Effects of Terminology on Health Queries: An Analysis by User's Health Literacy and Topic Familiarity

Carla Teixeira Lopes and Cristina Ribeiro

Faculty of Engineering, University of Porto / INESC TEC, Porto, Portugal

Abstract

Prior studies have shown that terminology support can improve health information retrieval but have not taken into account the characteristics of the user performing the search. In this chapter, the impact of translating queries' terms between lay and medico-scientific terminology, in users with different levels of health literacy and topic familiarity, is evaluated. Findings demonstrate that medico-scientific queries demand more from the users and are mostly aimed at health professionals. In addition, these queries retrieve documents that are less readable and less well understood by users. Despite this, medico-scientific queries are associated with higher precision in the top-10 retrieved documents results and tend slightly to generate knowledge with less incorrect contents, the researchers concluded that search engines should provide query suggestions with medico-scientific terminology, whenever the user is able to digest it, that is, in users above the lowest levels of health literacy and topic familiarity. On the other hand, retrieval systems should provide lay alternative queries in users with inadequate health literacy or in those unfamiliar with a topic. In fact, the quantity of incorrect contents in the knowledge that emerges from a medico-scientific session tends to decrease with topic familiarity and health literacy. In terms of topic familiarity, the opposite happens with Graded Average Precision. Moreover, users most familiar with a topic tend to have higher motivational relevance with medico-scientific queries than with lay queries. This work is the first to consider user context features while studying the impact of a query processing technique in several aspects of the retrieval process, including the medical accuracy of the acquired knowledge.

Keywords: health information retrieval; query formulation; terminology; health literacy; topic familiarity

Paper category: Research paper

Introduction

Patients, relatives and friends are increasingly using the Web to search for health information. In fact, this is the third most popular online activity following email and using a search engine (Fox, 2011), being done by 72% of American Internet users (Fox & Duggan, 2013). The importance of an easy access to online health information is recognized by the U.S. Department of Health and Human Services (2010) which set a goal for 2020 to increase the proportion of online health information seekers who report easily accessing health information.

Although most users are satisfied with their health searches, some get frustrated or confused (Fox, 2006; Petrock, 2010). This happens more in individuals with less education as showed by the Pew Internet report (Fox, 2006). Twenty-two percent feel frustrated by the inability to find what they want (27% in those without a college degree and 18% in those with a college degree) and 18% feel confused with what they did find online (24% in those without a college degree and 15% in those with a college degree). Since educational level has a strong impact on health literacy, this is not surprising. By health literacy is meant the "capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions" (Kutner, Greenberg, Jin, & Paulsen, 2006).

The widespread use of the Web to retrieve health information implies a large diversity of users performing this task. One characteristic that is expected to differ between users is their health literacy, a differentiation that can be caused by differences in age or education. A study that assessed the usability of 125 websites offering health resources reported that about one third of these sites required a college education to comprehend extracted health information (Becker, 2004).

The mismatch in languages used by health consumers and health professionals also poses a barrier to effective access to relevant information (Zielstorff, 2003). Since information may be presented at a high reading level and include medical jargon (Cline & Haynes, 2001), the ability to understand the retrieved information may fail and, if so, user's satisfaction may be at risk. In fact, this is one of the typical problems felt by consumers when performing health information searches (Kogan, Zeng, Ash, & Greenes, 2001). Other popular problems are: difficulty or inability to formulate a health query due to the lack of proper medical terms (Toms & Latter, 2007; Zhang, 2010) and; the difficulty with formulating it without misspellings or use of wrong medical terms (Kogan *et al.*, 2001; McCray & Tse, 2003).

In this research we study the effect of translating query terms between lay and medico-scientific terminologies, in users with different topic characteristics, namely, health literacy and topic familiarity. In our experiment the search engine is a "black box". We believe that a user model that considers the above context features may be used to improve health information retrieval through, for example, the suggestion of alternative queries or by re-ranking results. The work presented here is the first to consider user context features while studying the impact of a query processing technique in several aspects of the retrieval process. The evaluation considers not only users' relevance assessments, as considered in several previous works, but also the quality of the medical knowledge that emerges from the search session.

The chapter is structured as follows. A review is given of the existing literature on the exploration of medico-scientific terminologies and the use of health literacy and topic familiarity in health information retrieval. The research questions and the experimental settings of the study are then described afterwards. The following sections have a detailed description of the findings that will be discussed along with their implications in the final section.

Related Work

In this section, is a literature review of works that explore medico-scientific terminologies with the goal of improving IR. In a second stage, IR works that explore the two main context features used in this work - health literacy and topic familiarity, are also discussed.

Exploration of Medico-Scientific Terminologies

It is known there are mismatches between consumer terminology and the ones used in health documents and standard medical vocabularies (Eerola & Vakkari, 2008). To evaluate the impact of this mismatch, Plovnick and Zeng (2004) compared the performance of consumer queries with the performance of the same queries reformulated with terminology from the Unified Medical Language System (UMLS). Each query was submitted to Google and MedlinePlus and the relevance was assessed comparing results with a gold standard answer. The authors used P@30 to compare both type of queries and, through descriptive analysis, concluded that this type of reformulation may be a promising strategy to improve consumer health-information searches. Previous studies (Patrick, Monga, Sievert, Houston Hall, & Longo, 2001; Zeng *et al.*, 2002) reached similar conclusions. Patrick *et al.* (2001) compared the performance of lay and medico-scientific queries on the retrieval of diabetes web information. The evaluation was based on the number of sites maintained by non-profit healthcare professional organizations, academic organizations, or governmental organizations that appeared in the top-20 results. Authors found fewer sites of this type when using lay queries. While studying the characteristics of consumer terminology for health IR, Zeng *et al.* (2002) concluded that 51% of the lay queries returned no information although matching information existed in the database.

Considering the poorer results of lay queries and the fact that non-experts use medico-scientific terminology less often than experts (White, Dumais, & Teevan, 2008, 2009), it is expected that comprehensive terminology support improves health IR (Zeng *et al.*, 2002). Some works therefore propose and evaluate strategies to translate lay terms into medico-scientific ones (Lu, Ray, Chan, & Chen, 2006). Others go further and present query suggestion systems (Luo, 2009; Luo, Tang, Yang, & Wei, 2008; Zeng, Crowell, Plovnick, Kim, Ngo & Dibble, 2006) and others come up with ways to identify the mixture of terminologies in order to minimize the language gap and improve health IR (Crain, Yang, Zha, & Jiao, 2010). These works are briefly described next.

Lu *et al.* (2006) translated query terms from lay to professional ones in the context of cross language health IR (CLHIR). If the lay term appears in the Medical Subject Headings (MeSH) thesaurus, an immediate translation is made. If not, the authors propose an approximate string matching of the non-professional terms to the professional ones. In the other cases, they propose to use Web resources with the argument that an increasing number of sites contain lay terms and their corresponding professional terms. Their evaluation showed improvements on the performance of MeSH concept mapping and CLHIR.

Luo, Tang, Yang and Wei (2008) and Luo (2009) propose and evaluate two similar search engines for health IR: MedSearch and iMed. Both search engines accept long queries and transform them to shorter ones by extracting the most representative terms. Moreover, they suggest medical phrases to help the user digest the retrieved documents and refine the query. These phrases are extracted and ranked based on MeSH, the collection of crawled webpages and the query. In addition, to help users provide information about their medical situation, iMed uses a questionnaire-based query interface. MedSearch was evaluated with questions posted on medical discussion forums and assessments from five non-medical persons. iMed was evaluated with real medical case records from the Family Medicine Online Database (FMOD) and medical exam questions with corresponding answers as the ground truth. In both cases, the experiments showed that the search engines handle medical queries effectively and efficiently.

The Health Information Query Assistant (HIQuA) system, developed by Zeng *et al.* (2006), suggests alternative query terms, selected according to their semantic distance to the user's initial query terms. Queries are first mapped to one or more concepts of the UMLS and then the semantic distance between concepts is calculated based on co-occurrences in medical literature, log data and on UMLS semantic relations. Authors found statistically significant higher rates of successful queries, that is, queries with at least one relevant result on the top-10, but no statistical differences on user satisfaction or users' ability to complete the task.

Crain, Yang, Zha, & Jiao (2010) propose a Bayesian model[1] to overcome the language gap between lay and medico-scientific terminology. Given a document, this model can infer the mixture of topics and dialects (slang, common and technical) and the most likely topic and dialect of each word. Authors found a 25% improvement in normalized Discounted Cumulative Gain (nDCG)@5 when using this model to support health IR.

The interplay between user contextual features and the terminological aspects of health IR is less explored in the existing literature. From the works mentioned above, only the health search engine described by Luo (2009) collects and uses information about the user through the questionnaire-based interface. Another study investigates the effect of user factors on the familiarity with health terms and uses gender as a proxy for background knowledge about gender-specific illnesses (Keselman, Massengale, Ngo, Browne, & Zeng, 2006). Authors recruited a convenience sample of 50 users and designed an instrument to test users' familiarity with 27 health terms of different "familiarity likelihood scores" and three categories: "male", "female" and "neutral". This study's findings support the idea that background knowledge and experience affect users' familiarity with health terms. Moreover, authors conclude that health literacy is another variable expected to influence familiarity. A more recent article (Zeng-Treitler, Goryachev, Tse, Keselman, & Boxwala, 2008) uses context to estimate consumer familiarity with health terminology but the explored features are not related to the user. In the proposed method, the authors use a network in which each node represents a term and each term is connected with other terms that co-occur with it. The context of a term can be a query session, a sentence, a paragraph, or a document. The method was applied to query logs and was validated using results from previous consumer surveys. The authors concluded that this method is a good alternative to existing term familiarity assessment methods.

Health Literacy in Information Retrieval

Health literacy is defined as "the degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions" (Kutner, Greenberg, Jin, & Paulsen, 2006, p.iii). A 2003 assessment of adult literacy (Kutner *et al.*, 2006) found that 36% of adults in the United States have basic or below basic health literacy skills. A good review of the literature on health literacy was done by McCray (2005). According to this author, a substantial portion of the literature addresses the mismatch between the health literacy of the patient and the readability of the documents.

Health literacy can be assessed through several existing instruments like the Test of Functional Health Literacy in Adults (TOFHLA) (Parker, Baker, Williams, & Nurss, 1995) that takes up to 22 minutes to administer and Short Test of Functional Health Literacy in Adults (STOFHLA) (Baker *et al.*, 1999), a smaller version of TOFHLA. The Rapid Estimate

[1] Probabilistic model based on Bayes rule.

of Adult Literacy in Medicine (REALM) (Davis *et al.*, 1993) is another option, easier and quicker to administer. In non-English languages there are other tools like the Short Assessment of Health Literacy for Spanish-speaking Adults (SAHLSA) (Lee, Bender, Ruiz, & Cho, 2006) that was developed based on REALM and also incorporates a comprehension test using multiple-choice questions.

To the best of our knowledge, few IR studies consider user's health literacy. In the Health Information Query Assistant study previously described (Zeng *et al.*, 2006), the authors empirically concluded that query recommendations are not adequate for inadequate health literacy users. Another work (Wang & Liu, 2005) describes a personalized health IR system that adjusts results to users' health literacy level, but no evaluation was performed.

The Influence of Topic Familiarity in Information Retrieval

Topic familiarity, or domain knowledge as it is also frequently referred to, can be defined as the user's general knowledge about the topic of the search task. It is acknowledged that topic familiarity can be an important factor in IR (Capra & Pérez-Quiñones, 2006) and there are several research works that explore this feature.

Studies investigating the relationship between topic familiarity and information search behavior (Kelly & Cool, 2002; Liu & Belkin, 2010; Qu, Liu, & Lai, 2010; Wen, Ruthven, & Borlund, 2006) are based on user studies and all evaluate the familiarity with the topic through users' self-assessment. They differ on the type of analyzed behaviors and, typically, these behaviors are acquired through log records of the user study. The conclusions of these studies state that, as the familiarity with the topic increases, so does the search efficacy. Moreover, the resources the user values become more specialized, the user's effort (task completion time and number of queries) decreases and the importance given to certain relevance criteria change. As can be seen through the studies described in the rest of this section, performance conclusions are not always consensual.

Regarding the relation between topic familiarity and query formulation, Wildemuth (2004), in a longitudinal study, analyses the search terms used by medical students on six clinical problems. This is done in three occasions, one before students received any instruction on the topic, the second just after a course on the topic and the third occurred six months after the end of the course. Wildemuth concluded that, when domain knowledge was very low (first assessment), users made more moves, i.e., additions and deletions of concepts to the query. This is probably due to their initial inability to choose the appropriate terms and is in accordance with the conclusions of the study described above (Qu *et al.*, 2010). Finally, Wildemuth also concluded that, although it improved performance in all occasions, system assistance during query formulation is more useful when users have less knowledge on the topic. This work also has a good literature review on the effects of domain knowledge in IR.

Another study explores the influence of topic knowledge on the use of a thesaurus for query expansion (Sihvonen & Vakkari, 2004). The authors conducted a user study with 15 users with knowledge on the topic and 15 users without it. Results were acquired through search logs and interviews with the subjects. Authors concluded that the use of thesauri was helpful for experts but not for novices in order to improve search effectiveness. The number of documents that were judged relevant by external experts measured the search success. This conclusion contradicts the conclusions from the previous study (Wildemuth, 2004).

Studies analyzing the influence of topic familiarity on IR performance focus on different aspects. Liu and Belkin (2010) considered document usefulness, and their primary goal was to know if topic knowledge could be used to predict it. Kelly and Cool (2002) considered efficacy as the ratio between the number of documents saved and the number of documents viewed. Other authors (Al-Maskari & Sanderson, 2010) investigated factors influencing user satisfaction and found no relationship between familiarity and satisfaction. They also found no significant differences between familiar and unfamiliar users in the number of relevant documents identified by the users, the number of TREC relevant documents and the time taken by the user to locate the first relevant documents. The same authors conducted a user study (Al-Maskari & Sanderson, 2011) with 56 subjects and 56 topics from the TREC collection to analyze the influence of users' cognitive skills on user effectiveness. They asked users to assess familiarity after completing the search for each topic and found no significant correlation between familiarity and users' perceptual speed. Another study (Muresan, Cole, Smith, Liu, & Belkin, 2006) used the TREC HARD track (Allan, 2003) to examine the impact of document characteristics like readability and concreteness/abstractness on document relevance assessments by users with different levels of familiarity with the topic. Authors concluded that a higher readability has positive effects on retrieval performance, regardless of user's familiarity with the topic.

Only one study was found that considered users' topic familiarity in health IR (Lopes & Ribeiro, 2010). Its authors studied the impact of several context features on query formulation and relevance assessment in health searches and concluded that, in more familiar tasks, users employ medico-scientific terminology more often and formulate longer queries. Moreover, authors found that relevance decreases as the familiarity with the topic increases.

Methodology

Research Questions

Two research questions drove the research described in this chapter. The questions are similar in their aims but differ in the object of analysis:

- What is the impact on *the characteristics of the retrieved documents* of replacing lay query terms by medico-scientific ones? (RQ1)
- What is the impact on *search task precision* (RQ2), *users' comprehension of documents* (RQ3), *accuracy of the medical knowledge* (RQ4), *task completion status* (RQ5) of replacing lay query terms by medico-scientific ones, in users with different levels of health literacy and topic familiarity?

RQ1 does not consider Health Literacy (HL) or Terminology Familiarity (TF), because the characteristics of the retrieved documents are the only surveyed feature that does not depend on the user.

To answer the research questions, a laboratory user study with the following settings was conducted.

Information Needs and Queries

Eight health information needs were defined based on questions submitted to the health category of the Yahoo! Answers service. From the list of open questions in this category in decreasing order of popularity, the information needs satisfying the three following requirements were selected. Since most of the health searches on the Web continue to be about diseases (Fox, 2006; Fox, 2011), information needs were focused on questions about treatments to diseases/conditions. Because the goal of this study is to study the effects of lay and medico-scientific queries in users with different characteristics, it was ensured that, for each information need, queries were different. For that reason, the disease/condition also had to be associated with different syntaxes in both terminologies, as defined in a glossary of medico-scientific and popular medical terms developed in a European project (Stichele, 1995). For example, diabetes would be excluded because it is simultaneously a lay and a medico-scientific term. Moreover, each query had to have at least 30 results in each search system used in this study.

The selected information needs (IN) were:

- About 3 days ago, I started having a burning feeling every time I urinated. How should I treat this?
- For the past 5 days my head has been very itchy and I don't have lice. What can I do to stop the itching?
- I have high uric acid (8.0 mg/dL) with reference units 3.6 - 7.7. How can I lower my uric acid level?
- I am suffering with an inflammation on my lips and mouth area for more than a year. I have difficulties eating. What can I do to treat it?
- My father got bit by a dog and is in the hospital with a bone infection. How is this treated?
- I frequently get heartburn even when I stay away from spicy stuff. What can I do to prevent it?
- I have been noticing lots of hair coming out from my head. Usually I only comb my hair once a day. What can I do to stop losing my hair?
- I'm on the computer all day so I type a lot and use the mouse. My right pointing finger is starting to give me some joint pain. How I can treat my finger?

The researchers defined the queries for each information need whereas the users only assessed the documents retrieved with the queries. Medico-scientific and lay queries were built concatenating the symptom or disease in each terminology with the word “treatment”. For each disease/condition, the lay and medico-scientific terms were extracted from the glossary of medico-scientific and popular medical terms mentioned earlier. As an example, the medico-scientific query for the first information need would be *dysuria treatment* and the lay one would be *painful urination treatment*. Although smaller and less current than other existing consumer health vocabularies, this glossary was not restrictive in the selection of information needs. Moreover it is singular for its multilingual characteristics that were needed for a parallel study with different aims.

Retrieval Systems

Google was used as a “black box” search engine with two different collections, the entire Google index and the set of pages indexed by Google that belong to HONCode certified sites. Google custom search was used to limit the second collection to those specific sites. This certification is proposed by the Health On the Net Foundation (HON) to help assess the accuracy of health content and the credibility of the publishers. For each query, the top-30 results from each retrieval system were collected.

To reduce the risk of Google learning from the previous submitted queries, it was ensured that the returned links were never clicked. Further, to prevent changes in the search engine or in the HON collection, all queries were submitted within a very short time span.

Tasks

The combination of a query and a retrieval system led to a task that can be executed by a user. Each user was assigned a set of eight different tasks. In the assignment of the tasks to users a Latin square-like procedure was applied so that all users assess the relevance: (1) of all information needs, but only once each; (2) of queries of both types of terminology, the same number of times; and (3) in all the retrieval systems the same number of times. The order of tasks was permuted to avoid possible bias of relevance assessments owing to human behavior. Moreover, each iteration of relevance assessments (4) contained queries of both types of terminology and (5) had tasks in both retrieval systems the same number of times.

Search Procedure

Users started answering a quiz to evaluate their health literacy. They then answered a questionnaire where they were asked about their familiarity with the medico-scientific terms associated with the information needs. Although users did not assess documents retrieved with their own queries, they were asked to provide the query they would formulate for each information need. After this questionnaire, users enrolled in a sequence of eight tasks. Every task is associated with a single query defined on top of the associated information need and the type of query. In each task, users had to assess the relevance of the top-30 URL retrieved with that specific query and then fill a post-search questionnaire.

For each URL, the user had to indicate the type of the document; its relevance to the information need considering his own context; and how much he comprehended its content. Relevance and comprehension were assessed in a 3-value scale. For relevance, the three values were “not relevant”, “partially relevant” and “totally relevant”, denoted by 0, 1 and 2, respectively. For comprehension, the three values were “I did not understand the document's content”, “I partially understood the document's content” and “I understood the document”, denoted by 0, 1 and 2, respectively.

In the post-search questionnaire users are asked (1) if they have already searched for that topic, (2) to evaluate the task in terms of familiarity, (3) to evaluate their feeling of success with the task, and (4) to indicate treatments for the condition mentioned in the task.

Health Literacy Assessment

Since there isn't any Portuguese instrument to assess health literacy, SAHLSA was adapted to this language because, when compared to English, Spanish is closer to Portuguese. The fifty medical concepts used in SAHLSA were translated to Portuguese and users were asked to associate each concept to one of two terms, in less than four minutes. Users were instructed not to guess the answer. With SAHLSA, if users score less than 37, they have inadequate health literacy. Users were grouped in three classes (inadequate, elementary and good) based on the SAHLSA threshold and clusters obtained through hierarchical clustering.

Topic Familiarity Assessment

To evaluate topic familiarity, users were asked if they had previously searched for that topic. They also had to evaluate task familiarity in a 5-value scale and say if they knew the meaning of the medico-scientific concept behind the disease/condition associated with the information need. To compute a single measure to assess topic familiarity (Combined Topic Familiarity – CTF), the previous metrics was combined as follows:

$$CTF = TaskFam + 3 \times PreviousSearch + 2 \times KnewMSTerm$$

This formula considers that *TaskFam* is assessed in a 1 to 5 scale, *PreviousSearch* as 0 or 1 and *KnewMSTerm* as 0 or 1. The user's task familiarity assessment is considered the most important feature, followed by the existence of previous searches about the topic and the knowledge of the medical term. CTF is an integer that varies between 0 and 10. Since this is a discrete variable and 10 categories are not justifiable, CTF was grouped in three categories of familiarity: unfamiliar (CTF<=3), somehow familiar (3<CTF<7) and familiar (CTF>=7).

Medical Accuracy Assessment

In the post-search questionnaire, users had to write an answer to the information need that drove the task. A medical doctor evaluated this answer in relation to the correct and incorrect content it possessed. Answer correctness was evaluated in a scale of 0 (inappropriate answer) to 2 (appropriate answer). The middle value (1) was used for answers with "some value". In terms of answer's incorrectness, user's answer was classified with 0 (all or almost all content is incorrect), 1 (some incorrect content) or 2 (no incorrect content). To exemplify the independence of these characteristics, consider the answer "Reduce the ingestion of red meat. Increase weight." to the information need given as example. This answer has some correct content but it's not complete and would be classified with 1 in terms of correctness. In addition, it also has some incorrect content, being assessed with 1 in terms of incorrectness. If this answer did not contain the second sentence, its incorrectness assessment would change to 2. On the other hand if it had several other wrong suggestions, it could be classified with 0 in terms of incorrectness. Answer's correctness and incorrectness values were added into a single variable called "medical accuracy" that, therefore, varies between 0 (lowest accuracy) and 4 (highest accuracy).

To evaluate the reliability of the medical assessments, a second medical doctor judged 30% of the answers and the inter-rater reliability was estimated through the weighted Cohen's Kappa, an adaptation of Cohen's Kappa to ordinal scales that treats disagreements differently. The measured weighted Cohen's Kappa, with squared weights, for the correctness ratings is 0.68 (95% CI: [0.54, 0.77]), indicating a substantial agreement. For the incorrectness ratings, this measure is 0.7 (95% CI: [0.48, 0.84]), also pointing a substantial agreement. These inter-rater reliability results assure the quality of the initial ratings.

Readability Assessment

Document readability was automatically evaluated using the SMOG (Simple Measure of Gobbledygook) metric defined as:

$$SMOG = 1.043\sqrt{30\frac{\#\,polysyllables}{\#\,sentences}} + 3.1291$$

This metric was adopted because it has been recommended as a measure of readability in consumer-oriented healthcare documents (Fitzsimmons, Michael, Hulley, & Scott, 2010). To compute SMOG, the main content of the documents was extracted, excluding components like menus, advertising, footers and headers. Then, the HTML tags were excluded to obtain a text document with the main contents of the original one. With this document, SMOG was computed using a readability metrics API that is available at http://ipeirotis.appspot.com/readability-api.html.

Summary of Context Features

A summary of the context features used in this study can be seen in Table 1. One feature was only related to the user, two features related the user and the document, three features relate the user with the task and the others are only related to the document. Along with feature' categories and description, the scale of the associated variable is also presented. In the description of documents' features are distinguished the ones that are automatically computed and the ones that require human intervention. All these features were used in the analysis of the data.

TAKE IN TABLE 1

Users

Forty information science undergraduate students participated in this study (25 females; 15 males) with a mean age of 22.25 years (sd=6.42). In the health literacy test, evaluated in a 0 to 50 scale, users had in average 45.48 (sd=5.97). These results show that, globally, users have good health literacy. Users are distributed by health literacy classes as: Inadequate (9 users), Elementary (13 users) and Good Literacy (18 users).

User familiarity with a topic depends on the task's subject. A global analysis demonstrated that topic familiarity is mostly low. As said before, CTF varies between 0 and 10 and its mean value is 3.92 with a standard deviation of 2.18. Pairs "user, topic" are distributed by the proposed topic familiarity categories as follows: Unfamiliar (161 pairs), Somehow familiar (113 pairs) and Familiar (46 pairs). Through this distribution, it can be seen that the majority of tasks presented a topic unfamiliar to the user.

In an open question, users were asked about the difficulties they have when performing health web searches. Two of the most frequent pointed issues were "finding medical terminology to formulate the query" and "dealing with the quantity of medico-scientific terminology found in the retrieved documents", both found in 21% of the answers.

Data Analysis

The data analysis was done using descriptive and inferential statistics. Differences between populations were visualized using boxplots that graphically describe variables and their dispersion depicting the 25th percentile (Q1) subtracted of 1.5 of the interquartile range, Q1, the median (Q2), the 75th percentile (Q3), Q3 plus 1.5 of the interquartile range and outliers.

In terms of inferential statistics it was used the strategy presented in Figure 1. Whenever possible it was applied a parametric test instead of a non-parametric due to the former's greater statistical power. The selection of the hypothesis test depends on the number of groups to be compared and on the scale of the variable that is being compared. Whenever a nominal variable is involved, as happens in almost all documents' characteristics, the test of equal proportions with the chi-squared value was used. Note that, when comparing two samples, the chi-squared test for equality of two proportions is the same thing as a z-test since the chi-squared distribution with one degree of freedom is the square of a normal deviate one. In situations where ordinal variables are involved, the Mann-Whitney test was employed and used the W letter to indicate the test value. In variables with a ratio scale, whenever it was possible the t-test was applied. In the other situations the Mann-Whitney test was applied. The only exception occurs in the SMOG analysis, where the Welch t-test was applied, because there were differences in both groups' variance. The Welch t-test is an adaptation of the t-test intended for use with two samples with unequal variances. When more than two groups are being compared the one way ANOVA or the Kruskal-Wallis test (KW) was initially applied to verify if there were significant differences between the groups and, if so, either a Tukey's test or a pairwise comparison was applied. In the pairwise comparison the

Bonferroni correction was applied, dividing α by the total number of comparisons to minimize the type I error. These comparisons identified location of the differences.

TAKE IN FIGURE 1

Document Characteristics Analysis (RQ1)

This study involved the evaluation of 1652 URLs. From these, 879 were retrieved through queries with lay terminology and 886 through queries with medico-scientific terms, with 113 URLs being retrieved through both types of queries.

As can be seen in Table 2, queries with medico-scientific terminology led to more HTTP Errors and more "no content" errors than queries without it. However, none of these differences are statistically significant.

In terms of document type, both types of queries retrieved mostly webpages (Table 2). This proportion is significantly higher in the first type of queries. The Portable Document Format (pdf) is the second most common type of document in both types of queries and its proportion is significantly higher in queries with medico-scientific terminology. Just like pdf documents, PowerPoint and Word documents are more frequent in queries with medico-scientific terminology, yet neither of these proportions is significantly different between types of queries. The larger proportion of non-webpages retrieved by queries with medico-scientific terminology indicates that this type of queries retrieves more documents not specifically built for the dissemination of health information on the Web.

TAKE IN TABLE 2

The proportion of HONCode certified pages is very similar in both types of queries (Table 2). As expected, queries with lay terminology retrieved more consumer-oriented documents (classified as "for patients" by the HON), a difference that is statistically significant. In terms of medico-scientific documents the opposite happens, that is, queries with medico-scientific terminology retrieve more medico-scientific documents (classified as "for health professionals" by the HON), also a statistically significant difference.

As explained before, document readability was assessed through the SMOG metric. Overall, SMOG ranged from 3.71 to 33.09 with a mean of 7.94 (sd=2.35). As expected, documents retrieved with queries containing medico-scientific terminology are more difficult to read (Table 2). Since the two samples are not homogeneous in variance, the Welch's t test was used and showed that the difference between both medians was statistically significant.

Summary.

Our findings show that, replacing lay query terms by medico-scientific ones results in retrieving a smaller proportion of webpages, a large proportion of pdf, less consumer-oriented documents, more professional-oriented documents and less readable documents. The smaller proportion of webpages indicates medico-scientific queries retrieve more documents not specifically built for the dissemination of health information on the Web. Since the HON classifies the majority of the documents retrieved with medico-scientific queries as "for health professionals", it can be concluded that users have to be better prepared to access contents retrieved with these queries. This is confirmed by these documents' lower readability.

Precision Analysis (RQ2)

To evaluate precision Graded Average Precision (GAP), Graded Precision at 5 (gP5) and Graded Precision at 10 (gP10) were used. These measures were proposed by Robertson, Kanoulas & Yilmaz (2010), based on a probabilistic model that generalizes average precision to the case of multi-graded relevance. These measures consider a model in which the user has a binary view of relevance even when using a non-binary scale of relevance. In this model, each point of relevance in the scale has a probability (g_i) of being the grade from which the user considers the documents relevant. The GAP and gP@n measures are defined as:

$$GAP = \frac{\sum_{n=1}^{\infty} \frac{1}{n} \sum_{m=1}^{n} \delta_{m,n}}{\sum_{i=1}^{c} R_i \sum_{j=1}^{i} g_j}$$

g_i – Probability that the user sets the threshold at grade i, i.e., in a relevance scale of {0..c}, he considers grades $i...c$ as relevant and the others as non-relevant.

$$\delta_{m,n} = \begin{cases} \sum_{j=1}^{\min(i_m, i_n)} g_j & if \quad i_m > 0 \\ 0 & otherwise \end{cases}$$

R_i - Total number of documents in grade i for this query

i_n - Relevance grade of document at rank n

$$gP@n = \frac{1}{n} \sum_{m=1}^{n} \frac{\sum_{j=1}^{\min(i_n, i_m)} g_j}{\sum_{j=1}^{i_n} g_j}$$

If i_n>0, document at rank n will contribute to the calculations.

More details on these measures can be seen in the paper of Robertson, Kanoulas & Yilmaz (2010). Based on the evaluation results presented by GAP's proponents, an equally balanced g1 and g2, i.e., g1=g2=0.5 was used.

These measures are based on relevance assessments made by the participants. Like in Borlund (2003), it is assumed that these assessments represent the value of the documents for a particular user at a particular moment, and thus can only be made by the user at that time. Additionally, while the current practice in IR involves the use of a gold standard to compute precision, this was intentionally done this way because this work is not interested in topical relevance as classic works usually are. Instead it is interested in situational relevance that encompasses cognitive relevance and can only be assessed through user judgments. Saracevic (1996) distinguishes these types of relevance as:

- Topical relevance: the relation between the query's topic and the documents' topic;
- Cognitive relevance: relation between the state of knowledge and cognitive information need of a user, and the retrieved documents, being inferred from criteria like cognitive correspondence and informativeness;
- Situational relevance: the relation between the task at hand and the retrieved documents, being inferred by criteria like usefulness in decision making, appropriateness of information in resolution of a problem and reduction of uncertainty.

Only by studying situational relevance can the influence of health literacy and topic familiarity be fully explored. For example, documents about the topic that are not understood by the user are not considered useful for the situation at hand.

The initial analysis is global and does not consider user context features. In Figure 2 six boxplots are presented. For each of the three precision measures, a boxplot is presented for each type of query, with and without medico-scientific terms. It is possible to see that queries containing medico-scientific terms tend to have higher precision with every measure. However, the only significant difference was found with gP10 at α=0.05 (t(317.6)=-1.70, p=0.045). This means that, in the top-10 results, medico-scientific queries retrieve a higher proportion of relevant documents.

TAKE IN FIGURE 2

GAP distribution by health literacy and query type can be visualized in Figure 3. Similarly to the global tendency, GAP tends to be higher in queries with medico-scientific terminology in every level of health literacy. However, no significant differences in GAP between types of query in each level of health literacy were found. In each type of query, the differences between levels of health literacy were also tested and no significant differences were found. Although non-significant, the higher GAP of medico-scientific queries in the lowest level of health literacy surprised us. Comparably to what happens with GAP, no statistically significant differences were found with gP5 and gP10 between types of query in each level of health literacy and between health literacy levels in each type of query.

In Figure 4 GAP distributions per query type and topic familiarity are presented. As with health literacy, there is a tendency to have higher GAP in medico-scientific queries in all levels of topic familiarity. However this is just a tendency since none of the differences is statistically significant. Similarly, there are no significant differences in the mean GAP between levels of topic familiarity in each type of query. In terms of gP5 and gP10, the only statistically significant difference was found on the "somehow familiar" level using gP5. Users of this level were found to have sessions with higher gP5 mean in queries with medical terminology than without it (t(111)=-2.1, p=0.019).

TAKE IN FIGURE 3 **TAKE IN FIGURE 4**

Summary.

Medico-scientific queries show a higher precision in the top-10 retrieved results. This agrees with previous studies (Patrick et al., 2001; Plovnick & Zeng, 2004; Zeng *et al*., 2002) that, through descriptive statistics, conclude that this type of queries leads to better results. The analysis by users' health literacy revealed no significant differences in all the comparisons made. This was surprising because medico-scientific queries were expected to have lower precision than lay queries in users with inadequate health literacy levels. Regarding the topic familiarity analysis, medico-scientific queries were found to have a higher precision in the top-5 retrieved results than lay queries on users "somehow familiar" with the topic. No significant differences were found in the mean GAP between levels of topic familiarity in each type of query what agrees with Al-Maskari & Sanderson (2010) who found no significant differences between familiar and unfamiliar users in the number of relevant documents.

Comprehension Analysis (RQ3)

In general, users understand documents well because the comprehension median is 2 (Totally understood) in a scale of 0 to 2. However, if this analysis was repeated by query type, it could be seen that, in lay queries, the median is still the same but, in medico-scientific queries, it drops to 1. These medians are significantly different (W = 13025482, $p< 2.2x10^{-16}$).

In Figure 5 the proportion of documents by level of health literacy, query type and comprehension level is presented. In this figure it is possible to see that, when compared with medico-scientific queries, comprehension is higher in documents retrieved with lay queries in every level of health literacy. Not only "totally understood" appears more often in lay queries but "not understood" documents also appear less. As can be seen in Table 3, all these differences are statistically significant.

TAKE IN FIGURE 5 **TAKE IN TABLE 3**

Moreover, Figure 5 also shows that users with higher health literacy "totally understand" more documents than users with inadequate health literacy, in both types of queries. The opposite happens with "not understood" documents. Using the Kruskal-Wallis test, statistically significant differences in document's comprehension between levels of health literacy were found (KW $\chi^2(2)=440.36$, $p<2.2x10^{-16}$ in lay queries and KW $\chi^2(2)=247.96$, $p<2.2x10^{-16}$ in medico-scientific queries). In a pairwise comparison (Table 4), it was found that comprehension in users with inadequate health literacy is lower than comprehension in users with elementary or good health literacy. Moreover, and unexpectedly, the comprehension of elementary health literate users was found to be higher than the one in users with good literacy.

TAKE IN TABLE 4

In Figure 6 the proportion of documents by level of topic familiarity, comprehension level and query type is presented. As can be seen, the comprehension of documents by users with different topic familiarities changes with the type of query. In line with the previous results, the comprehension of the documents is always higher in sessions with lay queries. As can be seen in Table 5 these differences are statistically significant.

TAKE IN FIGURE 6 **TAKE IN TABLE 5**

In lay queries, as expected, the comprehension of documents tends to increase with topic familiarity. In terms of significant differences, as can be seen in Table 6, it was found that unfamiliar users understand the documents retrieved with lay queries less well than other users.

TAKE IN TABLE 6

Surprisingly, with medico-scientific queries, users "somehow familiar" with the topic find documents harder to understand than users "unfamiliar" with the topic. Also, but now as expected, users "somehow familiar" with the topic understand documents less well than familiar users. As seen in Table 6 both these differences are statistically significant. In this type of queries, no significant differences between unfamiliar and familiar users were found.

Summary.

Documents retrieved with lay queries are comprehended better than documents retrieved with medico-scientific queries. This happens in general and also at all levels of health literacy and topic familiarity. In terms of health literacy, users with inadequate health literacy understand documents less well than users with higher health literacy using both types of queries. This is in agreement with the definition of health literacy. Birru *et al.* (2004), who studied information literacy instead of health literacy, reached a stronger but similar conclusion, concluding that low literacy users were unable to interpret the retrieved information. Surprisingly, the same happens in users with good health literacy when compared to users of elementary literacy. In terms of topic familiarity, users unfamiliar with the topic understand the documents retrieved with lay queries less well than other users. With medico-scientific queries, users "somehow familiar" with the topic understand documents less well than users familiar with the topic but also less well than users unfamiliar with it. The latter result leads to the conclusion that topic familiarity is not a necessary condition to comprehend a medico-scientific document. Characteristics like health literacy or knowledge about medico-scientific terminology may be more preponderant.

Medical Accuracy Analysis (RQ4)

As previously explained, users had to provide an answer to the information need that led to the task. A medical doctor later evaluated the answers in terms of correct and incorrect contents. These two assessments were then combined into what is called medical accuracy.

Figures 7, 8 and 9 show the distributions of the medical accuracy, correct contents and incorrect contents of the answer in each type of query. In these figures, it is possible to see that query terminology does not strongly affect these variables. The proportion of answers in each level of classification is similar and no significant differences were found in the median of each variable between types of queries.

TAKE IN FIGURE 7

TAKE IN FIGURE 8 **TAKE IN FIGURE 9**

As can be seen in Figure 10, despite a slight improvement of medical accuracy with the level of health literacy in both types of queries, the median of this variable is always 2. However, just as with answer's correctness and incorrectness, differences are not significant. Significant differences between levels of health literacy in each type of query were not found either. The distributions of medical accuracy by topic familiarity and query type can be seen in Figure 11. The median of this variable is always 2 and, against our expectations, medico-scientific queries seem to result in more accurate answers in users who are not familiar with the topic. Between query types in each level of familiarity, and between levels of familiarity in each type of query, no significant differences were found.

TAKE IN FIGURE 10 **TAKE IN FIGURE 11**

The median of answer's correctness is always 1 ("answer with some value") except in users familiar with the topic using medico-scientific queries, in which case it is 0 ("inappropriate answer"). In users familiar with the topic, the median of answer's correctness is significantly lower in medico-scientific queries (W=337.5, p = 0.03) when compared with lay queries. In other users, no significant differences were found. In medico-scientific queries there are differences in answers' correctness between levels of topic familiarity (KW $\chi^2(2)=11.72$, p=0.003). Further analysis led to the conclusion that, with this type of queries, users non-familiar with the topic give answers more accurate than those familiar with the topic (W=1540, p=7.45x10^{-12}<0.01/3). In lay queries, no significant differences were found. These results are surprising because familiar users were expected to be better prepared for medico-scientific queries and also to give better answers than other users, independently of the query type.

In terms of incorrect contents the tendency is symmetric to the one described above, i.e., in medico-scientific queries there is a slight tendency to have answers with less incorrect content as the familiarity with the topic increases. In non-familiar users the median is 1, in those who are somehow familiar it is 1.5 and in familiar users it is 2. In spite of this tendency, no significant differences between query types in each level of topic familiarity were found. In both query types no significant differences between levels of topic familiarity were found.

Summary.

The type of query does not affect answer's correctness, incorrectness and global accuracy, neither in the general user nor in users with specific levels of health literacy. In terms of topic familiarity and with respect to answers' correctness, familiar users give answers with less correct content with medico-scientific queries than with lay queries. Moreover, these users give answers with less correct content than non-familiar users with medico-scientific queries. Concerning answers' incorrectness, in medico-scientific queries, there is a tendency to have answers with less incorrect content as the familiarity with the topic increases, yet a non-significant difference. In medical accuracy, that combines the above measures, no significant differences were found. These results probably mean that familiar users were more restrained, less verbose when giving their answers what leads to answers with simultaneously less correct and less incorrect contents.

Motivational Relevance Analysis (RQ5)

Motivational relevance was evaluated through users' assessment of the task completion status in a scale of 1 (completely unsatisfied) to 5 (completely satisfied).

Since the median of the task completion status is 4 in both types of queries, it is possible to say that users were globally satisfied with the search sessions. The distributions in both types of queries are very similar denoting that the type of query does not interfere with users' feeling of success.

An analysis of the motivational relevance by health literacy (Figure 12) reveals that users with inadequate health literacy feel less satisfied than elementary or good health literacy users. There are significant differences between health literacy levels in both types of queries (lay - $\chi^2(2)=8.18$, p=0.017; medico-scientific - $\chi^2(2)=6.26$, p=0.044). At $\alpha = 0.05$, in lay queries, users with inadequate health literacy feel less satisfied than those with elementary (W= 647.5, p=0.005<0.05/3) and good health literacy (W=961, p=0.0086<0.05/3). In medico-scientific queries, users with inadequate health literacy feel less satisfied than elementary health literate users (W=680, p=0.01<0.05/3). As can be seen in Figure 12 there are no visible differences between types of queries in each level of health literacy. Through hypothesis tests the same conclusion was reached, i.e., there are no significant differences in the median of the task completion status between query types in each level of health literacy.

Although the median of motivational relevance is always 4 (Figure 13), this variable slightly increases with topic familiarity, independently of the query type. However, this is only a tendency since there are no significant differences between levels of topic familiarity in each type of query. Users familiar with the topic tend to be more satisfied with medico-scientific queries than with lay ones, but this difference is also not significant.

TAKE IN FIGURE 12 TAKE IN FIGURE 13

Summary.

In general, the type of query does not affect motivational relevance. The analysis by health literacy revealed that inadequate health literate users feel less satisfied than elementary and good health literate users with lay queries and less satisfied than elementary health literate users in medico-scientific sessions. Through these results it is possible to see that low health literacy has a negative impact on users' feeling of success in health search sessions. In terms of familiarity, users tend to be more satisfied with medico-scientific queries than with lay ones, but this difference is not significant.

Discussion and Implications

As demonstrated, medico-scientific queries demand more knowledge from users. Documents retrieved with these queries are mostly aimed at health professionals thus requiring users to be better prepared in health subjects. Moreover, these documents are less readable and are less well understood than documents retrieved with lay queries. When compared with lay queries, medico-scientific queries have higher precision at the top-10 retrieved documents. Although non-significant, medico-scientific queries surpass lay queries in gP5 and GAP. Considering GAP, the same happens in all levels of health literacy (non-significant differences). Surprisingly, the same trend happens in inadequate health literate users but, since these users have higher GAP than other users on both types of queries, this probably happens because "less subject expertise seems to lead to more lenient and relatively higher relevance ratings" (Saracevic, 2007, p. 2136). This means inadequate health literate users may give higher relevance scores than users with more health literacy to documents that are less helpful to them. Regarding medical accuracy, no significant differences were found but medico-scientific sessions slightly tend to generate knowledge with less incorrect contents and equal correct contents than lay sessions.

Comparing users with different levels of health literacy, users with inadequate health literacy understand documents less well and feel less successful than users with higher health-literacy, with both types of queries. This corroborates a previous explanation stating that the former type of users assign higher relevance scores to documents that are not helpful to them. Although not significant, answers' medical accuracy tends to increase with user's health literacy. This is due to the presence of less incorrect knowledge in users with more health literacy, another trend that was found. This is true on both types of queries but is stronger in medico-scientific sessions, showing that users with higher levels of health literacy are more apt to assimilate medico-scientific documents. These findings indicate that search engines should detect inadequate health literate users and return documents with contents adequate to them.

Concerning topic familiarity, users not familiar with a topic, when compared with other users, understand the documents retrieved with lay queries less well. In medico-scientific queries, "somehow familiar" users understand documents less well than non-familiar users. This means that, in medico-scientific documents, health literacy may be more important to document's comprehension than topic familiarity. In terms of non-significant differences, the quantity of incorrect contents in the knowledge that emerges from a medico-scientific session tends to decrease with topic familiarity. Moreover, users who are familiar with the topic tend to have higher motivational relevance and a higher GAP with medico-scientific queries when compared to lay queries.

Our findings suggest that a personalized query suggestion system would improve the IR experience in the health domain. The usefulness of a query suggestion system has also been recognized by Toms & Latter (2007) who examined consumers searching for health information. These authors concluded that systems that provide assistance to query development are more helpful than specialized medical search engines. They infer that the key to successful queries, one of the major challenges in this type of search, is in the underlying infrastructure that supports the search process, which should be responsive to both consumers and experts. However, it is the authors' opinion that personalization should not be bipolar and distinguish only health consumers from health professionals. The personalization of the query suggestion system should be made by level of health literacy and level of familiarity with the health topic, which change with the topic and the health consumer. According to our results, users who have inadequate health literacy or are unfamiliar with the topic should be provided with recommendations of lay queries. On the other hand, users with higher health literacy or topic familiarity should be given alternative queries with medico-scientific terminology.

A previous study suggests that non-expert domain expertise is dynamic and may be developing over time (White *et al*., 2009). The approach proposed in this chapter, when compared to the bipolar personalization strategy mentioned previously, does not have the drawback of hindering learning over time for health consumers. In fact, in users who are not unfamiliar with a topic, the system, through the queries it suggests and the documents it might give access to, supports and encourages people to learn more about the topic. Yet, in users unfamiliar with the topic, the suggestion of lay queries reinforces behavior. To address this gap, either the query suggestion system, or the system that predicts the familiarity with the topic, should take into account the number of previous searches on the topic. This information might help assess if the user is prepared to receive medico-scientific queries or even if he can raise one level in the scale of topic familiarity. This should, however, be carefully studied as further work. Moreover, this approach can only be effective if the system has access to all previous searches each user has done on the topic.

Conclusions and Future Work

A user study to analyze how changes in query terminology affect the health retrieval experience of users with different levels of health literacy and topic familiarity was conducted. Several aspects related to the information retrieval experience were studied, namely documents' readability, documents' comprehension, sessions' precision, sessions' medical accuracy and motivational relevance.

Results suggest that a personalized query suggestion system would improve information retrieval experiences in the health domain. Depending on the user, namely on his health literacy and topic familiarity, the system should provide medico-scientific or lay alternative suggestions to the query inserted by the user. This would not only give access to new types of documents but would also foster the learning of terminology that can be used in future queries. Reresults also suggest that users with inadequate health literacy and users who are unfamiliar with the topic, should be provided with recommendations of lay queries. On the other hand, users with higher health literacy or higher topic familiarity should be given alternative queries with medico-scientific terminology.

Although this research was focused on web search, its conclusions are applicable to other types of retrieval systems. For example, when searching for health information in a library retrieval system, users have similar problems in query formulation. In addition, they continue to have different levels of health literacy and topic familiarity and therefore comprehend and assimilate documents differently. In this context, it is clear that these users can also benefit from a query suggestion system like the one described above. Moreover, this type of system could also be useful to librarians that help end-users formulating queries or finding documents.

In addition to the suggestion system, findings also suggest that search engines should detect users with inadequate health literacy and return documents with contents adequate to them, either with pictorial contents or with higher levels of readability. Moreover, readability should be incorporated in search engines ranking algorithms. In fact, it was found that readability is important to all health consumers in both types of queries and that the relevance of a document highly depends on its comprehension. Health websites developers who want to provide information to consumers should also be aware that, if they need to use medico-scientific terminology, they should, at least, simplify the remaining contents.

As a future work it would be interesting to conduct a study specifically designed to analyze how topic familiarity can be predicted by means of past queries (e.g.: proportion of queries containing medico-scientific terminology), other types of retrieval behaviors or even information pertaining the users’ clinical history (for example, the time elapsed since the diagnosis of a condition may imply higher background knowledge on that specific disease/condition). The dynamic nature of topic familiarity makes this a challenging goal.

ACKNOWLEDGMENTS

Thanks to Fundação para a Ciência e a Tecnologia for partially funding this work under the grant SFRH/BD/40982/2007. Thanks to Dagmara Paiva, M.D., for her contribution on the evaluation of the medical accuracy of users’ answers.

References

Al-Maskari, A., & Sanderson, M. (2010). A review of factors influencing user satisfaction in information retrieval. *Journal of the American Society for Information Science and Technology, 61*(5), 859-868. doi:10.1002/asi.21300

Al-Maskari, A., & Sanderson, M. (2011). The effect of user characteristics on search effectiveness in information retrieval. *Information Processing and Management, 47*(5), 719-729. doi:10.1016/j.ipm.2011.03.002

Allan, J., (2003). HARD track overview in TREC 2003 high accuracy retrieval from documents. In *Proceedings of TREC 2003, NIST, (Special publication, Notebook No. 500-255)*, (pp. 24-37). Boston, MA: University of Massachusetts

Baker, D. Williams, V.W., Parker, R.M., Gazmararian, J.A., & Nurss, J. (1999). Development of a brief test to measure functional health literacy. *Patient Education and Counseling, 38*(1), 33-42. doi:10.1016/S0738-3991(98)00116-5

Becker, S. A. (2004). A study of web usability for older adults seeking online health resources. *ACM Transactions on Computer-Human Interaction, 11*(4), 387-406. doi:10.1145/1035575.1035578

Birru, M., Monaco, V., Charles, L., Drew, H., Njie, V., Bierria, T., . . . Steinman, R. (2004). Internet usage by low-literacy adults seeking health information: an observational analysis. *Journal of Medical Internet Research, 6*(3), e25.

Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research, 8*(3). Retrieved from http://www.informationr.net/ir/8-3/paper152.html

Capra, R.G., & Pérez-Quiñones, M. (2006). Factors and evaluation of refinding behaviors. In *SIGIR 2006 Workshop on Personal Information Management,* August 10-11, Seattle, WA, (pp.16-19). Retrieved from http://pim.ischool.washington.edu/pim06/files/capra-paper.pdf

Cline, R. J. W., & Haynes, K. M. (2001). Consumer health information seeking on the Internet: the state of the art. *Health Education Research, 16*(6), 671-692. *doi:10.1093/her/16.6.661*

Crain, S.P., Yang, S-H., Zha, H., & Jiao, Y. (2010). Dialect topic modeling for improved consumer medical research. In *AMIA Annual Symposium Proceedings, 2010* (pp. 132-136). Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041409/

Davis, T. C., Long, S. W., Jackson, R. H., Mayeaux, E. J., George, R. B., Murphy, P. W., & Crouch, M. A. (1993). Rapid estimate of adult literacy in medicine: A shortened screening instrument. *Family Medicine, 25*(6), 391-395.

Eerola, J., & Vakkari, P. (2008). How a general and a specific thesaurus cover expressions in patients' questions and physicians' answers. *Journal of Documentation, 64*(1), 131-142.

Fitzsimmons, P. R., Michael, B. D., Hulley, J. L., & Scott, G. O. (2010). A readability assessment of online Parkinson's disease information. *The Journal of the Royal College of Physicians of Edinburgh, 40*(4), 292-296. doi:10.4997/JRCPE.2010.401

Fox, S. (2006, October 29). *Online Health Search 2006*. Washington, DC: Pew Research Center. Retrieved from http://www.pewinternet.org/2006/10/29/online-health-search-2006/

Fox, S. (2011, February 1). *Health topics.* Washington, DC: Pew Research Center. Retrieved from http://www.pewinternet.org/2011/02/01/health-topics-2/

Fox, S., & Duggan, M. (2013). *Health online 2013*. Washington, DC: Pew Research Center. Retrieved from http://www.pewinternet.org/2013/01/15/health-online-2013/

Kelly, D., & Cool, C. (2002). The effects of topic familiarity on information search behavior. In *JCDL '02 Proceedings of the 2nd ACM/IEEE Joint Conference on Digital Libraries*, (pp. 74-75). New York, NY: ACM

Keselman, A., Massengale, L., Ngo, L., Browne, A., & Zeng, Q. (2006). The effect of user factors on consumer familiarity with health terms: Using gender as a proxy for background knowledge about gender-specific illnesses. In N. Maglaveras, I. Chouvarda, V. Koutkias, & R. Brause (Eds). *Biological and medical data analysis: 7th international symposium, ISBMDA 2006, Thessaloniki, Greece, December 7-8, 2006 Proceeding,* (pp.472-481). Berlin, Germany: Springer.

Kogan, S., Zeng, Q., Ash, N., & Greenes, R. A. (2001). Problems and challenges in patient information retrieval: A descriptive study. In *Proceedings of the AMIA annual symposium*, (pp. 329-333).

Kutner, M., Greenberg, E., Jin, Y., & Paulsen, C. (2006). *The health literacy of America's adults: Results from the 2003 national assessment of adult literacy,* (NCES 2006–483).Washington, DC: U.S.Department of Education and National Center for Education Statistics.

Lee, S-Y. D., Bender, D.E., Ruiz, R.E., & Cho, Y.I. (2006). Development of an easy-to-use Spanish health literacy test. *Health Services Research, 41*(4 Pt 1), 1392-1412. doi: 10.1111/j.1475-6773.2006.00532.x

Liu, J., & Belkin, N.J. (2010). Personalizing information retrieval for people with different levels of topic knowledge. In *Proceedings of the 10th annual joint conference on Digital libraries, Gold Coast, Queensland, Australia, June 21-25,* (pp.385-386). doi:10.1145/1816123.1816125

Lopes, C., & Ribeiro, C. (2010). Context effect on query formulation and subjective relevance in health searches. In *IIiX '10: Proceeding of the third symposium on information interaction in context*, (pp. 205-214). New York, NY: ACM. doi:10.1145/1840784.1840815

Lu, W-H., Lin, R.S-J., Chan, Y-C., & Chen, K-H. (2006). Overcoming terminology barrier using web resources for cross-language medical information retrieval. In *AMIA Annual Symposium Proceedings*, *Washington, DC, November 11-15*, (pp. 519-523).

Luo, G. (2009). Design and evaluation of the iMed intelligent medical search engine. In *Proceedings of the IEEE 25th International Conference on Data Engineering, Shanghai, China, March 29-April 2, 2009*, (pp. 1379-1390). New York, NY: IEEE. doi: doi:10.1109/ICDE.2009.10

Luo, G., Tang, C., Yang, H., & Wei, X. (2008). MedSearch: a specialized search engine for medical information retrieval. In *Proceedings of the 16th Internation Conference on the world wide web,* (pp. 1175-1176). New York, NY: ACM. doi:10.1145/1242572.1242752

McCray, A. T. (2005). Promoting health literacy. *Journal of the American Medical Informatics Association, 12*(2), 152-163. *doi:10.1197/jamia.M1687*

McCray, A., & Tse, T. (2003). Understanding search failures in consumer health information systems. In *AMIA Annual Symposium Proceedings,* (pp. 430-434). Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1479930/

Muresan, G., Cole, M., Smith, C. L., Liu, L., & Belkin, N. J. (2006). Does familiarity breed content? Taking account of familiarity with a topic in personalizing information retrieval. In *Proceedings of the 39th Annual Hawiian International Conference on System Sciences,* (p. 53c). doi:10.1109/HICSS.2006.130

Parker, R., Baker, D.W., Williams, M.W.., & Nurss, J.R (1995). The test of functional health literacy in adults. *Journal of General Internal Medicine, 10*(10), 537-541. doi: 10.1007/BF02640361

Patrick, T. B., Monga, H. K., Sievert, M.R., Hall, J.H., & Longo, D. R. (2001). Evaluation of controlled vocabulary resources for development of a consumer entry vocabulary for diabetes. *Journal of Medical Internet Research, 3*(3), e24. doi:10.2196/jmir.3.3.e24

Petrock, V. (2010, August 10). Cyberchondriacs becoming empowered health information seekers. *eMarketer*™. Retrieved from http://www.webcitation.org/5ym7xLoYp

Plovnick, R.M., & Zeng, Q.T. (2004). Reformulation of consumer health queries with professional terminology: A pilot study. *Journal of Medical Internet Research, 6*(3), e27. doi:10.2196/jmir.6.3.e27

Qu, P., Liu, C., & Lai, M. (2010). The effect of task type and topic familiarity on information search behaviors. In *Proceeding of the third symposium on information interaction in context*, (pp. 371-376). New York, NY: ACM. DOI: 10.1145/1840784.1840841

Robertson, S.E., Kanoulas, E., & Yilmaz, E. (2010). Extending average precision to graded relevance judgments. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, July19-23, 2010, Geneva, Switzerland*, (pp. 603-610). New York, NY: ACM. doi:10.1145/1835449.1835550

Saracevic, T. (1996). Relevance reconsidered. In *Proceedings of the second conference on conceptions of library and information science (CoLIS 2)*, Copenhagen, Denmark, October 14-17, (pp.201-218).

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the Association for Information Science and Technology, 58*(13), 2126-2144. doi:10.1002/asi.20682

Sihvonen, A., & Vakkari, P. (2004). Subject knowledge, thesaurus-assisted query expansion and search success. In *Proceedings of RIAO 2004, the 7th International Conference, Avignon, France, April 26-28,* (pp. 393-404).

Stichele, R.V. (1995). *Multilingual glossary of technical and popular medical terms in nine European languages: Final report.* Gent, Belgium: University of Gent, Heymans Institute of Pharmacology. Retrieved from http://users.ugent.be/~rvdstich/eugloss/information.html

Toms, E.G., & Latter, C. (2007). How consumers search for health information. *Health Informatics Journal, 13*(3), 223-235. doi:10.1177/1460458207079901

Wang, Y., & Liu, Z. (2005). Personalized health information retrieval system. In *AMIA annual symposium proceedings,* October 22-26, Washington, DC (p. 1149).

Wen, L., Ruthven, I., & Borlund, P. (2006). The effects on topic familiarity on online search behaviour and use of relevance criteria. In M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika & A. Yavlinsky (Eds.), *Advances in Information Retrieval* (vol. 3936, pp. 456-459): Berlin, Germany: Springer-Verlag.

White, R.W, Dumais, S., & Teevan, J. (2008). How medical expertise influences web search interaction. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, July 20-24*, Singapore, Singapore. doi:10.1145/1390334.1390506

White, R.W, Dumais, S., & Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the second ACM international conference on web search and data mining, February 09-11, Barcelona, Spain* (pp. 137-141). doi:10.1145/1498759.1498819

Wildemuth, B. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology, 55*(3), 246-258. doi:10.1002/asi.10361

Zeng, Q.T, Crowell, J., Plovnick, R.M., Kim, E., Ngo, L., & Dibble, E. (2006). Assisting consumer health information retrieval with query recommendations. *Journal of the American Medical Informatics Association , 13*(1), 80-90. doi:10.1197/jamia.M1820

Zeng, Q., Kogan, S., Ash, N., Greenes, R. A., & Boxwala, A. A. (2002). Characteristics of consumer terminology for health information retrieval. *Methods of Information in Medicine, 41*(4), 289-298.

Zeng-Treitler, Q., Goryachev, S., Tse, T.T., Keselman, A., & Boxwala, A. (2008). Estimating consumer familiarity with health terminology: A context-based approach. *Journal of the American Medical Informatics Association, 15*(3), 349-356. doi:10.1197/jamia.M2592

Zhang, Y. (2010). Contextualizing consumer health information searching: an analysis of questions in a social Q&A community. In *Proceedings of the 1st ACM International Health Informatics - IHI '10, November 11-12, Washington, DC*. New York, NY: ACM.

Zielstorff, R.D. (2003). Controlled vocabularies for consumer health. *Journal of Biomedical Informatics, 36*(4-5), 326-333. doi:10.1016/j.jbi.2003.09.015

TABLES

Table 1 – Context features used in this study.

Category	Context feature	Scale	Description
User	Health literacy	SAHLSA score between 0 and 50. Grouped in Inadequate, Elementary and Good Health Literacy.	Grade obtained in the adapted SAHLSA health literacy assessment test.
User & Document	Relevance	Ordinal scale of 0 (not relevant) to 2 (totally relevant).	Obtained through users' assessment for each document.
	Comprehension	Ordinal scale of 0 (not understood) to 2 (totally understood).	Obtained through users' assessment for each document.
User & Task	Combined Topic Familiarity	Varies between 0 and 10. Grouped in three classes: unfamiliar, somehow familiar and familiar.	Combined three user's assessments: task familiarity, previous searches on the topic and knowledge on the medico-scientific term behind the topic.
	Answer's medical accuracy	Varies between 0 and 4.	Obtained through a medical evaluation of users' answers in terms of their correct and incorrect contents.
	Motivational relevance	Ordinal scale of 1 (disagree) to 5 (agree).	Motivational relevance relates the user's goals and motivations with the information objects. It is expressed by the user's feeling of success and his satisfaction (Saracevic, 1996). Obtained through the following post-search question "I believe I succeed in this task" as perceived by the user.
Document	Readability	Rational.	Obtained through the SMOG readability measure. Automatically computed.
	Type of document	Nominal: webpage, pdf, ppt, doc or other.	Identified by users and manually validated when inconsistencies were found.
	HONCode certification	Nominal: yes or no.	Positive if it is in the answer set of both retrieval systems. Automatic extraction.
	For consumers?	Nominal: yes or no.	Positive if it belongs to the consumers' category in the HONCode classification of documents. Automatically computed.

	For professionals?	Nominal: yes or no.	Positive if it belongs to the professionals’ category in the HONCode classification of documents. Automatically computed.

Table 2 – Differences in documents’ features for both types of queries.

	Feature	Lay queries	Medico-scientific queries	Significant differences?
Errors	% HTTP errors	0.57%	1.02%	No
	% No content errors	0.23%	0.68%	No
Document Type	% Webpages	95.8%	86.22%	$\chi^2(1) = 47.17$, $p=3.25x10^{-12}$
	% PDF	4.13%	12.7%	$\chi^2(1) = 40.78$, $p<8.54x10^{-11}$
	% PowerPoint docs	0.11%	0.57%	No
	% Word docs	0%	0.46%	No
HON Certification	% Certified pages	53.01%	52.71%	No
	% Consumer-oriented pages	33.9%	26.64%	$\chi^2(1) = 10.7$, $p=5x10^{-4}$
	% Professional oriented pages	5.12%	14.67%	$\chi^2(1) = 44.02$, $p=1.62x10^{-11}$
Readability	Mean SMOG	7.17	7.69	Welch’s t(8751.1)=-9.06, $p<2.2x10^{-16}$

Table 3 – Significant differences between the median of comprehension in both types of queries, by health literacy level.

	$Comp_{Lay} > Comp_{MS}$
Inadequate HL	W = 636653 $p = 1.104x10^{-7}$
Elementary HL	W = 1398119 $p < 2.2x10^{-16}$
Good HL	W = 2645866 $p < 2.2x10^{-16}$

Table 4 – Significant differences between medians of comprehension between levels of health literacy (I-Inadequate, E-Elementary, G-Good), by query type.

	Lay queries	Medico-scientific queries
$Comp_{hl=I} < Comp_{hl=E}$	W = 505318 $p < 2.2x10^{-16} < 0.01/3$	W = 566613 $p < 2.2x10^{-16} < 0.01/3$
$Comp_{hl=I} < Comp_{hl=G}$	W = 791604	W = 845470.5

	$p < 2.2\text{x}10^{-16} < 0.01/3$	$p < 2.2\text{x}10^{-16} < 0.01/3$
$\text{Comp}_{\text{hl=E}} > \text{Comp}_{\text{hl=G}}$	W = 1803308 $p = 6.366\text{x}10^{-9} < 0.01/3$	W = 1715570 $p = 13.06\text{x}10^{-5} < 0.01/3$

Table 5 - Significant differences between the median of comprehension in both types of queries, by topic familiarity level.

	$\text{Comp}_{\text{Lay}} > \text{Comp}_{\text{MS}}$
Unfamiliar	W = 3038410 $p = 1.583\text{x}10^{-6}$
Somehow familiar	W = 1785294 $p < 2.2\text{x}10^{-16}$
Familiar	W = 279186.5 $p = 9.079\text{x}10^{-16}$

Table 6 - Statistical differences between medians of comprehension between levels of topic familiarity (U-Unfamiliar, S-Somehow familiar, F-Familiar), by query type.

Lay queries		Medico-scientific queries	
$\text{Comp}_{\text{tf=U}} < \text{Comp}_{\text{tf=S}}$	W = 1796246 $p < 2.2\text{x}10^{-16} < 0.01/3$	$\text{Comp}_{\text{tf=U}} > \text{Comp}_{\text{tf=S}}$	W = 2068034 $p = 1.081\text{x}10^{-6} < 0.01/3$

$\text{Comp}_{\text{tf=U}} < \text{Comp}_{\text{tf=F}}$	W = 628375
	$p = 8.99\ \text{x}10^{-16} < 0.01/3$
		$\text{Comp}_{\text{tf=U}} < \text{Comp}_{\text{tf=F}}$	W = 827630.5
			p = 0.04279

$\text{Comp}_{\text{tf=S}} < \text{Comp}_{\text{tf=F}}$	W = 507711.5
	p = 0.01893
		$\text{Comp}_{\text{tf=S}} < \text{Comp}_{\text{tf=F}}$	W = 5345930
			$p = 1.757\text{x}10^{-7} < 0.01/3$

FIGURES

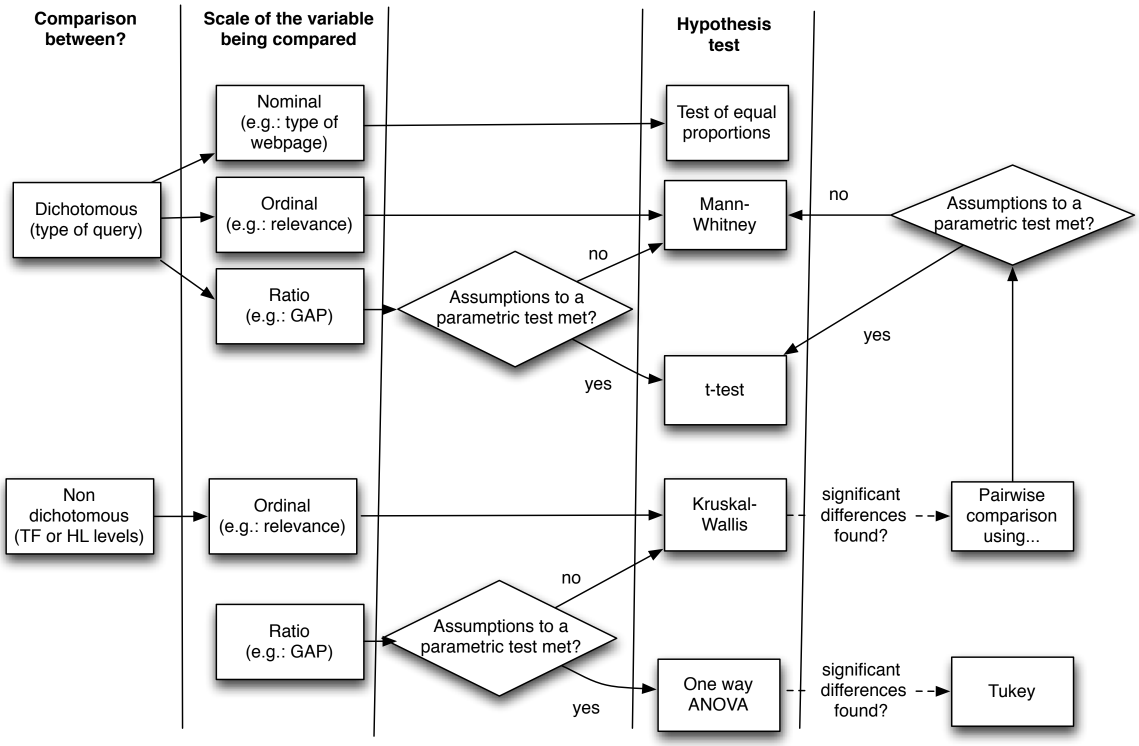
Comparison between?
Scale of the variable being compared
Hypothesis test
Dichotomous (type of query)
Nominal (e.g.: type of webpage)
Ordinal (e.g.: relevance)
Ratio (e.g.: GAP)
Assumptions to a parametric test met?
no
yes
Test of equal proportions
Mann-Whitney
t-test
no
Assumptions to a parametric test met?
yes
Non dichotomous (TF or HL levels)
Ordinal (e.g.: relevance)
Ratio (e.g.: GAP)
Assumptions to a parametric test met?
no
yes
Kruskal-Wallis
One way ANOVA
significant differences found?
Pairwise comparison using...
significant differences found?
Tukey

Figure 1 – Inferential statistical strategy.

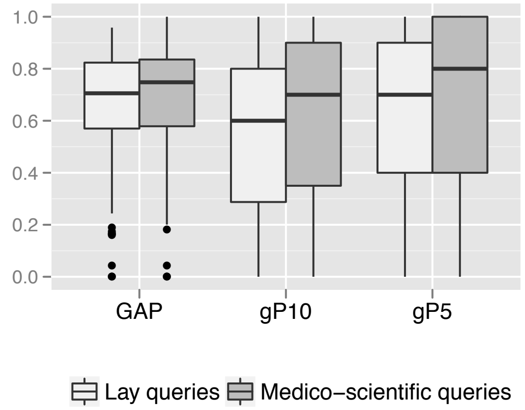

Figure 2 – GAP, gP10 and gP5 boxplots by type of query.

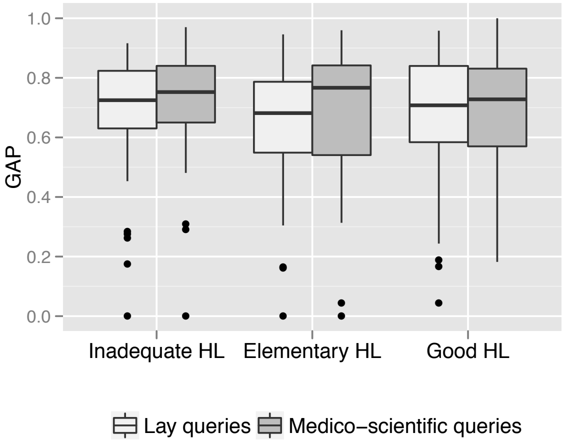

Figure 3 – GAP by type of query and health literacy level.

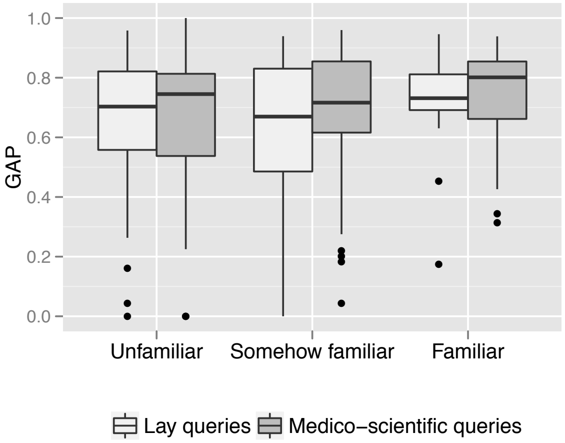

Figure 4 – GAP boxplots by type of query and topic familiarity level.

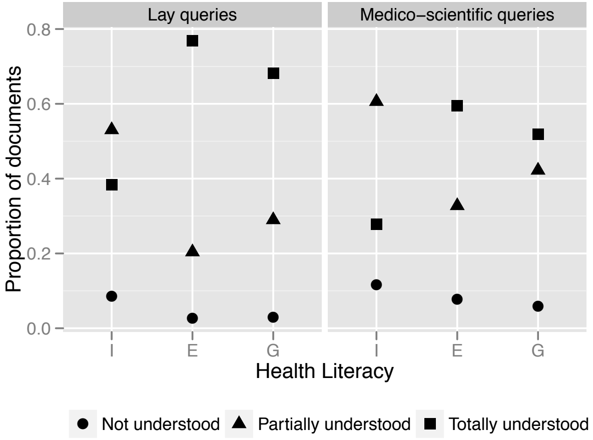

Figure 5 – Proportion of documents by health literacy (I-Inadequate, E-Elementary, G-Good), query type and comprehension level.

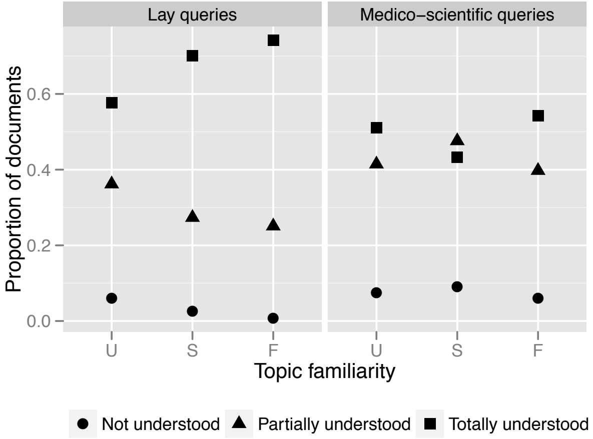

Figure 6 - Proportion of documents by topic familiarity (U-Unfamiliar, S-Somehow familiar, F-Familiar), query type and comprehension level.

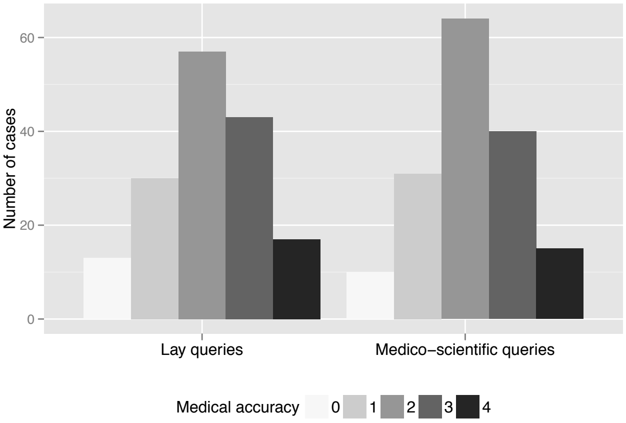

Figure 7 - Answer's medical accuracy by query type.

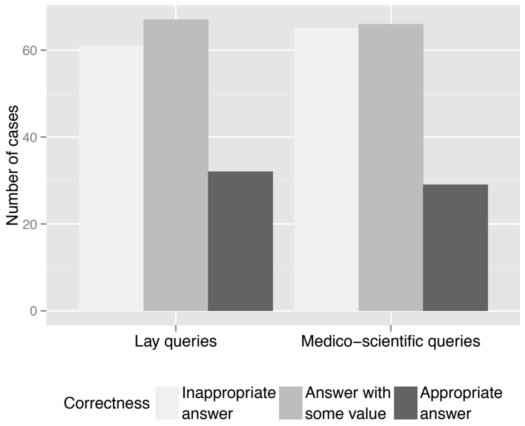

Figure 8 – Answer's correctness by query type.

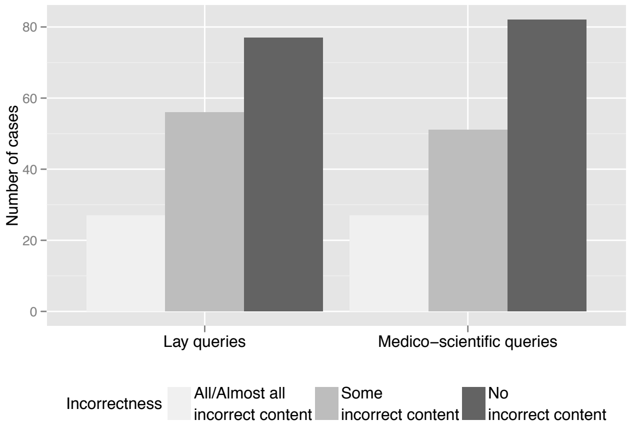

Figure 9 - Answer's incorrectness by query type.

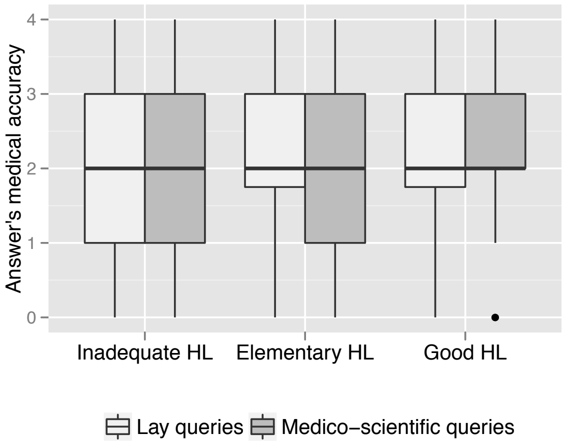

Figure 10 – Medical accuracy by health literacy and query type.

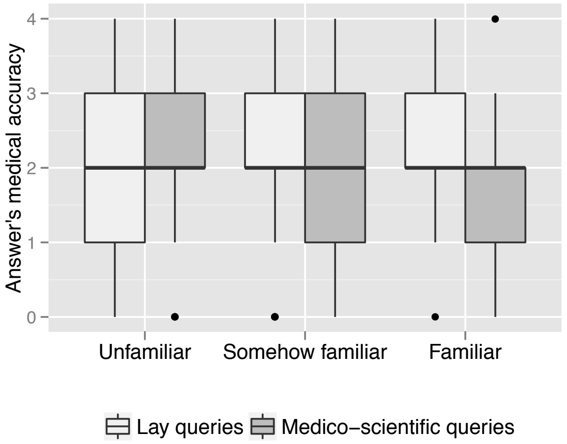

Figure 11 – Medical accuracy by topic familiarity and query type.

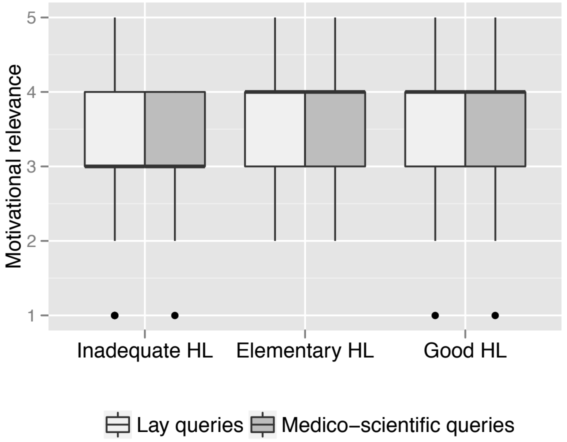

Figure 12 – Motivational Relevance by health literacy level and query type.

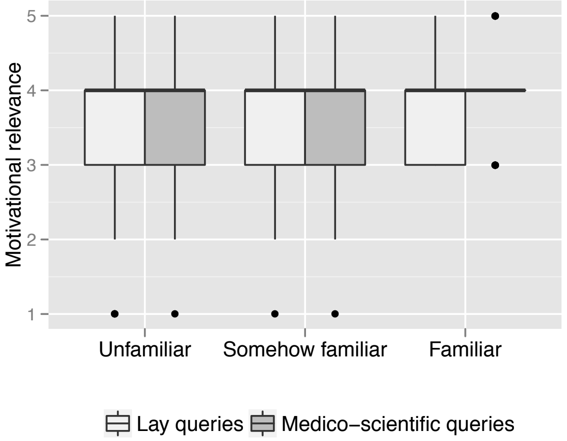

Figure 13 – Motivational Relevance by topic familiarity and query type.