

A Text Mining Approach to Detect Breast Cancer Risk Factors



Catherine Blake, Ph.D
cablake@email.unc.edu

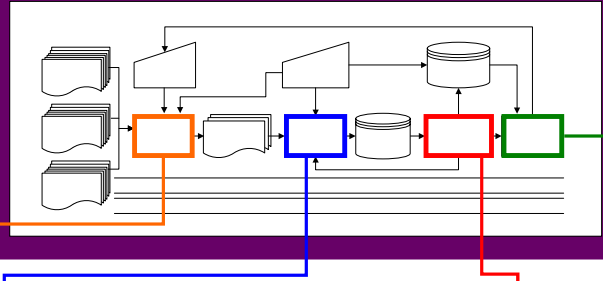
This research was supported by a dissertation award from the California Breast Cancer Research Program of the University of California #8GB-0175.

Background

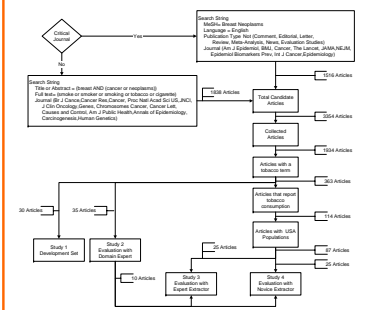
- Breast Cancer Risk factors are not well understood. Other than age and gender current risk factors explain only half of the breast cancer occurrence (DeVita, Hellman and Rosenberg, 2001)
 - MEDLINE currently comprises more than 115500 breast cancer articles¹. Each year, staff from the National Library of Medicine add 5400 new breast cancer articles to the MEDLINE corpus².
 - Premise** : Existing scientific literature captures candidate risk factors for breast cancer that should be further explored.
 - Hypothesis** : Automated techniques that synthesize typically unused evidence from scientific literature will enable scientists to identify new candidate breast cancer risk factors.
- 1 Pubmed query "breast neoplasms"[MH] issued August 30, 2004
2 Average number of breast cancer articles published each year in the last decade.

Information Synthesis

- Developed from a rich collection of qualitative data including interviews, observations, and information artifacts. User groups comprised experts in medicine and public health.
- Information Synthesis process comprises: (1) four critical tasks – retrieval, extraction, verification and analysis, (2) two user provided information constructs – hypothesis projection and context information, and (3) two process level behaviors – iteration and collaboration (Blake and Pratt, 2002) (Blake, 2003).
- The Multi-User Extraction for Information Synthesis (METIS) automates critical tasks within the Information Synthesis process.



METIS Retriever



METIS Information Extractor

- Semantic grammar developed from 30 breast cancer articles.
- Unified Medical Language System (UMLS) used to create extraction rules that generalize.
- Nineteen facts extracted from each article including
 - Number of subjects with the medical condition
 - Geographical location of study
 - Level of risk factor exposure
 - Age of subjects in the study (minimum, maximum, mean, and standard deviation)
 - Start and end date of the study

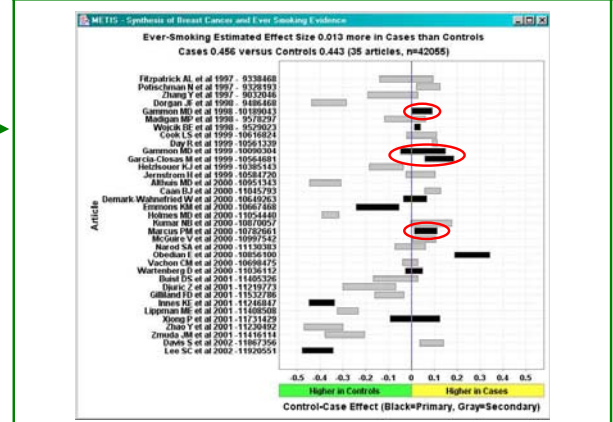
METIS Verifier

The screenshot shows the METIS Verifier interface with a text document. Callouts indicate 'Keyword in context', 'Electronic version of the article', and 'Add, remove, or update proposed facts'.

Findings and Future Work

- Users iterate and collaborate within and amongst the retrieval, extraction, verification and analysis tasks
- Users require the full-text of an article to collect all of the information required for their analysis.
- Automating process is possible. Analysis of tobacco and alcohol consumption on small of breast cancer corpus showed consistency with known attributable risk.
- Further research is required to evaluate new risk factors.

METIS Analyzer



Thirty-five articles capture tobacco consumption rates of unique breast cancer populations. Of those, twelve articles (shown in black) report consumption rates as primary information and four articles (circled) would be included in a traditional analysis.

References

Blake, C. & Pratt, W. (2002). Collaborative Information Synthesis. In *Proceedings of Annual Conference of the American Society for Information Science and Technology (ASIST 2002)*, Philadelphia, PA.
 Blake, C. (2008) Information Synthesis: A Mixed-Initiative Meta-Analytic Approach to Facilitate Knowledge Discovery from Scientific Text. *Doctoral Dissertation from the School of Informatics and Computer Science, University of California, Irvine.*
 Vincent TD, Samuel H, Steven AR (2001). *Cancer, Principles and Practice of Oncology*. Lippincott, Williams & Wilkins, Philadelphia.

Breast cancer corpus of ~2000 articles collected using a combination of automated and manual techniques.