article[15] or to the Khresmoi Report on automatic document categorization, trustability and readability (D1.6)[16].

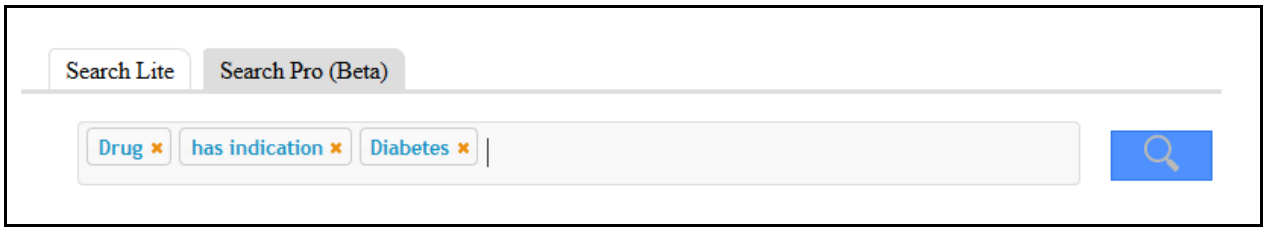
When the trustability level is below 100%, missing principles are identified and specified by the system. As mentioned above, all the resources indexed in Khresmoi for Everyone were manually selected and reviewed and hence, already have a certain degree of trustworthiness. Automatically measured trustability provides an additional tool to determine web site compliance to the HONcode principles.

[15] Available at http://thesai.org/Publications/ViewPaper?Volume=5&Issue=3&Code=IJACSA&SerialNo=9

[16] D1.6, Khresmoi Deliverable, http://www.khresmoi.eu/assets/Deliverables/WP1/KhresmoiD16.pdf

Search Pro

Along with Search Lite which offers a "typical" web search experience, for some queries you may wish to use Search Pro based on semantic search technology. The difference is that in this case instead of specific terms you can search for concepts. For example, if you wish to find the information about all drugs used in diabetes treatment, you may not know all such drug names, but Khresmoi for everyone will find it for you:



Tip: When you are typing a semantic query, you have to select each term from a drop-down menu.

Other possible queries are the following:

- "Disease or syndrome" + "treated by" + ADD YOUR TERM
- "Disease or syndrome" + "has symptom" + ADD YOUR TERM