# Effects of language and terminology of query suggestions on medical accuracy considering different user characteristics

Carla Teixeira Lopes[1,2*] Dagmara Paiva[3,4]
Cristina Ribeiro[1,2]

ctl@fe.up.pt, dpaiva@med.up.pt, mcr@fe.up.pt

[1]Faculdade de Engenharia, Universidade do Porto, Porto, Portugal
[2]INESC TEC, Porto, Portugal
[3]ISPUP-EPIUnit, Universidade do Porto, Porto, Portugal
[4]Unidade de Saúde Familiar Monte Murado, Vila Nova de Gaia, Portugal

**Abstract**

Searching for health information is one of the most popular activities on the Web. In this domain, users often misspell or lack knowledge of the proper medical terms to use in queries. To overcome these difficulties and attempt to retrieve higher-quality content, we developed a query suggestion system that provides alternative queries combining the Portuguese or English language with lay or medico-scientific terminology. Here, we evaluate this system's impact on the medical accuracy of the knowledge acquired during the search. Evaluation shows that simply providing these suggestions contributes to reduce the quantity of incorrect contents. This indicates that even when suggestions are not clicked, they are useful either for subsequent queries' formulation or for interpreting search results. Clicking on suggestions, regardless of its type, leads to answers with more correct content. An analysis by type of suggestion and user characteristics showed that the benefits of certain languages and terminologies are more perceptible in users with certain levels of English proficiency and health literacy. This suggests a personalization of this suggestion system towards these characteristics. Overall, the effect of language is more preponderant than the effect of terminology. Clicks on English suggestions are clearly preferable to clicks in Portuguese ones.

*Keywords*: information retrieval, query suggestion, cross-lingual retrieval, health, personalization, medical accuracy, user study.

*Corresponding author.

# 1 Introduction

Searching for health information is the third most popular online activity, following email and using a search engine, being performed by 80% of U.S. Internet users (Fox, 2011). This domain poses specific challenges to health consumers, who frequently encounter additional difficulties in finding the correct terms to include in their queries (Zeng et al., 2006; Kriewel & Fuhr, 2010; Zhang, 2011). They not only lack knowledge of the proper medical terms (Zhang, 2010; Toms & Latter, 2007) but also often misspell medical terms (Kogan, Zeng, Ash, & Greenes, 2001; McCray & Tse, 2003).

Typically, Information Retrieval (IR) systems are evaluated regarding the topical relevance of their results. While this is obviously important in the health domain, other aspects also need to be considered to assure the accuracy of the knowledge obtained in the search session, essential to avoid health risks. IR systems should provide high-quality contents and, simultaneously, assure that users understand them. Previously we have concluded that English suggestions should be provided to users with higher levels of English proficiency (Lopes & Ribeiro, 2013), opening doors for higher-quality contents. We have also concluded that search engines should propose queries using lay and medico-scientific terminology (Lopes & Ribeiro, 2015) what may be used to adjust documents to the literacy of the users.

The difficulties in query formulation mentioned and our previous conclusions (Lopes & Ribeiro, 2013, 2015) motivated the development of a system that, based on an initial user query, suggests 4 different queries combining two languages (English and Portuguese) and two bodies of terminology (lay and medico-scientific). In this work, we evaluate the effect of the suggested queries on the medical accuracy of the obtained knowledge, in general and by query's language and terminology.

To the best of our knowledge, no previous works have explored cross-language query suggestions in the health domain. Moreover, our work also innovates because it considers different groups of users defined by their English proficiency, health literacy and topic familiarity. Note that, although previous studies have concluded that search assistance should be personalized to achieve its maximal outcome (Jansen & McNeese, 2005), little attention has been paid to how people perform query reformulations across different user groups.

The remainder of this article is structured as follows. Section 2 describes related work reported in the literature. In Section 3, we present the prototype we have developed to suggest alternative health queries. Section 4 describes the user study we conducted to evaluate whether the query suggestion prototype contributes to a more successful search experience. In Section 5 we detail the statistical strategy followed during the analysis of the data. We describe the findings of our investigation in Section 6 and discuss them in Section 7. We summarize and conclude in Section 8.

## 2 Related Work

In the following subsections, we begin by describing works that report efforts to support the formulation of health queries by laypeople. Given the nature of our query suggestion system, we also describe one work that, like ours, proposes suggestions in a language different from the one used in the original query. Considering our focus on medical accuracy, the last subsection details how others have considered medical accuracy in the evaluation of their systems.

### 2.1 Query formulation support in consumer health search

In consumer health information retrieval, there is an awareness that several difficulties can emerge due to the terminology gap between medical experts and lay people (Zielstorff, 2003). To overcome these difficulties in query formulation, some authors have proposed query expansion approaches. The Health Information Query Assistant proposed by Zeng et al. (2006) suggests terms based on their semantic distance from the original query. To compute this distance, the authors use co-occurrences in medical literature and log data as well as the semantic relations in medical vocabularies. A user study with 213 subjects randomized into 2 groups, one receiving suggestions and the other not receiving them, showed that recommendations resulted in higher rates of successful queries, i.e., queries with at least one relevant result among the top 10, but not in higher rates of satisfaction neither higher scores on the answer given to the predefined task.

Liu, Zhenyu, Chu, and Wesley (2007) propose a query expansion method exploiting the UMLS (Unified Medical Language System) to append the original query with terms that are relevant to the query's scenario. The evaluation was done with two test-collections, OHSUMED and McMaster Clinical HEDGES Database, and showed that a scenario-specific expansion is preferable than a statistical-based one in terms of precision and recall.

Two proposed search engines for health information retrieval — iMed (Luo & Tang, 2008) and MedSearch (Luo, Tang, Yang, & Wei, 2008) — provide suggestions of medical phrases to assist users in refining their queries. In these systems, the phrases are extracted and ranked based on MeSH (Medical Subject Headings), the collection of crawled webpages, and the query. Zarro and Lin (2011) presented a search system that also uses MeSH together with social tagging to provide users with lay and medico-scientific terms. To evaluate the impact of these suggestions, the authors conducted a user study with 10 lay subjects and 10 expert subjects. They found no differences in the behaviour of the two groups. Both groups preferred MeSH terms because their quality was considered superior to the quality of social tags.

Using three synonym mappings, Soldaini, Yates, Yom-Tov, Frieder, and Goharian (2016) proposed a system that clarifies queries formulated in lay terminology with medico-scientific terminology. With two task-based studies, authors studied the utility of this technique to the ability to correctly answer a medical question. Fattahi, Wilson, and Cole (2008) proposed a query expansion method

that uses non-topical terms (terms that occur before or after topical terms to represent a specific aspect of the theme, such as 'about' in 'about breast cancer') and semi-topical terms (terms that do not occur alone, such as 'risk of' in 'risk of breast cancer') in conjunction with topical terms (terms that represent the subject content of documents, such as 'breast cancer'). The authors found that web searches can be enhanced by the combination of these three types of terms.

Other studies approach query formulation using machine-learning techniques (Stanton, Ieong, & Mishra, 2014) or log-based ones (Dang, Kumaran, & Troy, 2012). Stanton et al. (2014) studied circumlocution, that is, using more words than necessary to describe something, in health queries. Given an informal query, authors identify the underlying professional concept. Dang et al. (2012) focused on domain dependent query reformulation using two different domains: health and commerce. Using two large query logs, authors show that a dependent approach outperforms an independent one.

## 2.2 Cross-language query suggestions

Although not in the specific area of health information retrieval, we identified only one work involving the proposal of query suggestions in a language different from the original query's language, namely, a study performed by Gao et al. (2010). The authors proposed a method to translate generalist queries using query logs and then estimate the cross-lingual query similarity using information such as word translation relations and word co-occurrence statistics. The evaluation was performed on French-English and Chinese-English tasks. Authors found that these suggestions, when used in combination with pseudo-relevance feedback, improved the effectiveness of cross-language information retrieval.

## 2.3 Evaluation of medical accuracy in consumer health search

Typically, IR systems are evaluated regarding the topical relevance of their results. In the health domain, besides relevance, it is also important to consider the medical accuracy of the knowledge obtained in the search session, a feature that is not associated with recall and precision (Hersh et al., 2002). From the studies described in the previous section, only three consider the medical accuracy in the evaluation of the systems.

Zeng et al. (2006) assessed the Health Information Query Assistant considering three perspectives: user satisfaction, relevance of the results and the answer given by users to the predefined task. To analyze the third outcome, authors graded the answers given by users according to a gold standard. A correct answer was given a score of 1, an incorrect answer was given a score of -1 and the absence of answer was graded as 0. Since users were asked to find 5 risk factors for heart disease or 3 treatments for baldness, all scores to a question were summed up and divided by 5 or 3. Zeng et al. (2006) argues that "being misinformed may be more harmful than being uninformed" what agrees

with the rationale underlying our decision of assessing answers correctness and incorrectness separately.

iMed (Luo & Tang, 2008), one of the systems described in the previous section, was evaluated with objective (the success rate, the number of search iterations, the number of search result Web pages viewed, and the time spent on the search process) and subjective (ease of using the system, ease of understanding the system, and overall satisfaction) performance measures. Authors used real medical case records and questions from the United States Medical Licensing Examination to assess the system. At the end of the search, users listed up to three diseases that best match the medical case. If one of the diseases is one of the correct diagnoses, the search is considered successful. The success rate is measured by the number of successful searches.

Similarly to what we do in this work, Soldaini et al. (2016) assessed their system considering users' question answering accuracy. However, the methodology to assess medical accuracy is different from ours. In their study, authors formulated a multiple-choice question with one correct and three wrong answers, to which users had to answer. In our view, our methodology is more natural since users are not required to answer a multiple-choice question and are not influenced by it in any way. To eliminate knowledge differences between users, in our work we assess the accuracy of the answers before and after the search. Moreover, we also consider the proportion of incorrect contents on the answer.

# 3 Suggestion Tool

We designed and developed a prototype for a suggestion tool that can be integrated into IR systems. Given a health query, our tool suggests alternative queries in two languages, Portuguese and English, using medico-scientific and lay terminology. As an example, for the Portuguese lay query 'tumor abdominal', the system suggests the following queries: 'abdominal tumor' (English lay query), 'abdominal neoplasm' (English and medico-scientific query) and 'neoplasia abdominal' (Portuguese and medico-scientific query). In Figure 1, we present the architecture of the suggestion tool, which will be further detailed in the following paragraphs.

## 3.1 Data structures

The system uses the open-access and collaborative Consumer Health Vocabulary (OAC CHV), which is available from the Unified Medical Language System (UMLS) and is intended to connect "informal, common words and phrases about health to technical terms used by health care professionals" (NLM, 2012). Efthimiadis (1996) classifies this knowledge structure as collection-independent, and because it is based on a thesaurus, it is considered to be a global method (Manning, Raghavan, & Schütze, 2008). The latest version of OAC CHV contains 57,795 health concepts and 146,324 English concept strings.

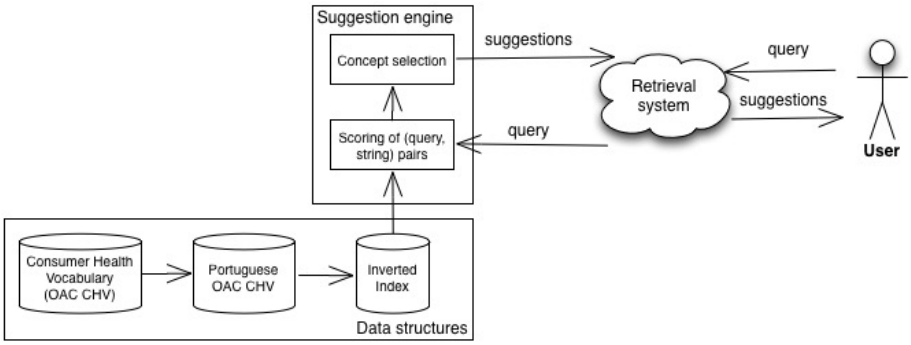


Figure 1: Architecture of the suggestion tool.

An OAC CHV concept is identified by a unique UMLS identifier and may be associated with several synonym strings that express that concept. Each OAC CHV concept is also associated with a OAC CHV preferred name and a UMLS preferred name. The OAC CHV preferred name is the string that best represents that concept for health consumers and is defined by the OAC CHV, whereas the UMLS preferred name is the preferred string for that concept as defined by the UMLS.

The OAC CHV vocabulary was translated into Portuguese using the Google Translator API (Application Programming Interface). To evaluate the translation process, 1620 randomly selected strings were assessed, that is, approximately 1% of the total number of strings in the OAC CHV vocabulary. Two researchers participated in this evaluation to classify each translation into one of the following categories: "correct translation", "incorrect translation", "translation to Brazilian Portuguese (pt-BR)", or "insufficient knowledge to assess the translation". To assess the inter-annotator agreement, we computed Cohen's kappa; we obtained a value of 0.94 (95% CI: [0.92, 0.96]), indicating almost perfect agreement. At the end of the evaluation, the researchers met again to make a final decision about the classifications. The researchers discussed each disagreement until consensus was reached. We found that 84.2% (95% CI: [82.3%, 85.9%]) of the translations were correct, 9.7% (95% CI: [8.3%, 11.3%]) of the strings were translated into pt-BR, and 3.6% (95% CI: [2.8%, 4.6%]) were incorrect translations; for the remaining 2.5% (95% CI: [1.8%, 3.5%]) of the strings, doubts regarding the translation persisted.

Using the Portuguese translation of the OAC CHV, we created an inverted index in which each stemmed term is associated with an inverse string frequency ($\mathrm{isf_t}$) and a postings list, i.e., a list of the strings in which the stemmed term appears. The computation of the inverse string frequency is similar to the computation of the inverse document frequency that is traditionally performed in IR, that is, $\mathrm{isf_t} = \log(\mathrm{N}/\mathrm{sf_t})$, where $\mathrm{sf_t}$ is the number of strings in which the term appears and $N$ is the total number of strings. Because strings are typically small, the probability of finding multiple occurrences of the same term

in a string is very small. For this reason, we decided to weight each term based only on its $\mathrm{isf_t}$, ignoring its frequency in each string ($\mathrm{tf_{t,s}}$). To determine the vocabulary of terms, namely, the list of terms in our inverted index, the strings were tokenized and stop words were removed. In the terms, all letters were converted to lower case and the accents were removed, and the terms were also stemmed.

## 3.2 Suggestion engine

The score assigned to each (query, string) pair is defined by $\mathrm{score(q, s)} = \sum_{\mathrm{t \in q}} \mathrm{isf_t}$. Because the length of strings and queries has a very small variance, we found that the additional computational power required to normalize the above score formula would not be justified by the gains thus achieved.

In this stage of the prototype development, to limit the number of suggestions, we decided to select only the string with the maximum score for each input query. For this string, we identify the associated concept and then return its OAC CHV and UMLS preferred names in English and Portuguese. If a suggestion is identical to the query or to any other suggestion, it will not be presented, i.e., the output of the system will contain only unique suggestions different from the query.

## 3.3 Retrieval systems

Two retrieval systems were used in this study: one that incorporated the developed suggestion tool and presents query suggestions (SYS+) and another that did not (SYS). Both of them used the Bing Search API to obtain web results for users' queries. To increase the usability of the interface with regard to learning, we decided to keep the interfaces very simple and similar to those used in the most popular search engines (Figure 2). The interface of the system with suggestions was presented in hues of blue, and the other, in hues of pink.

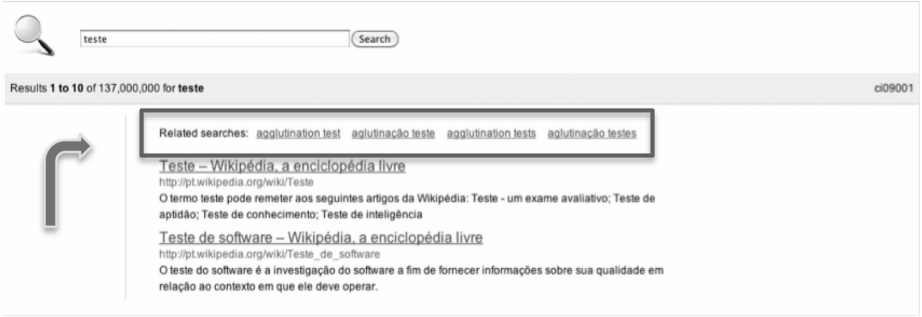


Figure 2: Results page for the system presenting query suggestions.

Users interact with the retrieval system by issuing a query, which may be an initial query or a subsequent query and may or may not be influenced by the

suggestions presented in the previous iteration. The suggestions are integrated into the search results as shown in Figure 2, that is, before the first result. The label *Related searches* is used to point to suggestions.

In both retrieval systems, we monitored the users' behavior through access logs. We developed a log mechanism that registered every action made on the system, namely, issued queries, the search results pages (SERP) presented, the URLs included in the SERP, the suggestions included in the SERP and user clicks. Each action on the system was associated with a specific user, because we required them to input an identification code on the system's homepage prior to starting the task.

# 4 Study

In this section, we describe the experiment that was conducted to evaluate if, how and which of the presented suggestions are useful to users with different levels of English proficiency, health literacy and topic familiarity. The Ethics Committee of the University of Porto approved this experiment. Research data pertaining this study is available on `http://hdl.handle.net/11304/54a2f494-ccf1-4ba6-9b0a-49e063793a3c`.

## 4.1 Research questions

Our experiment was motivated by the following research questions.

**RQ1**: Does a system that includes this suggestion tool lead to a more successful search experience in terms of medical accuracy?

**RQ2**: How does clicking on a suggestion affect the medical accuracy of the knowledge obtained in the search session?

**RQ3**: Does this effect differ with the language and terminology used to formulate the suggestions?

**RQ4**: Does this effect differ with the users' English proficiency, health literacy and topic familiarity?

## 4.2 Participants

We conducted a user study with 40 information science undergraduate students from whom we obtained informed consent. The participants were all native Portuguese speakers and, as seen in Section 6.1, were heterogeneous in terms of English proficiency, health literacy and topic familiarity.

## 4.3 Information situations

Each user was assigned a set of 8 tasks, which were equally distributed on both systems and each associated with one of 8 simulated work task situations, defined as "a short 'cover story' that describes a situation that leads to an individual requiring to use an IR system" (Borlund, 2003). To define the simulated situations, we selected 20 persons with no medical expertise and spanning a wide range of ages (from 30 to 68) and education levels (from high school to PhD degrees). These individuals were asked to state the health topic for which they had most recently searched on the Web. From these topics, we randomly selected 8 and, in collaboration with a medical doctor, created a scenario for each, in which we included one or more specific questions to be answered after the task. We quantified the number of items to include in the answer to facilitate the subsequent medical evaluation of the answers. The information situations are included in the research data file. These information situations were described to the users in the Portuguese language.

## 4.4 Task assignment

We applied a Latin-square-like procedure to ensure that each user was assigned each information situation exactly once and to simultaneously ensure that the information situations were equally balanced among the sequences of 8 tasks performed by each user. Therefore, overall, the same number of users performed the task associated with each information situation in each step of the task sequence (1st to 8th). Each user performed 4 tasks in SYS and 4 tasks in SYS+, switching between systems only once. We also guaranteed that, overall, each step of the task sequence (1st to 8th) was associated with the same number of assignments to both retrieval systems and that, overall, each retrieval system was used to address each information situation the same number of times. The described task assignment method resulted in a counterbalanced within-subject design.

## 4.5 Procedure

The users were first administered two quizzes, one to assess their health literacy and the other to assess their English proficiency. More details regarding the acquisition of these context features are provided in the following section. They were then administered an initial questionnaire, without being allowed to consult any information sources, in which, they were presented with the questions included in each simulated task. After completing this questionnaire, each user performed the assigned sequence of 8 tasks.

Each task was composed of 3 iterations, i.e., 3 similar stages in which the user was able to define or redefine a query and assess the top 10 results. The resulting set of 3 iterations constituted a search session. In each search session, the user visited and assessed 30 documents (corresponding to the 30 results returned by the system in 3 iterations and not necessarily webpages), (re)defined

3 queries and, while defining the last two queries in the system including query suggestions, had 2 opportunities to click on suggestions. The URL of the 9598 assessed documents (one of the iterations returned only 8 results) are available at the research data file.

The participants chose the language of the queries and were free to take any notes they needed to answer the questions posed in the information situation. After the third iteration, they were administered a post-search questionnaire in which they were asked to evaluate the task's topic in terms of familiarity and to, once again, answer the questions posed in the simulated situation.

## 4.6 Acquisition of user context features

To evaluate the users' English proficiency, we used an instrument developed by the European Council that grades English proficiency in the Common European Framework of Reference for Languages (CEFR), a widely accepted European standard for this purpose. Outside the scope of this work, the Faculty of Arts of the University of Porto validated the use of this instrument in the Portuguese community. The CEFR classifies individuals into three broad divisions (basic, independent and proficient), and each of these groups is further divided into two levels. Basic users can be classified as "beginner" or as "elementary", independent users as "intermediate" or "upper intermediate" and proficient users as "advanced" or as possessing "proficiency". To ensure a reasonable number of users in each group, we decided to analyze the data using only the broader divisions of the CEFR, i.e., *basic*, *independent* and *proficient* users.

To evaluate the users' health literacy, we have used the Medical Term Recognition Test (METER), a brief and self-administered instrument proposed by Rawson et al. (2010), with the suggested cutoff points of 0-20, 21-34, and 35-40 to distinguish *low*, *marginal*, and *functional* health literacy levels.

The users' familiarity with each topic was self-assessed on a five-level scale: *extremely unfamiliar* (1), *unfamiliar* (2), *neutral* (3), *familiar* (4) and *extremely familiar* (5). In the data analysis, the 5-point user familiarity ratings were further grouped into three categories: *extremely familiar*, *familiar* and *not familiar* (including *extremely unfamiliar*, *unfamiliar*, and *neutral*).

## 4.7 Medical accuracy assessment

The answers given, before and after the tasks, to the questions posed in the simulated situations were evaluated in terms of medical accuracy. A committee of two medical doctors defined a list of correct answers to each question. Where applicable, this committee also defined items that should be ignored. All elements that did not belong to either of these lists were considered incorrect. To define these lists, the medical jury used UpToDate®, a peer-reviewed evidence-based clinical decision support resource (Collins, Laronga, & Wong, 2012; Levin, Hsu, & Armon, 2012; Albrecht, 2012; Wald, 2012; Weston & Howe, 2012; Castells, 2012; Surks, 2012; R. M. Schwartzstein, 2012), and additionally the list of differential diagnoses of Harrison's Principles of Internal Medicine, 18th Edition,

for information situation 8, related to shortness of breath (R. Schwartzstein, 2011). The assessment criteria for each information situation is available on the research data file mentioned above.

To assess the reliability of this classification procedure, the second author - a medical doctor - and the first author of the paper assessed a random set of 30% of the total number of answers (40*8*2=640), including an approximately equal number of pre-search and post-search answers. At this point, the inter-rater agreement was computed, and for simulated situations for which the weighted Cohen's kappa between the assessments was below 0.8, the criteria were further detailed. The weighted Cohen's kappa is an adaptation of Cohen's kappa to ordinal scales that treats disagreements differently.

With the redefined criteria, the reliability was assessed once again, and a weighted Cohen's kappa (with squared weights) of 0.90 (95% CI: [0.83, 0.93]) was obtained for the correctness ratings, indicating almost perfect agreement. For the incorrectness ratings, the value of this measure was 0.75 (95% CI: [0.60, 0.83]), indicating substantial agreement. Since these inter-rater reliability results ensure the quality of the assessment procedure, the first author alone then assessed the remaining 70% of the answers. The first author's assessments were those that were used for the data analysis.

# 5 Data Analysis

We evaluated the impact of the suggestions on medical accuracy on the session level, that is, we compared sessions in which suggestions were used with sessions in which suggestions were not used. We analyzed the general use of suggestions and the specific use of lay, medico-scientific, English and Portuguese suggestions. Moreover, when analyzing specific types of suggestions, we considered the users' English proficiency, health literacy and topic familiarity. In addition to comparing sessions where suggestions were, or not, used, we compared different groups of users who used or did not use each type of suggestion.

As explained in the description of the study, the users were required to answer the questions posed in the simulated situations both before and after the task. This allowed for an evaluation of the knowledge that was acquired during the search task through the use of a $\Delta$ correctness measure and a $\Delta$ incorrectness measure. These metrics represent the value after the task minus the value before the task. If the value of one of these measures is positive, it indicates that the search task contributed to increasing either the correct or incorrect content in the response. Before the computation of $\Delta$ correctness and $\Delta$ incorrectness, we transformed the correct and incorrect values into the [0, 1] range. To accomplish this, in the correctness assessments, we divided the score by the maximum possible score achievable in that simulated situation. For example, if the simulated situation asked for 3 treatments, the maximum possible score would be 3. The incorrectness score represents simply the number of incorrect items included in the answer, so because of the absence of a pre-defined maximum, we opted to divide the incorrect score by the maximum incorrect score obtained by any user

in that simulated information situation.

We compared the means between groups (with and without the use of suggestions) using Student's t-test. When the assumption of homogeneity of the variances was not verified, we applied the Welch t-test. To compare groups of users, we applied one-way ANOVA and Tukey's test to identify the differences whenever significant differences were found. When reporting our results, we use * to indicate results significant at $\alpha = 0.05$ and ** to indicate results significant at $\alpha = 0.01$.

# 6 Results

This section begins with a description of the participants involved in the study, followed by several subsections, one for each research question. In the following sections, when we refer to suggestion usage, we mean clicking on a suggestion at any time during the session, that is, in the set of the three iterations that composed the task.

## 6.1 Participants

Forty participants were included in this study (24 female; 16 male), with a mean age of 23.48 years (standard deviation (sd)=7.66). The English proficiency assessment revealed a heterogeneous sample of users who, on average, had a proficiency of 19.93 (sd=8.86) on a scale of 0 to 40. In terms of CEFR classes, 16 users had *basic* English proficiency, 17 were *independent*, and 7 were *proficient* users.

In the health literacy test, evaluated on a scale of 0 to 40, the users had an average literacy of 25.55 (sd=7.40). The distribution of the users among the health literacy classes was as follows: *low* (7 users), *marginal* (28 users) and *functional* (5 users). The users' familiarity with each topic depends on the theme of the task. The pairs "user, topic" were distributed as follows: *not familiar* (86 pairs), *familiar* (114 pairs) and *extremely familiar* (120 pairs).

## 6.2 Systems comparison (RQ1)

In general, the quantity of correct response content after a search task increased significantly (t(319)=-21.62, p<2.2e-16**). The mean of $\Delta$ correctness was 0.45, and its standard deviation was 0.37. The quantity of incorrect content does not significantly differ (t(319)=0.24, p=0.81) before and after the search task. In fact, the mean of the $\Delta$ incorrectness measure was 0, and its standard deviation was 0.32. Contrary to the case for correct content, a lower value of the measure is better for incorrect content because a negative value indicates a reduction in the quantity of incorrect content after the search task. Henceforth, when we refer simply to correctness, we mean $\Delta$ correctness, and when we refer simply to incorrectness, we mean $\Delta$ incorrectness.

A comparison between SYS+ and SYS revealed that the system with suggestions outperformed the other system. In terms of $\Delta$correctness, the difference was not significant (0.47 for SYS+ vs. 0.43 for SYS) but with regard to $\Delta$incorrectness, we found that SYS+ contributed to reducing the incorrect content in the answers significantly more than did SYS (-0.04 for SYS+; 0.03 for SYS; t(312)=-1.7, p=0.04*).

## 6.3 Effects of clicking on suggestions (RQ2)

As seen in Table 1, the use of suggestions tended to improve the medical accuracy of the knowledge obtained in the search task, simultaneously contributing to an increase in the quantity of correct content and a decrease in the quantity of incorrect content. Moreover, we found that clicking on suggestions significantly increased the quantity of correct content acquired in the session, and using the system with suggestions significantly decreased the quantity of incorrect content.

Table 1: $\Delta$correctness and $\Delta$incorrectness comparisons and one-sided differences between systems and the use of suggestions.

| | Mean $\Delta$ | | Differences | |
|---|---|---|---|---|
| | No | Yes | test value | p-value |
| **Correctness** | | | | |
| System with suggestions? | 0.43 | **0.47** | t(316.5)=1.1 | 0.12 |
| Click on suggestion? | 0.42 | **0.52** | t(161.6)=-2.0 | 0.02* |
| **Incorrectness** | | | | |
| System with suggestions? | 0.03 | **-0.03** | t(312)=-1.7 | 0.04* |
| Click on suggestion? | 0 | **-0.02** | t(168.3)=0.6 | 0.28 |

No and Yes are the answers to the questions in the first column. Boldface indicates the higher quality improvement.

## 6.4 Language and terminology effects (RQ3)

In addition to this general analysis, we also performed an analysis by type of suggestion. In Table 2, we present the results by suggestion language and terminology. As shown in these tables, in general, all suggestions tended to improve the medical accuracy of the answers. In terms of significant differences, English suggestions contributed to the acquisition of more correct content. Regarding the suggestion terminology, the use of either lay or medico-scientific suggestions increased the quantity of correct content.

Table 2: Δcorrectness and Δincorrectness comparisons and one-sided differences by suggestion language.

| | Mean Δ | | Differences | |
|---|---|---|---|---|
| | without | with | test value | p-value |
| **Portuguese** | | | | |
| Correctness | 0.44 | **0.51** | t(60.8)=-1.2 | 0.12 |
| Incorrectness | 0 | **-0.04** | t(72.3)=0.88 | 0.19 |
| **English** | | | | |
| Correctness | 0.42 | **0.56** | t(94.6)=-2.6 | 0.005** |
| Incorrectness | **-0.01** | 0 | t(85.8)=-0.18 | 0.43 |
| **Lay** | | | | |
| Correctness | 0.43 | **0.54** | t(87.9)=-2.3 | 0.01* |
| Incorrectness | **-0.01** | 0 | t(78.7)=-0.2 | 0.4 |
| **Medico-scientific** | | | | |
| Correctness | 0.43 | **0.51** | t(123.3)=-1.8 | 0.04* |
| Incorrectness | 0 | **-0.02** | t(129.2)=0.6 | 0.3 |

Boldface indicates the higher quality improvement.

## 6.5 Analysis by English proficiency (RQ4)

### 6.5.1 Language comparison

The use of suggestions, in both languages, tended to increase the quantity of correct response content in every English proficiency level. We found that the use of English suggestions significantly increased the quantity of correct content among users with *basic* proficiency (without: 0.4, with: 0.56, t(43.0)=-2.9, p=0.01*).

We also found that in general, Portuguese suggestions tended to decrease the quantity of incorrect content in the *basic* and *proficient* groups. Meanwhile, English suggestions tended to have the same effect, but only in the *independent* and *proficient* groups. In Table 3, we present the significant differences found with respect to incorrectness. As seen from this table, the use of suggestions, either in Portuguese or in English, significantly decreased the quantity of incorrect content among *proficient* users.

### 6.5.2 Group comparison

Pertaining to the significant differences between levels of English proficiency, for which Tukey's adjusted p-values are presented in Table 4, we found that without the use of Portuguese suggestions, *proficient* users provided answers with

Table 3: Δincorrectness comparisons and one-sided significant differences by language and English Proficiency (EP).

| | Mean Δ | | Differences | |
|---|---|---|---|---|
| | without | with | test value | p-value |
| Proficient EP/Portuguese | 0.07 | **-0.02** | t(51.5)=1.9 | 0.03* |
| Proficient EP/English | 0.09 | **-0.03** | t(39.4)=1.8 | 0.04* |

Boldface indicates the higher quality improvement.

more correct content than did *basic* and *independent* users. When Portuguese suggestions were used, we only found differences between them and *independent* users. Because we did not observe such differences with the use of English suggestions and, with these suggestions, the quantity of correct content increased in all groups, we hypothesize that English suggestions contributed to reducing the differences between these groups of users.

Table 4: Tukey's adjusted p-values for one-sided significant comparisons of the *proficient* group with the other groups in terms of Δcorrectness and Δincorrectness.

| | Correctness | | | Incorrectness |
|---|---|---|---|---|
| **Proficient EP >** | without PT | with PT | without EN | without EN |
| Basic EP | 0.008** | - | 0.006** | 0.03* |
| Independent EP | 0.003** | 0.02* | 0.002** | - |

EP stands for English proficiency; EN, for English; and PT, for Portuguese.

Surprisingly, we also found that *proficient* users, when not using English suggestions, exhibited a greater increase in incorrect content than *basic* users. In part, this can be attributed to their different behavior in the pre-search answers, in which the users with *basic* proficiency provided answers with significantly more incorrect content than did the *proficient* users (0.54 vs. 0.12, t (99.7) = 3.17, p= 0.001**). In the post-search answers, we found no significant differences between the two groups in terms of incorrectness. Nevertheless, the quantity of incorrect content decreased for the users with *basic* proficiency when they did not use English suggestions, whereas that for *proficient* users increased.

When English suggestions were used, we found no significant differences in Δincorrectness among different levels of English proficiency. When using English suggestions, users with *basic* proficiency tended to increase the incorrectness of their answers, whereas that for *independent* and *proficient* users decreased. As can be seen in Table 3 this last difference is significant.

## 6.6 Analysis by health literacy (RQ4)

### 6.6.1 Terminology comparison

An analysis by health literacy and terminology, for which the significant results are presented in Table 5, revealed that users in the *marginal* health literacy group were able to provide answers with significantly more correct content when they clicked on lay suggestions. In terms of incorrect content, we found that this same group benefited from the use of medico-scientific suggestions. Contrary to our expectations, users with *functional* health literacy provided answers with more incorrect content when clicking on lay suggestions.

Table 5: Δcorrectness and Δincorrectness comparisons and one-sided significant differences by terminology (Lay/MS) and health literacy (HL).

| | Mean Δ | | Differences | |
|---|---|---|---|---|
| | without | with | test value | p-value |
| **Correctness** | | | | |
| Marginal HL/Lay | 0.45 | **0.58** | t(43.3)=-1.9 | 0.03* |
| **Incorrectness** | | | | |
| Marginal HL/Medico-scientific | 0 | **-0.11** | t(79.8)=2.3 | 0.01* |
| Functional HL/Lay | **0.02** | 0.26 | t(13.5)=-1.8 | 0.05* |

Boldface indicates the higher quality improvement.

### 6.6.2 Group comparison

Regarding the correctness differences between the health literacy groups, as shown in Table 6, we found that when lay suggestions were not used, users with *low* health literacy provided answers with less correct content than did users with higher health literacy. Because we found that suggestions tended to increase the quantity of correct content at every level of health literacy, we concluded that lay suggestions assisted in eliminating the differences between these groups of users, improving the ability of the low-literacy users to answer with a higher quantity of correct content.

When using medico-scientific suggestions, users with *functional* and *marginal* health literacy were distinguished from the low-literacy group by a larger quantity of correct content. When using terms from medico-scientific suggestions, users with *functional* health literacy also provided more correct answers than did *marginal* users (Tukey's adjusted p=0.032*).

Surprisingly, we found that with suggestions, the *functional* health literacy group provided answers with more incorrect content than did the marginally literate group (Tukey's adjusted p=0.003** with lay suggestions; Tukey's adjusted p=0.001** with medico-scientific suggestions). These findings led us to

Table 6: Tukey's adjusted p-values for one-sided significant comparisons of the *low* health literacy (HL) group with the other groups in terms of Δcorrectness.

| **Low HL <** | without lay | with medico-scientific |
|---|---|---|
| Marginal HL | 0.008** | 0.04* |
| Functional HL | 0.03* | 0.02* |

Table 7: Tukey's adjusted p-values for one-sided significant comparisons of the *functional* health literacy (HL) group with the other groups in terms of answer length variation.

| **Functional HL >** | with medico-scientific | without lay | without medico-scientific |
|---|---|---|---|
| Low HL | 0.02* | 0.02* | 0.03* |
| Marginal HL | - | - | 0.02* |

investigate further. We believe that the higher quantity of incorrect content provided by these users when using suggestions and their inferior results compared with users with lower levels of health literacy can be attributed to the fact that they tended to provide longer answers. Although this tendency may also have contributed to their higher quantity of correct answer content, the effect of long answers on the correctness analysis was not as strong as the corresponding effect on the incorrectness analysis. In fact, as explained in Section 5, the number of correct items in an answer had a predefined maximum, whereas the number of incorrect items did not. With this hypothesis in mind, we performed an analysis similar to that for incorrectness, but for the variation in answer length instead. Similar to what was found for incorrectness, when *functional* users used medico-scientific terms from suggestions, the lengths of their answers increased more than they did without these suggestions (with: 843.1; without: 387.4; t (18.4)=-1.9, p=0.04*). Upon comparing groups of users, we found several significant differences. Table 7 presents the p-values of the significant differences between the *functional* group and the other groups. In addition to these differences, we also found that with medico-scientific suggestions, the low-literacy group provided answers with a smaller variation in length than did the *marginal* group when they clicked on a suggestion (p=0.02*). These results corroborate our initial conjecture that the higher quantity of incorrect content may be attributable to longer answers.

## 6.7 Analysis by topic familiarity (RQ4)

### 6.7.1 Terminology comparison

The significant results of the analysis by topic familiarity are presented in Table 8. These results show that *non-familiar* users provided answers with a larger

quantity of correct content when they clicked either lay or medico-scientific suggestions. Meanwhile, users who were *extremely familiar* with a topic benefited from lay suggestions in terms of incorrect content.

Table 8: Δcorrectness and Δincorrectness comparisons and one-sided significant differences by terminology (Lay/MS) and topic familiarity.

| | Mean Δ | | Differences | |
|---|---|---|---|---|
| | without | with | test value | p-value |
| **Correctness** | | | | |
| Not Familiar/Lay | 0.4 | **0.69** | t(17.6)=-3.1 | 0.003** |
| Not Familiar/Medico-scientific | 0.39 | **0.57** | t(41.9)=-1.9 | 0.03* |
| **Incorrectness** | | | | |
| Extremely Familiar/Lay | 0 | **-0.19** | t(24.7)=2.5 | 0.01** |

Boldface indicates the higher quality improvement.

#### 6.7.2 Group comparison

In a comparison of the familiarity groups, we found that when using lay suggestions, the *extremely familiar* group provided answers with less incorrect content than did the *non-familiar* (Tukey's adjusted p=0.04*) and *familiar* (Tukey's adjusted p=0.006**) users.

## 7 Discussion of Results

The system with suggestions (SYS+) tended to demonstrate better performance in terms of correctness and incorrectness of the answers. However, the only significant difference was found with regard to incorrect content, namely, SYS+ reduced the quantity of incorrect content ($< 0.05$ as seen in Table 1). From these findings, we can conclude that a retrieval system that includes the proposed suggestion tool contributes to a better retrieval experience as measured by an outcome that is particularly important in health searches, namely, assistance in reducing the quantity of incorrect content in the acquired knowledge.

In Table 9, we present a general summary of the significant findings previously reported by language and by terminology. This table shows that clicking on suggestions, regardless of their type, leads to answers with more correct content than those obtained when no suggestions are clicked.

A deeper analysis reveals that only Portuguese suggestions offer no significant benefit in terms of answer correctness.

English suggestions are advantageous for *proficient* users in terms of the incorrectness of the answers. Surprisingly, *basic* English proficiency users pro-

Table 9: Summary of the significant findings.

| | **Correctness** | **Incorrectness** |
|---|---|---|
| **All** | ↑ | |
| **English** | general (↑)**<br>basic EP (↑) | proficient EP (↓) |
| **Portuguese** | | proficient EP (↓) |
| **Lay** | general (↑)<br>marginal HL (↑)<br>not familiar (↑)** | functional HL (↑)<br>extremely familiar (↓)** |
| **Medico-scientific** | general (↑)<br>not familiar (↑) | marginal HL (↓) |

↑ denotes an increase and ↓ a decrease in each outcome. ** denotes a result that is significant at $\alpha$=0.01. All other results are significant at $\alpha$=0.05. EP stands for English proficiency; and HL, for health literacy.

vided more correct content with English suggestions than without them. Despite their lack of English skills, these users were still capable of extracting accurate information from English documents. Although English suggestions tended to increase the incorrectness outcomes of *basic* proficiency users, we did not find this effect to be significant. Because these users were not significantly affected in any retrieval outcome by suggestions of this type, we hypothesize that although English queries yielded worse results than Portuguese queries in a previous study (Lopes & Ribeiro, 2013), it might be better to use them at least once than not to use them at all.

Excluding the unexpected increase in incorrect content in the *functional* health literacy group with lay suggestions, the use of lay suggestions was found to have a beneficial effect on medical accuracy. The higher degree of incorrectness of the answers given by users with *functional* health literacy when using lay queries is surprising and is not easy to explain. One possible reason may be the length of their answers, but this issue must be further explored. Medico-scientific suggestions are beneficial in terms of medical accuracy. Users who were not familiar with the topic provided more correct answers with both lay and medico-scientific suggestions, but the evidence for increased correctness was stronger for lay queries.

In Table 10, we present the significant differences found when comparing groups of users who either clicked or did not click on certain types of suggestions. In general, groups with higher English proficiency, health literacy or topic familiarity had a better retrieval experience than users below their level in terms

of the acquisition of correct content.

Table 10: Summary of the significant differences found in the group comparisons.

| | | **Correctness** | **Incorrectness** |
|---|---|---|---|
| **English** | **w/o** | EP3>{EP1**,EP2**} | EP3>EP1 |
| | **w/** | | |
| **Portuguese** | **w/o** | EP3>{EP1**,EP2**} | |
| | **w/** | EP3>EP2 | |
| **Lay** | **w/o** | {HL3,HL2**}>HL1 | |
| | **w/** | | HL3>HL2** |
| | | | TF3<{TF2**,TF1} |
| **Medico-scientific** | **w/o** | | |
| | **w/** | {HL3,HL2}>HL1 | HL3>HL2** |

** denotes a result significant at $\alpha$=0.01. All other results are significant at $\alpha$=0.05. EP1 stands for basic English proficiency (EP); EP2, for independent EP and EP3, for proficient EP. HL1 stands for low health literacy (HL); HL2, for marginal HL; and HL3, for functional HL. TF1 stands for Not Familiar; TF2, for Familiar; and TF3, for Extremely Familiar.

We can observe that when they did not click on suggestions, the English *proficient* users provided answers with more correct content than did the *basic* and *independent* users. This shows that in a retrieval system without suggestions, these users are better prepared to search for health information and/or to answer medical questions. However, when using English suggestions, the *proficient* users' superiority is no longer significant. In addition, when using Portuguese suggestions, the *proficient* group's superiority holds only in comparison with the *independent* proficiency group. Because the quantity of correct contents tended to increase in every level of English proficiency when suggestions were used, we hypothesize that English suggestions contribute to reducing the differences between groups of users. Regarding the incorrectness outcome, because the use of English suggestions reduces the incorrect content retrieved by *proficient* users ($< 0.05$ as seen in Table 3), suggestions of this type contribute to eliminating the difference that exists between these groups when they do not click on English suggestions.

When not clicking on lay suggestions, users with *low* health literacy provided answers with less correct content than did *marginal* and *functional* users, a difference that did not exist when these users did click on suggestions of this type. Because we found that clicking on lay suggestions increased the quantity of correct answer content for users with marginal health literacy ($< 0.05$ as seen in Table 5) and tended to increase it for users at the other levels of

health literacy, we conclude that these suggestions help reduce the differences between these groups of users, improving the ability of low-literacy users to answer with a higher quantity of correct content. However, although clicking on medico-scientific suggestions tended to increase the quantity of correct content at all levels of health literacy, the improvement was larger at higher levels of health literacy, causing a significant difference to emerge between these groups. This result corroborates the findings of a previous study showing that users with higher levels of health literacy are more apt to assimilate medico-scientific documents (Lopes & Ribeiro, 2015). Probably because of the length of their answers, the *functional* health literacy group performed worse than the marginally literate group in terms of incorrect content when using suggestions.

Regarding the analysis by topic familiarity, we found that both lay ($< 0.01$ as seen in Table 8) and medico-scientific ($< 0.05$ as seen in Table 8) suggestions increase the quantity of correct contents acquired by *non-familiar* users. In addition, lay suggestions contribute to reduce the quantity of incorrect contents in *extremely familiar* users. The usefulness of medico-scientific suggestions for *non-familiar* users and lay suggestions for *extremely familiar* users suggests that topic familiarity is not discriminant characteristic in terms of terminology selection.

While the OAC CHV might be considered limited in coverage when compared with medico-scientific vocabularies (e.g. SNOMED-CT), it is the most comprehensive vocabulary satisfying three requirements essential for this system: (1) containing lay terminology, (2) relating lay terminology with medico-scientific terminology and (3) being freely available. The Multilingual Glossary of Popular and Technical Medical Terms (Stichele, 1995) has fewer terms and is not related with medico-scientific terminology. The Mayo Consumer Health Vocabulary (Seedorff et al., 2012) and the Consumer Health Terminology Thesaurus, developed by the WellMed Medical Management[1], also have fewer terms and are not freely available.

Although we would have preferred to conduct this study with a more realistic sample, including a larger diversity of users, we think this sample of users has not affected our conclusions. Being aware that this set of users is typically younger than the general population, we don't consider this to be a problem because we assessed topic familiarity, mostly compensating the effect that a wider age distribution could have on knowledge acquisition. With a sample composed of general users, we would expect a more right-skewed distribution of English proficiency, health literacy and perhaps topic familiarity. However, that will not be a problem because we will only be doing comparisons between groups and the existing sample is heterogeneous regarding these characteristics.

# 8 Conclusions

In this work we study the effect of several types of health query suggestions on the medical accuracy of the knowledge obtained in the search session, more

[1] `https://www.wellmedmedicalgroup.com/`

specifically, in its correctness and incorrectness. Query suggestions combine two languages, Portuguese and English, and two types of terminologies, lay and medico-scientific. In the evaluation we also consider users' proficiency in the English language, their health literacy and topic familiarity.

A retrieval system that included the implemented suggestion system tended to perform better than a system without suggestions in terms of correctness and incorrectness. Of these differences, the only significant one was the incorrectness difference, which is an outcome that is extremely relevant in the health domain.

We demonstrated that clicking on suggestions of medical concepts related to an initial query using different languages and terminologies was beneficial with respect to the medical accuracy of the knowledge obtained in a search session, particularly for its correctness. In terms of language, English suggestions were found to be more effective than Portuguese ones, in general and even in the lowest level of English proficiency. Regarding the terminology, both were found to be effective, in general and in certain groups of users.

We found that the benefits of the suggestions vary with their type and the characteristics of users, with some types of suggestions being more useful to some users than others. This suggests that a personalisation of the suggestion system towards users' English proficiency and health literacy might be even more advantageous.

Although some of our findings can, indirectly, be useful in assisting health consumers in searching the Web more efficiently, our findings and conclusions are directed toward the design of search engines, either generalist or health-specific. It is important to draw conclusions that can be used in generalist search engines because these are typically the starting point for health searches. Currently, the major generalist search engines provide suggestions for queries related to those provided by the user. It is easy to visualize the inclusion of the types of suggestions considered in our study in this service this is already provided by generalist search engines, whenever a user enters a health query. Large search companies have ready access to data that can be used to infer users' native languages as well as their proficiency in English and their health literacy.

We believe that there are several opportunities to improve the query suggestion system developed in this work. We envision enhancements to the algorithm, the underlying data structures and the interface. Moreover, we plan to contribute to the development of the OAC CHV vocabulary through the validation of the translation of the OAC CHV into the Portuguese language. This will not only improve the performance of the query suggestion system but also contribute to the resources available in Portuguese. Envisioning the validated Portuguese OAC CHV, we have already begun with the development of a web application that will be disseminated in the medical community (Silva & Lopes, 2016).

Although our user study was conducted with Portuguese users, we believe our conclusions are still valid in Brazilians users, a much wider set of users. Note that we are using a version of the OAC CHV translated with the Google Translator API, that supports only Portuguese, not distinguishing between Portuguese-

Portugal and Portuguese-Brazil. Yet, we are aware that some expressions might differ (we noticed that during the evaluation of the automatic translation) and that it could be beneficial for the quality of the suggestions to work with a specific version of the Portuguese OAC CHV for the Brazilian users. This version could be built on top of the Portuguese validated version.

# 9 Acknowledgments

Thanks to Fundação para a Ciência e a Tecnologia for partially funding this work under the grant SFRH/BD/40982/2007 to Carla Teixeira Lopes, UID/DTP/04750/2013 to the Epidemiology Research Unit (EPIUnit) and the project UID/EEA/50014/2013 to the INESC TEC. The authors would also like to thank Andreia Ribeirinho Soares, MD, for her contributions on the definition of the criteria used to assess medical accuracy.

# References

Albrecht, M. A. (2012). Clinical manifestations of varicella-zoster virus infection: Herpes zoster. In T. W. Post, M. S. Hirsch, & B. H. McGovern (Eds.), *Uptodate.* Waltham, MA: Available from: http://www.uptodate.com/ (Accessed on September 17, 2012.).

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

Castells, M. C. (2012). Insect bites. In T. W. Post, D. B. Golden, D. F. Danzl, T. Rosen, A. M. Feldweg, & E. L. Baron (Eds.), *Uptodate.* Waltham, MA: Available from: http://www.uptodate.com/ (Accessed on September 17, 2012.).

Collins, L. C., Laronga, C., & Wong, J. S. (2012). Ductal carcinoma in situ: Treatment and prognosis. In T. W. Post, L. J. Pierce, D. F. Hayes, A. B. Chagpar, R. B. Duda, & D. S. Dizon (Eds.), *Uptodate.* Waltham, MA: Available from: http://www.uptodate.com/ (Accessed on September 17, 2012.).

Dang, V., Kumaran, G., & Troy, A. (2012). Domain dependent query reformulation for web search. In *Proceedings of the 21st acm international conference on information and knowledge management* (pp. 1045–1054). New York, NY, USA: ACM. Retrieved from `http://dx.doi.org/10.1145/2396761.2398401` doi: 10.1145/2396761.2398401

Efthimiadis, E. N. (1996). Query expansion. *Annual Review of Information Systems and Technology (ARIST)*, *31*, 121–187.

Fattahi, R., Wilson, C. S., & Cole, F. (2008, July). An alternative approach to natural language query expansion in search engines: Text analysis of non-topical terms in web documents. *Inf. Process. Manage.*, *44*(4), 1503–1516. Retrieved from `http://dx.doi.org/10.1016/j.ipm.2007.09.009` doi: 10.1016/j.ipm.2007.09.009

Fox, S. (2011). *Health topics* (Tech. Rep.). Washington, DC: Pew Internet & American Life Project.

Gao, W., Niu, C., Nie, J. Y., Zhou, M., Wong, K. F., & Hon, H. W. (2010). Exploiting query logs for cross-lingual query suggestions. *ACM Trans. Inf. Syst.*, *28*(2), 1–33. Retrieved from http://dx.doi.org/10.1145/1740592.1740594 doi: 10.1145/1740592.1740594

Hersh, W. R., Crabtree, M. K., Hickam, D. H., Sacherek, L., Friedman, C. P., Tidmarsh, P., ... Kraemer, D. (2002). Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions. *Journal of the American Medical Informatics Association : JAMIA*, *9*(3), 283–293. Retrieved from http://view.ncbi.nlm.nih.gov/pubmed/11971889

Jansen, B. J., & McNeese, M. D. (2005). Evaluating the effectiveness of and patterns of interactions with automated searching assistance. *J. Am. Soc. Inf. Sci.*, *56*(14), 1480–1503. Retrieved from http://dx.doi.org/10.1002/asi.20242 doi: 10.1002/asi.20242

Kogan, S., Zeng, Q., Ash, N., & Greenes, R. A. (2001). Problems and challenges in patient information retrieval: a descriptive study. In *Proceedings amia symposium* (pp. 329–333). Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243602/

Kriewel, S., & Fuhr, N. (2010). Evaluation of an adaptive search suggestion system. In *32nd european conference on information retrieval research (ecir 2010)* (pp. 544–555). Springer.

Levin, K., Hsu, P. S., & Armon, C. (2012). Acute lumbosacral radiculopathy: Prognosis and treatment. In T. W. Post, J. M. Shefner, & J. F. Dashe (Eds.), *Uptodate.* Waltham, MA: Available from: http://www.uptodate.com/ (Accessed on September 17, 2012.).

Liu, Zhenyu, Chu, & Wesley. (2007, April). Knowledge-based query expansion to support scenario-specific retrieval of medical free text. *Information Retrieval*, *10*(2), 173–202. Retrieved from http://dx.doi.org/10.1007/s10791-006-9020-6 doi: 10.1007/s10791-006-9020-6

Lopes, C. T., & Ribeiro, C. (2013, May). Measuring the value of health query translation: An analysis by user language proficiency. *Journal of the American Society for Information Science and Technology*, *64*(5), 951–963. Retrieved from http://dx.doi.org/10.1002/asi.22812 doi: 10.1002/asi.22812

Lopes, C. T., & Ribeiro, C. (2015). Effects of terminology on health queries: An analysis by user's health literacy and topic familiarity. In A. Woodsworth & W. D. Penniman (Eds.), *Current issues in libraries, information science and related fields* (Vol. 39, pp. 145–184). Emerald Group Publishing Limited. Retrieved from http://www.emeraldinsight.com/doi/abs/10.1108/S0065-283020150000039013 doi: 10.1108/S0065-283020150000039013

Luo, G., & Tang, C. (2008). On iterative intelligent medical search. In *Sigir '08: Proceedings of the 31st annual international acm sigir conference on research and development in information retrieval* (pp. 3–10). New

York, NY, USA: ACM. Retrieved from http://dx.doi.org/10.1145/1390334.1390338 doi: 10.1145/1390334.1390338

Luo, G., Tang, C., Yang, H., & Wei, X. (2008). MedSearch: a specialized search engine for medical information retrieval. In *Cikm '08: Proceeding of the 17th acm conference on information and knowledge mining* (pp. 143–152). New York, NY, USA: ACM. Retrieved from http://dx.doi.org/10.1145/1458082.1458104 doi: 10.1145/1458082.1458104

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (First ed.). Cambridge: Cambridge University Press. Hardcover. Retrieved from http://www.worldcat.org/isbn/0521865719

McCray, A. T., & 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/

NLM. (2012). *2012AA consumer health vocabulary source information.* Retrieved from http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CHV/index.html

Rawson, K. A., Gunstad, J., Hughes, J., Spitznagel, M. B. B., Potter, V., Waechter, D., & Rosneck, J. (2010, January 1). The METER: a brief, self-administered measure of health literacy. *Journal of general internal medicine*, *25*(1), 67–71. Retrieved from http://dx.doi.org/10.1007/s11606-009-1158-7 doi: 10.1007/s11606-009-1158-7

Schwartzstein, R. (2011). Section 5: Alterations in circulatory and respiratory functions: Dyspnea. In D. Longo, A. Fauci, D. Kasper, S. Hauser, J. Jameson, & J. Loscalzo (Eds.), *Harrison's principles of internal medicine (18th ed.).* New York: McGraw-Hill Medical.

Schwartzstein, R. M. (2012). Approach to the patient with dyspnea. In T. W. Post, T. E. King, & H. Hollingsworth (Eds.), *Uptodate.* Waltham, MA: Available from: http://www.uptodate.com/ (Accessed on September 17, 2012.).

Seedorff, M., Peterson, K. J., Nelsen, L. A., Cocos, C., McCormick, J. B., Chute, C. G., & Pathak, J. (2012, November 1). Incorporating expert terminology and disease risk factors into consumer health vocabularies. In *Biocomputing 2013* (pp. 421–432). WORLD SCIENTIFIC. Retrieved from http://dx.doi.org/10.1142/9789814447973_0041 doi: 10.1142/9789814447973_0041

Silva, A. C., & Lopes, C. T. (2016). Health translations: A crowdsourced, gamified approach to translate large vocabulary databases. In *Cisti'2016 - 11th iberian conference on information systems and technologies.*

Soldaini, L., Yates, A., Yom-Tov, E., Frieder, O., & Goharian, N. (2016). Enhancing web search in the medical domain via query clarification. *Information Retrieval Journal*, *19*(1-2), 149–173. Retrieved from http://dx.doi.org/10.1007/s10791-015-9258-y doi: 10.1007/s10791-015-9258-y

Stanton, I., Ieong, S., & Mishra, N. (2014). Circumlocution in diagnostic medical queries. In *Proceedings of the 37th international acm sigir conference on research & development in information retrieval* (pp. 133–142). New

York, NY, USA: ACM. Retrieved from http://dx.doi.org/10.1145/2600428.2609589 doi: 10.1145/2600428.2609589

Stichele, R. V. (1995, December). *Multilingual glossary of technical and popular medical terms in nine european languages* (Tech. Rep.). Gent: Heymans Institute of Pharmacology, University of Gent. Retrieved from http://users.ugent.be/\~{}rvdstich/eugloss/welcome.html

Surks, M. I. (2012). Clinical manifestations of hypothyroidism. In T. W. Post, D. S. Ross, & J. E. Mulder (Eds.), *Uptodate.* Waltham, MA: Available from: http://www.uptodate.com/ (Accessed on September 17, 2012.).

Toms, E. G., & Latter, C. (2007, September 01). How consumers search for health information. *Health informatics journal*, *13*(3), 223–235. Retrieved from http://dx.doi.org/10.1177/1460458207079901 doi: 10.1177/1460458207079901

Wald, A. (2012). Treatment of irritable bowel syndrome. In T. W. Post, N. J. Talley, & S. Grover (Eds.), *Uptodate.* Waltham, MA: Available from: http://www.uptodate.com/ (Accessed on September 17, 2012.).

Weston, W. L., & Howe, W. (2012). Treatment of atopic dermatitis (eczema). In T. W. Post, R. P. Dellavalle, M. L. Levy, J. Fowler, & R. Corona (Eds.), *Uptodate.* Waltham, MA: Available from: http://www.uptodate.com/ (Accessed on September 17, 2012.).

Zarro, M., & Lin, X. (2011, October). Using social tags and controlled vocabularies as filters for searching and browsing: A health science experiment. In *The fifth workshop on human-computer interaction and information retrieval.*

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 : JAMIA*, *13*(1), 80–90. Retrieved from http://dx.doi.org/10.1197/jamia.m1820 doi: 10.1197/jamia.m1820

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 symposium* (pp. 210–219).

Zhang, Y. (2011, October). A review of search interfaces in consumer health websites. In *The fifth workshop on human-computer interaction and information retrieval.*

Zielstorff, R. (2003, October). Controlled vocabularies for consumer health. *Journal of Biomedical Informatics*, *36*(4-5), 326–333. Retrieved from http://dx.doi.org/10.1016/j.jbi.2003.09.015 doi: 10.1016/j.jbi.2003.09.015