CHAPTER

Online Health Information Retrieval by Consumers and the Challenge of Personal Genomics

Mark S. Boguski

INTRODUCTION

In an age of personalized medicine, nothing represents the zeitgeist more than individual consumers using the Internet and World Wide Web to seek medical and health information. According to surveys by the Jupiter organization and Harris Interactive, 71% of people who use the Internet also used it to seek health information in 2007 and this percentage, which represents an estimated 160 million people in the United States, had increased by 37% since 2005 (Anonymous 2007; Levy, 2007). Consumers perform health information search (HIS) and retrieval (HIR) for themselves as well as for friends and family. Studies have shown that most of these consumers do not later discuss the information with a health-care provider and that, for many people, the Internet may be the primary or even sole source of health information. (Fox and Fallows, 2003) Seventy percent of people who obtain health information online say that it has influenced a decision about their treatment. Clearly, it is important for health care professionals to understand how their patients find health information and the pitfalls associated with this activity. Indeed, given the challenges that consumers face in obtaining quality health care information from Internet sources, health-care providers will increasingly be in a position where they have to act as reviewers of information and as educators of patients who bring this information to their attention.

CHARACTERISTICS OF CONSUMER SEARCHES FOR HEALTH INFORMATION

A 2007 study by the Pew Internet & American Life Project (Fox, 2006) showed that the top 10 reasons American adults search online for health information are:

- a specific disease or medical condition;
- a particular medical treatment or procedure;
- diet, nutrition and supplements;
- exercise or fitness;
- prescription and non-prescription drugs;
- a particular doctor or hospital;
- health insurance, alternative medicine or treatments;
- mental health issues and environmental health hazards.

Zeng and Tse (Zeng and Tse, 2006) reported that, in performing searches, consumers use query terms that consist of "everyday language, technical terms (with or without knowledge of..."
the underlying concepts) and various explanatory models, all influenced by psychosocial and cultural variations..." According to Lorenz and Spink (Lorenz and Spink, 2004), lay terminology is only partly effective in retrieving useful health information and often produces irrelevant or misleading information because, as reported by Zeng et al. (Zeng et al., 2006), "... the terms and concepts used by consumers often do not accurately reflect their information needs and therefore do not constitute effective queries."

Problems with such consumer health vocabularies may occur at various levels (Zeng et al., 2002):

- **Lexical (form-level)** mismatches (e.g., misspellings)
- **Semantic misunderstanding** (e.g., incomplete or misinterpretation of abbreviations or acronyms, over-generalization, redundancy)
- **Misleading mental models** are conceptual models that consumers employ either incorrectly and/or in a manner different than health care professionals.

Form-level (lexical) mismatches need no further explanation, as examples are easy to imagine. In the case of semantic mismatches, searches with general search engines may yield a large amount of irrelevant information. Consider the abbreviation "MS", which a consumer might use to mean multiple sclerosis but could also be interpreted to mean Microsoft mass spectrometry or the state of Mississippi (Romacker et al., 2006). In another example, "CHF" would be interpreted as congestive heart failure by a physician whereas an international banker would interpret CHF as a unit of currency, the Swiss franc.

Other groups (Tse and Soergel, 2003) have similarly characterized problems with consumer health vocabularies on multiple levels, namely:

- **Shared forms/different concepts** (e.g., the phrase "the results of the [diagnostic] test were negative" could mean "unfavorable" in the mind of a patient but is usually a positive step forward in the mind of a physician.)
- **Different forms/shared concepts** (e.g., "blood cancer" is a lay term that a physician might refer to as a "hematologic malignancy." Likewise when a physician refers to "metastatic disease" the patient would say "the cancer has spread.")
- **Different forms/different concepts** (e.g., "Miracle cure" rather than a "statistically unlikely" or "idiopathic" remission of a disease, emergence from coma, etc.).

The most challenging HIS/HIR problems of all undoubtedly arise at the conceptual level because it is intimately tied to an individual's level of health literacy (McCray, 2008). Concept-level mismatches might include notions outside the framework of mainstream medicine such as homeopathy, acupuncture and other alternative medicine concepts. Another example is when a patient may focus on a body part (anatomical location) or symptoms when a clinician might view the same problem in terms of pathophysiology or from a sub-speciality perspective. Even mainstream medical concepts are subject to significant cultural influences (Payer, 1996).

There have been a number of attempts and exploratory studies to map lay terminology to controlled vocabularies in a number of formal coding systems and comprehensive collections of medical terms such as CTHS/SNOMED, MeSH, ICD-9 or the UMLS Metathesaurus (Shortliffe et al., 2003). Furthermore, ontologies have been employed to perform semantic expansion on queries that may be too general or simplistic (see Spasic et al., 2005) and below. The idea is to take queries consisting of consumer health terms and to translate or reformulate them into professional terms and qualifiers with the aim of improving query meaning, precision and recall. So far such experiments have met with only limited success (Pownick and Zeng, 2004; Zeng et al., 2006). There is also the challenge of “back translating” from professional to consumer language the results of a search using a reverse medical dictionary.

Two additional reasons for suboptimal performance of consumer HIS/HIR is short query length and low complexity: Consumer HIS/HIR queries tend to be very short and too general to be effective. Indeed one study found that 63% of consumer queries consisted of only a single word and only 10% contained more than two words (Zeng et al., 2002). Also, consumers are generally unaware of stop words that add little or nothing to search performance. Thus, limited knowledge of both medical vocabulary and query string search principles contributes to the construction of simplistic, ineffective queries. Another important aspect is the inability or unwillingness to construct complex queries (such as Boolean combinations of terms) even when such “advanced search” options are provided via simple check boxes. Users seem to prefer (or at least have become accustomed to) search and retrieval as an iterative process during which their goals are refined, focused or revised again and again (Fredin and Prabu, 1998) until either a satisfactory answer is obtained or they abandon the effort altogether due to frustration or fatigue. A precise, unambiguous and well-formed query would ideally produce the answer(s) one is looking for in the first iteration. This ideal assumes that the information sources containing the information exist and are accessible.

**WHAT AND WHERE ARE CONSUMERS SEARCHING?**

Equally important to how searches are formulated by consumers are the sources of the information that are being searched. As an editorial in the British Medical Journal put it: "...the internet has vastly increased the availability of information, but often what it offers is unattained, incomplete, irrelevant, and plain wrong" (Jones, 2003). Online health information is problematic as it encompasses everything from evidence-based, peer-reviewed content in professional journals to advertiser-sponsored content and personal testimonials (Table 21.1). A consumer’s online HIS/HIR experience may even be greatly influenced by non-substantive characteristics of a site such as an attractive and professional-looking appearance even though the quality of the underlying information is unknown.

, although actual cases of harm have been difficult to document (Crocco et al., 2002).
Sixty-six percent of consumer HIS/HIR begin with a general search engine such as Google or Yahoo! (Fox, 2006) and the rest will start with one of a number of emerging or established specialty sites that are surveyed in Table 21.1. These sites fall into four general categories (online libraries, e-magazines, search and social networking) that largely reflect their historical origins and/or stages of development. Although a complete description of the origins and development of these sites is beyond the scope of this chapter, one pioneering organization deserves special mention.

Online medical libraries date back to mid-1960s, when the US National Library of Medicine (NLM) experimented with typewriter machines and satellite communications (Schoolman and Lindberg, 1988). Searches at this time were not interactive but rather conducted by “batch processing.” In the early 1970s, interactive searches of NLM's MEDLINE became possible through dedicated remote terminals in 10 regional and 14 large academic medical libraries (Atlas, 2000). During this era, users of these systems were trained librarians who learned specific query commands and retrieval protocols. It was not until after the advent of personal computers and satellite communications (Schoolman and Lindberg, 1988) that online search of the medical literature became accessible to consumers. In 1986 a computer application named “Grateful Med,” that ran on IBM-compatible DOS computers was produced by NLM, was distributed at a cost of $9.95 through the National Technical Information Service of the US Department of Commerce (Schoolman and Lindberg, 1988). Using Grateful Med and a 300 bits-per-second modem, users could use dial-up telephone lines to NLM computers and receive literature citations. The current descendents of these pioneering efforts are PubMed (Wheeler et al., 2007) for the professional literature and MedlinePlus (Miller et al., 2004) for consumer information.

The Dot-com era from 1995 to 2001 (2007) was characterized by experimentation with new sites and services including commercial entities supported by sponsored advertising. Compared with online “libraries,” which maintained content and user interfaces more oriented toward health care professionals than consumers, Dot-com era sites introduced magazine-type articles and layouts focused on non-professional users, enabled by the development of web browsers such as Netscape Navigator (later named Netscape Navigator (2007)). Somewhat prior to this, sites providing comprehensive (“horizontal”) Internet search were mainly being used by academic researchers and computer professionals (e.g., (Boguski and Ouellette, 1995)). But gradually, at first, search engines of the time (e.g., AltaVista (2007)) began to index content that expanded into the consumer realm. The ascendance of horizontal Internet search by companies like Yahoo! (2007) and Google (2007) has recently led to the emergence of several start-up companies applying a similar approach to deep vertical (specialized) health care content on the web (Table 21.1). The latest experimentation with consumer-oriented, medically-oriented websites involves social-networking, following the example of more horizontal (although demographically differentiated) social-networking communities and sites (e.g., MySpace (2007) and Facebook (2007)).

In July 14, 2007 article in the Wall Street Journal, Borzo reported running a variety of searches on several vertical health sites and noted that many features were not self-evident and required repeated experimentation to uncover. Borzo’s overall conclusion from this limited, non-scientific study was that conducting a useful search requires the consumer to run multiple queries on several sites and to then compare the results and reach their own consensus on the adequacy and usefulness of the retrieved information. These observations are consistent with more formal research (Fox, 2006). Borzo did not examine, however, the underlying sourcing and quality of the information. When one examines

---

**Table 21.1** Comparisons of Selected Online Health Information Sites

<table>
<thead>
<tr>
<th>Libraries Publishers</th>
<th>e-Magazines</th>
<th>Search portals</th>
<th>Social Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content Curated Aggregated</td>
<td>Commissioned Aggregated</td>
<td>Distributed</td>
<td>User-generated</td>
</tr>
<tr>
<td>Audience Healthcare Professionals</td>
<td>General public</td>
<td>General public</td>
<td>General public</td>
</tr>
<tr>
<td>Stage Pre-Internet</td>
<td>Dot-com era</td>
<td>Beta</td>
<td>Pre-alpha or alpha</td>
</tr>
<tr>
<td>Revenue Gov't-sponsored Subscription</td>
<td>Gov't-sponsored Advertising</td>
<td>Advertising</td>
<td>Advertising Services?</td>
</tr>
</tbody>
</table>
this aspect, a perplexing assortment of both public and propri- 
ety information from both consumer-oriented sources as well as 
sources designed for medical professionals is revealed (Table 21.2).

Only a tiny fraction of Internet health information sites 
publish any sourcing and date-stamped information or other 
information quality indicators (Anonymous, 2007). Published 
guidelines for quality assurance and quality control are rare and 
consumers are left to evaluate the quality of the information 
based on whether or not they consider the site to be a trusted 
"brand." Given this situation, and the difficulty non-professionals 
have in constructing effective queries (see previous section), 
the best advice on Internet HIS/HIR that one can give to the 
consumer at the present time is to validate their fi-


<table>
<thead>
<tr>
<th>Site</th>
<th>Primary content</th>
</tr>
</thead>
</table>

*Survey date was November 16, 2007

PERSONALIZED GENOMICS FOR CONSUMERS

Personalized genomics involving DNA polymorphism scans (e.g., see (Weber, 2006)) is currently in a tumultuous gestational 
stage that will ultimately lead to a new way to teach and practice medicine (Childs et al., 2005). Consider the following examples:

- A father uses commodity DNA sequencing technologies and publicly-available medical databases to investigate the 
elusive cause of his daughter’s genetic illness (Maher, 2007).
- A controversial scientist publishes his autobiography which 
is partially based on the complete sequence of his own 
genome (Venter, 2007).
- A Harvard geneticist launches a Personal Genome Project 
to encourage medical altruism and self-knowledge, makes his 
own biological material available for study and encourages 
others to volunteer to do likewise (Church, 2005).
- Several start-up companies (one of which is backed by 
Google) form to provide genotyping services as direct-to-
consumer businesses (Winslow, 2007).
- The cover story of WIRED magazine asserts that “a new 
$1000 DNA test can tell how you’ll live – and die” (Goetz, 
2007).

The ethical, legal and social issues arising in this milieu 
certainly signify a brave new world well beyond the scope of 
this chapter. But what are the medical issues? Kohane and col-
leagues (Kohane et al., 2006) consider the situation “a threat 
to genomic medicine” because the volume and complexity
of genotypes and their statistical, phenotypic associations will undoubtedly lead to a plethora of incidental, “abnormal” findings that will be pursued at great cost but little benefit by the patients and their physicians. A person’s genotype will become an “incidentalome” – analogous to the incidentaloma (Mirilas and Skandalakis, 2002) recognized by a previous generation of physicians but unbelievably more complex. Kohane and colleagues call for several key actions including the creation of information systems for estimating and explaining the risks associated with incidental genomic findings (Kohane et al., 2006). They envisioned such systems as being by medical professionals… in the clinic and at the bedside. However, it is likely that consumer versions will lead the way as patients “surf the web” trying to understand the implications of their genotypes and manage their personal health accordingly. Dr. Google has office hours 24/7.

**SUMMARY AND CONCLUSIONS**

Major challenges exist in both educating and assisting consumers with HIS/HIR queries, assessing the sources and quality of the information they find, and helping them decide how to act on it. Much more work is needed on the theory and practice of health information search and retrieval by consumers as health care systems become more patient-centric and consumers are expected to make informed choices and exert more control over the management of their personal health. Health care professionals need to become familiar with the challenges consumers face. Health care systems will need to devise a mechanism for joint consumer-provider review of information, and decision support for the development of personalized health care. Personal genomes are an extreme example of the challenges we all face.

**REFERENCES**


M.C. and Valle, D. (2005). A science of the individual: ‘s genotype will become an ‘incidentalome’ – analogous to the incidentaloma (Mirilas and Skandalakis, 2002) recognized by a previous generation of physicians but unbelievably more complex. Kohane and colleagues call for several key actions including the creation of information systems for estimating and explaining the risks associated with incidental genomic findings (Kohane et al., 2006). They envisioned such systems as being by medical professionals… in the clinic and at the bedside. However, it is likely that consumer versions will lead the way as patients “surf the web” trying to understand the implications of their genotypes and manage their personal health accordingly. Dr. Google has office hours 24/7.

**TABLE 21.3** Selected Sites of Interest for Personal Genomics*

<table>
<thead>
<tr>
<th>Site</th>
<th>Services Offered</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not-for-Profit</td>
<td>‘a forum for those searching for explanations and the help of the interested community of geneticists, patients, physicians, scientists and family members’</td>
<td>Not applicable</td>
</tr>
<tr>
<td><a href="http://www.MyDaughtersDNA.org">www.MyDaughtersDNA.org</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personal Genome Project</td>
<td>‘focuses on the practical issues of recruiting and informing volunteers… a test bed for personalized medicine and new ways of interfacing with the research subjects’</td>
<td>Not applicable</td>
</tr>
<tr>
<td><a href="http://www.pgen.us">www.pgen.us</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commercial Businesses</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23andMe Personal Genome Service</td>
<td>‘personal insight into ancestry, genealogy, and inherited traits’</td>
<td>Not available</td>
</tr>
<tr>
<td><a href="http://www.23andme.com">www.23andme.com</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Navigenics</td>
<td>‘a personalized genetic analysis that, combined with relevant health and wellness information, enables a far more personalized health strategy for each individual’</td>
<td>$2500</td>
</tr>
<tr>
<td><a href="http://www.navigenics.com">www.navigenics.com</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>decODE genetics</td>
<td>‘Subscribers… can take their genome and examine it in the context of the literature’</td>
<td>$985 (Introductory, promotional price)</td>
</tr>
<tr>
<td><a href="http://www.decode.com">www.decode.com</a></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

custom/porlettes/recordDetails/detailmini.jsp?_nfb=true&_ &ERICCestSearch_SearchValue=7&JFS69971&ERICCestSearch_&SearchType=0&n=0&accno=JFS69971

WIKIPEDIA RESOURCES


Author Queries

AQ1: Please specify whether Anonymous a or b.
AQ2: Only 9 reasons have been listed under the heading “Characteristics of consumer searches for health information”. Please check.
AQ3: Please confirm the citation of Table 21.3.

DELTE CITATION AS INDICATED

insert two new citations in bracketed list as indicated

split current bullet # 7 as indicated