Enhancing web Search in the Medical Domain via Query Clarification

Luca Soldaini · Andrew Yates
Elad Yom-Tov · Ophir Frieder
Nazli Goharian

Abstract The majority of Internet users search for medical information online; however, many do not have an adequate medical vocabulary. Users might have difficulties finding the most authoritative and useful information because they are unfamiliar with the appropriate medical expressions describing their condition; consequently, they are unable to adequately satisfy their information need. We investigate the utility of bridging the gap between layperson and expert vocabularies; our approach adds the most appropriate expert expression to queries submitted by users, a task we call query clarification. We evaluated the impact of query clarification. Using three different synonym mappings and conducting two task-based retrieval studies, users were asked to answer medically-related questions using interleaved results from a major search engine. Our results show that the proposed system was preferred by users and helped them answer medical concerns correctly more often, with up to a 7% increase in correct answers over an unmodified query. Finally, we introduce a supervised classifier to select the most appropriate synonym mapping for each query, which further increased the fraction of correct answers (12%).

Keywords query clarification · personalized search · medical informatics · health search

1 Introduction

According to a May 2013 report from the Pew Research Center, 85% of U.S. adults use the Internet (Zickuhr 2013); per another survey from the Pew Research

Luca Soldaini · Andrew Yates · Ophir Frieder · Nazli Goharian
Information Retrieval Lab
Computer Science Department
Georgetown University
E-mail: {luca, andrew, ophir, nazli}@ir.cs.georgetown.edu

Elad Yom-Tov
Microsoft Research
Herzeliya, Israel
E-mail: eladyt@microsoft.com

Center's Internet & American Life Project, 72% of them have looked for health information online within the past year (Fox and Duggan 2013). Simply put, the Internet has become a primary source for health information dissemination and potential decision making. The key question we address is: "*When non-professionals search using lay terms, can medically oriented queries be clarified (i.e., improved by adding the most relevant expert expression) to return more useful and authoritative results?*"

Trustworthy health care resources, even those addressed to consumers, employ appropriate medical terminology; yet laypeople do not have the necessary knowledge to express their information need using such vocabulary, thus struggling to satisfy their information needs (Zeng et al. 2004). This represents a language gap which is difficult to overcome either by searchers, who need to learn a specialized vocabulary to describe their information need, or by experts, who are required to speculate on the ways in which laypeople will phrase their intent. This language gap was noted as one of the primary reasons for failures of retrieval engines (Carmel et al. 2006).

We demonstrate that such a gap can be effectively overcome in an automated way. We introduce a system that clarifies queries formulated in layperson terms with expert terminology. Our system takes advantage of three laypeople-to-expert synonym mappings; each map associates one or more layperson expressions to one or more expressions used by medical professionals.

We evaluated our approach via two task-based user studies. We simulated a health search scenario in which users interacted with a search engine to answer health-related questions. That is, we gave users the task of answering a question using a website that implicitly modifies their queries before submitting them to Bing[1] and displaying the retrieved results. We analyzed how users answered the health-related questions in relation to the query clarification technique used. Our studies involved two types of users: laypeople with limited—if any—medical knowledge and expert users with medical training. We find that lay users prefer and benefit from queries clarified with expert medical terms, whereas medically trained users prefer the original queries. Thus, our experiments illustrate the importance of bridging the language gap between experts and laypeople.

The *Health On the Net (HON) Foundation*[2] is an organization that certifies those health-related websites that meet specific reliability standards ("HONcode" of conduct). While such certification is useful in identifying authoritative medical resources, it does not address the language gap between laypeople vocabulary and expert terminology. In other words, even when only retrieving HON-certified web pages, users might not successfully satisfy their information need due to poor query formulation. Nevertheless, in our experiments, users who clicked websites that had HONcode certification were 7.7% more likely to answer a health-related question correctly, illustrating the importance of clarifying medical queries to increase the likelihood of retrieving trustworthy web pages.

In summary, our contributions are:

– We demonstrate the feasibility of clarifying health-related queries by adding expert medical terms; such clarification retrieves more useful medical resources as evidenced by the users' performance when answering medical questions;

[1] `http://www.bing.com`

[2] `http://www.healthonnet.org`

- We evaluate the effectiveness of three medical synonym mappings when used to bridge the language gap between laypeople and experts;
- We show that users exposed to results retrieved using a clarified query visit more health-related websites that meet the quality criteria established by the HON foundation;
- We introduce a classifier that predicts, for each query, the best synonym mapping for clarification;
- We propose a system that only relies on resources, such as query logs and medical synonym mappings, that are available a priori (i.e., before a query is submitted to a search engine): thus, our approach is suitable to high-traffic search engines, due to its negligible computation cost per query.

The reminder of the paper is organized as follows: Section 2 contains a summary of previous work that relates to our effort; in Section 3, the three synonym mappings used for query clarification are introduced; we describe our experimental setup in Section 4 and present the results in Section 5; we detail a classifier that selects the optimal synonym mapping for each query in Section 6; finally, we discuss the implications of this work in Section 7.

2 Related Work

Interest in medical search is steadily increasing, and many approaches to improve its accuracy have been proposed. Some prior efforts introduced systems to help health professionals searching medical literature; more recently, researchers have focused on helping laypeople retrieving accurate health information on the web.

We categorized previous research efforts that relate to query clarification into four groups. First, in Section 2.1, we highlight query expansion approaches for medical literature retrieval that share similarities with our contribution. Then, in Section 2.2, we provide an overview of those research efforts aimed at characterizing the behavior of laypeople when seeking health information on the web. In Section 2.3, we present recent research contributions concerned with studying how domain expertises influence the behavior of health information seekers. Finally, in Section 2.4, we survey the most relevant efforts that have been proposed in recent years to improve consumer health search.

2.1 Query Expansion for Medical Literature Retrieval

Document-based query expansion techniques, such as pseudo-relevance feedback, have been extensively studied in the context of medical literature retrieval. For example, Abdou and Savoy (2008) found that blind-query expansion improved performance retrieving MEDLINE[3] documents. Similarly, Jalali and Matash Borujerdi (2010) studied the effect of using concept based pseudo-relevance feedback in the medical domain.

Domain specific resources, such as the Unified Medical Language System (UMLS) Metathesaurus[4] and Medical Subject Headings (MeSH) terms[5], have also been ex-

[3] http://www.nlm.nih.gov/pubs/factsheets/medline.html

[4] http://www.nlm.nih.gov/research/umls/

[5] http://www.ncbi.nlm.nih.gov/mesh/

tensively employed for query expansion in the medical domain. For example, Liu and Chu (2007) used UMLS to perform query expansion on the OHSUMED test collection (Hersh et al. 1994). Griffon et al. (2012) expanded PubMed Boolean queries using synonyms from UMLS. Mu et al. (2014) introduced and compared two search engines—SimpleMed and MeshMed—on OHSUMED. SimpleMed augmented search results with additional fields for articles' abstracts and MeSH terms, while MeshMed added a term definition and MeSH ontology browser. Users were asked to define biomedical terms and demonstrate knowledge of medical terms' relations to each other; their performance improved when using MeshMed. Other approaches have used MeSH terms to expand queries in ImageCLEFmed[6] (Díaz-Galiano et al. 2009) and TREC Genomics[7] (Jalali and Borujerdi 2008; Lu et al. 2009).

Since poor formulation of the query submitted by consumers might affect the quality of results retrieved by a search engine, we found pseudo-relevance feedback to be not suitable for query clarification. Instead, we used three laypeople-to-expert synonym mappings to enhance health queries. One of them, similarly to some efforts in the medical literature retrieval domain, takes advantage of a portion of UMLS to define synonyms.

2.2 Laypeople as Health Information Seekers

Interaction between consumer seeking health information and web search engines has been extensively studied in recent years. Early on, Eysenbach and Köhler (2002) noticed that consumers' query formulation is often suboptimal. Moreover, they observed that laypeople struggle with identifying trustworthy websites. Spink et al. (2004) examined a large query log from Excite[8] and AlltheWeb[9]. Their findings suggest that most consumers fail to understand the limitations of web search when searching medical advices; furthermore, they rarely reformulate queries to include synonyms or alternate health expressions that could increase the quality of retrieved results. Toms and Latter (2007) also noticed that consumers are often unable to properly formulate queries when looking for health resources.

More recently, Cartright et al. (2011) studied the behavior of consumers when searching for health information. Their findings suggest that users perform evidence-directed and hypothesis-directed exploratory health searches. Powell et al. (2011) conducted a comparative study between popular search engines (Google, Bing, Yahoo! and Ask.com) in retrieving health information about breast cancer. They noticed that, while all the search engines were able to provide somewhat satisfactory results, the rankings of retrieved web page was often suboptimal, therefore leaving room for improvement to help users get more accurate information.

Lastly, Zuccon et al. (2015) analyzed the results retrieved by two commercial web search engine (Google and Bing) on a set of queries formulated by laypeople describing medical symptoms. For both engines, only three of the top ten retrieved results were both relevant and from trustworthy websites. Their analysis suggests

[6] `http://imageclef.org/`
[7] `http://ir.ohsu.edu/genomics/`
[8] `http://www.excite.com/`
[9] `http://www.alltheweb.com/`

that current search engines are not sufficiently equipped to satisfy the information need associated with the laymen queries in their dataset.

Query clarification, introduced in this work, was designed to improve queries submitted by users with limited medical vocabulary to retrieve more relevant and trustworthy web pages.

2.3 Influence of Domain Expertise in Health Search Behaviors

Researches have also studied the differences between experts and laypeople when performing health-related searches. White et al. (2008) analyzed interaction logs from Google, Yahoo!, and Microsoft Live Search. Based on their analysis, the authors concluded that health experts—compared to laypeople—are more likely to visit authoritative medical websites, issue long queries, use domain appropriate terms, spend more time searching, and reformulate queries often. Palotti et al. (2014) proposed a set of features that could help discern queries issued by health professionals from queries issued by laypeople.

While our experiments confirm some of the aforementioned findings, our work focuses on how to bridge the gap between laypeople and medical experts rather than analyzing the differences between the two groups.

2.4 Efforts in Improving Consumer Health Search

The interest in helping laypeople access reliable medical resources has increased in the last few years. Zeng et al. (2006) started the Consumer Health Vocabulary (CHV) initiative, a resource designed to link medical terms and expression used by consumers to terms health care professionals use. CHV is included in UMLS since version 2011AA. Yates et al. (2014) proposed a system to programmatically extract synonyms from a corpus of medical forum posts. Can and Baykal (2007) created MedicoPort, a retrieval engine that enhances health queries using UMLS. Luo et al. (2008) built MedSearch, a search engine designed to process long, discursive queries and retrieve trustworthy results from a set of hand picked sources. The proposed system increased search results diversity, as well as suggesting new queries.

More recently, Goeuriot et al. (2013, 2014b) introduced, in the ShARe/CLEF eHealth Evaluation Lab, a task concerned with improving systems designed to help laypeople seeking health information online. Participating systems were asked to retrieve relevant documents from a set of certified websites by the HON foundation and other hand-picked trusted resources. In a subsequent work, Goeuriot et al. (2014a) provided a more detailed analysis of the impact of query complexity on the performance of the participating systems. Query complexity was estimated by the number of medical concepts in the query (manually annotated). Their findings suggest that the increase in query complexity affected the retrieval performances of all systems under examination.

Stanton et al. (2014) studied the use of circumlocution in diagnostic medical queries (i.e., situations in which a non-expert uses many words to describe a symptom in place of the appropriate medical term). The authors proposed a supervised approach to link circumlocutory queries to medical concepts.

Nie et al. (2014) introduced a local/global learning approach to question answering in the medical domain. Their system is designed to match questions written by laypeople with answers provided by experts. First, medical concepts are extracted from questions and mapped to to SNOMED CT[10] terms; then, questions are matched to answers.

Query clarification, unlike other efforts, directly addresses the limited vocabulary used by laypeople in health searches, which has been identified as one of the major shortcomings in consumer health search. Compared to other systems that are also designed to improve consumer health search, we do not propose an end-to-end custom search solution; rather, we introduce a methodology that improves existing search engines that consumers are already familiar with. Finally, rather than relying on hand picked, trusted sources (which would require continuous human curation), our approach automatically improves queries using three a priori synonym mappings. Therefore, our system can be easily and affordably integrated into existing search solutions.

3 Methodology

We bridge the gap between laypeople and experts in the health search domain to improve users' ability to answer medical questions. As such, we investigated using three different synonym mappings to perform query clarification.

For each query, we generated three clarified queries using the synonym mappings described in Section 3.1. Each mapping associates an expression from layperson's vocabulary (i.e., a word or phrase a non-expert would use to describe a health concept) to one or more expressions used by medical professionals, which we refer to as "clarification candidates". Section 3.2 describes the algorithm used to select the most appropriate expression among clarification candidates.

For each of the four query versions (the original and the three derived from clarification), we used Bing to retrieve relevant search results. In Section 3.3, we discuss the overlap between each synonym mapping, as well as the overlap between the retrieved results.

3.1 Medical Synonym Mappings

3.1.1 Behavioral

Based on Yom-Tov and Gabrilovich (2013), this mapping links expressions commonly used by laypeople to describe their medical condition to 195 symptoms listed in the International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10)[11]. The synonyms were generated in two ways. First, the most frequent search terms that led users to click on Wikipedia pages describing symptoms were selected. Second, frequently occurring lexical affinities (Carmel et al. 2002) were added to the list. Lexical affinities are word pairs appearing in close proximity in the 50 highest ranked search results retrieved when

[10] `http://www.ihtsdo.org/snomed-ct/`
[11] `http://www.who.int/whosis/icd10/`

symptoms were used as queries. The list was validated by medical professionals, and 88% of terms were found to be appropriate expansion terms for the symptoms. The list was generated using search information from the Yahoo! search engine collected in 2010. A detailed description of this mapping can be found in (Yom-Tov and Gabrilovich 2013).

3.1.2 MedSyn

Based on Yates and Goharian (2013), this synonym mapping focuses on diseases and symptoms. It was generated from a subset of UMLS filtered to remove irrelevant terms types. SIDER 2 (Kuhn et al. 2010) was used to keep only terms with UMLS semantic types that were assigned to side effects listed on drug labels. Synonyms of these terms were identified using UMLS' semantic network and added to the map. Finally, relevant common terms from a drug review data set (Yates and Goharian 2013) were added to the map as synonyms of the appropriate terms. To ensure that only expert terms were added to queries, we kept only terms designated as *preferred terms*[12] in UMLS as candidate expressions (i.e., expressions used to clarify a query).

3.1.3 DBpedia

This mapping takes advantage of Wikipedia redirect pages as a mean to map laypeople expressions to expert terminology. Redirect pages are meant to route users to the most appropriate expression for a concept. For example, the Wikipedia page for "acid reflux"[13] redirects to "gastroesophageal reflux disease"[14]. Wikipedia redirect pages have been successfully employed in building general ontologies (Suchanek et al. 2008), creating domain specific thesauri (Milne et al. 2006), and improving query reformulation (Milne et al. 2007; Xu et al. 2008). We took advantage of DBpedia[15], a project aimed at extracting structured information from Wikipedia, to parse redirect pages. Through this knowledge base, we label two expressions X and Y as synonyms if there exits a redirect from page X to page Y. To prevent query drift, we only kept those redirect terms which led to a Wikipedia page describing a medical symptom, drug, or disease. This ensures that those terms in the query that are not health-related are not attempted to be clarified.

3.2 Candidate Selection

In some instances, a synonym mapping associates an expression (which could be either a word or a phrase) in a query with more than one clarification candidate $\{c_1, \ldots, c_m\}$. However, not all clarification candidates are equally suitable for expansion: some are more apt at representing the medical concept in the query and are therefore preferred in trustworthy resources. Therefore, our goal is to **select**

[12] In UMLS, an expression is labeled as *preferred term* if it is found to be the most appropriate to represent a concept.

[13] `http://en.wikipedia.org/wiki/Acid_reflux/`

[14] `http://en.wikipedia.org/wiki/Gastroesophageal_reflux_disease/`

[15] `http://dbpedia.org/`, accessed July 2013

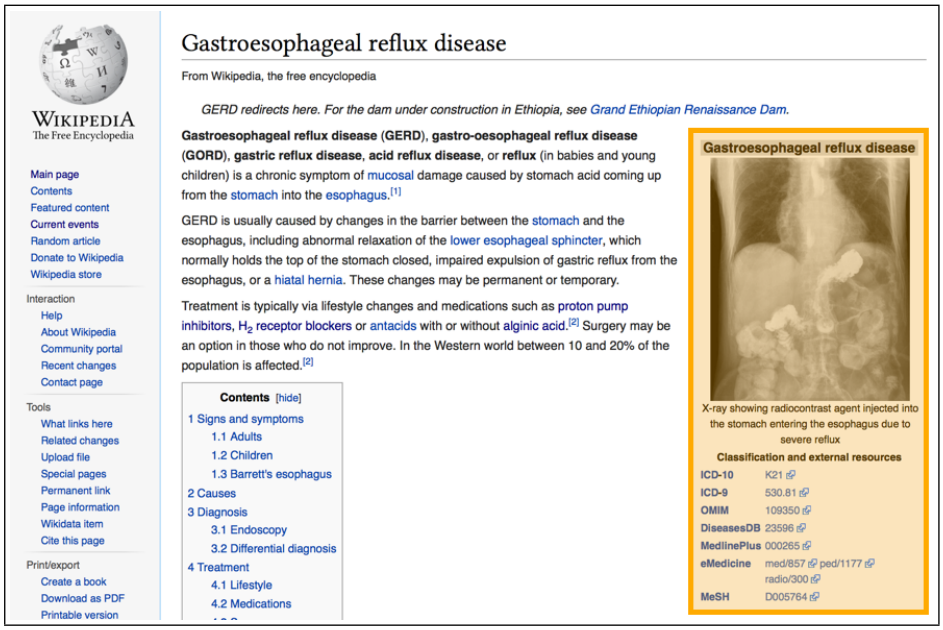

Fig. 1 A screen shot of the Wikipedia entry for "Gastroesophageal reflux disease". The information box is displayed on the right side of the page, highlighted in orange. Because it contains several medically-related identification codes, this page was identified as health-related.

the clarification candidate c_k that better represents the medical concept expressed by consumers in the query. The following heuristic was considered to achieve this goal: when multiple clarification candidates are identified by a mapping, we choose the candidate c_k whose probability of appearing in health-related Wikipedia pages is maximized. Wikipedia was deemed appropriate to determine the best clarification candidate because of its strict manual of style[16] and the expertise of the editors curating the Medicine Portal[17] [more than half of the editors are medical practitioners, 85.5% holds a university degree (Heilman and West 2015)].

Formally, given the set $\mathbb{W}$ of all Wikipedia pages, and the set of health-related pages $H(\mathbb{W})$, $H(\mathbb{W}) \subset \mathbb{W}$, we estimate the probability of a page p being health-related given a candidate expression c_k as:

$$\Pr\{p \in H(\mathbb{W}) \mid c_k \in p\} = \frac{\Pr\{p \in H(\mathbb{W}) \wedge c_k \in p\}}{\Pr\{c_k \in p\}} \quad (1)$$

The numerator of the fraction on the right side of Equation 1 is calculated by dividing the number of pages in $H(\mathbb{W})$ containing c_k by the size of $H(\mathbb{W})$. The denominator is obtained by dividing the number of Wikipedia pages containing c_k by the total number of pages in Wikipedia.

[16] `http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Medicine-related_articles`
[17] `http://en.wikipedia.org/wiki/Portal:Medicine`

In accordance with the previously stated heuristic, the candidate maximizing the following equation is selected for clarification:

$$\underset{c_k \in \{c_1,\ldots,c_m\}}{\arg\max} \left(\Pr\{p \in H(\mathbb{W}) \mid c_k \in p\}\right) \tag{2}$$

Intuitively, the more a clarification candidate appears in health-related Wikipedia pages, the more likely it is that the candidate is the most appropriate expression to describe the concept in the query. Therefore, we clarify a query with the expression c_k that maximizes Equation 1.

The set $\mathbb{W}$ was defined over a snapshot of Wikipedia obtained on November 4, 2013. We took advantage of the content of the information box (e.g., Figure 1) of each Wikipedia entry to define the set $H(\mathbb{W})$ (i.e., to determine which pages should be considered health-related). In detail, any page whose information box contained one of the following medically-related identification codes was designated as health-related: *MedlinePlus*, *DiseasesDB*, *eMedicine*, *MeSH*, or *OMIM*. Of $2,794,145$ unique pages indexed, about 0.88% ($24,654$ pages) were identified as health-related.

We avoided augmenting a query with more than one clarification candidate to minimize the likelihood of query drift. If multiple expressions in a query can be mapped to an expert term using a synonym mapping, we consider the longest, as it fully captures the information need of the user. If multiple expressions of the same length can be clarified, we choose the one with the highest conditional probability.

3.3 Overlap Between Mappings

We compare and contrast the synonym mappings introduced in Section 3.1 as a means of providing a greater understanding of their differences and similarities. In detail, we examine the size of the mappings, as well as the overlap between each pair. Finally, we analyze the overlap of set of results retrieved for each query in our dataset before and after being clarified by each synonym mapping.

Table 1 Size of the synonym mappings.

	Unique expressions	**Synonym pairs**
Behavioral	593	611
MedSyn	$6,760$	$43,703$
DBpedia	$64,652$	$177,116$

Table 1 shows the size of each synonym mapping in terms of unique expressions and in terms of synonym pairs (i.e., pairs of non-expert expression X and expert expression Y). An expression may either be a single word ("GERD") or a multi word phrase ("gastroesophageal reflux disease"). *Behavioral* has the fewest number of expressions, whereas *DBpedia* has the most. In fact, *Behavioral* is much closer to a one-to-one mapping than *MedSyn* and *DBpedia*, as both include relationships between many more pairs of synonyms. Note, however, that *Behavioral* only includes medical symptoms, which may explain its size in comparison to the other

synonym mappings. The size difference shown in Table 1 unsurprisingly affects the number of clarification candidates of each mapping. *Behavioral* selected, on average, M=1.02 (SD=0.24) candidates per query, while *MedSyn* selected M=1.16 (SD=1.07) candidates. The difference between the two is not statistically significant (Mann-Whitney U test, $p = 0.243$). *DBpdia*, the largest mapping, consistently selected the largest number of candidates per query: M=2.46 (SD=4.42) (difference is statistically significant over *Behavioral* and *MedSyn*, $p < 0.05$).

Table 2 Percentage overlap between the lists of synonyms. Each cell (i, j) in the table represents the overlap of synonym mapping i with synonym mapping j as a percentage of the size of mapping i. To better understand the relative size of each overlap, the number of overlapping expressions is also reported.

	Behavioral	***MedSyn***	***DBpedia***
Behavioral	-	21.3% (126 expressions)	98.5% (584 expressions)
MedSyn	1.9% (126 expressions)	-	8.0% (540 expressions)
DBpedia	0.9% (584 expressions)	0.8% (540 expressions)	-

The overlap between each list of synonyms is shown in Table 2. For each cell (i, j) in the table, we report the overlap of synonym mapping i with synonym mapping j as a percentage of the size of mapping i. *Behavioral*, the mapping with the smallest synonym list (as shown in Table 1), is almost completely contained (98.5%) within *DBpedia*, the largest mapping. *Behavioral* and *MedSyn* have far fewer expressions in common, as about one fifth (21.3%) the expressions in *Behavioral* are also present in *MedSyn*.

Table 3 Query overlap between the unclarified query ("no clar.") and the queries clarified by each mapping. *MedSyn* is the most similar to the baseline, while *Behavioral* and *DBpedia* are the most similar synonym mappings. Unlike Table 2, this table is symmetrical, as all queries in the dataset were clarified using all synonym mappings.

	no clar.	***Behavioral***	***MedSyn***	***DBpedia***
no clar.	-	2%	30%	0%
Behavioral	2%	-	28%	74%
MedSyn	30%	28%	-	36%
DBpedia	0%	74%	36%	-

Table 3 shows the overlap between the unclarified queries the queries clarified by each mapping (as described in Sections 3.1 and 3.2). In cases where a synonym mapping had no clarification expression to add, we say that the *null* term was added; this allowed us to compute overlap between the unclarified query (which we refer to as "no clar.") and each synonym mapping. By definition, "no clar." adds the *null* term to each query. *MedSyn* added the *null* term (i.e., did not add any clarification expression to the query) 30% of the time, while both *Behavioral* and *DBpedia* added a expression to the vast majority of queries. *Behavioral* and *DBpedia* often lead to similar clarification (74% overlap), which is to be expected

given the high overlap between the two synonym lists. Finally we note that, despite the fact that only 8% of the synonyms found in *MedSyn* occurred in *DBpedia* (Table 2), the overlap in terms of expressions added to the queries by the two mapping was considerably higher (36%). This outcome is likely due to the fact that the queries in our dataset, which are among the 500 most common health queries on Bing (Section 4.1), contain health expressions that are very likely to be included in both synonym mappings.

Table 4 Overlap of the URLs of results retrieved by the unclarified query ("no clar.") and by the queries clarified by each mapping. *MedSyn* is most similar to the unexpanded baseline, but still adds a significant number of URLs.

	no clar.	***Behavioral***	***MedSyn***	***DBpedia***
no clar.	-	14%	38%	13%
Behavioral	14%	-	36%	74%
MedSyn	38%	36%	-	42%
DBpedia	13%	74%	42%	-

Table 5 Overlap of the snippets of results retrieved by the unclarified query ("no clar.") and by the queries clarified by each mapping. While these results closely resemble those presented in Table 4, they are not equivalent, as (*i*) different pages may share the same snippet and (*ii*) web search engines may alter result snippets based on the query.

	no clar.	***Behavioral***	***MedSyn***	***DBpedia***
no clar.	-	18%	38%	16%
Behavioral	18%	-	40%	75%
MedSyn	38%	39%	-	45%
DBpedia	16%	75%	45%	-

The overlap between the URLs of the retrieved results is shown in Table 4, while Table 5 contains the overlap between the snippets of the retrieved results. The results for two types of overlap are not equivalent. This behavior is expected, as (*i*) different web pages could potentially share the same snippet and (*ii*) web search engines may alter the text of snippets of results to better match the formulation of a query. Nevertheless, both types of overlap yield similar results on our dataset; we notice that the snippet overlaps tend to be slightly higher, as different—but similar—pages on the same website have, in some cases, equivalent snippets. Statistics shown in Tables 4–5 confirm that *Behavioral* and *DBpedia* are the most similar mappings. Both have little overlap with the URLs of results retrieved with the unclarified query (13% and 14%, respectively); a slight increase can be observed for both mappings when overlap is measured with respect to the snippets of retrieved results. Queries clarified with *MedSyn* retrieved, on average, 38% of the results retrieved by the unclarified query.

Summarizing our comparison, queries clarified using *Behavioral* and *DBpedia* retrieve the most similar set of results, even though the former mapping comprises

of only a small subset of the latter. Of all synonym mappings, *MedSyn* yields the most similar results to the baseline; yet, it still adds a significant number of clarification expressions and URLs over the unclarified query.

4 Experimental Plan

To evaluate the effectiveness of our clarification strategy, we used the three synonym lists introduced in Section 3.1 to clarify 50 queries from a Bing query log. Details regarding the set of queries are provided in Section 4.1. Laypeople and medical experts were enrolled to assess the impact of the proposed methodology. For each query, we created a multiple-choice question; participants were required to answer it to demonstrate their understanding of the retrieved results. We overview the query creation process in Section 4.2. Query clarification was evaluated using an online platform we introduce in Section 4.3.

All the resources detailed in this section (queries, questions, and anonymized user interaction reports) are publicly available at the authors' GitHub page[18].

4.1 Queries Dataset

As previously mentioned, we studied the impact of query clarification on a sample of common health-related queries for a Bing query log. To do this, we extracted the set of all English-language queries submitted to Bing by users in the United States during November 2013. This set was filtered to extract those queries which contained a symptom, drug name, or disease name, or one of their synonyms, as listed in Wikipedia. We randomly sampled 50 out of the 500 most common queries in the resulting list. Sampling was done to reduce the dimensionality of the dataset, thus making the experimentation more tractable.

The 50 queries in the dataset contain 93 unique terms and have an average length of 2.6 terms (median length is equal to 2). This is not statistically significantly different (rank-sum test) from the queries in the larger set of 500 queries, which have an average length of 2.5 (median is 2), and contain 463 unique terms. The list of the 50 queries is included in appendix A.

4.2 Evaluation Questions

The process laypeople follow while looking for medical information on the Internet is akin to a task-based retrieval scenario: consumers have a specific information need that they try to satisfy through web search engine. Thus, for our task-based experiment, we created, for each query, a question that would estimate the quality of the retrieved results in providing helpful information to a user. Users in our scenario are given a similar task to (Hersh et al. 1996), where medical students were asked to use a search system to gather information to answer a question. Such approach is also common in focus groups examining the behavior of laypeople seeking health information on the web (Eysenbach and Köhler 2002; Toms and

[18] `https://github.com/Georgetown-IR-Lab/query-clarification-data`

Latter 2007). Since a users' ability to correctly answer questions is uncorrelated with the number of relevant documents read (Hersh et al. 1996) or precision and recall (Hersh et al. 2002), we consider the users' question answering accuracy when we analyze our results.

Our design goal was to formulate questions that (a) were highly relevant to the query, (b) required reading at least one, if not many, of the links shown and (c) were not easily intelligible by reading the snippets provided with each search result. Each question was created using the following procedure: first, the authors read the query and content of the search results; then, they formulated a question based on the content of the retrieved web pages; finally, they generated four possible answers—one correct, three wrong. The volume of data needed by our study ruled out the option of proposing open questions.

4.3 Online Evaluation Platform

We developed a website (Figure 2) to determine the effectiveness of the proposed clarification methodology. Through this website, laypeople and medical experts answered a set of health-related, multiple-choice questions using a set of search results retrieved using Bing. For each query in the dataset, we showed participants the query itself and the question simulating the information need associated with the query. Users were asked to find the answer to the question presented to them by using the displayed search results. We required the participants to open (click) at least one link before choosing the correct answer among four possible choices to prevent bias in results selection. To minimize the number of factors involved in the study, users were not allowed to modify the displayed query. For each respondent and each query, an interaction report consisting of the links clicked and the answer given was created.

We interleaved search results to quantify the impact of each synonym mapping we used for query clarification. Interleaving, introduced by (Joachims 2002), is a technique designed to receive implicit user feedback about two retrieval methods without introducing bias due to the presentation of the results. Team draft interleaving (Radlinski et al. 2008) was chosen for the evaluation platform; as its name might suggest, it mimics how players are usually divided in teams at the beginning of friendly matches. Given two ranked lists A and B of retrieved results, $A = \{a_1, \ldots, a_n\}, B = \{b_1, \ldots, b_n\}$, we operate as follows: for each pair of results a_i and b_i of rank i, an unbiased coin is flipped; if heads, a_i is ranked before b_i

Table 6 An example of query in our dataset. The first mapping, "no clar.", represents the original unclarified query as extracted from the Bing query log. The last column contains the question formulated by the authors. In parentheses we report the four corresponding answers (the correct one is in **bold**).

Mapping	Query	Question
no clar.	excessive burping	"*Which of the following solution does NOT help with excessive ructus?*" (avoiding drinking through a straw, taking an antacid, eating slowly, **swallowing air**)
Behavioral	excessive burping belching	
MedSyn	excessive burping eructation	
DBpedia	excessive burping belching	

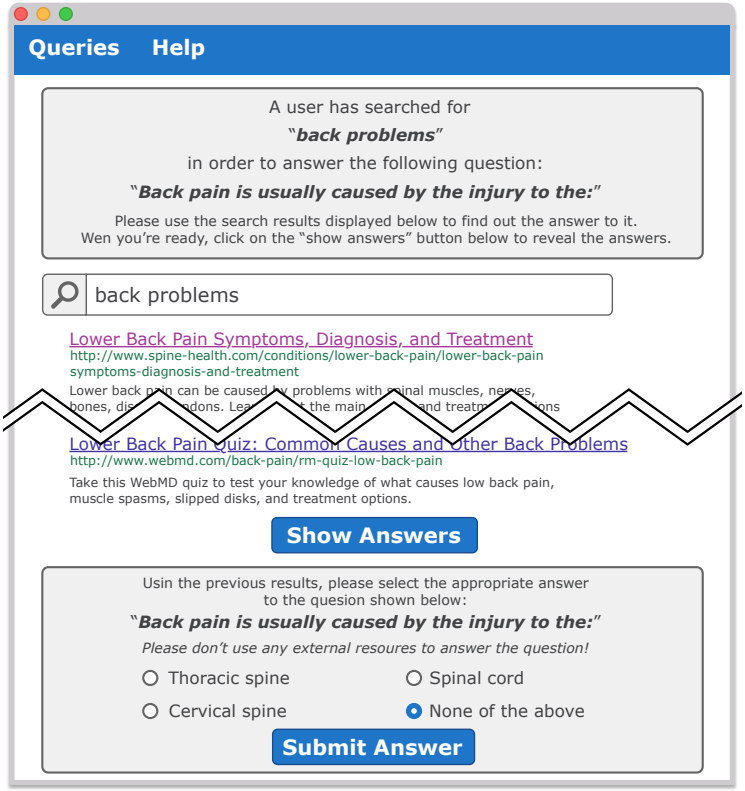

Fig. 2 The main interface of the website. The top third of the screen shows the question for the user, while the middle part displays the original query and ten interleaved results. The bottom section shows the question which the user is asked to answer. Even when results obtained via a clarified query are presented, the original query is shown; users are not allowed to reformulate the query at any point. The multiple choice options to the question are initially hidden and can be revealed by the user after opening (clicking) at least one result.

in the interleaved set of result; if tails, b_i is ranked first. As detailed in (Radlinski and Craswell 2013), team draft interleaving shows comparable levels of expert agreement to other interleaving methods, and it is less prone to introducing bias.

We tested query clarification among laypeople recruited using Amazon Mechanical Turk[19]. Each participant was asked to answer 20 medical questions. Workers were paid between \$2.00 and \$4.50 (M=\$3.53, SD=\$0.99), depending on when they accepted the task. We enrolled as many workers as needed to obtain at least 5 interaction reports per query per pair of methods. In total, 80 workers registered for the task.

We also enrolled 12 freelance medical experts using Elance[20]. These workers were paid \$20.00 for their efforts. We provided interleaved results retrieved using original queries and queries clarified by *MedSyn* to this group of participants. *MedSyn* was chosen because its promising results on preliminary tests. The size of

[19] `http://mturk.amazon.com/`
[20] `http://www.elance.com/`

this group was also determined by the need of at least 5 interaction reports for each query.

5 Results

We analyzed the results collected in the two task based experiments to determine (*i*) whether users preferred the results retrieved by a clarified query or not, and (*ii*) whether query clarification increased the likelihood of correctly answering the question associated with each query. After determining that query clarification improves task-based retrieval for lay users (as evidenced by the clarification methods' Kemeny rankings), we analyzed whether this improvement also holds for medical experts; finally, we investigated whether query clarification led to more trustworthy web pages (as identified by the HON certification) being returned.

Our findings are presented in the remaining of this section. In detail, we illustrate how the results retrieved using clarified questions are preferred by lay users in Section 5.1, while Section 5.2 quantifies the difference in fraction of correct answers for each synonym mapping and original query. Differences between the two groups of participants are described in Section 5.3 and 5.4. Finally, we show the users are more likely to answer correctly to the question associated with each query when visiting certified health pages in Section 5.5.

5.1 Kemeny Ranking

Team draft interleaving assumes that, for each query, the method preferred by a user is the one that retrieved the majority of web pages (s)he visited. Thus, we assigned a preference to synonym mapping i when compared with mapping j if a user clicked more results retrieved by a query clarified with mapping i than results retrieved by a query clarified with mapping j.

The Kemeny-Young method (Young and Levenglick 1978) was used to determine the users' preferred ranking among the three synonyms lists and original

Table 7 The best synonym mappings as determined by the Kemeny-Young method. "no clar." represents the set of retrieved results by the original (unclarified) query. The leftmost column indicates that results retrieved by queries clarified with *MedSyn* were the preferred over all queries. However, when only considering those instances where questions were correctly answered, *Behavioral* was the preferred mapping, shortly followed by *MedSyn* (central columns). When only preferences associated with incorrectly answered queries (rightmost column), *MedSyn* is, once again, the preferred mapping.

All questions	**Correctly answered** (tie between two rankings)		Incorrectly answered
1st: *MedSyn*	**1st: *Behavioral***	**1st: *MedSyn***	1st: *MedSyn*
2nd: no clar.	**2nd: *MedSyn***	**2nd: *Behavioral***	2nd: no clar.
3rd: *DBpedia*	**3rd: no clar.**	**3rd: no clar.**	3rd: *DBpedia*
4th: *Behavioral*	**4th: *DBpedia***	**4th: *DBpedia***	4th: *Behavioral*

query ("no clar."), which we will refer to as "candidates" throughout the rest of this section. The Kemeny-Young method was originally designed to combining prioritized/ranked votes; in information retrieval, it has been used to perform rank aggregation on search result sets (Dwork et al. 2001; Dasdan et al. 2009), on candidates in question answering tasks (Agarwal et al. 2012), and on short texts in social media (Subbian and Melville 2011).

The score for each ranking (which, in this context, is a permutation of the list {no clar., *Behavioral*, *MedSyn*, *DBpedia*}) is computed by summing the number of votes for each pair of candidates in the ranking. The ranking with the highest score is the Kemeny ranking.

Formally, given a ranking $r = \{c_1, \ldots, c_m\}$, the score $S(r)$ of the ranking is calculated as:

$$S(r) = \sum_{\substack{i,j \in \{1,\ldots,m\} \\ i<j}} (\text{count of user preferring candidate } i \text{ over candidate } j) \quad (3)$$

By definition, the Kemeny ranking maximizes the number of pairwise agreement between users, where two users agree if they have expressed preference of a candidate over another candidate. In other words, a ranking $r = \{c_1, \ldots, c_m\}$, will score high if, for all $i, j \in \{1, \ldots, m\}, i < j$ many users prefer candidate c_i over candidate c_j.

Table 7 shows the Kemeny rankings for the Mechanical Turk users with respect of the set of all questions (left), the set of questions which were answered correctly (center), and the set of questions which were answered incorrectly (right). When the set of all questions is considered, results retrieved by queries clarified via *MedSyn* are preferred by Mechanical Turk users, followed by web pages retrieved by unclarified queries. If only the set of correctly answered questions is considered, two rankings achieve the same Kemeny score; in both cases, results retrieved by clarified queries are preferred (*Behavioral* and *MedSyn*). When only the set of incorrectly answered questions is considered, an identical ranking to the set of all queries is observed. This symmetry, while perhaps counterintuitive, is due to the fact that the results retrieved by the unclarified queries ("no clar.") are preferred more highly in those cases when a question is incorrectly answered; this preference skews the results when all questions are considered, thus causing the symmetric behavior observable in Table 7.

Results retrieved by queries clarified through *MedSyn* are preferred more highly across all questions, regardless of whether questions were answered correctly or not. *Behavioral*, while being the preferred clarification mapping for correctly answered questions, ranks last when the set of all questions is considered. We hypothesize that such behavior is due to the skewness induced by the aforementioned preference expressed for unclarified queries. We observe that *Behavioral* does not exhibit such skewness with respect of the set of correctly answered questions; this could be caused by the fact that users seem to equally prefer queries clarified by *Behavioral* and *MedSyn*.

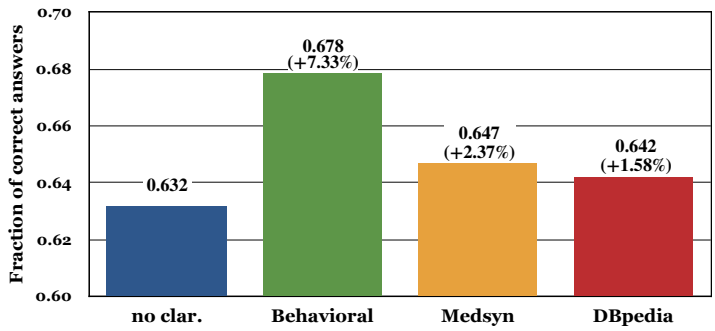

Fig. 3 Average fraction of correct answers for each clarification candidate. For each candidate, the fraction is calculated over all the query/user combinations where the candidate is preferred. *Behavioral*, the method with the highest fraction of correct answers, improves over the baseline (no clarification, leftmost bar in blue) by 7.33% (statistically significant, Welch's t-test, $p < 0.05$).

5.2 Fraction of Correct Answers for Each Mapping

While the Kemeny-Young method provides great insights about the preference expressed by participants towards results retrieved using clarified queries, its findings are insufficient to properly determine which synonym mapping is the most appropriate for query clarification. In particular, the Kemeny ranking does not measure the difference between *MedSyn* and *Behavioral*, the two most preferred mappings for the set of correctly answered questions (Table 7, center). To quantify such difference, we calculate the average fraction of correct answers for each clarification candidate when the query clarified by such candidate is preferred (Figure 3).

Behavioral had the highest fraction of correct answers (0.678). In other words, when users express a preference for results retrieved by a query clarified with *Behavioral*, they were able to correctly answer the question associated with the query 68% of the time. This results represent an improvement of 4.63% over *MedSyn*, an improvement of 5.38% over *DBpedia*, and an improvement of 7.33% over no query clarification (statistically significant, Welch's t-test, $p < 0.05$). This suggests that *Behavioral* is to be considered the best-performing synonym mapping, since it both achieves the highest Kemeny ranking for correctly answered questions and yields the highest fraction of correct to incorrect question answers.

The findings detailed in this subsection corroborate our observations regarding the Kemeny ranking: *MedSyn*, while being the most preferred synonyms mapping across all questions, is associated with a lower rate of correct answers, due to the strong preference expressed for it for the set of incorrectly answered questions. On the other hand, *Behavioral* achieves the highest fraction of correct answers; to the fact that it is one of the most preferred clarification mappings in the set of correctly answered questions, and the least preferred for the set of incorrectly answered questions.

5.3 Users Analysis

As previously mentioned, the synonym mappings were tested on two groups of users: laypeople, recruited via Amazon Mechanical Turk, and freelance medical professionals, enrolled on Elance. Given the differences between the members of the two sets, we compare the two groups. Descriptive statistics are reported in Table 8, while the distributions of users are represented in Figure 4. All users answered questions better than would be expected by chance (i.e., 25% of the time). Furthermore, the vast majority ($> 95\%$) of users answered questions correctly over 50% of the time.

As shown in Table 8, the expert group correctly answered a higher number of questions (statistically significant, Welch's t-test, $p < 0.05$). Moreover, experts were found to visit more web pages before answering to each question, which is consistent with the findings reported in previous studies (White et al. 2008). Users in both groups were found to click on more results before correctly answering a question, although the difference was not found to be significant (Welch's t-test, $p = 0.687$ for laypeople, $p = 0.556$ for experts).

We quantified the inter-agreement between the two sets of participants using Fleiss' kappa (Table 8). Experts were found to have a substantially higher agreement than laypeople. This observation, alongside the higher success rate, confirms the intuition that experts are more likely to correctly answer the proposed questions. This could be due to the fact that health professionals, thanks to their background, are able to successfully infer the necessary information from the retrieved results to satisfy their information need. We hypothesize that laypeople are instead more likely to randomly guess when they are presented with a difficult question, thus exhibiting both lower agreement and lower success rate.

For the laypeople group, we observed a moderate positive correlation between the average number of web pages visited and the fraction of correct answers (Spearman's correlation, $r_s = 0.228$, $p < 0.05$). In other words, those users who visited more web pages were more likely to correctly select the correct answer. For the expert group, we noticed a strong but not significant negative correlation between

Table 8 Overview of the differences between laypeople and experts. The significance of differences between the two groups were measured using Welch's t-test (2-tailed).

	Laypeople	**Experts**
Number of survey participants	80	12
Fraction of correct answers Sig. difference between groups, $p < 0.05$	M=0.655, SD=0.135	M=0.723, SD=0.116
Average clicks per correct answer Sig. difference between groups, $p < 0.05$	M=1.94, SD=0.84	M=3.19, SD=1.42
Average clicks per wrong answer Sig. difference between groups, $p < 0.01$	M=1.60, SD=0.93	M=2.86, SD=1.23
Intra-agreement within groups (Fleiss' kappa)	0.4477	0.6528

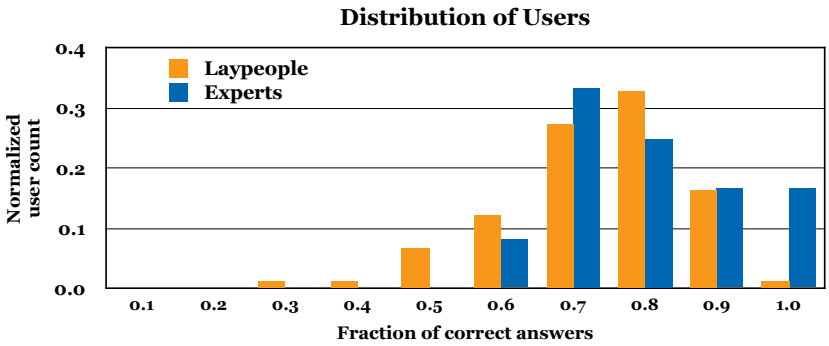

Fig. 4 Distributions of the fraction of correct answers by laypeople (orange, N=80 M=0.655, SD=0.135) and experts (blue, N=12, M=0.723, SD=0.116).

the average number of web pages visited and the fraction of correct answers (Spearman's correlation, $r_s = -0.558$, $p = 0.083$). This finding, while not conclusive, may suggest that more skilled experts—who have a higher success rate—may need to visit less web pages to correctly answer a question.

For both groups, a very strong correlation was found between the number of results clicked by a user before correctly answering a question and the number of results clicked before incorrectly answering a question (Spearman's correlation, $r_s = 0.780$ for experts, $r_s = 0.882$ for experts, $p < 0.01$ for both groups). This suggest that the number of visited results is unique to each user, and it is not influenced by the perceived difficulty of each question.

A fixed compensation was given to experts throughout the experiment; on the other end, the reward per task for laypeople increased over time to speed up data collection. To verify that higher compensation rates did not skew the performances of workers, we tested whether any relationship existed between retribution and fraction of questions correctly answered. However, no correlation was found between the two variables (Spearman's correlation, $r_s = 0.110$, $p = 0.405$).

Finally, we note that unlike laypeople, experts seem to prefer the unclarified queries over the clarified ones. Nevertheless, the difference in success rate between the two is not significant (Welch's t-test, $p = 0.409$). We hypothesize that such findings could be explained by the fact that experts are more likely to effectively determine those documents that could satisfy their information need from the text snippet, thus not benefiting from query clarification. Such hypothesis would be consistent with previous studies investigating the relationship between domain knowledge and search results click-through events (Cole et al. 2011).

5.4 Questions Analysis

We analyzed the fraction of each question's correct answers to assess query difficulty and determine eventual differences between the test groups. Figure 5 shows the distributions of correct answers for the two groups of users. The average fraction of correct answers is, as expected, higher for experts (M=0.732, SD=0.308 vs. M=0.579, SD=0.223). We observe that, for 19 out of 50 queries, all health professional correctly answered the proposed questions; in contrast, none of the

questions was correctly answered by laypeople. However, all expert users answered four questions completely incorrectly, whereas none of the questions shown to the lay user had a success rate lower than 0.11 (i.e., for each question, at least 11% of lay users answered the question correctly). This can be partially caused by the smaller size of medical experts we enrolled in our experimentation.

Finally, we observed a strong correlation of the success rate of each question between the two groups (statistically significant, Spearman's correlation, $r_s = 0.622$, $p < 0.01$). This finding suggests that some questions are, for both laypeople and experts, inheritely more difficult than others.

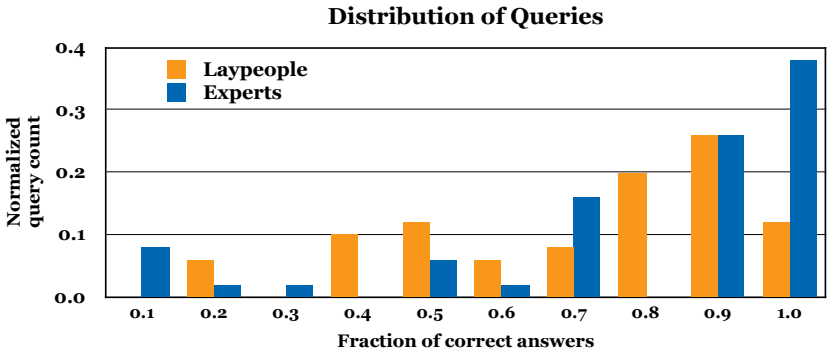

Fig. 5 Distributions of the fraction of correct answers by question for Mechanical Turk users (N=50 M=0.579, SD=0.223) and Elance users (N=50 M=0.732, SD=0.308).

5.5 Reliability of Results

As described in the introduction, the *Health On the Net Foundation* (HON)[21] is an organization that publishes a code of good conduct ("HONcode") for health-related online resources, issuing a certification for those websites that conform to it. The HONcode ensures that a website is reliable and useful in the medical information it provides. On average, M=3.43 interleaved results were certified by the HON foundation (SD=2.02, Mdn=3), while M=4.78 were not certified (SD=2.45, Mdn=4).

Table 9 Correct/incorrect number of answers when users clicked HON-certified websites. These resources led to an 7.7% statistically significant increase (Fisher's exact test, $p < 0.05$) in correct answers.

	Certified by HON	**Not Certified**
Questions answered correctly	426	566
Questions answered incorrectly	158	270

We studied the impact of HON-certified results on the fraction of correct answers given by Mechanical Turk workers. Table 9 shows the number of health-

[21] `http://www.healthonnet.org/`

related questions answered correctly and incorrectly when Mechanical Turk users clicked on and did not click on websites certified by HON. Users were 7.7% statistically more likely (significant at $p < 0.05$, Fisher's exact test) to answer the question correctly after visiting a website with HONcode certification. Such increase remains statistically significant ($p < 0.05$) when the performance of each user are normalized by the number of results visited. Therefore, we conclude that HON certified website helps laypeople answer medical questions, lending credence to the importance of such certification.

The majority (88%) of the clicks were on HON-certified websites returned by a clarified query, which again confirms the effectiveness of our system in connecting laypeople with trustworthy medical resources. Furthermore, the ratio of HON-certified vs. not certified websites remains constant at any rank position (Spearman's rank correlation coefficient $r_s = 0.921$, significant at $p < 0.01$), although the number of clicks exponentially decreased for lower ranked results. This bias toward higher ranked results is to be expected, as shown by previous research (Joachims et al. 2007).

6 Selecting the Optimal Synonym Mapping for Query Clarification

As previously mentioned in Section 5.2, query clarification increases the fraction of correctly answered questions. However, while all the mappings showed an overall improvement over the baseline, no single clarification technique consistently outperformed all others; moreover, for some queries, the unclarified query led to a higher success rate than any of the clarified queries. These observations are supported by the findings reported in Table 10. *Behavioral*, the best performing synonym mapping, improves over the baseline in 66% of the cases, while *MedSyn* and *DBpedia* outperform the baseline only in 62% and 50% of the cases, respectively. Finally, when considering any synonym mapping, we notice that, for 86% of the queries in the dataset, the baseline is outperformed; this implies that, for the remaining 14% of queries in our dataset, results retrieved by the unclarified query yield the highest rate of correctly answered questions. Motivated by these findings, we investigated whether the most appropriate mapping can be predicted to further increase the benefits of query clarification.

Table 10 Percentage of queries where the baseline (no clar.) is outperformed by each synonym mapping. Queries clarified using *Behavioral*—the best mapping—outperformed the unclarified query in 66% of the cases. The last row of the table contains the percentage of queries where any of the synonym mappings outperforms the baseline.

Synonym mapping	**Percentage of queries in which baseline ("no clar.") is outperformed**
Behavioral	66% (33 queries)
MedSyn	62% (31 queries)
DBpedia	50% (25 queries)
any synonym mapping	86% (43 queries)

Previous work on query performance prediction (Yom-Tov et al. 2005; Carmel and Yom-Tov 2010) has demonstrated that selective query expansion through a predictor achieves significant performance gains compared to either always expanding or always not expanding queries. In this section, we introduce a classifier that, given a query, either predicts which synonym mapping among *Behavioral*, *MedSyn*, and *DBpedia* should be used to clarify the query, or predicts to perform no clarification. For the reminder of this section, we will refer to the four possible outcome of the classifier as "clarification candidates".

The classifier was implemented as ensemble of four classifiers, one for each clarification candidate. In detail, four binary logistic regression models $\mathbb{M} = \{M_1, \ldots, M_4\}$ were trained as one-vs-the-rest classifiers: given a query q_i and its best clarification candidate $\mathcal{C}_k$, we trained model M_k with class label 1, and models $M_h \in \mathbb{M}, h \neq k$ with class label 0.

Two sets of features were used to train each model. The first one was defined over each query and each clarification candidate; it includes estimations of the likelihood of unigrams, bigrams, and trigrams in the query of appearing in any Wikipedia page, as well as their likelihood of appearing in health-related Wikipedia pages (as defined in Section 3.2). The longest common subsequence (LCS) between the clarified and unclarified (normalized by the length of the unclarified query) was also considered, as well as an indicator of the presence of the clarified query in any other clarification candidate. The second set of features was defined over each web page retrieved by a query q_i processed by a clarification candidate $\mathcal{C}_k$; in particular, we considered the domain name, LCS between the clarified query and the page title, LCS between the clarified query and the search snipped of the page, and the

Table 11 Features used as predictor variables for each logistic regression model M_k.

Features over query q_i and clarification candidate $\mathcal{C}_k$
Probability of bigrams and trigrams in q_i of appearing in Wikipedia
Probability of unigrams (stopwords excluded) in q_i of appearing in Wikipedia
Probability of bigrams and trigrams in q_i of appearing in health-related Wikipedia pages
Probability of unigrams (stopwords excluded) in q_i of appearing in health-related Wikipedia pages
Normalized longest common subsequence between clarified query $\mathcal{C}_k(q_i)$ and q_i
Presence of clarified query $\mathcal{C}_k(q_i)$ in any other clarification candidate $\mathcal{C}_h, h \neq k$ for query q_i
Features over query each web page p retrieved by clarified query $\mathcal{C}_k(q_i)$
Domain name of p (e.g., `nlm.nih.gov`)
Normalized longest common subsequence between page title of p and $\mathcal{C}_k(q_i)$
Normalized longest common subsequence between search result snippet of p and $\mathcal{C}_k(q_i)$
p is certified by HON

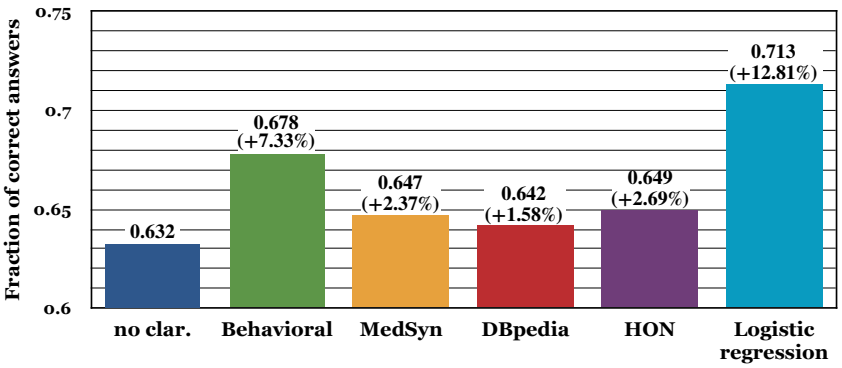

Fig. 6 Average fraction of correct answers by laypeople. Six approaches are compared: unclarified query (no clar.), three synonym mappings (*Behavioral*, *MedSyn*, *DBpedia*), a baseline classifier trained on the number of HON-certified pages retrieved (HON) and the proposed classifier (Logistic regression). Logistic regression outperforms the baseline by 12.81% (statistically significant, Welch's t-test, $p < 0.05$).

presence of the page in the Health on Net database as predictor variables. The detailed list of features is presented in Table 11.

To determine the optimal clarification mapping for a query q_i, we used each model M_k to calculate an estimation $p_{i,k}$ of the likelihood of clarification candidate $\mathcal{C}_k$ of being the optimal mapping for q_i. For each q_i, the system chose as clarification mapping the one with the highest likelihood, i.e., $\text{argmax}_k(p_{i,k})$.

The system was implemented using the *Scikit-learn* Python package (Pedregosa et al. 2011) and tested under ten-fold cross validation. The results are presented in Figure 6. We compared the performance of the logistic regression classifier with the results obtained by each individual synonym mapping. We also considered a simple multinomial logistic regression classifier trained on the fraction of retrieved results that are certified by HON as an additional baseline.

The logistic regression classifier performs well, improving over every individual synonym mapping. In detail, it achieves a 12.81% increase over the unclarified query, an 11.06% increase over *DBpedia* (Welch's t-test, $p < 0.05$), a 10.20% increase over *MedSyn* ($p < 0.05$) and a 5.16% increase over *Behavioral* ($p < 0.1$). Furthermore, it also outperforms (9.86% improvement, $p < 0.05$) the simple classifier trained on the number of HON-certified pages retrieved.

The positive results presented in this section confirm that query clarification can be further improved by selecting the most appropriate clarification candidate for each query.

7 Discussion

Seeking information on medical topics is a common task for search engine users. Arguably, this information need also has one of the most important and immediate effects on the well-being of users. However, the technical nature of this information makes it inaccessible to many users, partly because of the jargon used by medical professionals. A significant effort has been made by providers of information in the medical domain to make their content accessible to laypeople. Such accessibility

is required at several levels. At the semantic level this requires using terms that are likely to be used by non-specialists, both so that they can be retrieved when non-specialist terms are used in the search engine and so that when they are read, they can be understood by the non-specialist.

While many documents on the web use both layperson terms and medical terms, our results reveal that this effort is insufficient. We studied users' ability to complete a task-based retrieval task in which the users search to answer health-related questions. We found that by clarifying queries submitted by non-experts to a major Internet search engine, the likelihood that a user will answer health-related questions correctly increases significantly, even though the documents they read were, ostensibly, written for non-experts. Thus, our approach bridges the language gap between medical professionals and laypeople.

We compared three synonym mappings when used query clarification; our results show that all three are effective resources for such task. *Behavioral* seems to be the preferred mapping when questions are answered correctly. Furthermore, we proposed a supervised classifier that is able to select the most appropriate query clarification. The classifier outperformed every individual synonym mapping.

One interesting aspect of our results is that we did not explicitly provide the expert medical term corresponding to a layperson term: we implicitly added it to the query. Nevertheless, users found the results obtained using this approach superior for the purpose of answering their question. Moreover, even HON-certified pages, which are targeted at novice users, were better retrieved using clarification. This means that implicit query clarification is highly useful, and does not require making the user aware of the correct medical terminology.

Finally, we stress that our system minimally impacts the retrieval performances, as the query clarification terms can be computed before a query is submitted by using search logs and synonym mappings. This aspect makes it suitable to be deployed on high-traffic search engines.

8 Acknowledgments

This work was partially supported by the US National Science Foundation through grant CNS-1204347.

References

Samir Abdou and Jacques Savoy. Searching in MEDLINE: Query expansion and manual indexing evaluation. *Information Processing & Management*, 44(2):781–789, 2008.

Arvind Agarwal, Hema Raghavan, Karthik Subbian, Prem Melville, Richard D Lawrence, David C Gondek, and James Fan. Learning to rank for robust question answering. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 833–842. ACM, 2012.

Aysu Betin Can and Nazife Baykal. Medicoport: A medical search engine for all. *Computer methods and programs in biomedicine*, 86(1):73–86, 2007.

David Carmel and Elad Yom-Tov. Estimating the query difficulty for information retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2(1):1–89, 2010.

David Carmel, Eitan Farchi, Yael Petruschka, and Aya Soffer. Automatic query wefinement using lexical affinities with maximal information gain. In *Proceedings of SIGIR '02*, pages 283–290. ACM, 2002.

David Carmel, Elad Yom-Tov, Adam Darlow, and Dan Pelleg. What makes a query difficult? In *Proceedings of SIGIR '06*, pages 390–397. ACM, 2006.

Marc-Allen Cartright, Ryen W White, and Eric Horvitz. Intentions and attention in exploratory health search. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 65–74. ACM, 2011.

Michael J Cole, Xiangmin Zhang, Chang Liu, Nicholas J Belkin, and Jacek Gwizdka. Knowledge effects on document selection in search results pages. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1219–1220. ACM, 2011.

Ali Dasdan, Chris Drome, and Santanu Kolay. Thumbs-up: A game for playing to rank search results. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 1071–1072, New York, NY, USA, 2009. ACM.

M C Díaz-Galiano, M T Martín-Valdivia, and L A Ureña López. Query expansion with a medical ontology to improve a multimodal information retrieval system. *Computers in biology and medicine*, 39(4):396–403, April 2009.

Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 613–622, New York, NY, USA, 2001. ACM.

Gunther Eysenbach and Christian Köhler. How do consumers search for and appraise health information on the world wide web? qualitative study using focus groups, usability tests, and in-depth interviews. *Bmj*, 324(7337):573–577, 2002.

Susannah Fox and Maeve Duggan. Health online 2013. `http://www.pewinternet.org/Reports/2013/Health-online.aspx`, 2013.

Lorraine Goeuriot, Gareth JF Jones, Liadh Kelly, Johannes Leveling, Allan Hanbury, Henning Müller, Sanna Salanterä, Hanna Suominen, and Guido Zuccon. ShArE/CLEF eHealth evaluation lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports. 2013.

Lorraine Goeuriot, Liadh Kelly, and Johannes Leveling. An analysis of query difficulty for information retrieval in the medical domain. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 1007–1010. ACM, 2014a.

Lorraine Goeuriot, Liadh Kelly, Wei Li, João Palotti, Pavel Pecina, Guido Zuccon, Allan Hanbury, Gareth Jones, and Henning Mueller. ShArE/CLEF eHealth evaluation lab 2014, task 3: User-centred health information retrieval. In *Proceedings of CLEF*, volume 2014, 2014b.

Nicolas Griffon, Wiem Chebil, Laetitia Rollin, Gaetan Kerdelhue, Benoit Thirion, Jean-François Gehanno, and Stéfan J Darmoni. Performance evaluation of Unified Medical Language System's synonyms expansion to query PubMed. *BMC medical informatics and decision making*, 12(1):12, 2012.

James M Heilman and Andrew G West. Wikipedia and medicine: Quantifying readership, editors, and the significance of natural language. *Journal of medical Internet research*, 17 (3):e62, 2015.

William Hersh, Chris Buckley, T. J. Leone, and David Hickam. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of SIGIR '94*, pages 192–201, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

William Hersh, Jeffrey Pentecost, and David Hickam. A task-oriented approach to information retrieval evaluation. *J. Am. Soc. Inf. Sci.*, 47(1):50–56, January 1996.

William R Hersh, M. Katherine Crabtree, David H Hickam, Lynetta Sacherek, Charles P Friedman, Patricia Tidmarsh, Craig Mosbaek, and Dale Kraemer. Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions. *J Am Med Inform Assoc*, 9(3):283–293, 2002.

Vahid Jalali and MRM Borujerdi. The effect of using domain specific ontologies in query expansion in medical field. In *Innovations in Information Technology*, pages 277–281. IEEE, December 2008.

Vahid Jalali and Mohammad Reza Matash Borujerdi. Information retrieval with concept-based pseudo-relevance feedback in MEDLINE. *Knowledge and Information Systems*, 29 (1):237–248, July 2010.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, 25(2):7, 2007.

Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology*, 6:343, January 2010.

Zhenyu Liu and Wesley W. Chu. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. *Information Retrieval*, 10(2):173–202, January 2007.

Zhiyong Lu, Won Kim, and W John Wilbur. Evaluation of Query Expansion Using MeSH in PubMed. *Information retrieval*, 12(1):69–80, January 2009.

Gang Luo, Chunqiang Tang, Hao Yang, and Xing Wei. Medsearch: a specialized search engine for medical information retrieval. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 143–152. ACM, 2008.

David Milne, Olena Medelyan, and Ian H Witten. Mining domain-specific thesauri from Wikipedia: A case study. In *Proceedings of the 2006 IEEE/WIC/ACM international conference on web intelligence*, pages 442–448. IEEE Computer Society, 2006.

David N Milne, Ian H Witten, and David M Nichols. A knowledge-based search engine powered by Wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 445–454. ACM, 2007.

Xiangming Mu, Kun Lu, and Hohyon Ryu. Explicitly integrating MeSH thesaurus help into health information retrieval systems: An empirical user study. *Information Processing & Management*, 50(1):24–40, January 2014.

Liqiang Nie, Mohammad Akbari, Tao Li, and Tat-Seng Chua. A joint local-global approach for medical terminology assignment. *MedIR 2014*, page 17, 2014.

João Palotti, Allan Hanbury, and Henning Müller. Exploiting health related features to infer user expertise in the medical domain, 2014.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12: 2825–2830, 2011.

John Powell, Nadia Inglis, Jennifer Ronnie, and Shirley Large. The characteristics and motivations of online health information seekers: cross-sectional survey and qualitative interview study. *Journal of Medical Internet Research*, 13(1), 2011.

Filip Radlinski and Nick Craswell. Optimized interleaving for online retrieval evaluation. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 245–254. ACM, 2013.

Filip Radlinski, Madhu Kurup, and Thorsten Joachims. How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 43–52. ACM, 2008.

Amanda Spink, Yin Yang, Jim Jansen, Pirrko Nykanen, Daniel P Lorence, Seda Ozmutlu, and H Cenk Ozmutlu. A study of medical and health queries to web search engines. *Health Information & Libraries Journal*, 21(1):44–51, 2004.

Isabelle Stanton, Samuel Ieong, and Nina Mishra. Circumlocution in diagnostic medical queries. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 133–142. ACM, 2014.

Karthik Subbian and Prem Melville. Supervised rank aggregation for predicting influencers in twitter. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 661–665. IEEE, 2011.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A large ontology from Wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203–217, 2008.

Elaine G Toms and Celeste Latter. How consumers search for health information. *Health Informatics Journal*, 13(3):223–235, 2007.

Ryen W. White, Susan Dumais, and Jaime Teevan. How medical expertise influences web search interaction. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 791–792, New York, NY, USA, 2008. ACM.

Yang Xu, Fan Ding, and Bin Wang. Entity-based query reformulation using Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*,

CIKM '08, pages 1441–1442, New York, NY, USA, 2008. ACM.
Andrew Yates and Nazli Goharian. ADRTrace: Detecting Expected and Unexpected Adverse Drug Reactions from User Reviews on Social Media Sites. In *Proceedings of the 35th European conference on Advances in Information Retrieval (ECIR'13)*, 2013.
Andrew Yates, Nazli Goharian, and Ophir Frieder. Relevance-ranked domain-specific synonym discovery. In *Advances in Information Retrieval*, pages 124–135. Springer, 2014.
Elad Yom-Tov and Evgeniy Gabrilovich. Postmarket drug surveillance without trial costs: Discovery of adverse drug reactions through large-scale analysis of web search queries. *Journal of medical Internet research*, 15(6):e124, 2013.
Elad Yom-Tov, Shai Fine, David Carmel, and Adam Darlow. Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In *Proceedings of SIGIR '05*, pages 512–519. ACM, 2005.
H Peyton Young and Arthur Levenglick. A consistent extension of condorcet's election principle. *SIAM Journal on Applied Mathematics*, 35(2):285–300, 1978.
Qing T Zeng, Sandra Kogan, Robert M Plovnick, Jonathan Crowell, Eve-Marie Lacroix, and Robert A Greenes. Positive attitudes and failed queries: an exploration of the conundrums of consumer health information retrieval. *International journal of medical informatics*, 73 (1):45–55, 2004.
Qing T Zeng, Tony Tse, Guy Divita, Alla Keselman, Jon Crowell, and Allen C Browne. Exploring lexical forms: first-generation consumer health vocabularies. In *AMIA Annual Symposium*, 2006.
Kathryn Zickuhr. Who's not online and why. `http://www.pewinternet.org/2013/09/25/whos-not-online-and-why-2/`, 2013.
Guido Zuccon, Bevan Koopman, and João Palotti. Diagnose this if you can. In Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr, editors, *Advances in Information Retrieval*, volume 9022 of *Lecture Notes in Computer Science*, pages 562–567. Springer International Publishing, 2015.

A List of Unclarified Queries

The following list contains the 50 queries sampled out of the 500 most popular health queries that were used to evaluate the impact of query clarification. The set of questions associated with each query, the clarified queries, and the set of user interaction reports are available at `https://github.com/Georgetown-IR-Lab/query-clarification-data`.

acid reflux	acid reflux symptoms
back problems	bloated stomach
blood clot	bloody stools in adults
body odor	brown vaginal discharge
can't sleep	common cold symptoms
difficulty breathing	double vision causes
dropsy disease	erectile dysfunction remedies
excessive burping	excessive sweating
fear of heights	foods cause gout
foods to avoid with acid reflux	graves disease
hair loss in women causes	hairloss
heat stroke	hives
how to lose weight	how to stop a nosebleed
indigestion	indigestion symptoms
kidney failure symptoms	leg blood clot symptoms
memory loss	memory loss in women

nervousness	nosebleed
pressure ulcers	profuse sweating
ringing ears	salivary gland stones
shaking hands	slow heart rate
spontaneous abortion	stress incontinence
suicidal thoughts	sunlight causing hives
sweating sickness	sweet sweat
tooth ache	trouble swallowing
trouble swallowing when eating	what causes hair loss in women