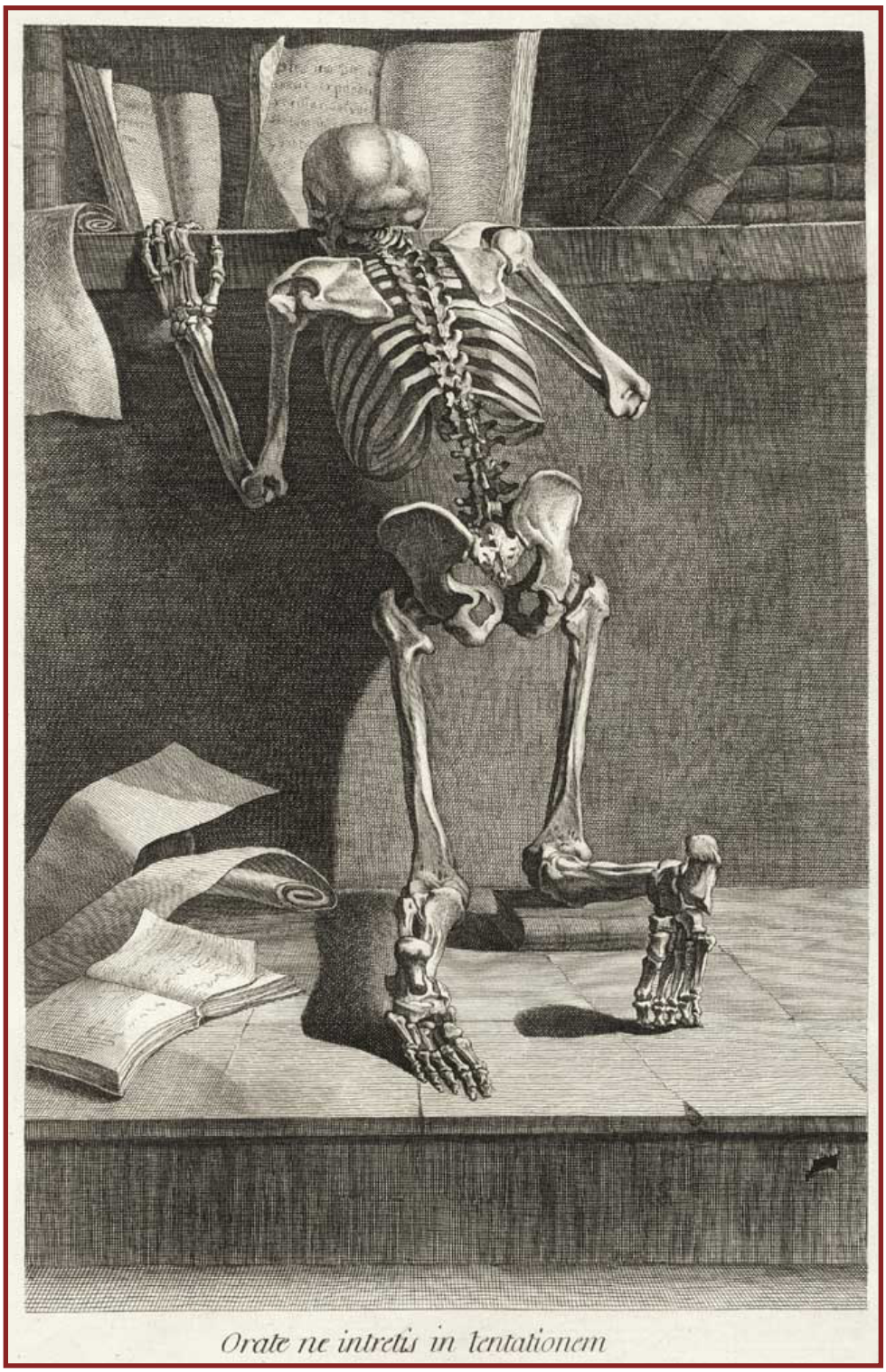
HYPOTHESIS

The Journal of the Research Section of MLA



Volume 20
Number 3
Fall 2008

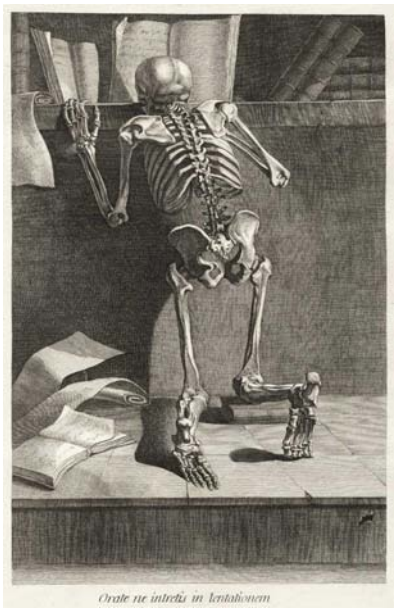
HYPOTHESIS

The Journal of the Research Section of MLA

COLUMNS

ARTICLES

NEWS



Cover Art: The image above is from Jacques Gamelin's *Nouveau Receuil d'osteologie et de Myologie* published in 1779. For more information and images visit the NLM's Historical Anatomies on the Web. This digital project includes numerous high quality images from the library's important anatomical atlases.

Have an image you'd like see on the cover? Please let Co-editor Lisa know!

HYPOTHESIS (ISSN 1093-5665) is the official journal of the Research Section of MLA. It is published three times a year by the Section: Spring (March), Summer (July/August) and Fall (November). Items to be included should be sent to the Co-Editors by the 15th of the preceding month (i.e., February 15th for Spring, June 15th for Summer, and October 15th for Fall). Copy is preferred by e-mail but will be accepted in other formats. HYPOTHESIS is indexed in the Cumulative Index to Nursing and Allied Health Literature™ and the CINAHL® database. HYPOTHESIS is available online at http://www.research.mlanet.org/hypothesis.

CHAIR'S COLUMN

Susan Lessick, MA MLS AHIP
Head, Grunigen Medical Library, UCI Medical Center

One of our goals this year is to highlight the research work of our MLA, Section, and Chapter colleagues. Toward this end I asked MLA Headquarters if they would post on the MLA Web site the titles of research grants and links to original research papers (if available) of all the past grant recipients both for the MLA Research, Development, and Demonstration Project Grant and the Donald Lindberg Research Fellowship. I asked because I think adding this research topic information, along with the names of the award winners, would be extremely helpful to librarians interested in submitting a grant application -- it would give them a better understanding of topics of successful proposals and it could spark ideas for future research. Well MLA generously agreed to try to do this. Lisa Fried who is in charge of Awards, Grants, and Scholarships at MLA Headquarters and I have been contacting past award winners to ask them for this information and Lisa has been able to update the web pages with this new research information. And from now on, Lisa is going to include the titles of research projects and links to research papers with the names of each past winner. I am so grateful to Lisa and Carla at MLA HQs for addressing this need in such a timely fashion and just doing it! So please check out the MLA Research Grant and the Lindberg Research Fellowship pages to see our progress so far.

In a similar vein, our section has also been putting forth much effort to identify MLA chapters and sections with active research committees and research grant and awards programs. The thought here is to gather and organize all MLA research-related information in one convenient web location to highlight all research activities that are going on at the grassroots level of MLA and the individual research contributions of each member. This chapter/section research information has now been compiled and will be available, along with some other new features and resources, on the new Research Section web site which is coming soon! I was so delighted to discover that three other sections in MLA have research committees (NAHRS, PH/HA, and TSS). PH/HA also offers a research award for best published research paper in the field of public health librarianship and TSS sponsors a continuing education grant for members who wish to develop research skills. Six chapters have active research committees (MLGSCA, MAC, MCMLA, NCNMLG, SCC, and SC); four chapters offer research grants to their chapter members (MLGSCA, NY-NJ, SCC, and SC); three chapters sponsor awards programs for best papers at annual chapter meetings (MAC, SCC and SC); and two chapters (SCC and SC) even have research mentor programs. It's so gratifying to see that these other sections and chapters have taken such an ardent interest in research and have committed dollars and energy to supporting research goals for their members.

While these are great strides, it is my fervent belief that our Section, along with our other MLA partners, need to do more to encourage librarians to become creators and proactive users of best evidence available to improve *libraries* and advance our profession. The Fall brings many opportunities to submit grant applications and undertake a research project, so check out MLA's Grant and Scholarships page and the grant opportunities of your local chapter. Let's all be challenged this year to start a research project – whether it be a 'small ball' evaluation or a grant funded project – it will improve your library – it will demonstrate your library's value -- and research is personally rewarding. JUST DO IT!

Let's work together to create and sustain a culture of inquiry throughout MLA and listen and learn from each other. Feel free to reach out and share your suggestions, concerns, and reflections with me about our section, research support in general, or your research experiences in particular. Please contact me at slessick@uci.edu or by phone at 714-456-6488. Have a wonderful autumn and holiday season.

LITERATURE REVIEW

Ruth Fenske, PhD AHIP
Reference Coordinator, Grasselli Library, John Carroll University

As health sciences librarians, we are all interested in consumers' health information seeking. There have been a number of articles having to do with this general topic in the last few months.

Buente W, Robbin A. Trends in Internet information behavior, 2000-2004. *J Am Soc Inf Sci Tech*. 2008 Sept;59(11):1743-60.

Yoo EY, Robbins LS. Understanding middle-aged women's health information seeking on the web: a theoretical approach. *J Am Soc Inf Sci Tech*. 2008 Feb15;59(4):577-90.

Chung DS, Kim S. Blogging activity among cancer patients and their companions: uses, gratifications, and predictors of outcomes. *J Am Soc Inf Sci Tech*. 2008 Jan 15;59(2):297-306.

Keselman A, Browne AC, Kaufman DR. Consumer health information seeking as hypothesis testing. *J Am Med Inform Assoc*. 2008, Jul-Aug;15(4):484-95.

Zeng-Treitler Q, Goryachev S, Tse T, Keselman A, Boxwala A. Estimating consumer familiarity with health terminology: a context-based approach. *J Am Med Inform Assoc*. 2008 May-June;15(3):349-56.

Leroy G, Miller T, Rosemblat G, Browne A. A balanced approach to health information evaluation: a vocabulary-based naïve Bayes classifier and readability formulas. *J Am Soc Inf Sci Tech*. 2008 Jul 15;59(9):297-306.

Keselman A, Smith CA, Divita G, Kim H, Browne AC, Leroy G, Zeng-Treitler Q. Consumer health concepts that do not map to the UMLS: where do they fit? *J Am Med Inform Assoc*. 2008 Jul-Aug;15(4):496-505.

Buente and Robbin look at the demographic and social and technological contexts of use of the Internet by doing a secondary analysis of the Pew Internet and American life telephone surveys conducted between 2000 to 2004. Number of respondents ranged from 1501 to 3453. Use was divided into use to communicate, inform, entertain, and shop. Searching for health information falls into the "inform" category. The authors specifically acknowledge their inability to assess cognitive and psychological attributes of individuals as predictors of types Internet use because those types of data were not collected by the Pew researchers.

Their factor analysis shows what the authors call a "secondary digital divide." Those who use the Internet to obtain information are younger, college-educated, and frequent Internet users. Going online to buy something is associated with higher socioeconomic status while using the Internet for entertainment with lower socioeconomic status. Nature of use was not related to race or ethnicity. Length and frequency of use were predictors of information use and online buying. Those who are online for fun tend to be inexperienced users.

The Pew data show that Internet users looking for health or medical information increased from 53.8% in early 2000 and 66.4% in late 2002. Specific questions about the use of the Internet for health information were not asked in subsequent surveys in the time period at hand.

The Yoo and Robbins article, described next, tells us about later studies that appear to update this study in regard to antecedents of looking for health information on the Internet. It is unclear why this article, based on old data, was accepted for publication. One is left wondering if the secondary digital divide they established continues in 2008.

Yoo and Robbins tell us that the later Pew data show that Internet use for health information has risen to 80% in 2006. This study focuses on "how and why middle-aged women use health-related Web sites" using theory from mass communication and social psychology. Previous studies showed middle-aged women to be frequent seekers of health information.

The authors carefully explain the development of their model, their hypotheses, sampling procedure, and variables. They mention their questionnaire only briefly and do not append a copy. Response rate was just over 50% (n=354) women. Comparison of the respondent's characteristics to known characteristics of the population showed that education and income levels among the respondents were higher than in the population as a whole and the respondents were more likely to be employed full-time or part-time.

The overall result was that middle-aged women are more influenced to use health-related web sites by attitude toward health-related web use and their motivations for using the web to find health –related information than by confidence in their ability to use the websites. Results may have been influenced by the relatively high socioeconomic status of the sample.

This article is difficult to read because the authors use abbreviated names for their variables throughout. In the last paragraph of their results they mention a "modified model" which is not shown. Although this study appears to have been carefully done, the results do not seem to be commensurate with the amount of effort expended.

Chung and Kim look at the use of blogs. Their focus is on cancer patients and their family and friends. Their theoretical framework is uses and gratifications, one of the theoretical frameworks used by Yoo and Robbins. Uses and gratifications refers to how people use media and the gratifications sought and obtained from the use of various media. Participants were 111 of 153 individual bloggers invited to participate in the study. Nearly 62% were patients and the remainder were friends and family of cancer patients. Over three- quarters were women. The respondents were disproportionally Caucasian. Nearly three- quarters had a bachelor's degree or higher. Most considered themselves skilled Internet users.

Patients were more likely to host blogs than companions. Factor analysis was performed on 16 possible perceived benefits of blogs. Four resulting factors—(1) emotion management, (2) information sharing, (3) problem-solving, and (4) prevention and care, in that order, accounted for 69% of the variance in perceived benefit to their conditions.

There were no significant differences between patients and companions on information sharing and prevention and care, but there were significance differences between the two groups on emotion management and problem-solving. Those who perceived blogs to be credible obtained more problem-solving, prevention and care, and information sharing gratification. Those who hosted their own blogs were more likely to obtain emotional management gratification. They authors speculate that the socially interactive nature of blogs may account for the primacy of emotion management and information-sharing benefits.

The authors point out that an important limitation of their study is that they did not assess the reliability of information on the blogs. They mention possibly having a trained cancer information specialist available to "monitor (or authenticate) the information flow and educate and train patients to evaluation the credibility of the information they use."

A common thread in these first three studies is that use of the Internet to obtain information is more the province of people of relatively higher socioeconomic status. As health sciences librarians, we need to be thinking about ways we could promote the use of the Internet to obtain health-related information among segments of the population that do not currently use the Internet to obtain information.

Keselman, Browne, and Kaufman look at the "process of consumer health information seeking in the absence of a diagnosis." Their theoretical framework was a combination of an existing cognitive framework for understanding information seeking and a critical reasoning and hypothesis testing framework used in education research. The cognitive framework they used sees task performance as an iterative process in which goals are set, action is taken, results are evaluated, and new goals are set. Success is dependent on the searcher's competencies in several areas. They posit that health information seekers start with a hypothesis in mind.

Evaluation of information found is done on the basis of agreement or disagreement with the hypothesis.

Twenty lay people, male and female employees of the National Library of Medicine, of mixed races and varying levels of education, were asked to read a hypothetical scenario, discuss possible causes of the symptoms, and then search MedlinePlus while thinking aloud. All were audio-recorded and searches were tracked. Each participant's understanding of the scenario and other competences was rated.

Eight participants started by attempting to verify a particular hypothesis about the causes of the described symptoms. Five participants, all with no more than a high school education, started by searching a broad area. The remaining seven chose what the authors call a bottoms-up strategy. Hardly anyone found the right answer. The authors describe the paths taken by the participants, in detail. Incorrect or imprecise domain knowledge was a problem. There was ample evidence of all kinds of problems, particularly confirmation bias, selective perception, and premature termination of the search for evidence. The authors suggest ways problems could be addressed at the various stages of the search process. For instance, query formulation support tools could be provided to help the searcher make the query more specific and more definitions of professional terms could be provided. At places where searchers form an hypothesis, alternative hypotheses could be suggested.

Not much readability research has been done on language in the medical domain. Zeng-Treitler, Goryachev et al set out to develop a measure of medical terminology difficulty, based on context. Most readability formulas are based on word length or lists of easy and difficult terms. The "contextual network method was designed based on the assumption that difficult terms tend to occur in the context that contains other difficult terms and easy terms tend to occur in the context that contain easy terms." They designed a contextual network algorithm based on term co-occurrence. Each node in the network represents a term which is connected to co-occurring terms. Some nodes, called root terms, have a preassigned familiarity value (easy or hard); others have familiarity values which are calculated based on the network structure and root term values. The value of unknown nodes is defined as the weighted average of its neighbors.

The contextual network method was applied to 12 million MedlinePlus query logs. First the queries were mapped to UMLS terms. The hundred most frequent connections to each single node were noted, resulting in 34,710 nodes with 777,456 connections. Easy and difficult root terms were identified using methods described in the article. Then familiarity values were calculated for 23,675 unknown nodes. The results were validated by comparison with two earlier surveys of consumer familiarity with health terms, which resulted in a frequency-based predicted familiarity score. There were positive correlations between the context scores and the earlier survey scores and negative correlations between the survey score and the number of letters and syllables per word. They suggest the context-based model and the frequency-based predictive model should be used together to develop health specific readability formulas.

Leroy et al look at consumers' use of online health information and suggest yet another method of assessing readability, which takes vocabulary into consideration. After reviewing three principal ways to evaluate online health information, they describe development of an automated document classifier that can distinguish among three levels of difficulty based only on the vocabulary used in the documents.

They used an automated classification algorithm based on Bayes theorem. This approach has been used by others to identify junk mail. The classifier calculates the probability of a specific hypothesis being true, given certain evidence. Base probabilities to use as evidence for classification into the three levels of difficulty were calculated using 100 medically themed blog entries written by lay people for the easy level, 100 consumer education documents written by professionals for the intermediate level, and 50 JAMA articles for the difficult level. Two different evaluations of the classifier on this group of 250 documents showed a 98.0 and 98.4% accuracy.

They then gathered 90 online documents on melanoma, depression, and prostate cancer from commercial, government, education, and consumer group websites. Alternative and complementary sites were included. Those 90 documents were then evaluated using the Readability Analyzer developed by NLM and five readability formulas, the authors' classifier, and, manually, by a medical librarian expert and a representative consumer. The general outcome is that, for the Readability Analyzer, all pages except those originating from consumer groups were written at too high a level for the general public. The classifier found that 90% of the commercial documents and 70% of the educational documents were written at a level appropriate for the general public. The medical librarian expert considered more documents to be too difficult for consumers than the representative consumer did, perhaps indicating that the librarian underestimated consumers or that the representative consumer overestimated herself. They say "the classifier results indicate the situation may not be as bleak as generally suggested" and suggest using their classifier in conjunction with readability formulas.

Keselman, Smith et al asked this question: To what extent are the health terms used by laypeople a reflection of a different set of concepts from those of professionals?

One thousand forty-six terms from the Open Access and Collaborative Consumer Health Vocabulary were manually mapped to the 2007 UMLS Metathesaurus. They outline four possible relationships between lay and professional terms. One is an exact match of both the term and the concept to which it refers. Second is what they call a "lay synonym". In this case the concept is the same but lay people call it by a common name not normally used by professionals. Third are terms used differently by lay people and professionals. Fourth are terms that cannot be mapped to the professional vocabulary for a variety of reasons. Sixty-four terms could not be mapped the UMLS. Of these, 47 denoted concepts that could be part of professional medical discourse and 17 were lay terms.

One of their observations is that "individuals' thinking of health issues is very specific to the details of the situation." Laymen tend to think of things in reference to their own specific situation and its effect on their life. They also say people tend to use different terms when talking among themselves than when talking with health professionals. This is especially true in the areas of sexual health and wellness/beauty/physical fitness. At times professionals use the same words but ascribe different meaningsto the words.

Despite the fact that just about all consumer health terms could be mapped to medical terms, we, as health sciences librarians, know that much is lost in the translation between doctor speak and patient speak. The authors suggest that they did not consider the case of "lay usage of professional terms, when lay individuals use existing professional terms, but ascribe mean that differs from their professional definition.: They say this case "may be as common as it is difficult to investigate." They further say that lay use of "almost any health term" involves some "vagueness or alternation of meaning." Lay concepts of a term are less detailed and "concepts in lay models are likely to differ from the professional ones in their organization and relationship to one another."

On the surface, this sounds very similar to the third possible relationship cited above: terms used differently by lay people and professionals. They say "in this situation, while the lexical term string is the same, the concept is different, and this concept inheres in more than simple conceptual unsophistication." Perhaps the unexplored case is a more advanced instance of their third possible relationship.

This group of seven articles considered who uses the Internet to look for health information (Buente and Robbin), how and why people look for health information on the Internet (Yoo and Robbins), the use of blogs by patients and their companions (Chung & Kim), consumer health information seeking as hypothesis testing (Keselman, Browne, and Kaufman), and finally there are three articles on various aspects of consumer health vocabulary. The general impression is that consumers of higher socioeconomic status do look for health information on the Internet. Consum-

ers encounter a variety of difficulties and gratifications. However, as they search for health information on the Internet, the three vocabulary articles suggest that vocabulary used in consumer health materials available on the Internet may not be as big a problem as we as health sciences librarians commonly think. Edelman, Smith, et al do suggest that the problems may be more complex than the three studies presented here reveal, perhaps explaining why these results don't conform to our practiced intuition.

THE RESEARCH MENTOR

Jonathan Eldredge, MLS PhD AHIP
Associate Professor, University of New Mexico

Interview with Library Researcher Jo Dorsch

Jo Dorsch has long been recognized as a leading published researcher within MLA. Jo holds the rank of Professor at the University of Illinois at Chicago in the Library and in the Department of Medicine. She has authored 25 peer-reviewed publications and been the recipient of 20 grants and sub-contracts. In this column Jo and I discuss the "Lessons Learned" in her career in the hope that her responses will provide helpful advice for all of us.

Question: Where do you find your ideas for research projects?

Jo Dorsch: I have to look no further than my work to find ideas about research projects. Most of my publications deal with evaluation of my instruction and outreach programs. It's important to me to measure the effectiveness of what I do. My work itself is very satisfying, but measuring the results of my work and sharing it with others, adds to the value of my contributions. As a library profession we've been preaching evidence-based practice to health professionals. I feel strongly that librarians should be contributing to the evidence base of our own profession.

Question: Could you please describe one of your most challenging research obstacles, and how you eventually succeeded in overcoming it?

Jo Dorsch: My biggest challenge is that I don't have formal training in research methodology and statistics. I've overcome the obstacle by taking advantage of professional development and continuing education opportunities. I've also come to realize that you don't have to be a statistics whiz to be able to do research -- you just have to know where to look for help. On my list of goals is to take a course in research that would address both quantitative and qualitative research methods.

Question: How do you find time for research?

Jo Dorsch: I'm not sure I even have time to answer this question! Seriously, the way I find time is by tying my research closely to my work. If I have to write a grant report, why not translate that work into a research project? If I'm developing an informatics curriculum, why not think about how I'm going to know if it's effective. Another way I keep myself on task is by working
with co-authors -- it helps you stick to timetables because you don't want to let down your colleagues.

Question: How do you identify potential collaborators for your research projects?

Jo Dorsch: I enjoy the collaborative research environment of the UIC Library. UIC Library colleagues are always willing to co-author, edit manuscripts, or simply listen to ideas. Beyond the library, I have collaborated with medical faculty interested in educational outcomes of our curriculum-based EBM efforts. Almost all of my collaborators have been co-instructors or co-investigators on outreach projects.

THE RESEARCH MENTOR, continued

Question: What's the most important advice that you have for the novice researcher?

Jo Dorsch: Write about what you know and draw from your experiences within librarianship. Stop and think before you start any new initiative and ask yourself some questions! Could what I'm doing add to the evidence base of the profession? Does it have the potential to be a research project? How might I evaluate my results? How will I know if what I do has any impact? What data do I need to collect? Do I need IRB approval to collect the data I want? Who else is a stakeholder and are they interested in co-investigating? Finally, look for a mentor who is interested in research and seek an opportunity to co-author with her -- you will learn so much from the process.

RESEARCH IN PROGRESS: Relating the Recommendation of *The Research Imperative* to Current Course Offerings in ALA Accredited MLIS Programs

Michelynn McKnight, PhD, AHIP
Assistant Professor, School of Library and Information Science
Louisiana State University, mmck@lsu.edu

Carol Rain Hagy, MFA
Graduate Assistant, School of Library and Information Science
Louisiana State University, chagy1@lsu.edu

MLA's new research policy statement, The Research Imperative includes several recommendations for research methods curricula in Master of Library and Information Science (MLIS) programs. In specific it recommends that such programs "ensure that opportunities to develop quantitative and qualitative research knowledge and skills appear throughout the curriculum" and "require master's degree students to undertake a research project in information science" [1].

Most recent literature on the subject of teaching research methods to master's degree students in LIS schools consists of case reports and opinion pieces. A few years ago, Soyeon Park conducted a significant study comparing the research requirements of some MLIS programs with those of other professional master's degrees (Master of Business Administration, Master of Social Work and Master of Education) offered at the same universities. She found that these MLIS programs had few research requirements, if any. She also found an inverse relationship between the rank of the LIS school in the contemporary US News & World Report ratings and research requirements. She found that the MBA and MSW programs often required quantitative research methods courses, but the ME programs did not. Park also differentiates between education for using published research and education for conducting original research [2].

Our research questions are "Do ALA Accredited MLIS programs offer the curricular opportunities specified in The Research Imperative?" and "Do they have the research project requirement for graduation also recommended in The Research Imperative?"
We will gather evidence of the current state of elements of The Research Imperative's curricular recommendations in course descriptions and graduation requirements available through the web sites of the fifty-seven ALA accredited MLIS programs. We will gather evidence from any relevant syllabi on the web sites. We will also map The Research Imperative's basic, advanced and specialized skills to elements of the course descriptions and we will observe what texts are most often required in LIS research methods courses avail-

able to MLIS students.

Two major drivers for MLIS curricular choices are, of course, the career aspirations of the students and requirements for program accreditation by the American Library Association. MLIS graduates pursue a wide variety of careers in corporations, institutions and as individual entrepreneurs. Graduates work in many kinds of libraries, in museums, in archives and in a variety of industries, and any given MLIS program may offer specialization tracks with differing requirements. Curricular requirements may also be imposed by other organizations, such as states with course requirements for the certification of school librarians.

Unlike the Doctor of Philosophy degree, the Master of Library and Information Science degree is primarily a professional degree and not a research degree. MLIS programs are brief, usually taking full-time students two years or less to complete. Some may continue their studies in advanced certificate programs and some may go on for the PhD, usually an additional three to five years beyond the MLIS. After completing course work and passing qualifying exams, doctoral candidates must successfully propose, implement and report on independent research projects before completing the degree. Research courses offered for PhD students may not be available to master's students in the same school.

References

1. Medical Library Association. The research imperative: the research policy statement of the Medical Library Association [web document]. Chicago, IL: The Association, 2007. [cited July 5, 2008] http://mlanet.org/research/policy/policy-08.html

2. Park S. Research methods as a core competency. Journal of Education in Library and Information Science. 2003 Winter; 44(1): 17-25

DE-IDENTIFICATION OF PATIENT DATA: Results of a Pilot Project at NLM

Melissa Resnick MS, MLS
Associate Fellow, National Library of Medicine

Introduction

De-identification is defined as the removal of personal health information (PHI) from clinical records. De-identification can unlock the research potential of long term clinical records but no well-supported and freely available de-identification tools exist. The Lister Hill Center for Biomedical Communications at the National Library of Medicine has developed a tool for recognizing sensitive information such as dates, person names and locations, text, numbers, and speech and initiated an effort to develop an open source text de-identification tool for application on clinical data. This project outlines the use of a tool to de-identify patient data, describes other de-identification tools available, outlines data format and presents preliminary results of the de-identification pilot project.

The de-identification tool – not yet named but referred to internally at Lister Hill as E-Scrubber - is currently being developed and tested with a set of data from the National Institutes of Health Clinical Center from studies done at the NIH and additional data sources are being sought. These clinical results contain information collected on each patient in a clinical study. Therefore, the goal for the NLM is to develop an algorithm that automatically detects and removes any identifying information while leaving the remainder of the results intact.

Background

Clinical records have recently been considered a great wealth of information for research, such as: (1) epidemiological investigations, (2) collection of data on drug interactions, and (3) natural language processing [1]. However, these clinical records contain personal health information (PHI), making it impossible to use

as raw data research. Investigators wishing to use clinical records for research have three methods by which to ensure protection of personally identifiable patient information: (1) Obtain permission from the patients, (2) obtain a waiver of informed consent from their Institutional Review Boards (IRB), or (3) use a data set that has had all (de-identified data set) or most (limited data set) of the identifiers removed [2].

Personal Health Information

In December 2000, Department of Health and Human Services released its standards (Privacy Rule) for Privacy of PHI during financial transactions, and sale or transfer of patient data and samples to databases, repositories, and researchers [3]. Three classes of entities are covered by the Privacy Rule. Thus, they are referred to as "covered entities," and include: (1) health care providers, (2) health care plans, and (3) health clearinghouses [4]. Employers, insurers, schools, or other entities that may have health information are exempt, except when providing and billing for health services. Rothstein (2005) explained that the Privacy Rule does not require that covered entities obtain patient consent for treatment, payment, and other health care operations. However, it does require that they provide a copy of their notice of privacy [5].

In addition to defining the protection of PHI during transactions involving patient data, the Privacy Rule defines PHI. PHI is individually identifiable health information [6]. The Privacy Rule lists 18 specific identifiers [7]. These identifiers include, but are not limited to: names; addresses; dates; telephone numbers; fax numbers; e- mail addresses; social security numbers; clinical record numbers; health plan beneficiary numbers; account numbers; certificate/license numbers; vehicle identifiers and serial numbers, including license plate numbers; device identifiers and serial numbers; URL's; IP address numbers; and biometric identifiers, including finger and voice prints, full-face photographic images, and any comparable images (45 CFR 164.514(b)(2)).

In cases in which covered entities perform transactions on patient data for research, the Privacy Rule requires that they protect PHI in one of two ways: (1) informed consent, and (2) de-identification, which will be defined later [8].

De-Identification

"The Privacy Rule protects PHI, but not the data and samples that are excluded by the very definition of PHI," [9]. Covered entities and other researchers only need to obtain patient consent when using individually personal identifiable information. However, they can always use and provide others with any patient data without prior authorization, as long as no PHI is present. In these cases, all PHI must be removed from the data. The process of de-identification makes this possible. De-identification is defined as the removal of personal health information (PHI) from clinical records. This can be performed in two different ways. The researcher can elect to remove PHI manually. This involves reading the clinical records and removing PHI. Dorr, Phillips, Phansalkar, Sims, and Hurdle (2006) note that manual de-identification of clinical records is a difficult and time-consuming task. On the other hand, the researcher can use a computer program to perform the de-identification process. A list of currently available De-Identification programs are shown in the table below.

De-Identification Programs

Name	Source
Scrubber	MIT/Harvard
Share Pathology Informatics Network (SPIN)	Regenstrief Institute in conjunction with the National Cancer Institute for the Indiana Network for Patient Care
De-id (region-specific to Pennsylvania)	University of Pittsburgh Medical Center

Formats

Most of the clinical notes received from the NIH Clinical Center were in HL7 or Health Level-7 format. Health Level-7 is a standard protocol which allows transfer of clinical records between computers in the same system or between computers in different systems. This format consists of messages, with each message con-

taining several segments, and each segment containing several fields. Each message is one clinical record. The segments contain different parts of the clinical record. For instance, the PID segment contains patient identification including but not limited to: (1) patient name, (2) date/time of birth, (3) gender, and (4) patient address [10]. To help with tagging PHI, two extra fields were placed in each line of information.

The second important format is the one that NLM uses for this project. In this format there are five different fields separated by vertical lines or pipes followed by a dollar sign. The first and third fields contain numbers for the computer programmer. The second field contains the information from the clinical record. Each line contains one word and/or punctuation. The fourth and fifth fields are used for tagging the PHI and making notes, respectively. The dollar sign delineates the end of each set of fields. Finally, an "end of sentence (EOS)" "beginning of sentence (BOS)" set of notations is placed after each semicolon, colon, and period.

Methods

The goals of the project were to: (1) identify and tag data fields and data elements that contain information that must be removed in order to comply with Health Insurance Portability and Accountability Act (HIPAA), (2) identify and tag possible strings of letters, words, and numbers that could possibly be personal identifiers. This review and tagging of the data is intended to ensure that as many as possible of the personal identifiers can be accounted for and programmed into the algorithm, and (3) create a gold standard for de-identification of documents that will be used to test the computer program under development.

In the first part of the project, the segments of the HL7 format of the clinical records were reviewed and all identified PHI tagged. This was accomplished by: (1) identifying one of the eighteen items of PHI defined by HIPAA, (2) placing a number one ("1") in the first extra empty field, and (3) a note in the second extra empty field describing the type of PHI. For instance, for the name "John," a number one (1) was placed in the first extra empty field. Then a note such as "name" was placed in the second extra empty field.

In the second part of the project, a data file containing last names was prepared for use. This name file was obtained from the NIH list of staff. To use the file, items such as addresses, (e-mail, IP, and other), names of institutions, and duplicate names had to be removed. The data file was first placed into an Access database. Once this was accomplished, all of the duplicate names were removed. Next, the file was placed into an Excel spreadsheet so that each item occupied the first cell of each row. Finally, the Excel file was reviewed, removing all unwanted items, leaving a clean spreadsheet containing all last names.

During the final part of the project, additional clinical records were reformatted into the NLM format. Data were reviewed and tagged in the same manner as were the records in HL7 format.

Results

Over a period of five months, 82,769 lines (about 4 megabytes of data) or 308 clinical records have been reviewed. Of these, 25 accession number, 49 account number, 316 address, 42 age, 50 city, 11 country, 1462 date, 2146 date/time, 49 date of birth, 1897 id number, 334 institution name, 189 location, 1869 name, 375 nonspecific number, 255 phone number, 44 prefix, 850 protocol number, 48 social security number, 39 state, 865 suffix, 653 time, 4 year, and 42 zip code elements of PHI have been identified and tagged.

Discussion and Conclusion

Even though these results only show a small portion of this project, these data do show the need for an automated de-identification program. It is time-consuming, tedious, and costly to manually remove PHI. The author worked full-time on the project for four months and tagged a total of 308 clinical records. The team would need a total of 10,000 tagged clinical records to determine the error rate of the system. To obtain this number, there would need to be many more staff devoted to the effort. Since this is not possible, the team will use the data that they have and use a mathematical process of extrapolation to determine this error rate. The next tasks include: (1) manually tagging more clinical data, (2) refining the algorithm, (3) testing the program, and (4) releasing the finished product. Manually

tagging as much data as possible gives the team more data to use in refining the algorithm and testing the program in order to ensure that the program has the lowest error rate possible. After all possible refining and testing, E-Scrubber will be made available on the Internet as an open-source de-identification program.

References

1. Wellner B, Huyck M, Mardis S, Hirschman L. Rapidly Retargetable Approaches to De-identification in Medical Records. *J Am Med Inform Assoc.* 2007 Sep-Oct;14(5):564-73.

2. Xiong L, Boronda K, Flowers C, Graiser M. De-identification of Medical Text. [Internet]. Emory University Department of Mathematics and Computer Science. 2007. [cited 2008 Aug 15]. Technical Report TR-2007-012-A. [19p]. Available from: http:www.mathcs.emory.edu/uploaded-files/RPT-#jjaji.pdf.

3. Goldstein G, Gordon J. Don't Ask, Don't Tell? Transfer and Sale of De-Identified Patient Data. *Journal of Clinical Research Best Practices* [Internet]. 2008 Feb [cited 2008 Aug 15];4(4)[5p]. Available from: http:www.firstclinical.com/journal/#bjjh/jhjd_HIPAA.pdf.

4. Rothstein M. "Research Privacy Under HIPAA and the Common Rule. *The Journal of Law, Medicine and Ethics.* 2005 April;33:154-58.

5. Ibid.

6. Ibid.

7. Goldstein G.; Krishna K, Kelleher K, Stahlberg E. Patient Confidentiality in the Research Use of Clinical Medical Databases. *Am J Public Health.* 2007 April;97(4):654-58.; Rothstein M.

8. Goldstein G.

9. Ibid.

10. Hoekman, T. Specifications for Version 2.3 of the Health Level Seven (HL7) Standard for Electronic Data Exchange in Healthcare Environments, with Special Emphasis on Inpatient Acute Care Facilities. Ann Arbor, MI: Health Level Seven, Inc.; c. 1997. Available from: http://www.med.mun.ca/tedhoekman/medinfo/hl#g/httoc.htm.

Other Resources

1. Dorr D, Phillips W, Phansalkar S, Sims S, Hurdle J. Assessing the Difficulty and Time Cost of De-identification in Clinical Narratives. Inf Med. 2006;45:246-52.

This research was supported in part by an appointment to the NLM Associate Fellowship Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education.

Acknowledgements

Allen Browne, Guy Divita of the Lister Hill National Center for Biomedical Communication proposed the project and served as mentors and guides on the work.

Fifth International EBLIP Conference in 2009

The program planners for the Fifth International Evidence Based Library and Information Practice (EBLIP 5) Conference have been developing the program over the past few months. MLA Research Section members might wish to submit research papers, reports of innovative practices, or hot topic discussion papers by the October 31, 2008 deadline.

The central theme of "Bridging the Gap" for EBLIP 5 includes the three dimensions of the Who, the What, and the How for making EBLIP a reality. These three dimensions are summarized below:

Bridging the gap……the Who:
between disciplines, between sectors, and between cultures.

Bridging the gap……the What:
the implementation gap, expectation gap, experimentation gap and the skills gap.

Bridging the gap……the How:
Using marketing/advocacy, data mining & management tools, collaborative working, communication, management support & leadership, and through international co-operation.

Beyond these broad themes, MLA Research Section members should consider submitting any applied research project report with relevance to library or information practice by October 31st. The author has noted the great flexibility in programming in past EBLIP conferences so members are encouraged to think broadly when designing their submissions.

The EBLIP 5 website provides details about the conference at: http://eblip5.kib.ki.se. Readers should feel free to contact the author at jeldredge@salud.unm.edu if they have any specific questions about EBLIP 5, or if they would like to receive individualized advice on how to make most of this exciting international conference.

Submitted by Jon Eldredge

HYPOTHESIS

The Journal of the Research Section of MLA