TECHNICAL REPORT
LHNCBC-TR-2004-003

Consumer Health Information Seeking;
A Report to the Board of Scientific Counselors

May 2004

Tony Tse, Ph.D.
Darren Gemoets
Graciela Rosemblat, Ph.D.

Consumer Health Information Seeking
A report to the Board of Scientific Counselors
May 13, 2004

Index

1. Introduction

Former Surgeon General C. Everett Koop once remarked, "Even though nothing can substitute for the expertise of your own doctor, no prescription is more valuable than knowledge" (Pifalo et al., 1997, p. 21). Recent surveys suggest that many Internet-using consumers concur. According to a Pew Internet and American Life Project study (Madden & Rainie, 2003), 66% of the online population surveyed in December 2002 reported having sought "health or medical information" on the Internet, an increase of 59% from March 2000.

However, despite intense online health information-seeking activity among consumers, fundamental questions remain about the nature of their needs, information access (and barriers), and health outcomes — the actual impact of health information on everyday life. As the recent report, *Setting the Public Agenda for Online Health Search* (Greenberg, D'Andrea & Lorence, 2003), asserts: "It is of great importance to understand factors influencing online searches for health information and to develop technical and educational approaches for maximizing quality and benefit of health searches" (p. 1). Some of the "roadblocks, bumpy roads and hazards on the information superhighway" (p. 676) enumerated by Cline & Haynes (2001) include:

- Disparities: e.g., financial, literacy, computer skills;
- Navigational challenges: e.g., information overload, lack of a central organizing principle, searching difficulties, jargon, usability and accessibility; and
- Questionable quality: e.g., lack of peer review, inaccurate/misleading information.

Another potential barrier is health literacy, the ability to understand health-related documents, such as instructions and directions, for making informed decisions about personal health care choices. Davis et al. (1996) estimate that 75% of patients with chronic health conditions have limited health literacy. As *Healthy People 2010* (HHS, 2000) points out, "Health literacy is increasingly vital to help people navigate a complex health system and better manage their own health" (ch. 11, p. 9).

Despite rapidly growing interest, the role of consumer health information and information systems in the evolving health care services environment is not well understood. For example, the "expert patient" concept proposed by England's National Health Service has triggered mixed reactions among health care professionals (Shaw & Baker, 2004). Surveys indicate that some fear that informed patients will require significantly greater resources, while others embrace this initiative as one that better optimizes distribution of resources, including increased patient responsibility and self-efficacy.

However, there has been little research on the proper role of interactive consumer health systems and their effectiveness. As Lindberg and Humphreys (1999) observe, "There is a general move in the U.S. to make every effort to facilitate the public's access to scientific literature — despite a paucity of evidence that such access improves health outcomes" (p. 55). Thus, more investigation is needed on how health information systems might enable patients and improve outcomes while accounting for the concerns of health care providers (e.g., the demand for unrealistic, inappropriate, or unproved treatments). Building better communication and understanding between patients and physicians is clearly one objective. As Taylor (1996) notes, "Information systems do not and cannot replace visits with physicians; they can, however, make such encounters more productive, for both doctor and patient" (p. 5).

While information mediators continue to help people find appropriate materials, many consumers are using disintermediated technologies (e.g., general search engines), thereby bypassing traditional gatekeepers. Without expert guidance, consumers face challenges such as defining information needs: "Health information seekers often must have a fairly precise understanding of what information they are looking for in order to find the results they want" (Greenberg, D'Andrea, & Lorence, 2003; p. 7). Even when information is retrieved, interpreting it and determining its relevance to a particular situation may also be challenging for consumers.

In this report, we first present a multidimensional framework for conceptualizing consumer health information seeking from a user perspective. Because of the complexity of the entire framework, we subsequently limit the scope of the CHIS project to a well-defined subsystem of interest — the interaction between consumers and online health information systems. We then describe evaluation methods for exploring multiple dimensions of this interaction, drawing examples from our recent and ongoing research activities. Next, a plan for evaluating the credibility of the results and integrating the findings, including an example, is discussed. Finally, future directions for the CHIS project are proposed.

2 Project Objectives

The Consumer Health Information Seeking (CHIS) project aims to investigate and support effective interactions between consumers and online health information systems. Understanding and characterizing the types of information-seeking barriers consumers face is a first step in identifying potential solutions. The ultimate project objective is to develop and evaluate potential solutions for assisting consumers in overcoming obstacles to accessing relevant health information — where and when they need it. Specifically, we aim to investigate the multidimensional factors of consumer health information seeking through:

- Research: explore information-seeking needs and barriers;
- Development: build tools to assist health information seekers; and
- Formative Evaluation: measure outcomes resulting from consumer interaction with health information systems.

3 Project Significance

While the National Library of Medicine® (NLM®) has traditionally served health professionals and biomedical researchers, the availability of health information has opened new opportunities to serve consumers (Miller, Lacroix, & Backus, 2000). For example, after MEDLINE® became freely accessible on the Web, nearly a third of the users were found to be members of the lay public (NLM, 1998). Soon afterward, the NLM launched MedlinePlus®, a comprehensive and authoritative consumer health information system, and continues to develop and maintain consumer health information services for the public: "Since 1998, NLM has expanded its mission beyond serving health professionals and researchers to encompass providing high quality electronic health information services for the public" (Lindberg, 2003).

The Lister Hill National Center for Biomedical Communications (LHNCBC) is well positioned to investigate consumer health information seeking. Its active research programs in areas such as natural language processing, medical knowledge representation, health communication, and information systems development provide access to relevant expertise locally. In addition, various informatics tools and consumer health information systems developed and maintained by NLM and LHNCBC are readily available as test-beds.

4 Background

In this section, we first describe the interdisciplinary nature of the Consumer Health Information Seeking (CHIS) project. We then present an idealized conceptual framework for consumer health information seeking that integrates components derived from numerous fields. We conclude by focusing attention on a particular area of interest within the broader framework addressed by our research — human-computer interactions during health information seeking.

4.1 Related Disciplines

Our work in consumer health information seeking draws from a number of disciplines, described below.

Information Science. Our information-seeking model, derived from traditional information science literature, views the information-seeking process as a dynamic and iterative series of cognitive processes, decision points, and physical actions required for satisfying information needs (e.g., Taylor, 1962; Dervin, 1977; Belkin, 1980; and Kuhlthau, 1988). A simplified conceptualization (Figure 1) abstracts several "core" non-linear, dynamic, and iterative information-seeking states common to these models:

- Need: identifying and expressing an information need;
- Access: finding relevant information; and
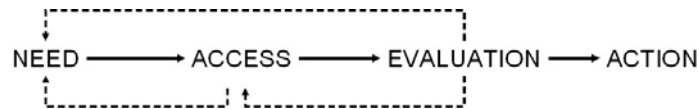- Evaluation: assessing information that is found.

Figure 1. Schematic of information seeking as a dynamic, iterative process

Briefly, information needs represent gaps in knowledge. Obtaining "missing" knowledge is anticipated to help resolve the need. However, consumers may have trouble identifying the missing knowledge (i.e., knowing what is not known), subsequently expressing it as a query, and ultimately recognizing it due to a lack of knowledge about medical concepts and terminologies. Additional challenges include finding credible information resources, constructing well-formed mental representations of a need, and navigating online systems to obtain information. Even then, lack of content coverage, limited system functionality, and devising useful information-seeking strategies may pose problems. The retrieved information must be evaluated within the context of the original problem, another potential barrier for consumers in the medical domain. While retrieved information may help fill a knowledge gap, it may also create new questions (gaps) and/or change the nature of the information need.

Medical Informatics. A goal of medical informatics is "to provide the theoretical and scientific basis for the application of computer and automated information systems to biomedicine and health affairs" (Lindberg, 1987; p. 31). As Friedman (2004) asserts in the Fundamental Theorem of Informatics, well-designed informatics tools and user-friendly information systems may enhance human capabilities for information processing. We use informatics techniques extensively, both to study information seeking and to assist consumers. Our research and development projects take advantage of available informatics tools, methods, standards, and evaluation techniques. As part of the underlying information infrastructure, these technologies typically are not visible to end-users, thereby minimizing cognitive load required for information seeking and allowing consumers to focus on their primary tasks.

Consumer Health. The goal of consumer health activities is to enable individuals to manage their own as well as others' health and wellness (e.g., as caregivers for family members). These activities include providing information for making informed decisions about health care choices and increasing self-efficacy in chronic disease management and health status monitoring (IOM, 2003). As consumers play a more active role in their own health care, they will assume greater responsibilities, such as learning about relevant medical and health care topics. While the

Internet provides an unparalleled opportunity to facilitate communicating and learning about personal health issues, the best way to use this medium is uncertain. The CHIS project investigates how consumers currently access health information online and explores ways to facilitate information seeking.

4.2 Conceptual Framework

Literature from various disciplines suggests that health information seeking consists of many components and their dynamic interactions along multiple dimensions. For example, a new generation of health communication research provides possible models and theories to draw upon. As Logan (in press) notes:

> …both [consumer health informatics and health communication] focus on how communication methods have an impact on consumer decisions. In areas of consumer informatics evaluation where the desired outcomes are shifts in public knowledge, motivations, attitudes and clinical behaviors, adoption of health communication campaign research … might yield more interactive, flexible and multidimensional evaluation tools.

Similarly, Cline and Haynes (2001) suggest that viewing online information seeking as an "interactive health communication" process may yield new insights:

> *We may improve our understanding, investigation, and ability to influence processes of health information seeking on the Internet by framing them as communication processes rather than information dissemination or educational processes.* … the Internet reflects a paradigm shift by offering interactivity and reciprocal influence, pointing toward transactional rather than one-way processes, and blending interpersonal and mass communication processes. (p. 687)

The information science field, which itself draws from diverse disciplines, provides a situated, contextual view of the information seeking. For example, Marchionini's (1995) personal information infrastructure accounts for interactions among several factors:

- Information seeker: e.g., knowledge, experiences, skills, values
- Task: "the manifestation of an information seeker's problem … what drives information-seeking actions" (p. 36)
- Search system, including the user interface
- Domain: field or discipline of knowledge
- Setting: environment in which searching takes place "having physical and conceptual/social components" (p. 46)
- Outcomes

Building on this notion of personal information infrastructures, the online consumer health information seeking framework in Greenberg, D'Andrea, & Lorence (2003), and the Health Information National Trends Survey (HINTS) model (NCI, 2004), we have developed a conceptual framework (Figure 2) for visualizing high-level health information seeking components and their relationships.
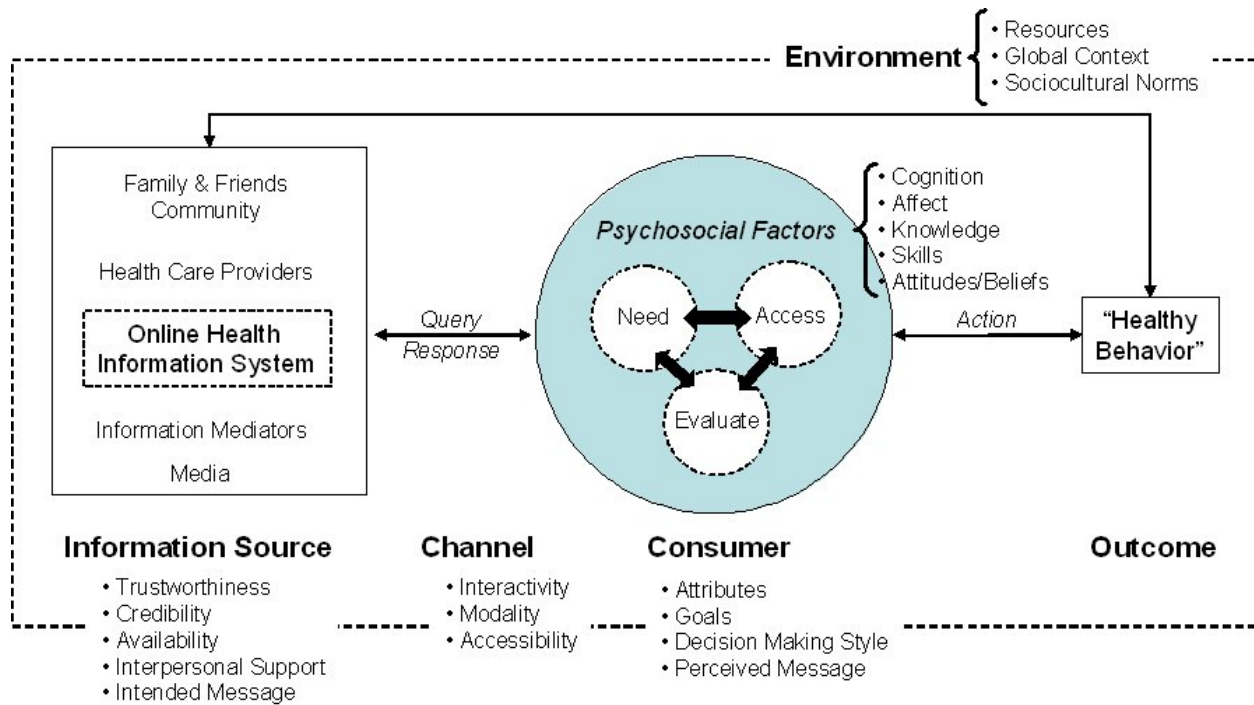
Figure 2. Conceptual framework for consumer health information seeking

While we recognize the complexity of the model and the research challenges posed by multiple interacting components and dimensions, we believe that the framework is useful for presenting a high-level map of areas to be explored in researching consumer health information seeking. As Greenberg and colleagues (2003) state, "the more we know about [online health information seeking] variables, the better we can design educational and technical strategies that help consumers get to the information they seek" (p. 1). We present the larger model as background to frame and situate the subsystem we propose to study in the CHIS project.

Consumer. Psychosocial factors may influence the basic information-seeking process (described previously). For example, an individual consumer's cognitive abilities, affective state, and existing knowledge of the information problem, domain, or information source could interact with the perception of information need, motivation in pursuing information, or amount of effort spent evaluating retrieved information. Further, other variables include demographic attributes, goals, and the perceived message received from the information source(s). For example, a message that is perceived as inconsistent with an individual's personal beliefs (i.e., cognitive dissonance) may contribute to ending an information-seeking session before an information need has been met.

Information Source. Many formal and informal sources of health information are available to consumers. Typically, information sources are characterized by perceptions of trustworthiness, credibility, availability, intrapersonal support provided, and the intended message. For example, during health information seeking, consumers may need to balance perceived quality of a source with available resources.

Channel. The communication channel includes attributes such as the level of interactivity (e.g., unidirectional versus bidirectional), modality (visual, audio, or multimodal), and accessibility (e.g., due to noise or insufficient bandwidth). For example, some consumers may prefer interpersonal communication over mass media, multimedia versus a single modality.

Outcome. Although health information seeking results in many outcomes (e.g., satisfying an information need), a highly significant one that has remained elusive is linking information access directly to better healthy behaviors. As Cline and Haynes (2001) point out:

> Despite abundant speculation regarding the consequences of consumer participation in interactive health communication, little research has investigated these issues… Ultimately interest and research on *effects* should focus on quality of health and health care. (pp. 686-7)

Environment. Overall, the context, such as available resources due to market and economic forces and sociocultural trends influence all of the components (e.g., the "digital divide").

Each of these components and its attributes potentially affect health information seeking. People do not seek health information in a vacuum — it is integrated into their lives. How and when the need to seek information (perceived or real) overcomes other competing needs, interests, and activities may be based on predictable factors, such as risk-benefit analysis, or simple serendipity. Nevertheless, the design of consumer health information systems will likely improve substantially only after we better understand the practical role and nature of health information seeking in people's lives. Our exploratory research begins to address some fundamental aspects of consumer health information seeking.

4.3 Scope of Project

Within the context of the conceptual framework, the Consumer Health Information Seeking (CHIS) project focuses on the role of online interactive technologies (Figure 3).
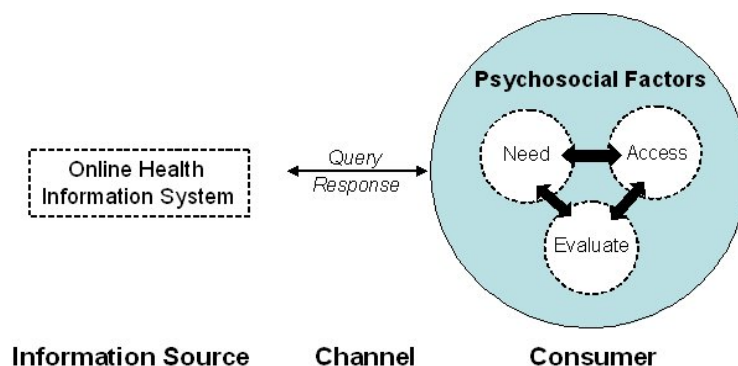
Figure 3. Focus of the Consumer Health Information Seeking project

Specifically, our scope is the interaction between consumers and online health information systems, mediated by various communication channels. We have selected this particular focus not only because it is an increasingly important aspect of information seeking, but it allows us to direct our resources and expertise toward problems that may result in both useful products and theoretical contributions.

5 Methods and Procedures

A brief description of the challenges facing online health information-seeking research is followed by a discussion of the categories of approaches for exploring different aspects of consumer health information seeking. We then illustrate procedures from our recent research.

Research on online information-seeking behavior and outcomes poses many methodological challenges. As illustrated in the conceptual framework (Figure 2), identifying, measuring, and integrating the relationships among numerous complex and dynamic components and their interactions is a daunting task. While focusing on a subsystem (Figure 3) decreases the total number of variables and interactions, other methodological issues still remain. For example, health communication researchers and behavioral scientists convened by the National Cancer Institute and the Robert Wood Johnson Foundation to discuss "eHealth" evaluation issues identified "gaps in research methods" (NCI, 2001) including:

- Difficulty in obtaining representative samples online, including underserved populations
- Challenges in identifying, operationalizing, and isolating key variables, such as "information exposure"
- Lack of techniques for tracking users over time and across space while maintaining confidentiality and respecting privacy
- Lack of standard assessment methods, variables, and operational definitions across research studies

Furthermore, the high degree of variation in consumer strategies, psychosocial factors, and general attributes influence the information-seeking process. For example, Greenberg, D'Andrea, and Lorence (2003) list a number of topics that would "help illuminate consumer behavior (p. 8)," such as:

- Identifying health information needs by population segment
- Understanding how consumers search for information and what health terms they use
- Learning what information is important to consumers and why
- Determining how consumers define "success" (or "satisfaction") during information seeking

These larger methodological challenges not withstanding, our intention is to address the multidimensionality of consumer health information seeking, such as the impact of psychosocial factors on health information seeking online, by using a number of different approaches. Conceptually, we envision the methods as forming a continuum:

- Artifactual Evaluation: data captured during information seeking *in situ*
    - Transaction Log Analysis: e.g., information preserved in Web log files about users' time and length of visit, navigation and search strategies
    - Content Analysis: e.g., coding/classifying queries that retrieve no documents
    - Web Usage Statistics: e.g., number of hits, page views, unique visitors
    - Simulation: e.g., running test queries to assess information retrieval performance

- Structured Assessment: data collected from participants during information seeking or while conducting preplanned tasks using directed questions/instruments
  - Survey/Questionnaire: e.g., intercept surveys while online
  - Structured Interviews
  - Usability Studies
- Direct Observation: data from real-time observation or interaction with participants
  - Field Studies
  - Open-ended Interviews
  - Focus Groups
  - Think Aloud

These evaluation methods complement each other and are mutually reinforcing. Artifactual evaluation gives researchers a window into *what* users did during information seeking in naturalistic settings. Additionally, artifacts collected automatically in machine-readable format (e.g., Web logs) facilitate the analysis of aggregated data. Structured assessments provide insight into *how* users act and feel. Direct observation, including think aloud, diaries, and other introspective methods, delve into motivation – *why* users behavior or feel in particular ways.

To date, we have been exploring ways to understand online health information-seeking behavior primarily through artifactual evaluation. As we begin to study the effects of psychosocial and other factors, we intend to incorporate other approaches. To illustrate some of the methods we have been using, we highlight recent activity in four areas:

- Online search behavior – understanding why query strings submitted to consumer health information systems during online searching failed to retrieve documents;
- Readability – exploring how simple measures of writing style (classic readability formulas) predict reading difficulty of health-related text written for consumers;
- Consumer health vocabulary – setting the foundation for bridging lay–technical terms and concepts; and
- Cross-language information retrieval – exploring ways to allow Spanish-speakers to find information from an English-language consumer health information system.

5.1 Online Search Behavior

We have been investigating online information-seeking behavior through transaction log analysis (TLA) of user actions recorded in anonymized system log files. A type of artifactual evaluation, TLA only provides a part of the picture: *what* searchers did at the site. *Why* they behaved in a certain way may only be inferred. Problems may range from identifying information needs to formulating effective queries to express those needs. Underspecified queries appear to be a common problem: queries submitted to Web search engines consist of 1.7 words on average (Brewer, 2001). Query errors, such as misuse of Boolean operators, use of unsupported operators (when any are used at all), and infrequent use of relevance feedback, were observed in an analysis of log files from Excite, a search engine (Jansen, Spink, & Saracevic, 2000).

Lack of domain knowledge presents an additional barrier. In an analysis of queries submitted to the NLM homepage, McCray et al. (1999) observed misspellings and otherwise "ill-formed" terms, such as ellipses (e.g., *Alzheimer* for *Alzheimer Disease*); truncations (e.g., *cranio*); and "consumer terms" (e.g., *virtual human, visual human,* and *virtual body* for *visible human*). Based on log analysis and patient interviews, Zeng and colleagues (2002) also observed frequent search failures at a hospital's Web site designed for consumers. They concluded that term mismatches occur at the lexical (words or phrases), semantic (meanings or concepts), and mental model (explanations) levels. Lack of domain knowledge also affects the determination of document relevance: in observing 17 consumers seeking health information online, Eysenbach and Kohler (2002) noted, "In some cases, participants continued the search even after finding a page with an answer, generally because they did not understand the information encountered…" (p. 575). Thus, users were unable to grasp the information they found and recognize its relevance to their information-seeking tasks.

Initially, we reviewed a sample of search failures — queries that retrieved no documents — from ClinicalTrials.gov and MedlinePlus to understand the types of problems consumers have in formulating queries (McCray & Tse, 2003). Each failed query represents a single instant during the information-seeking process where a mismatch has occurred between a searcher's expression of an information need (e.g., ill-formed query string) and the system (e.g., information retrieval mechanism, scope, and content). Overall, slightly over a third of the failed queries were "in scope," but otherwise "error free." We interpreted these results to indicate that no matching data in the system were available for these queries (e.g., *menstrual synchrony* in ClinicalTrials.gov and *Jimson Weed* in MedlinePlus). Another result was that we coded only a relatively small number of queries containing "consumer terms." Finally, it is interesting to note the relatively small differences in the distribution of errors between failed queries at ClinicalTrials.gov, a site containing specialized medical information, and MedlinePlus, a general health information site. The resulting analysis provides researchers with a way to classify individual query failures, thereby revealing trends and opportunities for developing system-based interventions.

The failure analysis has helped us prioritize research and development on strategies to support information seeking at ClinicalTrials.gov and has led to a study on the effectiveness of different system-based search algorithms for addressing some of the problems we uncovered (McCray, Ide, Loane, & Tse, in press). We tested the efficacy of these techniques by submitting the original user queries to our new search engine developed in-house. While the number of query failures was reduced, the largest improvement was found in the system's query suggestion capability. We also observed that these techniques are tolerant of certain types of errors and provide feedback to assist users in formulating their queries.

However, isolated queries provide only a limited perspective on information needs and information access: what consumers are looking for and how they go about finding it online. A broader view requires an understanding of psychosocial factors.

We have started to extend our TLA work to include user search moves and navigation patterns at the session level (i.e., user behavior recorded in the log files for an entire information-seeking episode). Because sessions represent a series of search activities (moves) over time, they provide additional context for analyzing online search behavior. Like "footprints," session-level

data may be pieced together to provide information or "narratives" about where users started searching on the site, what they searched for, where they went, and what else they did during a session. Such information will provide insight into how searches evolve over time, tactics that users adapt or abandon, and whether features designed to support user searching are effective.

Our goal is to develop an automated mechanism to capture sequences of search moves by session. These data may then be analyzed as individual sessions or in aggregate to discern common search patterns over multiple sessions. We also intend to explore different methods for representing moves to facilitate session level analyses, such as traditional navigation-based analyses (e.g., Rozic-Hristovsk, Hristovski, Todorovski, 2002), query word cooccurrence frequencies (Wang, Berry, & Yang, 2003), as well as zero-order state transition matrices (Figure 4) and maximal repeating patterns (Wildemuth, 2004). We also plan to integrate TLA data with those derived from usability studies, user interviews, and field studies. We intend to learn more about information seeking for experimental interventions, such as:

- Who searches the site and how may search behavior differ among groups?
- What information is most useful under various circumstances?
- Where may there be barriers to navigation, search, or comprehension?
- How do consumers use the information?
- What do consumers feel about the site (e.g., search experience and information found)?
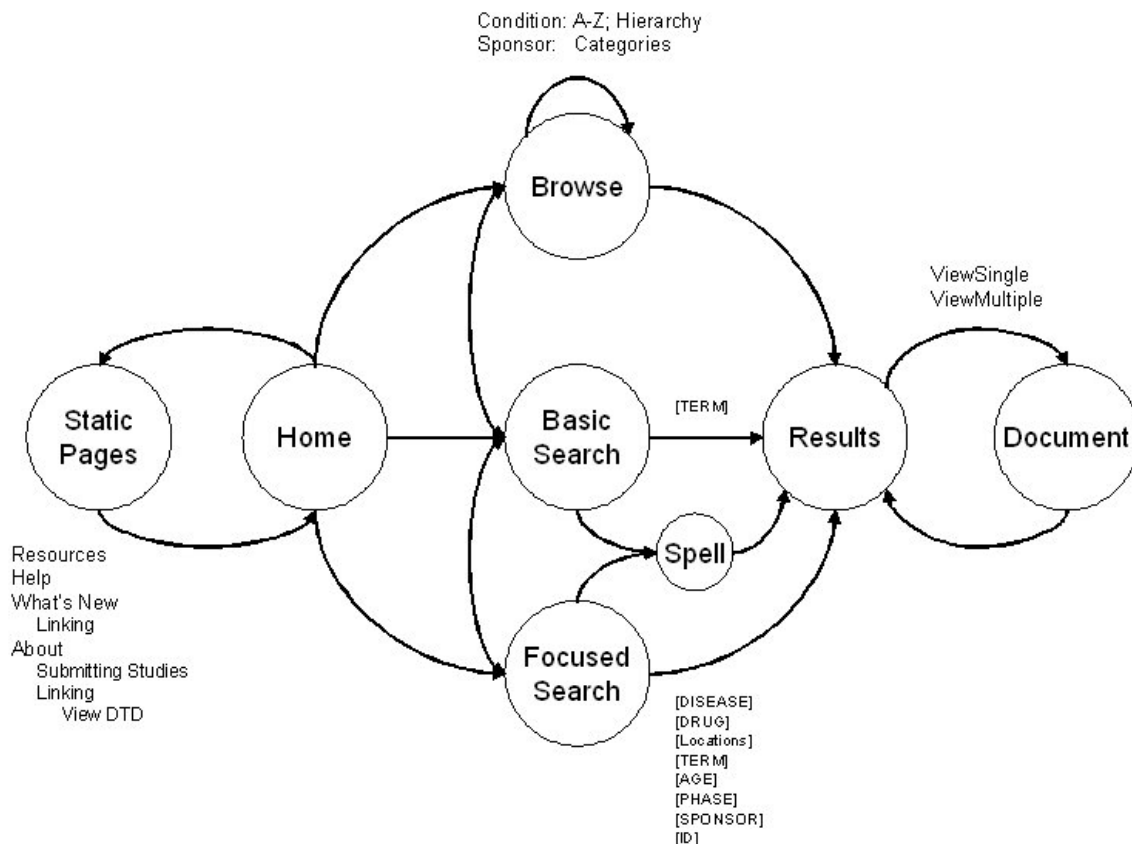

Figure 4. Sample state transition diagram of ClinicalTrials.gov for tracking navigation

5.2 Readability and Health Literacy

Once consumers find health information, they may have difficulty evaluating and comprehending it because of the lack of domain knowledge, health literacy level, text style, vocabulary, prose, or an array of other considerations. Our preliminary research on readability, a traditional measure of text difficulty, used artifactual evaluation and structured assessment methods to look at the baseline effectiveness of classic readability formulas for assessing consumer health text (Gemoets, Rosemblat, Tse, & Logan, in press). In a pilot study, we compared the predicted reading grade level of a small set of documents written by health communicators for consumers, based on two methods: classic readability formulas and the Cloze procedure (Taylor, 1958). The Cloze procedure is a structured assessment – it requires participants to read text with every 5th word removed and to fill in the missing words, based on the gist of the remaining text. Thus, the Cloze procedure starts to address reader-text interactions.

The readability formula, which focuses on particular quantifiable aspects of *writing style*, is an example of artifactual evaluation: reading difficulty is measured as a property of the text — an "artifact" of an author's thoughts. In actuality, the reading process represents a complex interaction among the text, the reader, his/her task, and situational factors:

> …texts are not difficult or easy in and of themselves, but they become difficult or easy at the interface with readers and the purpose of the activity (Snow, 2002; p. 113).

For example, anxious, newly diagnosed patients who have little background knowledge and want to learn about their condition will interact differently with the same text than adult care providers who have experience taking care of family members having the condition.

Historically, readability formulas were designed as "a screening device that provides probability scores for a quality in text … where difficulty of [writing] style overshadows difficulty of content" (Klare, 2000; pp. 151-2). That is, the formulas served as simple "filters" to gauge the likelihood that writing style would not interfere with comprehension for the majority of readers at a particular grade level. They do not, however, measure the likelihood that text will be understood, which requires information about the topic, concepts represented, and, importantly, the reader and her/his purpose for reading the text. As Klare (1963) pointed out in the classic book, *The Measurement of Readability*:

> …formulas measure only one aspect of writing — style. Perhaps certain formulas…approach a measure of content, but they touch on this only indirectly. … Formulas do not touch on organization, word order, format, or imagery in writing: they do not take into account the differing purposes, maturity, and intelligence of readers. …formulas measure only one aspect of style — difficulty. … Thus style difficulty is shown to be only one characteristic of a piece of writing. It is often an important one, however, as shown by its relationship to reading speed, acceptability, understanding, and learning. (pp. 24-5)

The formulas were developed using multiple regression equations and are based on empirical studies. For example, text assessed at the 9th grade level using the Flesch-Kincaid formula indicates that 50% of students in the 9th grade would answer comprehension questions correctly

75% of the time (Ley & Florio, 1996). Used primarily in education, the formulas were calibrated on grade level (Chall & Dale, 1995). Readability formulas rely on two factors:

$$\text{Assessed Reading Grade Level} = K_1 + K_2 (\textbf{Word Factor}) + K_3 (\textbf{Sentence Factor})$$

More recent readability assessments attempt to incorporate other factors, such as prior knowledge required, degree to which difficult terms or concepts are explained, and text layout (Chall & Dale, 1995). While the inclusion of such variables may have face validity, they are difficult to measure objectively. Hence, classic formulas remain popular in practice and research (Doak, Doak, & Root, 1996).

In addition, many systematic reviews of consumer health information and sites include readability components (e.g., Eysenbach, Powell, Kuss, & Sa, 2002; Berland et al., 2001). Despite their widespread use, questions remain about the appropriateness and effectiveness of the classic formulas for health texts, including:

- Applicability: formulas were originally developed for the classroom environment.
- Linguistic Factors:
  o Style: e.g., active/passive voice; first/third person
  o Linguistic structure: e.g., text coherence, discourse structure
  o Extratextual features: e.g., format, layout
- Health Domain Factors (Eysenbach et al, 2002; Roberts, Fletcher, & Fletcher, 1994):
  o Reader's Health Knowledge: e.g., medical knowledge and vocabulary, experience
  o Reader's Psychological State: e.g., motivation, emotional state, bias

Note that the last two factors address the psychosocial dimensions, the contributions of readers to creating meaning from information encoded in text by an author.

Readability formulas are widely used and computationally tractable. Therefore, we chose them to start our investigation on matching health texts with consumers' abilities and needs. The results of this pilot study suggest that classic readability formulas provide a reasonable estimate or "first approximation" of consumer health text difficulty for *highly educated* readers, as corroborated by the Cloze procedure. In addition, we observed differences in readability among texts from different journalistic genres (e.g., feature story vs. press release). We have begun to explore the effects of linguistic factors not accounted for by readability formulas.

We analyzed a number of linguistic factors that may affect comprehension (e.g., Russell-Pinson, 2002; Biesenbach-Lucas, 1995). For example, the most readable texts yielded the following distinctive linguistic factors:

- Short average word and sentence length;
- Frequent use of words associated with hedging devices;
- Fewer relative clauses and prepositional phrases;
- Infrequent use of in-text statistics;
- Frequent use of personal pronouns, but infrequent use of proper nouns; and
- Frequent use of active voice.

These preliminary results suggest that it would be useful to study the suitability of health texts by building upon readability formulas. We have developed an in-house tool, the Readability Analyzer, which implements the most common formulas. We intend to begin by thoroughly assessing linguistic factors associated with representative consumer health texts. We also plan to explore ways to measure various reader-text interactions and have begun experimenting with measures of reader affect and constructed meaning of the author's intended message (e.g., "take home" message).

We also recognize that reader abilities such as health literacy — "...the ability to read and understand materials related to personal health as well as navigate the health system" (HHS, 2000; ch. 11, p. 9) — substantially influence readability and comprehension. The consequences of mismatches between readability and health literacy are substantial. As Rudd and colleagues (1999) point out:

> Patients' literacy directly influences their access to crucial information about their rights and their health care, whether it involves following instructions for care, taking medicine, comprehending disease-related information, or learning about disease prevention and health promotion. Because consent procedures contain complex legal and medical jargon, a patient's literacy may influence his or her opportunities for inclusion in research and exposure to a variety of procedures. (p. 162)

We intend to extend and expand the notion of *readability* and the implementation of *readability formulas* to include significant linguistic and psychosocial factors (e.g., health literacy level of readers) tailored specifically for the consumer health domain. Our purpose is to develop a mechanism to predict the "suitability" of texts for particular audiences who have specific goals. Our definition of suitability not only encompasses readability and comprehension, but also psychosocial factors, such as beliefs, attitudes, and motivation, as well as purpose, such as scanning text to verify a particular fact versus reading for learning, and health literacy. For example, information about the benefits of life-style changes on cardiovascular health may be prescribed to three patients, each with different health literacy levels, needs, preferences, and motivation levels. We would like to provide a mechanism to help select the most appropriate text for each patient to consider the benefits of the recommended changes.

5.3 Consumer Health Vocabulary

We have used artifactual evaluation to identify and extract health terms from documents written by and for consumers (Tse & Soergel, 2003). We also compared automated and manual approaches for mapping such terms to existing concepts in the Unified Medical Language System® (UMLS®) (Divita, Tse, & Roth, in press). This work contributes to an understanding of how consumers and health communicators express health-related concepts. Such information could ultimately be used to help bridge the vocabulary gap between consumers and health care professionals.

Consumers often have difficulty expressing and understanding medical concepts. Further, terms may be interpreted differently by consumers, based on personal and cultural experiences, education level, and cognitive and affective states of mind (Stavri, 2001). Thus, identifying

common consumer expressions and mapping them to professional medical concepts is one way to support consumer health information seeking and comprehension of retrieved information.

Terminology is a bottleneck in consumer health information seeking that extends across all phases of the information-seeking process. Not knowing or not understanding technical terms hinders both information retrieval and making sense of any information that is found. Finding approaches to bridging the gap between professional and lay language presents a considerable challenge:

> In order to be effective, a standardized consumer health vocabulary will need to consist of "normal standard ways of expressing things" (in everyday life), and will also likely need to contain "informal" terminology. (Lewis et al. 2001; p. 1530)

We explored a corpus-based approach to obtain consumer health terms and examined characteristics of words and phrases used by and for consumers to describe health-related concepts. Student participants extracted health terms from anonymized archival postings on public online discussion forums and documents from the mass media, based on personal experience, knowledge, judgment, and context in the documents. Various lexical tools were applied to improve the likelihood of matching terms in the UMLS.

Although substantial "apparent" conceptual overlap was observed, approximately half of the extracted forms did not map to any concepts in the UMLS. These results support the notion that consumers use expressions different from those of domain specialists to represent similar medical concepts. Because the terms were extracted from written documents, the intended meaning could only be inferred from context — a limitation of artifactual evaluation. Other approaches, such as asking consumers to define medical terms using a structured assessment instrument, are needed to validate the "apparent" matches. Note that contrary to the results of our query failure analysis, many consumer terms were extracted. One reason might be the difference in sources of terms: queries and full text. The writing in full text documents tends to be more expressive, describing and explaining health-related concepts, than in queries.

While we have identified and characterized a small set of health-related words and phrases used by consumers manually (with some automated support), scaling up the process will require the use of informatics and natural language processing techniques. MetaMap Transfer or MMTx (Browne, Divita, Aronson, McCray, 2003; Aronson, 2001), a tool that automatically maps terms found in free text to UMLS concepts, is essential for large-scale extraction of terms to build a consumer health vocabulary. Brennan and Aronson (2003) propose to use it for linking patients to high-quality health resources in a dynamic, "just-in-time" manner. Another application is the automated indexing of biomedical research documents (Aronson et al., 2000).

To investigate the less effective performance of MMTx on text from consumer health documents, we compared suggestions from MMTx with human judgment (manual mapping) on two disease summaries from the Genetics Home Reference (GHR) system, a consumer health site. The participants were asked to identify medical terms, loosely defined as "something that would be of interest to a medical professional," and then find the closest matching concept in the UMLS. To ensure that appropriate terms were found, they were encouraged to review definitions, semantic locality information, and hierarchical context of each concept, even when

the surface forms were exact matches. A reference standard was generated based on a combination of adjudication and consensus. The documents were run through MMTx and the top suggestions were compared with terms in the reference standard. A successful match occurred when MMTx identified exactly the same UMLS concept as the participants.

MMTx identified and suggested concepts for over half of the terms identified by humans: 169 of the 316 terms in the reference standard. The terms MMTx missed were not due to lack of concept coverage in the UMLS, as the participants were able to map all but two terms. Of the 13 MMTx mapping failure categories, the most common reasons for failure included:

- Implicit Meaning (30% of all failures): Deep semantic knowledge that is available to humans is not available explicitly in the surface form of the term. For example, *family history* in the context of retinoblastoma was mapped to *FH: Eye disorder NOS* (C0455396) by humans and *Family history of* (C0241889) by MMTx.
- Narrower Term not in the UMLS (13%): Because the exact concept was not present in the UMLS, the term was mapped to a slightly broader concept: *Infantile-onset ascending hereditary spastic paralysis* was mapped to the "nearest" UMLS concept (C0751603) represented by a broader term, *Spastic Paraplegia, Hereditary, Autosomal Recessive.*
- Definitional Phrase: Definitions or explanations of concepts are often included in consumer health text as in the phrase *part of the eye that detects light and color*, mapped by humans to *Retina* (C0035298).

This resulting list of failure categories will be used to improve the algorithm. The next steps are to develop ways to identify, collect, and map terms quasi-automatically from consumer health materials to UMLS.

Knowing the forms used by consumers and how often they map to medical concepts could assist health consumers in (1) reformulating queries and (2) understanding retrieved medical documents. Procedures for automating the collection and mapping of consumer health expressions are prerequisites. The methodology used in this study, while requiring considerable manual involvement, represents a first step toward greater automation.

We are exploring different methods to create an interdisciplinary and flexible data scaffold for identifying, collecting, and maintaining consumer terms. One approach, for example, is to map existing consumer health vocabularies (see Zielstorff, 2003) to professional terminologies within the UMLS, such as:

- MedlinePlus "Go Local" vocabulary terms mapped to MeSH terms (MedlinePlus, 2003)
- Apelon's Consumer Health Vocabulary (Nath, 2003)
- WellMed's Consumer Health Terminology Thesaurus (Marshall, 2000)

Additionally, lay synonyms from general ontologies (e.g., WordNet) and vocabularies (e.g., Taxonomy of Human Services) that include lay words/phrases in the health domain may be mapped to UMLS concepts (e.g., Bodenreider, Burgun, Mitchell, 2003).

We also intend to use the other methodological approaches (i.e., structured assessment; direct observation/introspection) to identify and collect consumer terms used by consumers from a variety of discourse groups. Mapping such terms to reference concepts (e.g., UMLS) will be challenging, as consumers may associate different concepts with a term than professionals do (e.g., *cancer* and *neoplasm*). Thus, "consumer concepts" that map to sets of professional concepts or currently do not exist in the reference set (e.g., UMLS Metathesaurus) may need to be added to the consumer health vocabulary scaffold.

In developing a consumer health vocabulary and linking it to professional terminology, we will learn how to identify words or phrases that refer to "health-related concepts" and gain a preliminary understanding of the mental models or concepts that users associate with those words or phrases. We will also learn more about consumer discourse groups and the situational contexts in which laypersons use a particular term (e.g., *stomachache* in adults but *tummyache* with children). We believe that even a small-scale study for a particular health topic among well-defined discourse groups will yield fundamental insights into the nature and, ultimately, the bridging of the vocabulary gap between laypersons and professionals.

5.4 Cross-Language Information Retrieval

Language is a common barrier to consumer health information seeking. Many English-language Web sites are inaccessible to non-English speaking consumers. At present, manual translation between languages remains the most accurate method. However, it is a resource-intensive process, especially for a rapidly evolving field such as medicine.

We have been exploring automated strategies to allow Spanish-speaking consumers to query and retrieve relevant health information in English. While we recognize that such a cross-language information retrieval (CLIR) mechanism is only a partial solution, CLIR still may be useful for bringing consumers one step closer to satisfying their health information needs. To further bridge the gap between English and Spanish, we are also investigating Spanish language information surrogates, or "doclets," that may be created quasi-automatically using controlled vocabulary. As with abstracts, the information contained in a doclet is intended to help Spanish-speaking consumers judge the relevance of retrieved documents for their needs. To date, our CLIR studies have used artifactual evaluation methods: CLIR has been compared to the information retrieval performance of the English monolingual system. We plan to involve Spanish-speaking participants in future research (e.g., structured assessments).

In an initial study, we compared the performance of two CLIR methods, query translation and document translation, using ClinicalTrials.gov as a test bed (Rosemblat, Gemoets, Browne & Tse, 2003). In the first method, we translated queries from Spanish to English. In the second, all documents were translated and information retrieval was carried out in Spanish. For both, machine translation software was used. Although full-text documents provide more linguistic cues (e.g., part of speech information) than typical queries and, hence, more accurate translations, query translation demonstrated better CLIR performance in our test system. An analysis of the results showed two primary reasons: (1) prevalence of noun phrases among the queries and (2) availability of information retrieval enhancements in English only (e.g., synonym expansion using the UMLS Terminology Server).

A follow-up study (Rosemblat, Tse & Gemoets, in press) compared query translation-based CLIR using machine translation and a bilingual term table developed from various publicly available sources. Given that queries submitted to online systems typically contain few terms and are commonly noun phrases, we hypothesized that a Spanish-English look-up table customized for the consumer health/medical domain would be more effective and efficient than a full-blown machine translation system. In fact, we found that the bilingual term list roughly approximates CLIR performance with machine translation under experimental conditions using the ClinicalTrials.gov test bed. We anticipate that performance will continue to improve as we fine-tune the bilingual term list, add Spanish-English medical term records, and develop a Spanish language lexical variant generator (LVG).

A next step is to investigate the effectiveness of an information surrogate or "doclet" for helping point Spanish-speakers to relevant documents. Since ClinicalTrials.gov records contain many fields with controlled vocabularies, we have manually translated these terms into Spanish. Thus, doclets in Spanish could be created automatically from their full record counterparts in English. Notable exceptions are the free-text fields, including study title, purpose, and eligibility criteria. Currently, we are manually translating the study titles, but not the other free-text fields. We have also built a prototype system where all ClinicalTrials.gov static pages have been translated into Spanish. We intend to use the prototype to conduct structured assessments and observational/introspective evaluations with bilingual and monolingual Spanish-speaking consumers to explore the usefulness of the doclet concept and design, including user satisfaction. Depending on the results, we may consider alternative designs for doclets.

In the long term, a machine translation system tailored to terms in the clinical trial domain may be desirable. Although machine translation is not likely to replace expert human translation in the foreseeable future, especially for highly sensitive health-related information, high quality machine translation could substantially reduce the workload of experienced post-editors. Only then would the goal of providing comparable features and information (e.g., full-text records) in other languages be realized. Such a development process could serve as a model for making other consumer health information systems available in different languages.

6 Evaluation Plan

Because health information seeking is a complex process with multiple variables, it is difficult to evaluate the credibility of study results based on a single method. Thus, we plan to use a number of approaches for each study and look for a convergence of results. That is, data collected from artifactual evaluation, structured assessment, and direct observation/introspection procedures, all pointing toward a common, consistent explanation, will provide independent validation of our hypotheses. We intend to use such evidence both for filling in the details of the conceptual framework and building new tools to assist consumer health information seeking.

To illustrate the notion of data convergence resulting from different approaches, we present the following example.

Classifications are useful for organizing large amounts of information. However, while there are many types of technical medical classifications for various purposes, few are intended explicitly for consumers. One question we might explore is how do consumers categorize different diseases or conditions? That is, what are the different factors consumers use to classify diseases and conditions and how could common associations among laypersons be captured in high-level classification structure(s)? For example, do consumers tend to favor anatomical location (pancreas), physiological dysfunction (glucose metabolism), symptomatology (increased blood glucose), treatment (insulin), some combination, or other factors (e.g., etiology) when placing diseases into categories? The answer would be useful for many applications, from developing browse structures and organizing retrieved documents to designing consumer health information systems, such as portals.

To start, we would use an artifactual approach to explore how large numbers of consumers currently navigate existing disease categories. For example, the Browse by Condition by Disease Heading feature in ClinicalTrials.gov is modeled on a flattened MeSH® vocabulary structure from the "Diseases and Conditions" (C) and "Psychiatry and Psychology" (F) trees. While the high-level category names have been modified for consumers (e.g., "Neoplasms" changed to "Cancers and other Neoplasms"), the names of the lower level nodes have not been modified and are listed alphabetically at the same level (e.g., "Diabetes Mellitus," "Diabetes Mellitus, Insulin-Dependent," and "Diabetes Mellitus, Non-Insulin-Dependent"). Session-level transaction log analysis could be used to analyze sessions where users clicked on the Browse by Condition by Disease Heading feature any time during the session. The manner in which users navigated the browse tree and whether they ultimately requested leaf nodes (i.e., a particular condition or studies listed under that condition) or displayed other types of activity (e.g., frequently moved up and down branches or "jumped" across branches), could provide preliminary clues. Such high level data over thousands of user sessions might point to particular problem areas or trends. Further, correlating query terms used to search the site during the same session as browsing by condition may shed additional light on the disease category being sought.

Next, to understand how consumers conceptualize disease categories, we would try to elicit the characteristics or facets they use most frequently during classification and the relative weights given to each. This kind of open-ended analysis would be suitable for the direct observation/introspection approach. We would use purposive sampling to select participants who represent as wide a range of viewpoints as possible along different dimensions (e.g., demographics, psychosocial factor, and environmental context). Some participants might be interviewed and asked to "free associate" with different disease names, providing descriptors that come to their minds. Others might be asked to draw concept maps to show the relationships among the diseases and other health-related concepts. Still others might be given stacks of cards with disease names and asked to sort them into related clusters.

Using the results of these exploratory studies, we might begin to build disease classification frameworks for consumers. First, we might want to survey a large stratified sample of consumers, using structured assessments. For example, a number of diseases might be listed, each with several category selections in a multiple-choice question format. Survey participants would be asked to select the category they feel is most appropriate – where they would expect the disease to be classified. The results of such a survey might reveal classification differences

among different age groups, for instance. Based on these results, several prototype classifications would be subjected to usability testing, where participants would be provided with specific tasks (e.g., finding specific diseases, given a classification structure). Outcome measures would likely include time for completion, error rate, and user satisfaction. Other participants might be asked to "think aloud" so that their decision making and reasoning might be apparent to the researcher.

Once a classification framework (or set of frameworks) is selected as having provided the best outcomes during testing, an implementation could be deployed. Subsequently, use of the new browse structure would be monitored through TLA and reinvestigated, as necessary.

7 Project Schedule

During the first year of the Consumer Health Information Seeking (CHIS) project, we intend to continue building on the projects described previously. We plan to develop session-level transaction log analysis (TLA) tools, assess linguistic factors associated with the readability of consumer health text, and refine our prototype of a cross-language information retrieval system for ClinicalTrials.gov. We will also collaborate with other researchers to develop methods for building a consumer health vocabulary from artifactual data, such as preferred terms for UMLS concepts based on user warrant.

In the following two years, we plan to address the impact of psychosocial factors and environmental context on consumer health information seeking across the project activities. We would like to apply structured assessment and direct observation/introspection approaches. For example, the TLA tool will provide data on *what* people did to navigate and search ClinicalTrials.gov. An online survey will tell us *who* is using the site, *what* they are looking for, *how* they feel about their online interactions, and *whether* they found useful information (e.g., satisfaction). In addition, if possible, we would like to partner with a major medical center to survey people who inquire about clinical trials asking where they sought information, how they heard about the study, and what motivated them to contact the study recruiters.

Further, we plan to devise structured assessments for measuring different types of interactions between readers at different health literacy levels and text intended for consumers, such as comprehension, affective response, perceived message (versus intended message), and usefulness. Ideally, we could focus on different sample populations controlled for demographics as well as psychosocial factors, such as health literacy and domain knowledge. The goal would be to build models to predict the "suitability" of consumer health texts for particular consumer needs, tasks, and interests. We also intend to evaluate the ClinicalTrials.gov cross-language information retrieval prototype system by means of usability testing methods and focus groups with Spanish-speaking participants. Finally, we would like to validate consumer health terms — assembled from artifactual evidence (e.g., log files and documents) — using actual patients.

8 Summary

We are investigating a number of project areas related to online consumer health information seeking. Learning what information consumers need, how they look for it, and what they do with the information they find are fundamental questions for consumer health informatics research.

As more consumers seek health information online and take an active role in managing their own and their family's health care, understanding and developing a model of consumer health information seeking becomes increasingly important.

While our conceptual framework includes a number of consumer health information-seeking components and interactions, the Consumer Health Information Seeking (CHIS) project focuses on the consumer-online information system interaction along multiple dimensions. We intend to use a number of approaches to investigate the information-search process in the broader context of psychosocial factors, consumer interactions with information through different channels, and, ultimately, health outcomes. Convergence of results from different approaches for a single phenomenon will be used to evaluate the validity of the research.

9 References

Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001;17-21.

Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, Rindflesch TC, Wilbur WJ. The NLM indexing initiative. Proc AMIA Symp. 2000:17-21.

Belkin NJ. Anomalous states of knowledge as a basis for information retrieval. Canadian Journal of Information Science 1980;5:133-143.

Berland GK et al. Health information on the Internet: accessibility, quality, and readability in English and Spanish. JAMA. 2001;285(20):2612-21.

Biesenbach-Lucas S. A comparative genre analysis: the research article and its popularization Unpublished Ph.D. dissertation. Washington, DC: Georgetown University; 1995.

Bodenreider O, Burgun A, Mitchell A. Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: a feasibility study. Medical Informatics Europe (MIE) 2003:379-84.

Brennan PF, Aronson AR. Towards linking patients and clinical information: detecting UMLS concepts in e-mail. J Biomed Inform. 2003;36(4-5):334-41.

Brewer EA. The consumer side of search. Communications of the ACM 2001;45(9):41.

Browne AC, Divita G, Aronson AR, McCray AT. UMLS language and vocabulary tools. Proc AMIA Symp. 2003:798.

Chall JS, Dale E. Readability revisited: the new Dale-Chall readability formula. Cambridge, MA: Brookline Books:1995

Cline RJ, Haynes KM. Consumer health information seeking on the Internet: the state of the art. Health Educ Res. 2001 Dec;16(6):671-92.

Davis TC, Meldrum H, Tippy PKP, Weiss BD, Williams MV. How poor literacy leads to poor health care. Patient Care 1996:94-127.

Dervin B. Useful theory for librarianship: Communication, not information. Drexel Library Quarterly 1977;13(3):16-32.

Divita G, Tse T, Roth L. Failure analysis of MetaMap Transfer (MMTx). Medinfo '04: in press.

Doak CC, Doak LG, Root JH. Teaching patients with low literacy skills. 2nd ed. Philadelphia: Lippincott Williams & Wilkins; 1996.

Eysenbach G, Kohler C. 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. 2002 Mar 9;324(7337):573-7.

Eysenbach G, Powell J, Kuss O, Sa ER. Empirical studies assessing the quality of health information for consumers on the World Wide Web: a systematic review. JAMA. 2002 May;287(20):2691-700.

Friedman, CP. Evaluation Methods in Medical Informatics course. Lecture 1. 2004 Jan 5.

Gemoets D, Rosemblat G, Tse T, Logan R. Assessing readability of consumer health information: an exploratory study. Medinfo '04: in press.

Greenberg L, D'Andrea G, Lorence D. Setting the public agenda for online public health: a white paper and action agenda. Washington, DC: URAC [Utilization Review Accreditation Commission], 2003.

HHS (Health & Human Services). Healthy people 2010. 2nd ed. With understanding and improving health and objectives for improving health. 2 vols. Washington, DC: U.S. Government Printing Office, November 2000.

IOM (Institute of Medicine). The future of the public's health in the 21st century. Washington, DC: The National Academy Press. 2003;509 p.

Jansen BJ, Spink A, Saracevic T. Real life, real users, and real needs: a study and analysis of user queries on the web. Information Processing and Management. 2000;36:207-27.

Klare GR. The measurement of readability. Ames IA: Iowa State University Press. 1963.

Klare GR. Readable computer documentation. ACM Journal of Computer Documentation 2000;24(3):148-68.

Kuhlthau C. Longitudinal case studies of the information search process of users in libraries. Library and Information Science Research 1988;10:257-304.

Lewis D, Brennan PF, McCray AT, Tuttle M, Bachman J. If we build it, they will come: standardized consumer vocabularies. Medinfo 2001;10:1530.

Ley P, Florio T. The use of readability formulas in health care. Psychology, Health & Medicine 1996;1(1):7-28.

Lindberg DAB. Opening statement on fiscal year 2004 President's budget request for the National Library of Medicine. 2003 Apr. [Accessed 3/5/04] <http://www.nlm.nih.gov/about/2004openst.pdf>

Lindberg DAB. NLM Long Range Plan. Report of the Board of Regents, 1987.

Lindberg DAB, Humphreys BL. A time of change for medical informatics in the USA. In Yearbook of Medical Informatics 1999. Berlin: Springer-Verlag.1999:53-7.

Logan RA. Evaluating consumer informatics: learning from health campaign research. Medinfo '04: in press.

Madden M, Rainie L (Eds). America's online pursuits: the changing picture of who's online and what they do. Pew Internet & American Life Project. 2003:89 p.

Marchionini G. Information seeking in electronic environments. Cambridge UK: Cambridge University Press. 1995. 224 p.

Marshall PD. Bridging the terminology gap between health care professionals and patients with the Consumer Health Terminology (CHT). Proc AMIA Symp. 2000:1082.

McCray AT, Ide NC, Loane RF, Tse T. Strategies for supporting consumer health information seeking. Medinfo '04: in press.

McCray AT, Loane RF, Browne AC, Bangalore AK. Terminology issues in user access to Web-based medical information. Proc AMIA Symp. 1999:107-11.

McCray AT, Tse T. Understanding search failures in consumer health information systems. Proc AMIA Symp 2003: 430-4.

MedlinePlus Go Local XML Files. [Accessed 3/5/04]
<http://www.nlm.nih.gov/medlineplus/glftp/golocalxml.html>

Miller N, Lacroix EM, Backus JE. MEDLINEplus: building and maintaining the National Library of Medicine's consumer health Web service. Bull Med Libr Assoc. 2000 Jan;88(1):11-7.

Nath R. Consumer Health Vocabularies. Presentation and testimony at the Meeting of the Workgroup on National Health Information Infrastructure. Chicago: National Committee on Vital and Health Statistics (NCVHS), 2002 July 24. [Accessed 3/5/04]
<http://www.ncvhs.hhs.gov/020724p7.pdf> and http://www.ncvhs.hhs.gov/020724tr.htm>

NCI (National Cancer Institute). Health Information National Trends Survey (HINTS) I Final Report. Bethesda, MD: Division of Cancer Control and Population Sciences, National Cancer Institute. 2004 Feb 17. [Accessed 4/12/04]
<http://cancercontrol.cancer.gov/hints/hints_docs/final_report.pdf>

NCI. Online behavior change and disease management: a research dialogue. National Cancer Institute and the Robert Wood Johnson Foundation. 2001 Aug 2. [Accessed 4/12/04]
<http://www.hetinitiative.org/content/Behav_Chng_and_Disease_Mgt.pdf>.

NLM (National Library of Medicine). Online usage statistics smashed. NLM Newsline 1998 Jan-Mar;53(1):2. [Accessed 3/5/04]
<http://www.nlm.nih.gov/pubs/nlmnews/janmar98.html#Online>

Pifalo V, Hollander S, Henderson CL, DeSalvo P, Gill GP. The impact of consumer health information provided by libraries: the Delaware experience. Bull Med Libr Assoc. 1997 Jan;85(1):16-22.

Roberts JC, Fletcher RH, Fletcher SW. Effects of peer review and editing on the readability of articles published in Annals of Internal Medicine. JAMA. 1994;272:119-21.

Rosemblat G, Gemoets D, Browne AC, Tse T. Machine translation-supported cross language information retrieval for a consumer health resource. 2003. Proc AMIA: 564-8.

Rosemblat G, Tse T, Gemoets D. Adapting a monolingual consumer health system for Spanish cross-language information retrieval. ISKO '04. In press.

Rozic-Hristovsk A, Hristovski D, Todorovski L. Users' information-seeking behavior on a medical library Website. J Med Libr Assoc. 2002 Apr;90(2):210-7.

Rudd RE, Moeykens BA, Colton TC. Health and literacy: a review of medical and public health literature. In: Annual Review of Adult Learning and Literacy. San Francisco CA: Jossey-Bass 1999:158-199.

Russell-Pinson L. Linguistic and extratextual variation in medical English writings: a comparative genre analysis. Unpublished Ph.D. dissertation. Washington, DC: Georgetown University; 2002.

Shaw J, Baker M. "Expert patient"—dream or nightmare? BMJ 2004;328:723-4.

Snow CE. Reading for understanding: toward a research and development program in reading comprehension. RAND Reading Study Group. Santa Monica CA: RAND Education. 2002. 156 p.

Stavri PZ. Personal health information-seeking: a qualitative review of the literature. Medinfo 2001;10:1484-8.

Taylor PT. Consumer health informatics: emerging issues. Testimony before the Subcommittee on Human Resources and Intergovernmental Relations. Washington DC: General Accounting Office. GAO/T-AIMD-96-134. 1996. p. 14.

Taylor R. The process of asking questions. American Documentation 1962 Oct, 391-397.

Taylor WL. Recent developments in the use of "cloze procedure." Journal Q. 1958:42-8.

Tse T, Soergel D. Exploring medical expressions used by consumers and the media: an emerging view of consumer health vocabularies. Proc AMIA Symp 2003:674-8.

Wang P, Berry MW, Yang Y. Mining longitudinal Web queries: trends and patterns. . Journal of the American Society for Information Science and Technology 2003;54(8):743-58.

Wildemuth BM. The effects of domain knowledge on search tactic formulation. Journal of the American Society for Information Science and Technology 2004;55(3):246-58.

Zeng Q, Kogan S, Ash N, Greenes RA, Boxwala AA. Characteristics of consumer terminology for health information retrieval. Methods Inf Med. 2002;41(4):289-98.

Zielstorff RD. Controlled vocabularies for consumer health. J Biomed Inform 2003;36:326-33.

Tony Tse
Research Fellow

Education and Training
Dartmouth College, Hanover, NH   AB   1983-87   Biochemistry
Yale School of Medicine, New Haven, CT   1987-89
George Washington University, Wash, DC   MS   1992-96   Information Systems Technology
University of Maryland, College Park, MD   PhD 1996-03   Library and Information Services

Research and Professional Experience

Employment
Research Fellow, Lister Hill National Center for Biomedical Communications,
　　National Library of Medicine   1999-present
Senior Program Analyst, Howard Hughes Medical Institute   1991-1997; 1998-1999
Graduate Research/Teaching Assistant, University of Maryland   1997-1998
Biology Science Assistant, National Science Foundation   1990-1991

Honors
NLM Staff Recognition Awards   2000-2003
Eugene Garfield Doctoral Dissertation Fellowship, Beta Phi Mu   2001
ClinicalTrials.gov Team Group Award   2000

Publications
McCray AT, Ide NC, Loane RF, **Tse T**. Strategies for supporting consumer health information seeking. *Medinfo 04*, in press.

Soergel D, **Tse T**, Slaughter L. Helping healthcare consumers understand: an "interpretive layer" for finding and making sense of medical information. *Medinfo 04*, in press.

Gemoets D, Rosemblat G, **Tse T**, Logan R. Assessing readability of consumer health information: an exploratory study. *Medinfo 04*, in press.

Gillen JE, **Tse T**, Ide NC, McCray AT. Design, Implementation and management of a Web-based data entry system for ClinicalTrials.gov. *Medinfo 04*, in press.

Divita G, **Tse T**, Roth L. Failure analysis of MetaMap Transfer (MMTx). *Medinfo 04*, in press.

Rosemblat G, Gemoets D, **Tse T**. Adapting a monolingual consumer health system for Spanish cross-language information retrieval. *ISKO 04*, in press.

Tse T, Soergel D. Exploring medical expressions used by consumers and the media: an emerging view of consumer health vocabularies. *Proc AMIA Symp* 2003: 674-8.

McCray AT, **Tse T**. Understanding search failures in consumer health information systems. *Proc AMIA Symp* 2003: 430-4.

Tse T, Soergel D. Procedures for mapping vocabularies from non-professional discourse and a case study from the medical domain. *Proceedings of the American Society for Information Science and Technology (ASIS&T) Annual Meeting* 2003:174-83.

Rosemblat G, Gemoets D, Browne AC, **Tse T**. Machine translation-supported cross language information retrieval for a consumer health resource. *Proc AMIA Symp* 2003: 564-8.

Tse T, Johnson SO, and Ripple A. Update: expanding ClinicalTrials.gov search capabilities. *NLM Tech Bull*. 2002 Jul-Aug;327: e5.

Darren A. Gemoets

Senior Technical Consultant

Aquilent, Inc.
1100 West Street
Laurel, MD 20707

Education:

Lafayette College, Easton, PA   B.S.   1989-1993   Computer Science

Contractor Experience:

Developer / Technical Consultant / Researcher, National Library of Medicine, 2001-present.

Software Developer / Task Lead, NASA, Goddard Space Flight Center, 1999-2001.

Subsystem Developer / Lead, Intelsat, 1998-1999.

Software Developer, National Library of Medicine, 1996-1998.

Software Developer, NASA, Goddard Space Flight Center, 1993-1996.

Honors:

AppNet Virtuoso Award, AppNet, Inc., 1999.
Phi Beta Kappa, Lafayette College, 1993.

Publications:

"Adapting a Monolingual Consumer Health System for Spanish Cross-Language Information Retrieval" by G. Rosemblat, T. Tse and D. Gemoets, Proceedings of the Eighth International ISKO Conference (In press).

"Assessing Readability of Consumer Health Information: An Exploratory Study" by D. Gemoets, G. Rosemblat, T. Tse and R. L. Logan, Proceedings of the MEDINFO 04 Conference (In press).

"Machine Translation-Supported Cross-Language Information Retrieval for a Consumer Health Resource" by G. Rosemblat, D. Gemoets, A. C. Browne and T. Tse, Proceedings of The AMIA 2003 Symposium, pp. 564-568.

"Web-Altairis: An Internet-Enabled Ground System" by J. Coleman, D. Gemoets, K. Hughes and P. Miller, Proceedings of SpaceOps 2000.

Graciela B Rosemblat
Linguistics Specialist
under contract from TAJ Technologies, Inc.
Lister Hill National Center for Biomedical Communications
National Library of Medicine
8600 Rockville Pike
Bethesda, Maryland 20894
rosem@nlm.nih.gov

Education

Ph.D. Linguistics. Georgetown University, Washington, D.C., 1993
M.S. Linguistics. Georgetown University, Washington, D.C., 1986
TESOL. Teaching of English for Speakers of Other Languages. National Teachers'
Training in Living Languages. Buenos Aires, Argentina, 1972-1976

Languages

Fully proficient in Spanish, English, and Italian
Working knowledge of Portuguese

Computational Experience

- Natural Language Processing
- Machine Translation
- C language
- Part-of-Speech Tagging

Professional Experience

2002-Present	Linguistic Contractor. NIH. National Library of Medicine, 8600 Rockville Pike, Building 38A, B1N30A14. Bethesda, Maryland
1995-02	Computational Linguist and Lexicographer-Terminologist under Contract for the Pan American Health Organization (PAHO), Washington D.C.
1994-95	Spanish Linguist Consultant. SRA Corporation. Arlington, Virginia
1994-96	Free-lance Interpreter/Translator. CACI. Arlington, Virginia
1989	Lecturer, Spanish Composition and Style I. Georgetown University.
1986	Field Reader, Department of Education and Human Resources. Office of Bilingual Education. Washington, D.C.
1978-84	Executive Legal Administration (English-Spanish). Court Interpreter and Translator. Luis Vidana, Attorney at Law. Washington, D.C.

Honors

Ph.D. Dissertation, titled "The Semantics of Aspect in Spanish Participial Constructions", defended "with distinction". Georgetown University, Nov. 1993
Associate Editor, Georgetown Journal of Languages and Linguistics, 1990

Publications and Conference Presentations

"Assessing Readability of Consumer Health Information: An Exploratory Study", D. Gemoets, G.Rosemblat, T.Tse and R. Logan. To be presented at the *Medinfo Conference*. San Franscisco, California, September 2004.
"Adapting a Monolingual Consumer Health System for Spanish Cross-Language Information Retrieval", G. Rosemblat, T. Tse and D. Gemoets. To be presented at the 8th *International ISKO Conference*, London, England. July 2004.
"Machine Translation in Cross Language Information Retrieval", G. Rosemblat, D. Gemoets, T. Tse and A. C. Browne. In *Proceedings of the American Medical Informatics Association Annual Symposium*. November 2003. pp. 564-568.
"On Small Clauses in Spanish", *Journal of the School of Languages and Linguistics*, Georgetown University, Washington, D.C. Spring 1990. Vol.1, No 2., pp 209-233.