Retrieving *e*-Health Research:

The Challenge of Accessing the Knowledge[1]

By

Richard Davis, M.S.
Mia Liza A. Lustria, Ph.D.
Linda Lockett Brown, M.S.

of

Florida State University

[1] Paper presented at the 2006 American Society for Information Science and Technology (ASIS&T) Annual Meeting held in Austin, Texas on November 3-8, 2006.

Abstract

The Internet has increasingly become a viable tool for delivering interventions to promote better health behaviors and manage diseases. As an information and communication tool, it leverages the broad reach of mass media with the interactive features of interpersonal channels. As such, the past 10 years has seen a proliferation of pilot studies, evaluative studies and randomized controlled trials designed to test the efficacy of web-based interventions on health outcomes. These efficacy studies look at a variety of ways Internet-based interventions (*e*-health) can be used across a number of health conditions and in different contexts. In particular, the last two to three years has seen a significant increase in efficacy studies of web-based interventions published in the literature across different disciplines. The challenge of this from a scholarly perspective is that the sheer number of different ways the literature has been described and indexed has made it increasingly difficult for scholars to locate relevant articles and draw clear conclusions about the efficacy of web-based interventions across disparate fields of interest.

This study explores how this field of inquiry has grown and matured in the past 10 years by examining the growth of the literature specific to Internet-based health interventions and to explore the changes in terminologies used to describe this field. A frequency analysis of keywords used to describe web-based health intervention studies was conducted on over 150 peer-reviewed journal articles and a listing of high-value keywords were generated. A large number of unique keywords were discovered to be associated with the body of publications located according to the methodology described.

The results of this study will be of specific use to researchers interested in web-based health interventions and will be of practical value in aiding their literature searches.

This keyword analysis also reveals interesting insights into the development of an emerging interdisciplinary field of inquiry based on how scholarly activities are being described and/or represented across disciplines and specializations.

Background

Internet health interventions are increasingly being utilized to deliver health information and services to patients. These approaches differ from typical health interventions either because these are purely delivered over the Internet, or these are combined with more traditional intervention delivery approaches (e.g. print, face-to-face, computer-based, etc.). Enthusiasm for the use of the Internet for health information and services delivery has grown out of its perceived relative advantage over more traditional media. In particular, the Internet, is considered a hybrid channel, one that not only has the broad reach of the mass media, but also has the persuasive capabilities of interpersonal channels (Cassell, Jackson, & Cheuvront, 1998).

Empirical research measuring the efficacy of Internet-based health interventions has increased in the past three to four years. Despite the work being performed and reported, we are nowhere near understanding whether these Internet-based, interactive, health interventions are making real differences in health outcomes. The difficulty of measuring effects of this approach on patient outcomes per se is compounded by the sheer number of iterations in treatment and methodologies tested across different fields and specializations.

To address this issue, the authors sought to conduct a systematic review of efficacy studies of Internet-based health interventions. In particular, the authors initiated a comprehensive database search for scholarly publications reporting the results of studies of Internet-based health intervention outcomes. That objective continues to be the target of inquiry. However, upon completion of the database search and retrieval of the eligible studies, the investigators felt that the challenging nature of retrieving these citations warranted further investigation. This paper documents experiences conducting the database query for Internet-based health intervention research conducted in the past 10 years and as retrieved from 28 academic databases.

Research Questions

Bibliographic search technologies implemented by database providers make a variety of query methodologies available to researchers. Researchers have a number of search options by which they can locate relevant literature, by: keyword, index, abstract and full text searches. Queries may be constructed utilizing one of these methods alone or in combination with the others.

Keyword-only queries have been the dominant method of information retrieval for centuries (Sridhar, M. 2004). More recently, full-text searching, where the entirety of the stored document is the index, has become widely available. Full-text-based information retrieval utilizes automation to make publications available based not simply upon keywords assigned by authors or information scientist.

For this particular study, initial attempts to retrieve relevant literature presented significant challenges for the authors. Keyword-only searches, subject searches and finally full-text searches were all utilized to locate the literature.

As such, several questions regarding how this literature was being indexed were raised. The main challenge was: how could we successfully retrieve publications that matched specific criteria, without being forced to go through volumes of extraneous query results? Specifically, we were interested in discovering what keywords or search term(s) were associated with efficacy studies of Internet-based health interventions. And, might an investigator refine such a process by being informed of the highest-frequency keywords or search term(s) by which literature may be located that was stored in various bibliographic databases?

Given the fact that this is an emerging interdisciplinary field, understanding what keywords and/or search term(s) has been associated with the totality of e-health research publications could be informative to all investigators interested in this area. Individual search terms, as well as the general category of those terms, might be useful. Was each article's individual keyword or keywords clinical in nature, methodological in nature or neither?

As our database searches unfolded, we also became curious about the following questions: Based on the keyword analysis, what classification terms appeared and in what frequency? How are these disciplines or specializations describing this literature so that interested parties might easily retrieve appropriate citations? These latter two questions may bring to the fore issues regarding the challenge of indexing research literature in emerging interdisciplinary fields of study, such as e-health.

Literature

Scholars investigate, and upon completion of investigation, produce scholarly publications. Many of these publications are made available in juried scholarly journals. Within the academic realm, academic libraries make publications available to students, scholars and interested parties. Since the early years of the twenty-first century, most literature searches have been performed using computerized academic databases.

Historical research informs us that the earliest attempts at indexing of documents occurred an estimated 5,000 years ago in the ancient city of Sumer. Archivists placed the first few words of a text on the outside edge of the clay texts (Weidman, M. & Stumpfer, C., 2004). In modern terms, such indexing would be considered crude and inefficient.

Keywords are a mature and respected tool in the kit of the indexer and cataloger of today; albeit utilizing the latest in electronic storage and information retrieval systems.

Researchers are taught from the beginning of their education to use keywords to locate literature appropriate to the research question or questions at hand.

Digital systems, and the indexing based thereon, have extended the indexing process to the full text. As many scholars have noted, the best index of a text is the entirety of that text (Lin, F., Huang, S. & Chen, N., 2004). Keywords continue to be assigned to texts in the indexing process at the same time that electronic cataloging systems utilize sophisticated algorithms to capture significant terminology from the entirety of a particular text.

Bibliographic databases permit searching and retrieving based upon keyword, title, author(s), full text and numerous subdivisions in between. Keyword searching is, quite predictably, the most system-efficient method of retrieving citations from academic bibliographic databases. Scanning the indices of 100,000 journal publications by keyword is measurably faster than searching the full text of that same set of publications.

The bulk of indexing research in recent decades has revolved around the various technologies available for searching full text documents. Algorithms that scan the full text of a document and build search term matrices based upon word frequency have added to the assigned-keyword processes that are standard in information retrieval services. Digital storage, full text indexing and large data stores have yielded an unintended consequence: queries often return large numbers of inappropriate citations when full text search is utilized. Keyword-based queries, while returning far fewer citations, return far more appropriate citations (Weideman & Stumpfer, 2004). While efficiency and accuracy in information retrieval are critically important, systems that return a large percentage of inappropriate citations are counter-productive to the end user.

Aboutness is the essential nature of a document. Full text indexing does not provide a high degree of aboutness, therefore considerable effort has been expended in creating proximity matching algorithms (Lin, F., Huang, S. & Chen, N., 2004). Keyword assignment, as performed by professional catalogers and indexers, is able to capture more of the essential nature of a text. This adherence to the aboutness of a text is able to provide significant return to inquiries, both in system efficiency as well as high counts of appropriate citations after an initial query is complete (Hjerland, B. 2001).

Sridhar tells us that despite the technological advances of recent decades, moving from the venerable card catalogue to online full text databases has not measurably improved information retrieval (Sridhar, M. 2004).

Method

As part of a larger study, the investigators did an extensive search for peer-reviewed articles focusing on efficacy studies of Internet-based health interventions and their effects on health outcomes published within the last 10 years. In order to locate relevant literature, explicit inclusion criteria were formulated to avoid selection bias. In general, relevant studies were defined as studies that measured the effects of health interventions, delivered either via the Internet or with an Internet component on various patient outcomes (including but not limited to health behavior change; technology acceptance; health care cost savings; satisfaction with treatment; etc.). A total of 28 scholarly databases subscribed to by Florida State University's libraries were searched (See Attachment A for a list of the databases).

Appropriate search terms were agreed upon and the databases were divided up amongst the research team. Initial search parameters were:

E-health Electronic health E-medicine Electronic medicine
Internet World Wide Web Online Web computer-based network-based information technology communication technology
Health Medicine Healthcare
patient outcomes Intervention Communication

Promotion
Education

Informatics
health informatics
medical informatics
consumer health

Queries were placed utilizing each of the terms in white (above) in combination with a term in the shaded area. For example, a query was entered to find all juried publications listing (Internet OR e-health) AND (intervention OR outcome). This process was repeated in a systematic fashion.

Total returned results numbered in excess of 400,000 publications. The vast majority of returned citations were not appropriate to the investigation. A modified search strategy was created utilizing the following search criteria:

(internet AND health AND intervention AND outcome)

Slightly more than 400 citations were retrieved from the databases using this refined query.

A manual review was performed to judge the eligibility of each publication for the larger systematic review. Publications were categorized as eligible, ineligible and review articles. Ineligible citations were discarded. Eligible citations were included in an Endnote database for inclusion in the larger study. The reference lists of the review articles were likewise examined to identify additional studies that were not retrieved during the initial search. These additional studies were then added to the Endnote database. This step also served to verify the effectiveness of the initial search and in fact validated the thoroughness by which the team was able to conduct it.

A total of 161 scholarly publications were found to be eligible for the larger systematic review.

A reference manager software, Endnote (version 9), was used to manage the citations. Once the collection was purged of ineligible studies, the eligible citations were exported to a Microsoft Access database. The database contained two fields: the article title and an article number. In addition to the citations, each publication's keyword export file was imported into a separate Microsoft Access database. Each individual keyword yielded a separate record in the keywords database. The two databases were related logically, providing a 1:N relationship between article title and keyword(s) records. Utilizing the various functions of the database management system (DBMS), analysis was performed on the keyword counts for the occurrence of various unique and non-unique keywords. Articles without keywords were flagged with the keyword 'none'.

Keywords were then classified as (a) clinical, (b) methodological, (c) intervention type or (d)other. Clinical keywords were those keys that described an actual disease process such as asthma, HIV/AIDS, back pain, breast cancer and so forth. Methodological described the actual study method, including describers associated with the main population focus or foci of the study. Examples of methodological keys includes gender keys (female, male), age-cohort specifications (child, elderly, adult) as well as those keys that describe the study method. Intervention type was categorized as the actual treatment intervention described in the research. Interventions involved a gamut of treatments including college student alcohol use reduction, computer-assisted patient education, nutrition counseling online, etc. Keywords that did not fall within the three specific categories were classified 'other'. A separate database field was created to contain categorization data.

Frequency analysis was performed on the non-categorized keywords, the categorized keywords and the article titles for the initial search terms.

Journals were also categorized according to main foci or specialization. The following categories were assigned to the publications retrieved:

- Psychology/Psychiatry
- General medicine
- Nursing
- Diabetes
- Eating Disorders
- Cancer
- Internet/Telemedicine Research
- Other Disease-specific publication
- Education

Results

An initial analysis of the keywords extracted from the Endnote citation database yielded 1386 non-unique keywords. Keywords were present for 128 (77%) of 161 articles. Articles with keywords associated, had a mean of 11.0 keywords per article.

Table 1 summarizes the occurrence of specific keywords suggested in the initial general search strategy.

Table 1. Summary of main keyword occurrence

Keyword	**Count**	**Percent of articles containing the key**
Internet	72	42.86%
Health	0	0.00%
Intervention	0	0.00%
Outcome	0	0.00%

Used individually, the initial search keywords were not able to return a lot of relevant literature. Composite keys, as stored and utilized by the various databases and the associated query screens, contained one or more of the targeted keywords. Composite keys are keys that utilize more than one term, and may include ‘natural language’ structure. A composite key is found as a single key in the database’s index, for example the key sequence ‘Internet delivered intervention’, is a single key that contains two of the four keyword targets of the study. Counts of composite keys containing a combination of any two of the four keywords of interest were counted and is summarized in Table 2.

Table 2. Summary of composite keys

Keyword	**Count**	**Percent of articles containing the key**
Internet	93	57.06%
Health	52	31.90%
Intervention	3	1.84%
Outcome	8	4.91%

Using the keywords actually associated with each article, we analyzed the most popular keywords used to described the literature. Specifically, keywords, and composite keys were analyzed for frequency of appearance in the keys listed for each article. Excluding the keyword null entry (None), 14 keywords or composite keys appeared in 10 or more articles (Table 3).

Table 3. Summary of the most popular keywords used to described the literature

Keyword	**Count**	**Percent of articles containing the key**
Humans	72	44.72%
Internet	70	43.48%
Female	65	40.37%
Male	58	36.02%
Adult	42	26.09%
Middle Aged	38	23.60%
Research Support, Non-U.S. Gov't	36	22.36%

Research Support, U.S. Gov't, P.H.S.	16	9.94%
Treatment Outcome	15	9.32%
Pilot Projects	12	7.45%
Questionnaires	11	6.83%
Program Evaluation	10	6.21%
Comparative Study	10	6.21%
Social Support	10	6.21%

Keyword categories were analyzed as described in the previous section. Categories were assigned based on the informational content of the keyword or the composite key. Table 4 shows that most of the keywords used to index the citations dealt with research methods, followed by intervention type, and then by a description of the disease focus. Further analysis of unique keywords was performed. Table 4 illustrates that in addition to a large number of total keywords, there were also a large number of unique keywords used, most of which described intervention type.

Table 4. Summary of total and unique keywords by categories.

Keyword Category	**Total Keywords**	**Unique Keywords**
Clinical	89	81
Intervention type	498	324
Research Methods	613	146
Unspecified	186	73
Total	1386	624

As discussed in previous sections, Internet-based health interventions have been studied in a variety of fields and specializations. Results of the database search, and a cursory examination of the types of journals in which these were published reveal the multidisciplinary nature of this emerging field of study. Table 5 summarizes the specialized fields within which Internet-based intervention research has been conducted.

The most number of empirical research on this type of intervention has been conducted in the field of psychology followed by general medicine.

Table 5. Summary of keyword frequencies by journal category

Journal Category	**Count**	**Percent of total journals**
Psychology/Psychiatry	39	24.22%
General Medicine	21	13.04%
Nursing	13	8.07%
Diabetes	8	4.97%
Eating Disorders & Nutrition	14	8.70%
Cancer	5	3.11%
Internet/Telemedicine Research	20	12.42%
Other Disease-Specific Publication	22	13.66%
Education	19	11.80%

Discussion

We sought to identify keywords, search term(s) to provide guidance in retrieving relevant e-health intervention publications. Additionally, we sought to understand the broad categories into which keywords and search term(s) had been placed to assist in the location of appropriate citations. Expectations were that a number of keywords and search term(s) commonly associated with the field, e-health, would result from the analysis of keyword data retrieved from the various bibliographic databases. That expectation was not met by the data collected.

When total keywords, by category, were further analyzed for uniqueness, interesting information emerged. Specifically, within the category of intervention type, keywords were unique 324 out of 498 times. We believe that this wide range of keyword assignment to a limited number of articles would lead searchers to expend significant amounts of time in accessing the research. Within disciplines, various terms were used to index citations that had overlapping meaning and created difficulties in locating citations.

Using keywords or Boolean logic that access either keyword lists or full-text returned in excess of 400,000 citations, most of which were not relevant. Single keyword searches utilizing the four selected terms returned 72 citations with the keyword Internet. No citations were returned with the three additional keywords ‘health’, ‘outcome’ and ‘intervention’. Although a high percentage of articles within the bibliographic databases (75%) had keywords assigned and a large number of keywords were associated with each article (11), keyword-only searching accounted for zero citations appropriate to the larger meta-analysis project.

Composite keys were productive in terms of citations returned to queries. Composite keys, meaning search terms composed of two or more of the target terms, were far more prevalent in the keyword data than as individual keys. 'Health', as an individual keyword, was not found in a single publication. 'Health', as a contributor to a composite key, was found in 52 citations. The frequencies of Internet, intervention and outcome increased significantly when counted as individual keyword and as a component of a composite key.

Full-text searching with composite keys, using Boolean AND instructions, yielded appropriate citations at a rate of 35.5% (142/400). Thirty five percent appropriate citation return is a tractable literature search although such a return requires considerable effort on the part of the inquirer to glean appropriate citations from the total. Combining the high number of unique keywords discovered (624) with the very low recall rate of citations returned by the various databases, we believe that the development of a taxonomy for e-health research reporting would be a worthwhile effort.

A closer examination of the specific keyword categories by which these studies were indexed revealed that nearly half would be categorized as methodological. Intervention descriptor keywords were found in a large number of articles. Thirteen percent of keywords could not be categorized according to the research method utilized in the particular investigation and covered numerous interest areas. We find this fact to be illustrative of the multi-disciplinary nature of clinical treatment methods moving, in part, to an Internet-enabled modality.

Further evidence of differences in keyword assignment across disciplines is the fact that of 89 clinical keywords utilized, 81 of those keywords were unique to a particular article. Nearly every single article in this category had unique keywords assigned, further challenging individuals seeking a comprehensive bibliography of scholarship reported in the field of e-health intervention.

The results of the keyword analysis by journal category provides further insights into the multidisciplinary nature of this emerging area of research. Psychology and associated disciplines accounted for nearly one quarter of the publications; the remaining 74.78% of publications fell within eight broad categories of classification ranging from specific diseases (diabetes, cancer, eating disorders) to 20 publications found in journals associated with delivery methodologies and 19 journals associated with patient education. Specifically, 49 citations out of the total number of citations appeared in disease-specific journals.

Again, this observation, in addition to the numerous ways by which the individual articles were indexed, reveal the challenge of coming up with a common indexing scheme that would work across specializations. Then again, is this really necessary considering that most of the relevant citations were returned via full text searching rather than via keyword searching?

These issues bring to the fore questions for further research: How what is the current use of keyword identification or indexing? Who benefits the most from keyword identification? How might disparate but interdisciplinary fields be informed about how to better index particular publications? Would the development of a specialized taxonomy covering this interdisciplinary emerging field be useful for indexers and/or scholars interested in this field?

Limitations of the Study

The authors recognize that the results of the current research are limited by the fact that we have no knowledge about how these various publications were actually indexed or how the keywords were chosen. Some publications require authors to suggest certain keywords for their manuscripts and this may explain the disparities in keyword identification within and between specializations. Due to the proprietary nature of commercial database vendors, the investigators were not able to discern the methods utilized for keyword creation. We are unable to report if keyword assignment to a publication was performed by the publication author(s), the journal which contained the publication or the database vendor.

Additional limitations to the data analysis were potentially created by the coding of categories for keyword and journal. With a single investigator performing coding, bias may have been introduced into the categorization process. While the information generated by these data is informative, additional investigation is on-going incorporating a multiple coder methodology.

References

Cassell, M. M., Jackson, C., & Cheuvront, B. (1998). Health Communication on the Internet: An Effective Channel for Health Behavior Change? *Journal of Health Communication, 3*(1), 71-79.

Hjerland, B. (2001). Towards a theory of aboutness, subject, topicality, theme, domain, filed, content…and relevance. *Journal of the American Society for Information Science*, 52(9):774-778

Len, F., Huang, S. & Chen, N. (2004). Incremental revision of recommendation rules for information services. *E-Services Journal*. 2004 p. 85-109.

Shen, M., Calbretta, N., Cavanaugh, M., Datwani, N., Lew, C. & Dadhania, M. (2003). Analysis of current nuclear cardiology literature in MEDLINE database: A study of gated SPECT imaging using PubMed. *Journal of Nuclear Cardiology.* 10(6):650-655.

Sridhar, M. (2004). Subject searching in the OPAC of a special library: problems and issues. *OCLC Sytstems and Services: International Digital Library Perspectives* 20(4):183-191.

Weideman, D. & Strumpfer, C. (2004). The effects of search engine keyword choice and demographic features on internet searching success. *Information Technology and Libraries,* June 2004 p. 56-65.

Appendix A

Databases Searched

DATABASES
Academic Index is now Expanded Academic Index ASAP
AIDS & Cancer Research
ArticleFirst
CINAHL and Pre-CINAHL
Communication Abstracts
ContentsFirst
Educational Research Abstracts Online (ERA)
Elsevier Science Direct
Emerald Library E-Journals (MCB Publications)
ERIC (FirstSearch)
Expanded Academic ASAP
IngentaSelect
ISI Web of Knowledge
JSTOR
Kluwer (E-Journals)
MEDLINE (FirstSearch)
NLM Gateway
PsycARTICLES
PsycINFO
PubMed
ScienceDirect
Social Sciences Citation Index
Sociological Abstracts
Springer LINK
Synergy
Wiley Interscience Journals
Wilson Science Complete