11 Finding and understanding medical information online

The internet provides a wealth of information brought together by millions of people. Searching for information is one of the most frequent online activities. As a result, search engines have become tremendously efficient in gathering more information and presenting it quickly and efficiently. The information is available at the click of a button and presented in bite-size text snippets.

There are two potential problems with today’s use of search engines that have an especially disturbing impact in health care. First of all, not all information is trustworthy, much is incomplete, biased or subjective, and finding a complete answer in response to a query is not facilitated by the search engine interface or the results presentation. Second, searching objectively is difficult. Only a few words are used to search among millions of documents, those are used by the search engine to guess, estimate, or decide what is relevant, and only matching information is returned even if the stated query is erroneous. Users, in turn, are often unaware of missing information. Many, especially younger generations, do not read entire texts but limit themselves to scanning snippets provided in the results list.[1] These limitations are problematic since these search engines inform and influence entire families and communities.

This chapter will review query options for searching and presenting results and how these impact searching for and understanding of medical information by patients, caregivers and medical professionals.

11.1 Search in Medicine – Current State and Future Developments

11.1.1 Introduction

The internet is the main provider of information for many people, and it continues to grow both in content and in reach with an estimated 2.7 billion people now having access.[2] Within this sea of information is a large amount of medical, healthcare or fitness information providing advice, background knowledge and opinions on treatments, diseases and conditions. Unfortunately, not all information is trustworthy, not all sources provide complete comprehensive and objective information, and many readers do not have the necessary skills to understand it. Inadequate health literacy, defined as the “limited ability to obtain, process, and understand basic health information and services needed to make appropriate health decisions and follow instructions for treatment”[3], affects an estimated 89 million people in the United
States alone; this impedes their ability to understand and follow preventative care or treatments.[1] The cost of this lack of health literacy is estimated to be between $50 and $73 billion per year.[4] Comparable results are found internationally. In particular, low health literacy correlated with less understanding of their own condition[5, 6] makes people less likely to search for information[7] and less able to learn from information that they do read.[8]

In the medical field, much attention has been given to understanding information and less to searching for it. However, searching for information is one of the most frequent online activities. We rely on search engines to find most of our information, with an estimated 19.3 billion searches conducted in the United States in October 2013 alone.[9] Non-mobile searching consists of navigational, transactional, but mostly (50–80%) informational queries,[4, 10, 11] while a recent study on European mobile search showed a much higher percentage (60%) of transactional queries.[12] 35% of U.S. adults, representing 72% of internet users, have gone online to research a medical concern, and 77% of those searchers started at a search engine.[13] The majority of medical searches by laymen are focused on finding information on a specific medical problem (55%) or a specific medical treatment (43%).[13] Overall, the quality of the medical information available on the internet is high; however, quality does vary widely with more general queries concerning preventative and social health issues returning lower quality information.[14] Additionally, Web 2.0 and the social web have introduced new challenges to the transparency and disclosure that are required to evaluate the trustworthiness of medical information available on the internet.[15]

There are both dangers and advantages of these developments for patients and providers alike. The growing influence of search engines in how people gather and understand medical information makes it more of a concern that consumers have a limited grasp of healthcare vocabulary,[16] and younger people in particular often do not have the necessary skills to effectively use the internet to find health information.[17] Furthermore, consumer satisfaction with the internet as a source of medical information does not seem to be strongly related to their success in searching it.[18, 19] Furthermore, internet searches can cause unnecessary anxiety about common symptoms, cyberchondria, in individuals with little medical training.[20]

More positively, there is evidence that minorities with internet access are able to engage in information searching to overcome inequality in levels of access to healthcare information.[21] As usage of the internet among laymen has grown, patients have been bringing the information that they have found into their relationship with their medical provider. Although medical professionals continue to be the most sought after and most trusted sources of medical information,[22] the availability of health related information on the internet has fundamentally changed the dynamic between patients and doctors; changing patients from more passive consumers to more active partners.[23] This in turn requires that doctors become more “net-friendly” and be able to counsel these engaged patients to contextualize the information that they bring from the internet and then guide them towards more reliable and useful sources.
of information.[24] While the perceived quality and helpfulness of the patient’s communication with their doctor has no impact on how frequently patients engage in medical information seeking,[25] engaging in searching shows a strong positive impact on demand for healthcare services.[26]

11.1.2 Searching for health and medical information on the internet: needs, gaps and challenges

In the context of the importance of online health information as discussed above, there is a need to understand, support and improve online medical information gathering and facilitate searcher understanding. The design of the search-user interface is a crucial element in achieving this goal.

11.1.2.1 The need for better query expression

In traditional search engines, queries are limited to a few keywords, often guided by auto-completion suggestions provided by the search engine. Hitwise reports that in the Spring of 2012 about 50% of US searches were only two words long and a little over 80% of searches were four words or fewer.[27] This limits the expressiveness and precision of queries as well as the diversity of information returned. Furthermore, the traditional approach to search interfaces forces users to search for matching (confirming) information, while ignoring alternative and contradictory information. For example, the query “salty diet autism” would retrieve documents on the causal relationship between salt and autism, but would ignore documents discussing the age of the parents as a potential causal factor.

Query expansion, both interactive (IQE) and automatic (AQE) is a proposed solution to the limited amount of information users supply to the search engine. Although useful, most people do not like these automated methods[4] and thus many researchers try to coach more terms from them directly.[2, 11–13] IQE has been shown to be effective only if the user is making good query expansion decisions, which most users are unable to do.[28] Most IQE interfaces do not provide the user with enough context to enable them to make good decisions about the query expansion being offered. Without context, users tend to react conservatively and not choose expansion terms that they don’t fully understand.[29] In health informatics, this problem is compounded by the consumers’ limited grasp of medical vocabulary.[16]

The application of semantic search techniques is a cornerstone technology in many efforts to improve the user’s query and thus both the accuracy and completeness of the information being returned by the search engine. The term “semantic search” is quite broad; it most generally means that the search engine is aware of some or multiple aspects of the context of a search and is able to use that context to increase the accuracy of the results. In many ways semantic search can be thought of
a subset of the query expansion problem in which the search engine uses an ontology or other outside information to either modify the keywords in a search, expand or modify the search, or directly find documents using an ontological index.

Mangold[30] lays out several criteria for classifying semantic search engines’ architecture: is it a stand-alone system or a meta search engine query modification?; how does the method the system uses change the user’s query and ontology technology?; and how is the ontological data stored and represented internally? Another criterion is coupling, or how explicitly the concepts in the source documentation must be tied to the ontology. Tightly coupled systems require that the source documents be semantically annotated so that it becomes very easy to resolve homonymies. Loosely coupled systems do not require semantic annotation but this limits the search capabilities of the system. An additional important criterion is transparency or how much the user is required to interact with the ontology. A transparent system does not expose the ontological categories to the user whereas a hybrid or interactive system might require the user to validate or clarify the semantic tagging of the query or the results.

Regardless of the quality of terms, query expansion has a dramatic effect on the output of a search engine. For example, with Google, the number of distinct domains in the results as well as the ranking of results and domains changes significantly with the addition of suggested terms. Depending on the types of terms, expanded queries can become more specific or broader. A comparison of the social method and content-based method of query expansion, both using Google data, which involves expanding user’s queries with keywords from other successful searchers (social) or with words from the underlying available text corpus itself (content), shows that the social approach returns a smaller and significantly fewer number of distinct domains. [31] This suggests that the popular query suggestion tools may be narrowing the information being provided to the user.

11.1.2.2 The need for better information presentation
Search result presentation has changed little over the last few decades. Early improvements included the use of text snippets, highlighting matching keywords in the snippets and, more recently, clearer distinctions between sponsored and other results. Most search engines limit themselves to providing a listing of results with highlighting in the text snippets of matching words. A few exceptions exist in which results are further processed to help organize and digest the information. For example, some search engines cluster results. Yippy.com (formerly clusty.com) shows a listing of results and the labeled clusters can be used to focus on subcategories.

This uniform list-like format for reporting search results provides very little context for the results being returned. The search engine interface is not giving the user any feedback as to the appropriateness of her query in matching documents, quality of returned documents or missing information. In particular, no information
is presented on documents that contradict the query or suggestions of topics that are related to the query but would lead to an alternate result set. The cyberchondria phenomenon, which is the unfounded escalation of concerns over a common symptom after reading search results and material from the web, underscores the importance of providing context for medical searches.[20] Specific information given to those with little medical expertise can lead them to form incorrect conclusions as to their condition or the conditions of others.

11.1.2.3 Modern search challenge
The modern search engine needs to address these deficiencies both in query formation and in the display of results. It needs to provide a user interface that encourages users to provide more information to the search engine, and the output of the search engine needs to communicate both specific results and the context of those results. The interface should help the searcher gain a broad and complete understanding of the topic in which they are searching in addition to finding the specific answer they are looking for. It is important that this context be communicated in a visual way, as opposed to a purely textual representation, as searchers are no longer reading long sections of text; instead, they are making judgments based on small snippets.[10] It is also important that the search engine take a more active and intelligent role in filtering out misinformation and assisting the searcher in reasoning with the information that is returned.

With the rapid and increasing penetration of mobile platforms it becomes vital that the search interface be appropriate for the mobile experience; in particular, this means that the interface is even less dependent on keyboard usage and that it allows the query to be manipulated through touching or dragging. The interface must enable much more interactivity between the query and the search results all while remaining usable on the smaller screens of mobile devices.

11.1.3 Proposed solutions
There is a great deal of energy focused on addressing the above challenges. The efforts range from fairly narrow solutions aimed at improving information access for patients and caregivers to ambitious question-answering systems designed to enhance the ability of experts to reason through the ever-growing amounts of medical information.

11.1.3.1 Patient centric solutions
There are a number of efforts, many of which are in the commercial space, which have the explicit goal of helping patients and caregivers access, organize and understand medical information. They address many of the issues discussed above, particularly
problems of presenting only relevant medical information, by being very narrow in their scope and/or employing social networking techniques to enable users to work together to discover, filter and aggregate applicable content. Cake Health is focused on helping patients manage insurance and payment of their health care and alerting them to available health services that they are eligible for.[32] Meddik is a social networking startup that allows patients to organize a health timeline and then collaborate with one another around shared symptoms.[33] Smart Patients is another social networking startup that seeks to connect patients to each other and to information about medical trials. It includes a clinical trial search engine that allows patients to search for relevant clinical trials and then start discussions with others patients about the trial.[34]

One of the largest patient-oriented services is Microsoft’s Healthvault. It is a service designed to allow individuals to centralize all of their health information electronically and then share it with caregivers and medical professionals. Microsoft has put significant effort into making Healthvault the center of an electronic health record ecosystem. It launched in 2007 with partnerships that include the American Heart Association, Johnson & Johnson and the Mayo Clinic.[35] Microsoft has continued to expand this ecosystem as various vendors, such as Fitbit, introduce internet-enabled medical devices such as pedometers, scales, blood pressure monitors and blood glucose monitors. Spil and Klein[36] express doubts in the model and predict that Healthvault will succumb to the same factors that led to the failure of Google Health: primarily a lack of relevance to consumers, a lack of consumer trust and a perception of high risk.

11.1.3.2 General search engines

The general web search engines, in particular Google and Bing, continue to make efforts to improve the relevance accuracy of their results; this has an effect on medical information-seeking for many users. Bing acquired semantic search startup Powerset in 2007 and in 2009 announced integration of a semantically indexed version of Wikipedia and the ability to search using full questions.[37] In 2012, Google announced their Knowledge Graph which, as of 2012, was a 500 million object, 3.5 billion fact ontology.[38, 39] When Google announced the Hummingbird search algorithm, they implied that Google had begun to algorithmically expand the Knowledge Graph with web content in order to optimize the search engine for session-based conversational search.[40]

Both Bing and Google are using transparent semantic search, meaning that the user interacts with the system just as they would a typical keyword-based search engine; all of the semantic searching then occurs with no further user interaction, and the user is unaware of the ontology or ontologies being utilized. Neither Bing nor Google release detailed descriptions of the technology driving their respective search engines. This, as well as other confounding factors, makes it difficult to determine
how these changes are impacting general medical search. However, as these search engines are the most common starting point for search on the internet, even small changes can potentially have a large impact on medical search on the web.

11.1.3.3 Research IQ
There is significant ongoing work in academia to capitalize on the wealth of medical information now available in order to assist doctors and researchers. Research Integrative Query (Research-IQ), being developed at Ohio State University, is an attempt to construct a semantic search system that will allow non-technical medical practitioners to harness and reason with the large heterogeneous data sources that are now available. As described in 2011,[41] Research-IQ can be classified as a tightly coupled interactive semantic search system. An experienced knowledge engineer modeled the data dictionary from the Osteoarthritis Initiative (OAI) as a Unified Modeling Language (UML) class diagram. The data dictionary was then mapped to SNOMED-CT concepts and the resulting semantically annotated dictionary was modeled in the Apelon Distributed Terminology System. The Research-IQ system allows users to put in plain text queries that it then algorithmically annotates using MetaMap and presents the annotations to the users. The user then selects the valid annotations, at which point Research-IQ will return a list of OAI dictionary elements that were mapped to the user selected annotations. An evaluation of the relevance of the returned OAI elements by two subject matter experts showed that the majority of the returned elements were considered relevant to the expert's queries although the relevance was uneven.[42]

11.1.3.4 IBM Watson
In 2011 Watson, a question and answer system created by IBM, defeated the two highest scoring players of all time in the game show Jeopardy! Soon after, IBM announced its intention to leverage the technology developed to create Watson in order to assist medical doctors in diagnosing and treating patients. In March of 2014, IBM and the New York Genome Center announced a partnership to test a prototype of Watson designed for genomic research to assist oncologists in treating cancer patients.[43]

The system underlying Watson, DeepQA, is an abductive reasoning system. It is the essential process of the creating a hypothesis that would explain the observed facts if it is true.[44] DeepQA uses various natural language processing (NLP) techniques to create a hypothesis from a corpus of answer sources and then uses hundreds of NLP algorithms to either support or refute each hypothesis using a variety of content sources. The evidence is weighed using an ensemble of machine-learning algorithms that then present a ranked set of hypotheses to the user. A claimed advantage of Watson is that it is designed to do automated semantic reasoning on natural language content giving it the capabilities of a tightly coupled semantic search engine without requiring semantic annotation of the source documents. Another unique
ability of DeepQA is that it can alert users to facts that are missing from the current case which, if they were present, would allow the system to better diagnose the symptoms in question.[45]

As of 2012, many modifications have been made and evaluated so that Watson would work more effectively in the medical environment. These adaptations include adding a number of medical content sources, including standard medical texts such as the Merck Manual of Diagnosis and Therapy and the Medical Knowledge Self-Assessment Program (MKSAP); they were analyzed for medical concepts and semantic type using the Unified Medical Language System (UMLS). There were also numerous functional adaptations including adding the Medical Subject Heading (MeSH) system from U.S National Library of Medicine and SNOMED medical taxonomies from the International Health Terminology Standards Development Organization (IHTSDO), creating a rule-based annotator to recognize measurements and map them to concepts; this changed the way that Watson extracted supporting passages from the medical texts. DeepQA was trained on a random selection of 1322 questions from the American College of Physicians (ACP) Doctor's Dilemma competition. The system was then evaluated on 188 unseen Doctor's Dilemma questions. It selected the correct answer with an accuracy of 50.5% and the correct answer was in the top ten suggestions 77% of the time.[45]

11.1.3.5 Diagram-based search
Research on search engines and improving search and the understanding of results continues. We are proposing a search engine that leverages a new diagram-based interface for both search and depicting results. A basic node of the diagram consists of a box containing a word or short phrase. Two nodes can be linked together and the relationship between them described with a third term. In this fashion, each node can be related to one or more other nodes in the diagram. The diagram is a visual way of describing the relationship between multiple subject, predicate and object entities. When displaying results, the interface provides a list of documents from the corpus that best match the query, and it will modify the diagram by suggesting additional terms that the corpus suggests are relevant to the search near some nodes in the diagram. Additionally, the interface will thicken and/or color diagram lines to indicate the strength of the relationship between nodes as suggested by the corpus. The interface will also allow the user to highlight one or more of the nodes and dynamically change the ranking of the list of results being returned.

11.1.3.5.1 Backend Index
The diagram-based search interface allows the user to describe relationships between terms – but to be effective, the interface requires a document index that contains relationships extracted from documents, not just keywords. This index is built by a
triple parser that parses documents and then uses a combination of Support Vector Machines and Finite State Automata to identify the meaningful triples in the documents. This parser is able to achieve more than 91% precision and recall against a gold standard when parsing biomedical abstracts.[46–48]

### 11.1.3.5.2 Query formation

In order to search using the diagram interface, users will be able to touch/click to add a new node or drag additional nodes. As shown in Figure 1, each node can contain a search term, and the links between the nodes are directional and can contain additional search terms. In this example, the user is searching for information on salt as the cause of autism. Queries formed by cancer researchers using query diagrams had significantly higher precision and recall than a keyword search.[48] A pilot study performed with laymen showed that users can easily understand the search diagram interface and that – when using the interface – they include more keywords and relationship than when they perform a keyword search.[49]

![Figure 1: Screenshot of Query Diagram Interface](image)

### 11.2 Discussion

Watson and other sophisticated semantic search tools are aimed at helping professionals reason over the large amounts of medical information available. While there is a great deal of excitement and potential in these systems, proving their efficacy in the clinical setting is ongoing work. At present, systems like Watson or ResearchIQ are targeted almost exclusively at medical professionals; due to challenges in healthcare literacy, these systems would be unsuitable for many patients and caregivers. It is notable that systems like Watson also have a strong reliance on electronic health records (EMR), as it would be impractical to type in all of a patient’s symptoms and history before using Watson in a clinical setting. EMR continues to struggle with issues of standardization and consistency within the healthcare profession and challenges with consumer perceptions of relevance from laymen. Health Informatics systems designed to resource patients and caregivers tend to have a very strong social web component and often take the form of utilities that assist with the organization of
personal health care data. Systems such as ours attempt to assist patient and patient communities in reasoning over the data available to them in a comprehensible way while avoiding problems such as cyberchondria – but much work remains to be done to prove effectiveness for the general population. The large public search engines continue to improve their ability to provide more sophisticated answers and insights into user queries. However, given both the domain specificity of health informatics and the broad focus of the public search engines, it is reasonable to assume, in the short term, that the public search engines will continue to function primarily as methods of connecting patients and caregivers with social communities and domain-specific systems that address their specific context.

Where information is often incomplete or simply incorrect, systems designed specifically for presenting valid medical knowledge in a contextualized manner are necessary. This factor is the primary difference between systems designed for use by medical professionals in searching professional, vetted literature and broad systems used by laymen for searching broad heterogeneous data sources. Inefficient, incorrect or misguided queries created by users with a limited understanding of the medical field make information quality problems more damaging and difficulty to address. Retrieved health information needs to be trustworthy, and the technology should be helpful in displaying knowledge and guiding users to the information that is going to help them. This is a constantly evolving area that requires ongoing assessment.

Careful consideration must be made regarding the limitations of these technologies especially with the hype and aggressive marketing that often accompanies new medical technology initiatives. One danger is that people will become overly reliant upon the technology. For example, in earlier work, visual tables of contents led users to rely on the visualization while they ignored the underlying text, leading to reduced understanding of the information.[8] Especially with health information, a focus on safety must be maintained. Care must be taken in the presentation of results so that users are easily able to understand what has been displayed. Some of the dangers can be mitigated with proper evaluation, but the proper evaluation of information systems is a difficult, time-consuming and expensive task.

As with any up-and-coming technology, mobile implementation is a major consideration. Fast and convenient mobile access to information is quickly moving from a luxury to a need as consumer expectations change. It is already a key factor in user’s adoption of an information system. The proliferation of low cost mobile devices also represents an opportunity to provide additional healthcare information and resources to historically underserved populations, as minorities and low socioeconomic status groups are able to leverage online resources to compensate for their lack of access to healthcare resources.[21] This underscores the importance of providing a powerful but intuitive search interface for mobile users that enhances their ability to search for and understand medical information.
11.3 Conclusion

Given the correlation between health and health literacy, the ever-increasing amount of health and medical information available on the internet, and the impact of this information on both the patient’s understanding of their own health and the doctor/patient relationship, the importance of medical search has never been greater. And with increasing use and reliance on mobile internet, no solution should be put forward that cannot be used on different types of devices and by different consumers.

The systems discussed in this chapter use a wide variety of techniques to address issues such as poor term selection in queries and the user’s unwillingness to read large sections of text as they search. Through innovations in presentation, query assistance, semantic search, data storage and indexing, visualization and social networking technology, these systems attempt to minimize misunderstanding and misrepresentation and amplify the ability of professionals and laymen to understand and reason with the large amounts of information now available to them. Although much work remains to be done, particularly in evaluating the effectiveness of these systems, medical search technology continues to make strides towards being both easier to use and providing a richer context. Here patients can increase their health literacy and become more empowered healthcare consumers, and medical professionals can more effectively leverage the ever-growing body of research and professional literature in treating patients and moving the medical field forward.

References


[33] Learn and share with people who’ve been there | Meddik [Internet]. Meddik. [cited 2014 May 1]. Available from: https://www.meddik.com/#
[34] Smart Patients [Internet]. Smart Patients. [cited 2014 May 1]. Available from: https://www.smartpatients.com/
