

Accepted to the Doctoral Forum of the Fourth International Conference on Conceptions of Library and Information Science: Emerging Frameworks and Methods, Seattle, July 2002.

## **Information Synthesis: A Process used by Scientists in Medicine and Public Health to Overcome Information Overload**

Catherine Blake  
Dept Information and Computer Science  
444 Computer Science Building  
University of California, Irvine  
Irvine CA, 92697-3425  
Phone: (949) 824 8169  
FAX: (949) 824 4056  
cblake@ics.uci.edu

### **Abstract**

As the amount of peer-reviewed articles in public health and biomedicine continues to soar, scientists struggle to keep up new findings, even in narrow areas of expertise. Satisfied that an information retrieval system can identify a set of relevant articles; scientists now face the challenge of how to aggregate information from within the retrieved set.

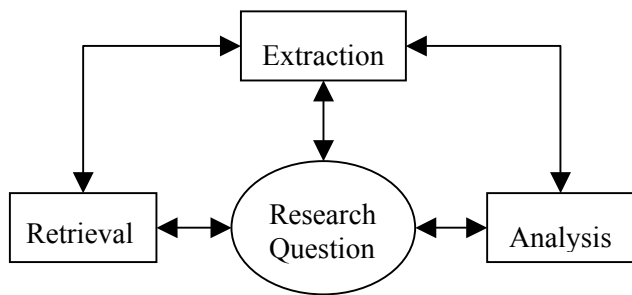
My dissertation will explore (1) how scientists in medicine and public health currently use biomedical literature in their work, specifically to identify new treatments or risk factors associated with a medical condition (2) the information requirements implied by their work practices, and (3) the development of computer systems that provide the required functionality.

I have observed that scientists in medicine and public health follow a process that I call information synthesis (IS). This process can be characterized by three distinct phases: collecting medical literature, extracting information from that literature and analyzing the extracted facts. The goal of my dissertation is to enable scientists to conduct this process faster and more comprehensively than current manual approaches allow. My semi-automated approach will also enable scientists to reduce the effect of publication bias by incorporating a broader search strategy. My work complements existing retrieval systems by focusing on the extraction and integration phases of the IS process.

I will demonstrate my approach by exploring the relationship between smoking and breast cancer. This relationship is important to scientists in public health because the currently known risk factors of breast cancer explain only 50% of the occurrences of the disease. Underpinning my approach is a modular design to support scientists in their exploration of other risk factors that may be associated with breast cancer.

## Introduction

Health and biomedical scientists have a long tradition of using peer-reviewed literature in their quest to answer research questions. When studies differ in their conclusions or sample sizes are small, scientists use a rigorous non-biased methodology to collect relevant articles, extract information from those articles and integrate the results. I call this process information synthesis. As displayed in figure 1, information synthesis is characterized by three phases: retrieval, extraction, and analysis.



**Figure 1 – The Systematic Review process.** Scientists collect articles, extract information from those articles, and analyze the extracted information. The result of each phase enables scientists to refine search criteria and extraction requirements. The research question is also refined when there is insufficient information available.

Identifying risk factors of a medical condition is an important problem in public health. Approximately 50% of people who get breast cancer have none of the known risk factors associated with the disease (Vincent, Samuel, & Steven, 2001). My dissertation will explore the development of computer tools that will enable scientists to explore potential risk factors. I will demonstrate my approach by exploring the relationship between breast cancer and the risk factor smoking. If successful this approach will: (1) reduce the amount of time to conduct a meta-analysis of potential risk-factors that are reported in scientific text related to breast cancer; (2) enable scientists to include more articles in a meta-analysis by removing the manual effort involved in extracting the information required; and (3) reduce the influence of publication bias when conducting an epidemiological meta-analysis.

## Background and Significance

Researchers in information science have demonstrated that the medical literature contains implicit connections that are useful when exploring new treatments strategies (Swanson, 1988; Swanson & Smalheiser, 1997). For example, Swanson identified eleven implicit connections in the literature, which suggested magnesium as a candidate treatment for patients suffering from migraines. Clinical trials later verified this hypothesis. In each of the seven new treatments identified by Swanson, connections were implicit in the literature.

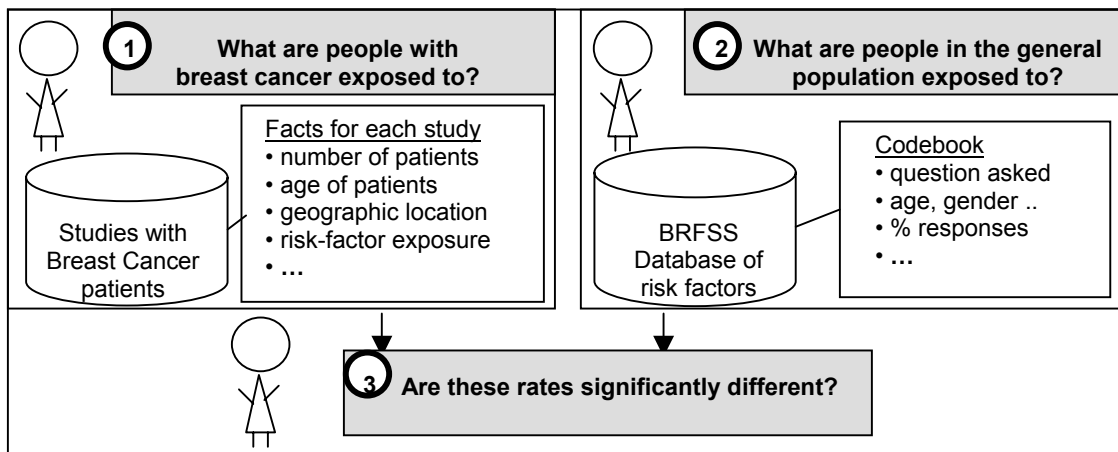
Although Swanson and his system identified high-level logical connections, scientists in medicine and public health often use precise information extracted from articles. They then integrate the extracted information using a statistical approach called meta-analysis (DerSimonian & Laird, 1986; Ingelfinger, Mosteller, Thibodeau, & Ware, 1994; Petitti, 2000). Critical to a meta-analysis is the articles initially selected. Current searching strategies employed by public health scientists often include both the medical condition and the risk factor. For example, a recent study that explored the relationship between breast cancer and alcohol used 42 published articles where 41 articles contained the term *alcohol* or a synonym in the title or abstract (Ellison, Zhang, McLennan, & Rothman, 2001). The search strategy used in this analysis did not include articles where the alcohol consumption was not the primary focus of the study. This means that publication bias (when results are published more often when a positive correlation is found

(Easterbrook, Berlin, Gopalan, & Matthews, 1991)) may be responsible for slightly elevated risk found by Ellison and his colleagues.

One approach to reduce the effect of publication bias is to consider all articles related to the disease, and then extract information related to the risk factor. Scientists in public health used such an approach to explore the relationship between smoking and impotence (Tengs & Osgood, 2001). Specifically they extracted the smoking rate of men in impotence studies and compared the rate with the smoking rate of a control population using the Behavioral Risk Factors Surveillance System (BRFSS). They found that men in impotence studies were significantly more likely to smoke than those in a population who were similar with respect to age, geographical location and time-period. The manual process to identify impotence articles that reported smoking from the 1008 original articles took two person months; it took an additional three hours per study to extract the required information (personal communication, N. Williams and T. Tengs, 2002). A similar selection criterion would identify approximately 70,000 articles related to breast cancer. Assuming that (i) it would take the same amount of time to identify breast cancer articles which reported smoking as impotence articles, and (ii) that UCI's medical and science libraries contain the same proportion of breast cancer articles as impotence articles, it would take 11.5 person years to identify the breast cancer articles that reported smoking and 1.8 person years to extract the information from these studies. Thus, the manual approach would take approximately 13.3 person years to complete.

### Approach

I will construct a computer system that extracts secondary information from published scientific studies to enable scientists to explore risk factors associated with breast cancer faster and more comprehensively than current manual techniques allow (this approach is outlined in figure 2.) This approach is based on a previous user study (Blake & Pratt, 2002) that indicated that scientists use a collaborative, iterative process throughout IS. The extraction techniques are based on a knowledge-based approach, which has been shown to improve precision when applied to the migraine-magnesium connections identified by Swanson (Blake & Pratt, 2002).



**Figure 2. Functionality Overview.** The system will extract secondary information, such as risk factor exposure, age and location from scientific studies related to breast cancer. After manual verification and manipulation of extracted facts, the system will identify a control population from the BRFSS and compare the rates from (1) and (2) using a random effects meta-analysis (3).

## Motivation to attend Doctoral Forum

As the amount of information available to scientists continues to increase, new methods of incorporating information into their work practices will become essential. I would like to attend the doctoral forum to explore research questions such as: How do scientists in medicine and public health use biomedical literature in their research? How do their work practices differ from other highly trained user groups who have specific information needs? What other services should digital libraries provide to expert users?

The information and computer sciences community has traditionally treated structured and unstructured data sources separately. I am also interested exploring how heterogeneous information sources can be integrated to provide users with a cohesive summary of the available information related to research question.

## References

- Blake, C., & Pratt, W. (2002). *Automatically Identifying Candidate Treatments from Existing Medical Literature*. Paper presented at the AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases, Stanford, California.
- Blake, C., & Pratt, W. (2002). *Collaborative Information Synthesis* (UCI Tech Report #02-04): Dept Info & Comp Sci, UC Irvine.
- DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7, 177-188.
- Easterbrook, P. J., Berlin, J. A., Gopalan, R., & Matthews, D. R. (1991). Publication bias in research. *Lancet*, 337, 867-872.
- Ellison, R. C., Zhang, Y., McLennan, C. E., & Rothman, K. J. (2001). Exploring the Relation of Alcohol Consumption to Risk of Breast Cancer. *American Journal of Epidemiology*, 154(8), 740-747.
- Ingelfinger, J. A., Mosteller, F., Thibodeau, L. A., & Ware, J. H. (1994). *Biostatistics in clinical medicine* (3rd ed.): McGraw-Hill Inc.
- Petitti, D. B. (2000). *Meta-Analysis, Decision Analysis and Cost-Effectiveness Analysis Methods for Quantitative Synthesis in Medicine* (2nd ed. Vol. 31). New York: Oxford University Press.
- Swanson, D. R. (1988). Migraine and magnesium: eleven neglected connections. *Perspect. Biol. Med.*, 31, 526-557.
- Swanson, D. R., & Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91, 183-203.
- Tengs, T., & Osgood, N. D. (2001). The link between smoking and Impotence : Two Decades of Evidence. *Preventive Medicine*, 32(6), 447-452.
- Vincent, T. D., Jr., Samuel, H., & Steven, A. R. (2001). *Cancer, Principles and Practice of Oncology* (6th ed.). Philadelphia: Lippincott, Williams & Wilkins.