Data Preparation for Biomedical Knowledge Domain Visualization:

A Probabilistic Record Linkage and Information Fusion Approach to Citation Data

A Thesis

Submitted to the Faculty

of

Drexel University

by

Marie B. Synnestvedt

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

December 2007

Acknowledgments

Completing a dissertation is like raising a child in that it takes a village to do it. This task would not have been completed without years of support from an extended village of family and friends, academic committee and colleagues, supervisors and co-workers. It would not have been possible to do this without the support of my husband and daughter, who have had to live with a doctoral student and all that entails for seven long years. I am also indebted to the dissertation committee who provided guidance in this work:

Xia Lin, Kate McCain, and Chaomei Chen from Drexel University

Nancy Roderer from The Johns Hopkins University

John Holmes from The University of Pennsylvania

Additional thanks go to Kate for helping me get an early start on the program, to Chaomei for supporting student use of his wonderful CiteSpace program for research, and to John for setting the example to follow and giving me my mantra.

There are many others who have provided support in different forms, giving everything from emotional support to job flexibility. My heart-felt thanks to all of you, and especially to Steve and Brenna, for helping me make it through to this day.

Table of Contents

List of Tables

List of Tables

List of Tables

List of Figures

List of Figures

Abstract

Data Preparation for Biomedical Knowledge Domain Visualization:
A Probabilistic Record Linkage and Information Fusion Approach to Citation Data

Marie B Synnestvedt
Xia Lin Ph.D.

This thesis presents a methodology of data preparation with probabilistic record linkage and information fusion for improving and enriching information visualizations of biomedical citation data. The problem of record linkage of citation databases where only non-unique identifiers such as author names and document titles are available as common identifiers to be linked was investigated. This problem in citation data parallels problems in clinical data and Knowledge Discovery in Databases (KDD) methods from clinical data mining are evaluated. Probabilistic and deterministic (exact-match) record linkage models were developed and compared through the use of a gold standard or truth dataset. Empirical comparison with ROC analysis of record linkage models showed a significant difference (p=.000) in performance of a probabilistic model over deterministic models. The methodology was evaluated with probabilistic linkage of records from the Web of Science, Medline, and CINAHL citation databases in the knowledge domains of medical informatics, HIV/AIDS, and nursing informatics. Data quality metrics for datasets prepared with probabilistic record linkage and information fusion showed improvement in completeness of key variables and reduction in sample bias. The resulting visualizations offered a richer information space for users through an increase in terms entering the visualization. The significant contributions of this work include the development of a novel model of probabilistic record linkage for biomedical citation databases which improves upon existing deterministic models. In addition a methodology for improving and enriching knowledge domain visualizations though a data preparation approach has been validated with analyses of multiple citation databases and knowledge domains. The data preparation

methodology of probabilistic record linkage with information fusion offers a remedy for data quality problems, and the opportunity to enrich visualizations with added content for user exploration, which in turn improves the utility of knowledge domain visualizations as a medium for assessing available evidence and forming hypotheses.

CHAPTER 1: INTRODUCTION

1.1 Background

1.1.1 The Parallel Problems of Citation Databases and Clinical Data Warehouses

This thesis develops a data-centered approach to improving visualizations of citation data through transfer of record linkage theory and methodology as used in the Knowledge Discovery in Databases (KDD) domain to Knowledge Domain Visualization (KDViz). The thesis is motivated by personal observations of problems with data quality encountered in prior knowledge domain visualization research using citation data drawn from bibliographic databases. A post-study analysis of a progressive knowledge domain visualization of medical informatics (Synnestvedt, et al 2005) revealed data that were previously thought to be representative of a forty-year time-period were in reality incomplete due to patterns of systematically missing data. The following figures show the correlation of missing abstract (Figure 1.1) and keyword (Figure 1.2) variables with publication year in a Web of Science (WOS) (Web of Science, 2007) dataset compared to a Medline dataset. The WOS dataset has a longer time period of missing abstracts compared to the Medline dataset, and low availability of keywords while MeSH term availability is at or near 100% complete throughout the entire time period. This is a problem because in the progressive knowledge domain visualization analysis process research front terms are determined by the sharp growth rate of their frequencies and the those front terms are derived from n-grams, or single words or phrases of up to four words, from titles, abstracts, descriptors, and identifiers of citing articles in WOS data (Chen, 2006). If there are data anomalies, i.e. systematic patterns of missing data by publication date or specialty within a knowledge domain, this could lead to a biased analysis. There is a correlation with year of publication in the pattern of missing data in the WOS data, and because of this anomaly any visualizations of this data may mislead a user in the cognitive process of mentally modeling the knowledge domain. While the issues of anomalous data in information visualization may not be recognized in the literature, a case for addressing the problem can be made by drawing parallels to the problems of data mining research with medical data drawn from clinical data warehouses.

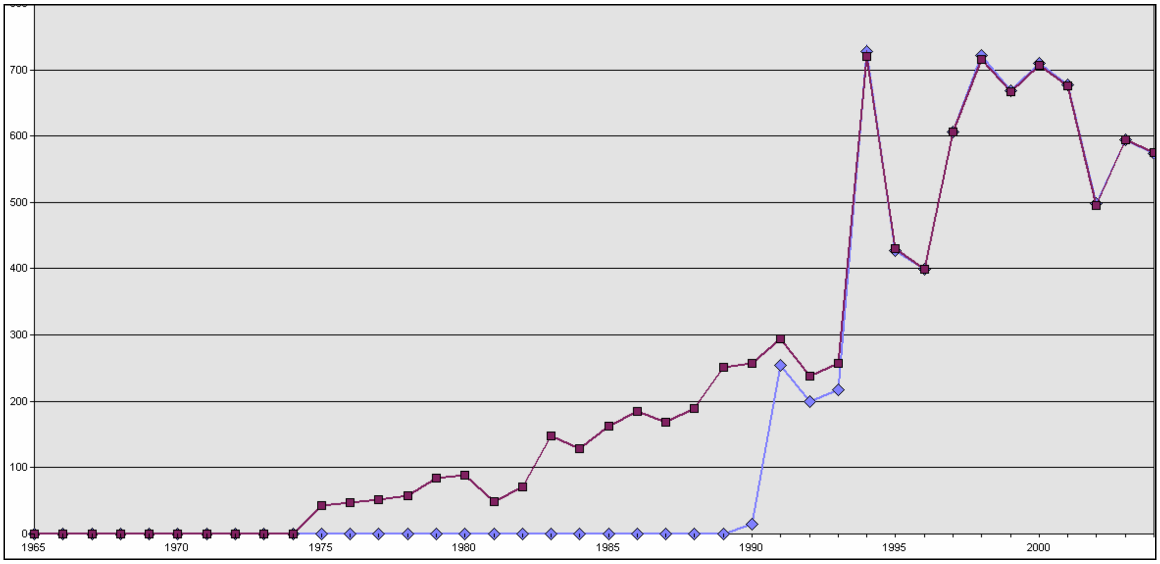


Figure 1.1 Documents With Abstracts By Year Of Publication, WOS (▲) Versus Medline (-■-)

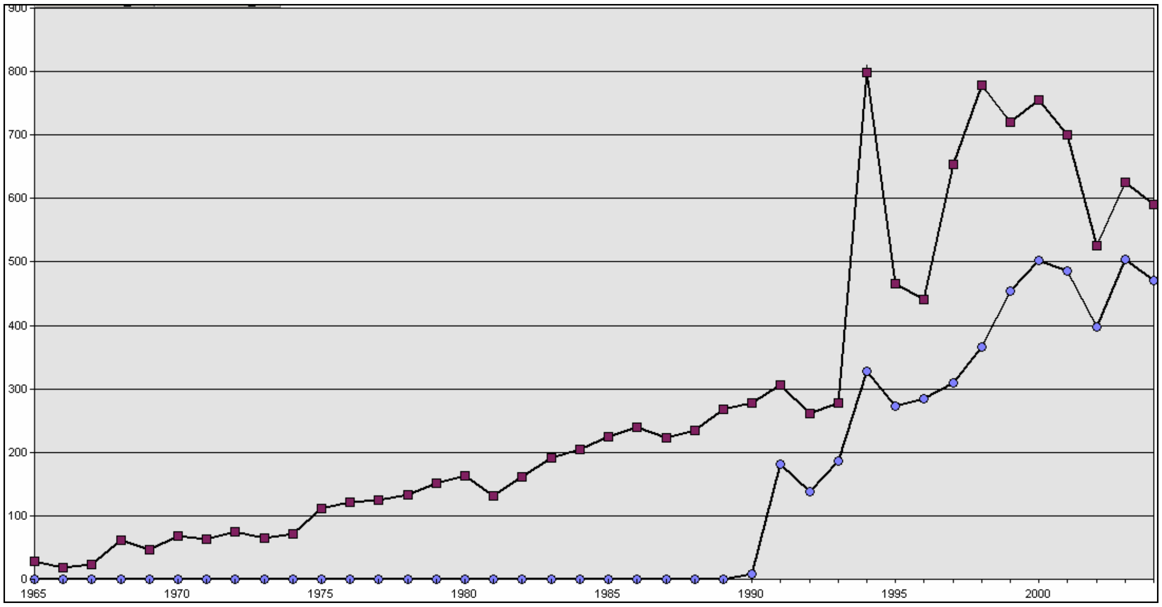
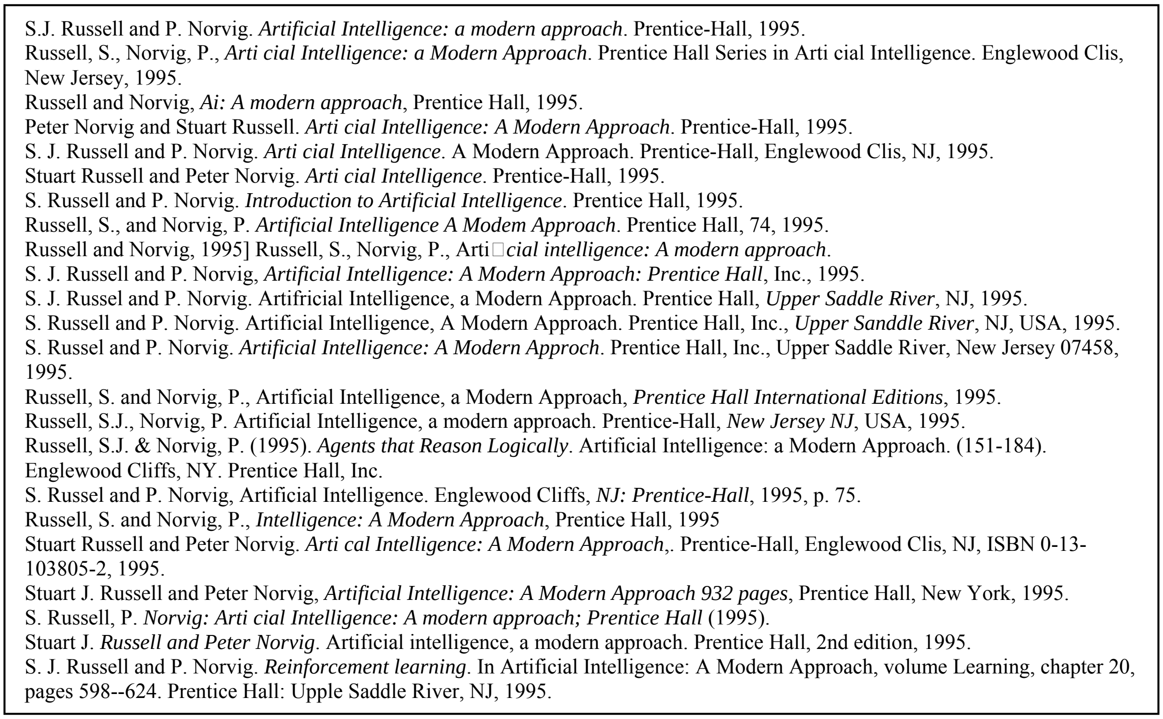


Figure 1.2 Documents With Key Words By Year Of Publication, WOS (●) Versus Medline (-■-)

There are strong parallels between citation databases and clinical data warehouses and several arguments can be made for the extension of a KDD approach to KDViz, including the similar objectives of KDD and KDViz analyses, the quality of data sources, and the potential benefits. Because the terminology of KDD and data mining (DM) are used interchangeably and have sometimes confused and overlapping definitions (Trybula, 1998) the definitions used for making this premise must first be established. The term data mining has at times been used derisively to describe questionable data analysis techniques used to misrepresent results of observational and experimental studies. That is not the meaning of a KDD approach to data analysis. A critical distinction to be made is that the data frequently come from transactional databases, i.e. the information was collected for other reasons than analysis, and the analyses seek to generate rather than confirm a hypothesis (Hobbs, 2001). KDD is usually a retrospective analysis of observational data and does not involve consideration of experimental design and related concepts (Smyth, 2000). Data mining has been described as an interdisciplinary approach which combines machine learning, statistical, and visualization techniques to gain insight into relationships and patterns hidden in data (Zupan, 1999). Han and Kamber (2001) define KDD as the automated or convenient extraction of patterns representing knowledge explicitly stored in large databases, data warehouses, or other large repositories. An emphasis is also placed on the knowledge discovery aspect of extracting unpredicted or previously unknown relationships or patterns (Trybula, 1998). The process of evaluating data, analyzing patterns, and extracting knowledge is analogous to the sorting, cleaning, and grading process involved in mining minerals. Data extracted and compiled from a repository becomes information, which is then developed into a collection of related inferences, then becoming knowledge. The extraction process is an iterative sequence of data cleaning, data integration, relevant data selection, data transformation, development of extracted patterns, and pattern evaluation (Han and Kamber, 2001). The knowledge discovery process is applied to explain existing data, make predictions or classifications, or summarize contents of large databases to support decision making (Babic, 1999).

Progressive knowledge domain visualization is an example of an analytic approach to citation data with objectives that parallel those of KDD. Visualizations in general are a medium for finding causality, forming hypotheses, and assessing available evidence through an exploratory process (Chen, 2006). As we have reported in a previous case study of a progressive knowledge domain visualization approach to analyses of the domain of medical informatics (Synnestvedt, et al 2005), the CiteSpace II application combines information visualization methods, bibliometrics, and data mining algorithms in an interactive visualization tool for extraction of patterns in citation data (Chen, 2006). Highly cited and pivotal documents, areas of specialization within a knowledge domain, and emergence of research topics are mapped for discovery through visual pattern recognition. The primary sources of data for CiteSpace analyses are the ISI Web of Science (WOS) citation databases, and a secondary source is the National Library of Medicine's Medline citation database via the PubMed system. The two data sources must be analyzed separately. The major distinction between the two sources of data from an analytic perspective is the availability of citation rate and cited reference data from WOS, and the availability of medical subject headings (MeSH) from Medline. Citation rates and cited references are the key to identifying pivotal documents and trends, and MeSH terms are useful for organizing documents by subject content according to a controlled vocabulary that is familiar and relevant to the medical community.

When viewed from a KDD perspective, the data drawn from citation databases can be characterized as having data quality issues as do the data from clinical data repositories. One of the challenges of working with clinical data repositories typically used in data mining is that real world data tend to be dirty, incomplete, noisy, and inconsistent (Hernandez & Stolfo, 1998; Han & Kamber, 2001). Citation data have characteristics that fit with this description of real world data. Garfield (1972) found that the inconsistency with which different authors abbreviate journal titles in references was an "immensely irksome problem". A recent description of citation references is that "they appear in many formats and are rife with errors of all kinds" (Pasula, 2003). Systems such as CiteSeer (Lawrence, 1999) were specifically designed to address the problem of matching variant citation formats, and an example of the variability is the reported finding by Pasula (2003) in CiteSeer of more then 100 distinct references from roughly 1000 citations to an AI textbook published by Russel and Norvig. A current search in CiteSeer for "(russell or russel) and norvig" found 329 citations with over 40 variations in citation format to the same 1995 book (Figure 1.3).



S.J. Russell and P. Norvig. *Artificial Intelligence: a modern approach*. Prentice-Hall, 1995.
Russell, S., Norvig, P., *Arti cial Intelligence: a Modern Approach*. Prentice Hall Series in Arti cial Intelligence. Englewood Clis, New Jersey, 1995.
Russell and Norvig, *Ai: A modern approach*, Prentice Hall, 1995.
Peter Norvig and Stuart Russell. *Arti cial Intelligence: A Modern Approach*. Prentice-Hall, 1995.
S. J. Russell and P. Norvig. *Arti cial Intelligence*. A Modern Approach. Prentice-Hall, Englewood Clis, NJ, 1995.
Stuart Russell and Peter Norvig. *Arti cial Intelligence*. Prentice-Hall, 1995.
S. Russell and P. Norvig. *Introduction to Artificial Intelligence*. Prentice Hall, 1995.
Russell, S., and Norvig, P. *Artificial Intelligence A Modem Approach*. Prentice Hall, 74, 1995.
Russell and Norvig, 1995] Russell, S., Norvig, P., Arti□*cial intelligence: A modern approach*.
S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach: Prentice Hall*, Inc., 1995.
S. J. Russel and P. Norvig. Artifricial Intelligence, a Modern Approach. Prentice Hall, *Upper Saddle River*, NJ, 1995.
S. Russell and P. Norvig. Artificial Intelligence, A Modern Approach. Prentice Hall, Inc., *Upper Sanddle River*, NJ, USA, 1995.
S. Russel and P. Norvig. *Artificial Intelligence: A Modern Approch*. Prentice Hall, Inc., Upper Saddle River, New Jersey 07458, 1995.
Russell, S. and Norvig, P., Artificial Intelligence, a Modern Approach, *Prentice Hall International Editions*, 1995.
Russell, S.J., Norvig, P. Artificial Intelligence, a modern approach. Prentice-Hall, *New Jersey NJ*, USA, 1995.
Russell, S.J. & Norvig, P. (1995). *Agents that Reason Logically*. Artificial Intelligence: a Modern Approach. (151-184). Englewood Cliffs, NY. Prentice Hall, Inc.
S. Russel and P. Norvig, Artificial Intelligence. Englewood Cliffs, *NJ: Prentice-Hall*, 1995, p. 75.
Russell, S. and Norvig, P., *Intelligence: A Modern Approach*, Prentice Hall, 1995
Stuart Russell and Peter Norvig. *Arti cal Intelligence: A Modern Approach*,. Prentice-Hall, Englewood Clis, NJ, ISBN 0-13-103805-2, 1995.
Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach 932 pages*, Prentice Hall, New York, 1995.
S. Russell, P. *Norvig: Arti cial Intelligence: A modern approach; Prentice Hall* (1995).
Stuart J. *Russell and Peter Norvig*. Artificial intelligence, a modern approach. Prentice Hall, 2nd edition, 1995.
S. J. Russell and P. Norvig. *Reinforcement learning*. In Artificial Intelligence: A Modern Approach, volume Learning, chapter 20, pages 598--624. Prentice Hall: Upple Saddle River, NJ, 1995.

Figure 1.3 Examples Of Forty Variations In Citations To The Same Book

The more standardized structural format of citation data available from citation databases reduces but does not eliminate the data quality problem. For example, Figure1.4 shows variations in citations to conference proceedings from a Web of Science (WOS) dataset used in a progressive knowledge domain visualization (PKDViz) study of the domain of medical informatics (Synnestvedt, et al 2005). In the context of citation analysis methods such as

PKDViz, this variability or noise in references will lead to an underestimation of citation and co-citation counts, and can result in a need to adjust the visualizations through a post-hoc process of aliasing.

```
1996 AMIA ANN FALL S
1997 AMIA ANN FALL S
AMIA ANN FALL S
AMIA P
IN PRESS JAMIA
IN PRESS P AMIA FALL
JAMIA S S
P 1996 AMIA FALL S
P 1997 AMIA ANN FALL
P 1998 AMIA FALL S N
P AMIA ANN FALL S
P AMIA ANN FALL S HA
P AMIA ANN FALL S PH
P AMIA ANN FALL S US
P AMIA ANN S
P AMIA ANN S AMIA S
P AMIA ANN S LOS ANG
P AMIA ANN S ORL FL
P AMIA FALL S
P AMIA S
P AMIA S S
```

Figure 1.4 Variations In Citations To Medical Informatics Conference Proceedings From WOS Data

1.1.2 Data Preparation for Data Mining, Record Linkage, and Citation Matching

One of the most important, time consuming, and difficult steps in the KDD process is data preparation or data preprocessing. The data preparation stage of the data exploration process has been estimated to require 60% of the total project time (Pyle, 1999). The advantage of data preprocessing is that it can substantially improve the overall quality of patterns mined (Han & Kamber, 2001). Two general techniques of data preparation are data cleaning and data integration. Data cleaning is undertaken to remove noise, correct inconsistencies, and address missing data. Data integration, also known as fusion, merges data from multiple sources into a coherent and enriched data store. Record linkage is a data preparation method of identifying database records that are syntactically different but refer to the same entity and lack a unique identifier. Various terms for the record linkage process are found in different user and research communities. The process that epidemiologists and statisticians refer to as record linkage is often referred to as data matching or the object identity problem by computer scientists and sometimes called merge/purge processing or list washing in commercial processing of customer databases or mailing lists (Christian and Churches, 2005). Deterministic record linkage is an ad-hoc process of exact matching based on one or more variables (Gomatam, 2002). Probabilistic record linkage methods are based on statistical and artificial intelligence techniques, and used to determine the matching (probabilistic matching) between records and for extracting a unique identifier or a set of variables acting as an identifier (Torra, 2003). The ideas of modern record linkage originated with geneticist Howard Newcombe who introduced odds ratios of frequencies and the decision rules for delineating matches and non-matches (Newcombe, 1959 and 1962). Newcombe's ideas have been implemented in software that is used in many epidemiological applications and often rely on odds-ratios of frequencies that have been computed a priori using large national health files. Fellegi and Sunter (1969) provided the formal mathematical foundations of probabilistic record linkage. Their theory demonstrated the optimality of the decision rules used by Newcombe and introduced ways of estimating crucial matching probabilities (parameters) directly from the files being matched (Winkler, 1999).

Many studies using probabilistic record linkage methodology can be found in the medical literature. However no work appears to exist on application of this theory and method to citation databases in the context of preparing citation data for subsequent analysis with methods such as progressive knowledge domain visualization. The general problem of duplicate detection has also been studied by library science research community as citation matching. While citation matching research has in common with duplicate detection the general issue of noise in citation data, there are differences in the problem and objectives. Citation matching addresses the problem of clustering many strings of text from source documents where the strings are very variable in structure, and the clustering is the final objective. The citation data available from citation databases are available in formats that are very well structured and tagged so it is possible to reformat and parse the data into a normalized relational database record format, and then linking records from two databases becomes a problem of standardizing on common fields and one-to-one linking of record pairs in the absence of a unique identifier while dealing with differences in spelling, punctuation, and abbreviation, and sometimes missing data within defined common fields (author last name, first name, middle initial, Publication date, volume, page, etc..). There are also differences in usage of standard numeric identifiers such as journal ISSN. For example WOS indexes the AMIA Symposium Proceedings as a supplement to the journal JAMIA, while in Medline the proceedings are indexed under a unique ISSN. There are also differences in the usage of print and electronic ISSN between the two databases. Two recent standards for unique document identifiers that theoretically could be used to link records are the publisher item identifier (PII) and digital object identifier (DOI), but the adoption and availability of these identifiers is limited and varies by journal publisher and database. In a sample of 18,197 records collected from the Medline database for a pilot study, 27% included a PII and 9% included a DOI. Neither identifier was available in an equivalent sample collected by direct export from the WOS database. A search of the online Bluesheets documentation for the DIALOG system indicated DOI availability only in non-medical databases (primarily engineering fields), and PII availability in the SCISEARCH and SOCIAL SCISEARCH databases from June, 2003 forward.

The Health Insurance Portability and Accountability Act (HIPAA) which took effect in 2003 in the United States does not place a restriction on the use of record linkage for linking citation data. HIPAA regulations specifically prohibit the use of names, social security numbers, or vehicle identification numbers, and mandate informed consent for research using medical records unless waived by an institutional review board. The risk to individuals is that linkage of one database to another creates not only new generalizable knowledge about cause-and-effect relationships but also more specific knowledge about some individuals (Clark, 2004). While there are now increased ethics and privacy considerations in the medical research domain with the use of record linkage in some settings, HIPAA regulations would not be a concern in the setting of linkage of citation data as publication data are non-medical and are public information.

1.2 Research Goals and Questions

The purpose of this research was to develop a specific model for record linkage of citation data, and to investigate the effects of the use of record linkage with information fusion data preparation methodology on biomedical knowledge domain visualizations. The problem of record linkage of citation databases where only non-unique identifiers such as author names and document titles are available as common identifiers in databases to be linked was investigated.

The research questions are:

1) Does a probabilistic record linkage model perform better than deterministic record linkage models in the linkage of citation data?

2) What are the effects of using record linkage with information fusion methodology to prepare citation data for knowledge domain visualization? .

Record linkage models were developed, and deterministic models compared with a probabilistic model in situations for which the truth is known through the manual development of gold standard or truth datasets. Performance for the two types of models was empirically compared with ROC analysis and a discussion of model failures presented. Data quality metrics were compared for datasets prepared without and with record linkage, and the effect on subsequent visualizations demonstrated. The methodology was carried out on linkages between records from the Web of Science, Medline, and CINAHL citation databases in the knowledge domains of medical informatics, HIV/AIDS, and nursing informatics.

The major contributions of this work are three fold. First, a connection has been established between the literature of probabilistic record linkage and the literature of knowledge domain visualization. Second, a novel model of probabilistic record linkage for biomedical citation databases that improves upon deterministic models is developed. Third, a methodology for improving and enriching knowledge domain visualizations though a data preparation approach is validated with analyses of multiple citation databases and knowledge domains.

1.3 Organization of Thesis

The remaining chapters of the thesis are organized as follows: Chapter 2 presents the background of record linkage theory and methodology, and reviews related work on citation matching from the library science literature. The methodologies used to evaluate record linkage models for citation data and the effects of information fusions on visualizations are presented in Chapter 3. Chapter 4 presents the results of ROC Analysis of deterministic record linkage models compared to a probabilistic model, and Chapter 5 present the results of Fusion studies on four sets of knowledge domain visualizations. The final chapter (Chapter 6) concludes the thesis with a summary and discussion of the major research findings, and areas for future studies.

CHAPTER 2: BACKGROUND AND LITERATURE REVIEW

This chapter presents a review of record linkage theory and methodology, comprehensively reviews related work on citation matching from the library science literature, and reviews related literature on probabilistic or merged data approaches to medical citation data.

2.1 Deterministic Record Linkage

The simplest deterministic record linkages are matches determined by 'all-or-nothing' comparisons of a collection of identifiers called the 'match key'. In this kind of matching when comparing two records the records are considered matches only if the matchkey on the two records agree on all characters. In a stepwise deterministic strategy (SDS) the records are linked in a sequence of steps each of which decides the linkage status (either match or non-match) of the record pair by considering exact agreement on a particular subset of identifiers. At each step the unique matches are extracted, the duplicates and the remaining unlinked observations in each of the two data sets (the residuals) form the input to the next step in the data linkage process, which continues with a different subset of identifiers. Steps that are implemented earlier in the procedure use collections of identifiers that are considered more reliable than those in later steps. (Roos & Wajda, 1991; Wajda et al, 1991; Gomatam, 2002).

2.2 Probabilistic Record Linkage

2.2.1 The Origins of Record Linkage

The term "record linkage" was first defined in 1946 as process which joins two separate pieces of information for a particular individual or family (Dunn, 1946). Howard Newcombe's insights led to computerized approaches for record linkage. The first insight was that the relative frequency of the occurrence of a value of a string such as a surname among matches and non-matches could be used in computing a binit weight (score) associated with the matching of two records. The second was that the scores over different fields such as surname, first name, age, etc.

could be added to obtain an overall matching score. He specifically considered odds ratios $\log_2(pL) - \log_2(pF)$ where pL is the relative frequency among links and pF is the relative frequency among non-links. Since the true matching status is often not known, he suggested approximating the above odds ratio with the ratio $\log_2(pR) - \log_2(pR)^2$ where pR is the frequency of a particular string (first, initial, birthplace, etc.). If a large universe file is matched with itself, then the second ratio is a good approximation of the first ratio (Winkler, 1999).

2.2.2 Fellegi-Sunter Theory of Record Linkage

Fellegi and Sunter provided a formal mathematical model for ideas that had been introduced by Newcombe and ways of estimating key parameters. To begin, notation is needed. Two files A and B are matched. The idea is to classify pairs in a product space A × B from two files A and B into M, the set of true matches, and U, the set of true non-matches. Fellegi and Sunter considered ratios of probabilities of the form:

Equation 1

$$R = P(\gamma \in \Gamma \mid M) / P(\gamma \in \Gamma \mid U)$$

where γ is an arbitrary agreement pattern in a comparison space Γ.

For instance, Γ might consist of eight patterns representing simple agreement or not on the largest name component, street name, and street number. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific values of name components such as "Smith", "Zabrinsky", "AAA", and "Capitol" occur. The ratio R or any monotonically increasing function of it such as the natural log is referred to as a matching weight (or score).

The decision rule is given by:

(1) If $R > T\ \mu$, then designate pair as a match.

(2) If $T\ \lambda \le R \le T\ \mu$, then designate pair as a possible match and hold for clerical review.

(3) If $R < T\ \lambda$, then designate pair as a non-match.

The cutoff thresholds T μ and T λ are determined by a priori error bounds on the rates of false matches and false non-matches. Rule (2) agrees with intuition. If $\gamma \in \Gamma$ consists primarily of agreements, then it is intuitive that $\gamma \in \Gamma$ would be more likely to occur among matches than non-matches and ratio (1) would be large. On the other hand, if $\gamma \in \Gamma$ consists primarily of disagreements, then ratio (1) would be small. Rule (2) partitions the set $\gamma \in \Gamma$ into three disjoint sub-regions. The region T $\lambda \leq$ R $\leq$ T μ is referred to as the no-decision region or clerical review region. This is an optional rule for situations where clerical review is desired. Pairs with weights above the upper cut-off are referred to as designated matches (or links). Pairs below the lower cut-off are referred to as designated non-matches (or non-links). The remaining pairs are referred to as designated potential matches (or potential links). The probabilities P(agree first |M), P(agree last | M), P(agree age | M), P(agree first | U), P(agreelast | U), and P(agree age | U) are called marginal probabilities. The probabilities P(| M) & P(| U) are called the m- and u-probabilities. The logarithms of the ratios of probabilities associated with individual fields (marginal probabilities) are called the individual agreement weights. The m- and u probabilities are also referred to as matching parameters. A false match is a pair that is designated as a match and is truly a non-match. A false non-match is pair that is designated as a non-match and is a truly a match (Winkler, 2003).

2.2.3 Automatic Parameter Estimation without Training Data

Fellegi and Sunter introduced methods for estimating optimal parameters (probabilities) in the likelihood ratio (1). They observed that

Equation 2

$$P(\gamma) = P(\gamma \mid M) P(M) + P(\gamma \mid U) P(U)$$

where "$\gamma \in \Gamma$ " is an arbitrary agreement pattern and M and U are two classes of matches and non-matches. If the agreement pattern $\gamma \in \Gamma$ is from three fields that satisfy a conditional independence assumption, then the system of seven equations and seven unknowns can be used to estimate the m-probabilities P(γ | M), the u-probabilities P (γ | U), and the proportion P(M). The conditional independence assumption corresponds exactly to the naïve Bayes assumption in machine learning (Winkler, 2003)

Machine learning algorithms that employ Bayesian networks are used to classify text into different groups. Bayesian networks are one of the standard tools in data mining and are also used for information retrieval methods such as web search engines. The EM-based algorithms (Expectation-Maximization algorithm) for finding maximum likelihood estimates in the latent classes models of record linkage are a direct generalization of ideas for automatically estimating parameters given in Fellegi and Sunter (Winkler 1999). Winkler (2000) showed how to estimate the probabilities in record linkage using the EM-Algorithm. Because of the additional structure available in record linkage, it is possible to obtain good matching results without subsets of training data. With general text, the algorithms of machine learning must create a structure for comparing that is facilitated by the training data. The advantage of training data is that it implicitly imposes additional structure on the learning with general text. With record linkage, additional structure is available with fields such as first name, last name, house number, and date-of-birth that have been parsed into separate components to be compared (Winkler, 1999). This is equivalent to the structured components of a field-tagged citation record such as author name, journal name, and publication date.

2.3 Methods for Matching and Duplicate Detection in the Library Science Literature

While no work appears to exist in the literature on application of deterministic or probabilistic record linkage (i.e., methods based on work of Newcombe and Fellegi-Sunter Theory of Record Linkage) to citation databases in the context of preparing citation data for subsequent analysis, the general problem of entity resolution or duplicate detection is present in the library science literature in several contexts. The problems of matching and merging duplicate records in library catalogs, bibliographies, and multi-database searches parallel record linkage problems. The methods used have similarities to record linkage methodology in several aspects, but only in the more recent work on citation clustering in the context of web databases such as CiteSeer are citations found to the work of Newcombe or Fellegi & Sunter. The library science methodologies have in common with record linkage a standard practice of "normalization" of the data (Coyle, 1985; Toney, 1992), which refers to the preparation of text data by converting case and removing punctuation and special characters. Normalization as the term is used here is equivalent to the data standardization step in a record linkage process and does not have any connection to the normalization of table structures in the relational data base sense of the word. Some of these techniques used to reduce the effects of minor typos, missing articles, and slight variations in wording are truncation, keywording, hashing, finding the Hamming distance between Harrisoned strings, Hamming and Harrissoning, soundex and similar techniques (Toney, 1992). Other commonalities with record linkage are the creation of a "match key" and an initial step to pool records into groups of potential matches. The algorithms sometimes include ad-hoc complex rules or weighting schemes as a work-around for data quality problems, and sometimes include the concept of thresholds and uncertainty zones, but lack the theoretical foundation of the work of Newcombe or Fellegi & Sunter. Much of the work on matching monographic records and bibliographic journal citation data took place prior to 1990 and may not be currently relevant in the light of advances in computer processing power, advances in string matching algorithms, the use of Z39.50 technology and related shift to virtual union catalogs. However the problem of duplicate detection continues to be a concern and challenge for matching in the context of virtual union catalogs (Cousins, 1999; Thornburg, 2005).

Related work on duplicate detection in the library science literature is presented in Tables 2.1 - 2.3, and some of the details or more notable aspects are discussed.

Table 2.1 Library Science Methods for Matching in Citation Databases (*=Method Includes Text Normalization Step)

Author	Context	Method/Match Key	*	Algorithm Performance	Comment
Giles et al, 1976	Oak Ridge Natl. Library Bibliographic journal Citation files	Year Initial Page Journal CODEN Journal Volume Author Name Sample Journal Name Sample Title Sample	?	"Performed quite well"	Combined Sorted and weighted matching Used Soundex on Author Name
Hawkins, 1981	Citation Databases	CODEN, year, pagination		"Identified large percentage of duplicates"	
Onorato et al, 1981	Citation Databases	First Author, Date, Title Sample	n/a	Not Tested	Proposal only
Slach, 1985	Upjohn Tech. Library Citation Databases	2-digit year+first four characters author name+beginning page	N	Duplicates incorrect : 1%	
Yannakoudakis, 1990	Citation databases, nonstandardized (untagged) ESA/IRS DIALOG	Data converted to standard format. USBC, 7 byte Title (1-5)+Author(1-2) Title : 8 least frequent characters in lexicographic order Author: 8 least frequent letters	Y	Precision = 97.9 Recall = 94.3 Relative Performance (RP) = 6.075 N=1191	
Toney, 1992	BCIN	Two stage :Bibliographic level (monograph or analytic) Personal or corporate author Analytic title (title of article or chapter) Title of main work (monograph or serial) Series title Date of publication Volume number Issue number Pagination	Y	Not reported	Uses weights and threshold values Discussion of rational for selecting fields and parsing data
Ayres, 1996	Multiple Projects	USBC	Y	Automatch Failure rate ~5%	

An early work on matching bibliographic journal citation files (Giles et al, 1976) (Table 2.1) used a combined sorting and matching scheme with fixed length keys:

Fixed length keys:
Year
Initial Page
Journal CODEN
Journal Volume
Author Name Sample: Soundex scheme (first letter author surname followed by up to 6 non-repeated consonants in surname and author's first initial)
Journal Name Sample: first 2 letters of first 4 words in journal title
Title Sample: first 4 and last 4 consonants in title
Three bits per field to indicate presence or error in key generation

Sorting and matching scheme:

1. Sort by page and year. If equal, match when
    a. Titles and authors are equal
    b. Authors are equal and journal or volume is equal, or
    c. Titles are equal and journal or volume is equal
2. Sort by author and title. If equal, match when
    a. Year, journal, and volume are equal
    b. Pages are equal and volume or journal is equal

The model of Slach (1985) is notable for the simplicity of the matchkey with a reported rate of incorrect duplicates similar to that of the more complex OCLC matchkey. The Universal Standard Book Code (USBC) is perhaps the most widely studied method (Goyal, 1987; Yannakoudakis, 1990; Ridley, 1992; Toney, 1992; Ayres, 1996). Alternate methods have been reported for creating a USBC, but a unique aspect of the USBC algorithm is a coding "signature" of longer elements such as author or title. The algorithm analyzes the frequency of alphanumeric characters in strings with the least common characters having most significance; the resulting code may be sorted by ascending frequency, title order, or reverse order (Ayres, 1996). The USBC seems to depend on clean data, especially on a clean title (Toney, 1992).

The work of Hickey and Ripka (1979) (Table 2.2) on matching monographic records from the OCLC appears to be the most highly cited single work. This method utilized a 52 byte key and a decision table of 16 different exact matches and partial matches and 14 keys. The Title key used specific character positions and would be sensitive to any to variability in title strings, and the actual duplicate detection reported was only between 54-60%. Coyle (1985) does not cite Newcombe or Felligi & Sunter, but the method employed an expert algorithm with matching based on a weighted evaluation of data elements to compensate for differences such as typos, missing data, and cataloging practice. The external table of weights was derived by experimentation and manual adjustment and included the concept of a "grey area" for unsure matches.

Table 2.2 Library Science Methods For Matching MARC Files And Monographic Records (*=Method Includes Text Normalization Step)

Author	Context	Method/Match Key	*	Algorithm Performance	Comment
Hickey et al, 1979	OCLC Monographic Records	52 byte key Variable length key Date Record Type Reproduction code Title (8 characters)	Y	54-69% of actual duplicates 1.3% incorrect N=1,000 to 214,000	Two-step, exact match grouping and partial match. Used decision table
Williams & MacLaury, 1979	Univ. of Illinois Monographic MARC II format files	2-step sort and match Date: last 2 digits Title Sample Name match 1st 5 characters Title hash with Harrison-Hamming test. Pagination	Y	?	Title match allowed for only minor variations such as simple typographical errors
Coyle et al, 1985	MELVYL Monographic Union Catalog	Exact match then Weighted matching over 10 elements. LCCN or ISBN, date, edition, truncated title = exact match	Y	Not reported	Weighted matching
Goyal, 1987	BNB, OCLC MARC files	USBC 17 character code Date, Edition, Language, Title length, publisher, title, volume	Y	20%-70% of very small samples	Compared minimum self-information and maximum entropy principle
Ridley, 1992	QUALCAT Monographic	Two stage match then rules USBC, 15 byte Date Volume Edition Author (least frequent characters) Title (least frequent characters)	Y	Not reported	Expert system Used weighted rules Of cert (certainties) and poss (possibilities) with thresholds for determining duplicates, non-duplicates and undetermined. Weights derived manually over time MARC records restructured to relational database
Cousins, 1998	COPAC union catalog	Two stage 1) ISBN or ISSN match or Author/title acronym match 2) Detailed field match	Y	Not reported	Used scoring system with threshold

The transition to a research focus on citation clustering in the context of web based citation databases begins with Hylton (1996) who created DIFWICS, a 240,000 record catalog of computer science literature (Table 2.3). Hylton focused on clustering intellectual works rather then matching documents but is important for the concept of linking citations to full-text documents on the Web. Recent work on citation matching has focused on clustering unstructured citation data from CiteSeer (Lawrence, 1999; Pasula, 2003; Wellner, 2004; Culotta 2005).

Table 2.3 Library Science Methods For Matching Web Based Unstructured Citations (*=Method Includes Text Normalization Step)

Author	Context	Method/Match Key	*	Algorithm Performance	Comment
Hylton. 1996	Web citation databases - Alf Christen Achilles - CS-TR project Bibtex and CS-TR formats	Clustering of Author-Title matches to identify intellectual works	Y	90% identified 1% inaccurate N=240,000	*n*-gram string comparison of randomly selected words from author and title field
Monge, 1997	Web citation databases - Alf Christen Achilles - CS-TR project Bibtex and CS-TR formats	Smith-Waterman algorithm Clustering of Author-Title matches to identify intellectual works	?	"Comparable to Hylton 1996" N=254,619	Cites Newcombe
Lawrence, 1999	CiteSeer	Machine learning algorithm for word and phrase matching, clustering citations obtained from full papers Focused on title-1st author name match	Y	5.3% incorrect clusters N=295 to 514	
Pasula, 2003	CiteSeer datasets from Lawrence, 1999	Relational probability model+Markoc Chain Monte Carlo method, clustering citations obtained from full papers	?	3% to 7% incorrect clusters N=295 to 514	
Wellner, 2004	CiteSeer datasets from Lawrence, 1999	Conditionally trained undirected graphical models	Y	4% to 7% incorrect clusters N=295 to 514	
Culotta, 2005	CiteSeer CORA	Clustering with Joint deduplication of papers and venues	Y	Pairwise F1 = 90.8 – 93.4 N=1500 to 1800	

An example of duplicate detection in practical everyday use is RefWorks (http://www.refworks.com/), a Web-based bibliography and database manager. RefWorks provides a "Close Match" and "Exact Match" comparison of a combination of Author Names, Title, and Year of Publication to locate duplicate records in RefWorks.

2.4 Related Work on Probabilistic or Merged Data Approaches to Medical Citation Data

Shaw (1991a, 1991b) investigated the clustering structure of composite representations in a cystic fibrosis document collection. The collection of 1,239 papers from years 1974-1979 included MeSH terms, the complete set of cited references, and a comprehensive set of citations to each paper from Science Citation Index. The process by which this composite database was created is not described, but a record linkage approach to creating fused datasets would readily enable the creation of larger datasets for investigation.

Torvik, Swanson, and Smalheiser (2005) have applied a "probabilistic similarity metric" to the Medline database with the objective of author name disambiguation for purposes of authority control and subsequent improvement in retrieval of papers by a given author. A model was developed for estimating the probability that a pair of author names (sharing last name and first initial) appearing on two different Medline articles refer to the same individual. The model used a similarity profile between pairs of articles based on title, journal name, coauthor names, medical subject headings (MeSH), language, affiliation, and name attributes (prevalence in the literature, middle initial, and suffix). The work is based on probabilistic information retrieval, but has similarities to probabilistic record linkage in that vectors of attributes are created for which a weighted probability of matching in pairwise comparison is estimated.

Bernstam, et al (2006) compared the effectiveness of citation-based algorithms to noncitation-based in identifying important articles. The study refers to "mapping" between Medline and WOS, but the methodology of "mapping" not described. Bernstam et al found that mapping between Medline and the WOS Science Citation Index (SCI) was difficult because incompatible article representations and multiple data entry errors made simple string matching inadequate. This is a not a record linkage study, however the study is relevant because it is related to the concept of the value of combining MeSH terms from Medline with citation data from WOS.

CHAPTER 3: METHODS

Methods are presented for two sets of studies: 1) The evaluation of the performance of record linkage models and 2) the evaluation of the effects of information fusion on visualizations. For each set of studies the sample selection process and variables are defined, and data analysis methods described.

3.1 Evaluation of Record Linkage Models

3.1.1 Sample Selection

A medical informatics dataset that was developed for prior visualization studies was the primary basis for this analysis (Synnestvedt et al, 2005). The dataset was defined by cross-referencing the Institute for Scientific Information's (ISI) Journal Citation Reports list of medical informatics journals for 2003 against a list of medical informatics journals from AMIA(AMIA, 2003). The twelve journals that both resources identified as important or relevant to medical informatics were selected for study. These twelve journals were also checked against the NCBI journals database for publication history, and the journals which were predecessors of some of the current journals were identified (Table 3.1).

Because ISI has indexed conference proceedings (including poster session abstracts) under journal names instead of conference proceeding names, meeting abstracts were excluded from the query on the WOS database. The Medline dataset has been regenerated for this study to include conference proceedings papers in the Medline dataset, and improve the overlap with the WOS dataset. In addition, the Medline dataset was supplemented with four additional medical informatics journals in order to increase the number of citations potentially available for matching during record linkage. The additional journals were selected by cross referencing the William H. Welch Medical Library of Johns Hopkins University School of Medicine's list of Informatics resources against the AMIA list. This resulted in a WOS dataset of 11,752 citation records, and

Table 3.1: Medical Informatics Datasets

ISSN	Journal Title	JCR 2003 Impact Factor	JCR 2003 I. F. Rank	Years Indexed in WOS	WOS #	Years Indexed in Pub Med	Pub Med #
0933-3657	Artificial Intelligence In Medicine	1.222	6	1992-2004	449	1993-2004	485
1538-2931	Cin-Computers Informatics Nursing	0.217	19	2002-2004	121	2002-2004	101
0169-2607	Computer Methods And Programs In Biomedicine	0.724	14	1985-2004	1609	1985-2004	1584
0010-468X	Computer Programs In Biomedicine(1)			1975-1985	437	1971-1985	512
0010-4809	Computers And Biomedical Research (2)			1968-2000	1403	1967-2000	1418
0736-8593	Computers In Nursing (3)			1992-2002	119	1983-2002	650
1089-7771	Ieee Transactions On Information Technology In Biomedicine	1.274	5	2000-2004	210	1997-2004	304
0020-7101	International Journal Of Bio-Medical Computing (4)			1975-1996	1021	1970-1996	1198
1386-5056	International Journal Of Medical Informatics	1.178	8	1997-2004	736	1997-2004	718
0266-4623	International Journal Of Technology Assessment In Health Care	0.754	12	1995-2004	742	1985-2004	1351
1532-0464	Journal Of Biomedical Informatics	0.855	11	2001-2004	152	2001-2004	157
1067-5027	Journal Of The American Medical Informatics Association	2.51	1	1994-2004	1674*	1994-2004	689
0195-4210	Proceedings / The Annual Symposium On Computer Application [Sic] In (5)					1991-1995	1009
1091-8280	Proceedings : A Conference Of The American Medical Informatics (5)					1996-1997	329
1531-605X	Proceedings / Amia Annual Symposium Amia Symposium (5)					1998-2002	946
-	Amia ... Annual Symposium Proceedings…					2003	458
0724-6811	M D Computing	0.500	17	1984-2001	500*	1984-2001	836
0272-989X	Medical Decision Making	1.718	3	1983-2004	871*	1981-2004	1145
1463-9238	Medical Informatics And The Internet In Medicine	0.915	10	1999-2004	136	1999-2004	134
0026-1270	Methods Of Information In Medicine	1.417	4	1964-2004	1572*	1962-2004	1895
Sub-total					11,752		15,919
Journals added to Medline Dataset Only							
1367-4803	Bioinformatics (Oxford, England)	6.701				1998-2004	2198
0010-4825	Computers In Biology And Medicine	0.973				1970-2004	1219
1357-633X	Journal Of Telemedicine And Telecare	1.094				1995-2004	1103
-	Medinfo					1995-2004	1332
Total					11,752		21,771

1: Continued by Computer Methods And Programs In Biomedicine; 2: Continued by Journal Of Biomedical Informatics; 3: Continued by Cin-Computers Informatics Nursing; 4: Continued by International Journal Of Medical Informatics; 5: WOS has AMIA Symposium Proceedings 1994 – 2002 indexed as supplement to JAMIA; *: Meeting abstracts excluded.

a dataset of 21,771 records from PubMed (Table 3.1) covering forty years from 1964-2004. While the WOS dataset is smaller in terms of total number of records, it is not a complete subset of the Medline dataset. This is primarily due to differences in selection of individual documents for indexing.

3.1.2 Variable Identification and Data Standardization

The first step taken to identify candidate variables for modeling was to compare the definitions of tagged field elements for the WOS and Medline export files. Table 3.2 shows the field elements found in common to the two record structures.

Table 3.2 Comparable Tagged Fields From WOS And Medline

WOSTag	WOSDesc	MedlTag	MedlDesc
AB	Abstract	AB	Abstract
AU	Authors	AU	Author
BP	Beginning page	PG	Pagination
DT	Document type	PT	Publication Type
EP	Ending page	PG	Pagination
ID	Keywords Plus®	MH	MeSH Terms
IS	Issue	IP	Issue
J9	29-character source abbreviation	TA	Journal Title Abbreviation
NR	Cited reference count	RF	Number of References
PI	Publisher city	PL	Place of Publication
PY	Publication year	DP	Date of Publication
SN	ISSN	IS	ISSN
SO	Full source title	JT	Journal Title
TI	Document title	TI	Title
VL	Volume	VI	Volume

The variables initially selected for standardization and evaluation for use in the linkage models are:

- First Author Last Name
- First Author First Initial
- First Author Middle Initial
- Journal ISSN
- Journal Abbreviation
- Year of Publication
- Volume
- Issue

- Begin Page
- EndPage
- Document title

These variables have been selected because of the common availability in both datasets, generally low rates of missing data, and likely ability to uniquely identify articles when used in combination. The following tables (Table 3.3 – 3.4) show the sample data survey for variables within each dataset.

Table 3.3 WOS Dataset Survey

WOS Field Tags	**Field Description**	**Sample Value**
AB	Abstract	In 1986, the National Library of Medicine began a long-term research
AU	Authors	LINDBERG, DAB
BP	Beginning page	281
DT	Document type	Article
EP	Ending page	291
ID	Keywords Plus®	INFORMATION; KNOWLEDGE
IS	Issue	4
J9	29-character source abbreviation	METHODS INFORM MED
PY	Publication year	1993
SN	ISSN	0026-1270
SO	Full source title	METHODS OF INFORMATION IN MEDICINE
TI	Document title	THE UNIFIED MEDICAL LANGUAGE SYSTEM
VL	Volume	32

Table 3.4 Medline Dataset Survey

Medline Field Tags	**Field Description**	**Sample Value**	**Equiv. WOS Field Tag**
AB	Abstract	In 1986, the National Library of Medicine began a long-term research and	AB
AU	Author	Lindberg DA	AU
PG	Pagination	281-91	BP, EP
PT	Publication Type	Journal Article	DT
MH	MeSH Terms	Information Storage and Retrieval/MEDLINE/National Library of Medicine (U.S.)/ *Unified Medical Language System/United States	ID
IP	Issue	4	IS
TA	Journal Title Abbreviation	Methods Inf Med	J9
DP	Date of Publication	1993 Aug	PY
IS	ISSN	0026-1270 (Print)	SN
JT	Journal Title	Methods of information in medicine.	SO
TI	Title	The Unified Medical Language System.	TI
VI	Volume	32	VL

While the two datasets have variables in common, the format of individual variables is not the same between the datasets. Standardization procedures of variables are necessary to increase performance of the record linkages (Torres, 2003). Procedures suggested by Torres (2003) are to:

1) Parse variables to build a uniform structure
2) Detect relevant keywords to help in the process of recognizing the components of variables
3) Replace all common forms of a word by single ones

All parsing and standardization routines were developed in a relational database form using Microsoft Office Access software. Citation data were exported from Web of Science in field tagged record format and from Medline via the PubMed database in Medline record format. The general standardization process was as follows:

1) Import the raw citation data into Access database as fixed length record of 3 fields
2) Use autonumber to create unique identifier for each line of record.
3) Create a working copy of table
4) Rename field1 FieldTag, Add a document ID field
5) Delete empty records
6) Run module to add unique document ID to all records and complete missing field tags
7) Use cross-tab query to pivot data to a normalized record structure
8) Parse and standardize variables to "least common denominator" format

The following tables show the raw data format with parsing code (Tables 3.5 – 3.6) and resulting standardized data formats for use in record linkage (Tables 3.7 – 3.8):

Table 3.5 Example of WOS Export Format and Parsing Code

Field Tag	Raw Data	Parsing to Standardized format	Creates Variables
AU	Aarts, J	WOS_working.Field3 AS FirstAuthor, IIf([FirstAuthor]="[Anon]",Null,IIf([FirstAuthor] Not Like "*,*",[FirstAuthor],IIf([FirstAuthor] Like "*,*",Left([FirstAuthor],(InStrRev([FirstAuthor],",")-1)),Left([FirstAuthor],(InStrRev([FirstAuthor],".")-1))))) AS [Last Name], IIf([FirstAuthor]="[Anon]",Null,IIf([FirstAuthor] Not Like "*,*",Null,Mid([FirstAuthor],(InStrRev([FirstAuthor]," ")+1),1))) AS [First Initial], IIf([FirstAuthor]="[Anon]",Null,IIf([FirstAuthor] Not Like "*,*",Null,Mid([FirstAuthor],(InStrRev([FirstAuthor]," ")+2),1))) AS [Mid Initial],	Last Name First Initial Mis Intial
BP	207	BP AS BeginPage	Begin Page
EP	216		End Page
IS	3	IIf([IS] Like "*-*",Left([IS],1),[IS]) AS Issue	Issue
PY	2004	PY AS Year	Year
SN	1067-5027	Left([SN],9) AS ISSN	ISSN
TI	Understanding implementation: The case of a computerized physician order entry system in a large dutch university medical center	Left([Title],50) AS Title50	TitleAbbrev
VL	11	VL AS Volume	Volume

Table 3.6 Example of Medline Export Format and Parsing Code

Field Tag	Raw Data	Parsing to Standardized format	Variables Created
AU	Aarts, Jos	Last Name: IIf([FirstAuthor] Is Null,Null,IIf([FirstAuthor] Not Like "* *",[FirstAuthor],IIf([FirstAuthor] Like "*#*" Or [FirstAuthor] Like "* * jr",Left([FirstAuthor],(InStr([FirstAuthor]," ")-1)),Left([FirstAuthor],(InStrRev([FirstAuthor]," ")-1))))) First Initial: IIf([FirstAuthor] Is Null,Null,IIf([FirstAuthor] Not Like "* *",Null,IIf([FirstAuthor] Like "*#*" Or [FirstAuthor] Like "* * jr",Mid([FirstAuthor],(InStr([FirstAuthor]," ")+1),1),Mid([FirstAuthor],(InStrRev([FirstAuthor]," ")+1),1)))) Mid Initial: IIf([FirstAuthor] Is Null,Null,IIf([FirstAuthor] Not Like "* *",Null,IIf([FirstAuthor] Like "*#*" Or [FirstAuthor] Like "* * jr",Mid([FirstAuthor],(InStr([FirstAuthor]," ")+2),1),Mid([FirstAuthor],(InStrRev([FirstAuthor]," ")+2),1))))	Last Name First Initial Mid Initial
IS	1067-5027 (Print)	Left([IS],9)	ISSN
VI	11	IIf([VI] Is Null Or [VI] Like "SUPPL",Null,IIf([VI] Like "* *",Left([VI],(InStr(1,[VI]," "))-1),IIf([VI] Like "*-*",Left([VI],(InStr(1,[VI],"-"))-1),[VI])))	Volume
IP	3	IIf([IP] Like "pt #",Right([IP],1),IIf([IP] Like "*-*" Or [IP] Like "* *",Left([IP],1),IIf([IP] Like "sup*#",Right([IP],1),[IP])))	Issue
DP	2004 May-Jun	Left([DP],4)	Year
TI	Understanding implementation: the case of a computerized physician order entry system in a large Dutch university medical center.	Title: IIf(Medline_Working_1.Field3 Like "*.",Left(Medline_Working_1.FIeld3,(InStrRev(Medline_Working_1.Field3,".")-1)),Medline_Working_1.FIeld3) TitleAbbrev: Left([Title],50)	TitleAbbrev

Table 3.6 (continued)

PG	207-16	I Do Until tbl.EOF tbl.Edit VarPageStringNum = "" i = 1 If IsNull(tbl!PG) Then VarPageStringLen = 0 VarPageString = "" Else VarPageString = tbl!PG VarPageStringLen = Len(VarPageString) End If For i = 1 To VarPageStringLen If Left(VarPageString, 1) Like "[!A-Z]" Then VarPageStringNum = VarPageStringNum + (Left(VarPageString, 1)) VarPageString = Mid(VarPageString, 2) Else VarPageString = Mid(VarPageString, 2) End If Next tbl!PGnum = Trim(VarPageStringNum) tbl.Update tbl.MoveNext Loop Do Until tbl.EOF tbl.Edit VarPageBeginPage = "" i = 1 If IsNull(tbl!PGnum) Then VarPageStringLen = 0 VarPageStringNum = "" Else VarPageStringNum = tbl!PGnum VarPageStringLen = Len(VarPageStringNum) End If For i = 1 To VarPageStringLen If Left(VarPageStringNum, 1) Like "[0-9]" Then VarPageBeginPage = VarPageBeginPage + (Left(VarPageStringNum, 1)) VarPageStringNum = Mid(VarPageStringNum, 2) Else i = VarPageStringLen End If Next tbl!BeginPage = VarPageBeginPage tbl.Update tbl.MoveNext Loop EPD: IIf([PG] Like "*;*",Mid([PG],InStr([PG],";")-1,1),Right([PG],1))	Begin Page End Page Digit

Table 3.7 Standardized WOS Record

Last Name	First Initial	Mid Initial	ISSN	Year	Volume	Issue	BeginPage	Journal Abbrev	Title Abbrev	WID
Aarts	J		1067-5027	2004	11	3	207	J AMER MED INFORM ASSOC	Understand	4561

Table 3.8 Standardized Medline Record

Last Name	First Initial	Mid Initial	ISSN	Year	Volume	Issue	BeginPage	Journal Abbrev	Title Abbrev	PMID
Aarts	J		1067-5027	2004	11	3	207	J Am Med Inform Assoc	Understand	14764612

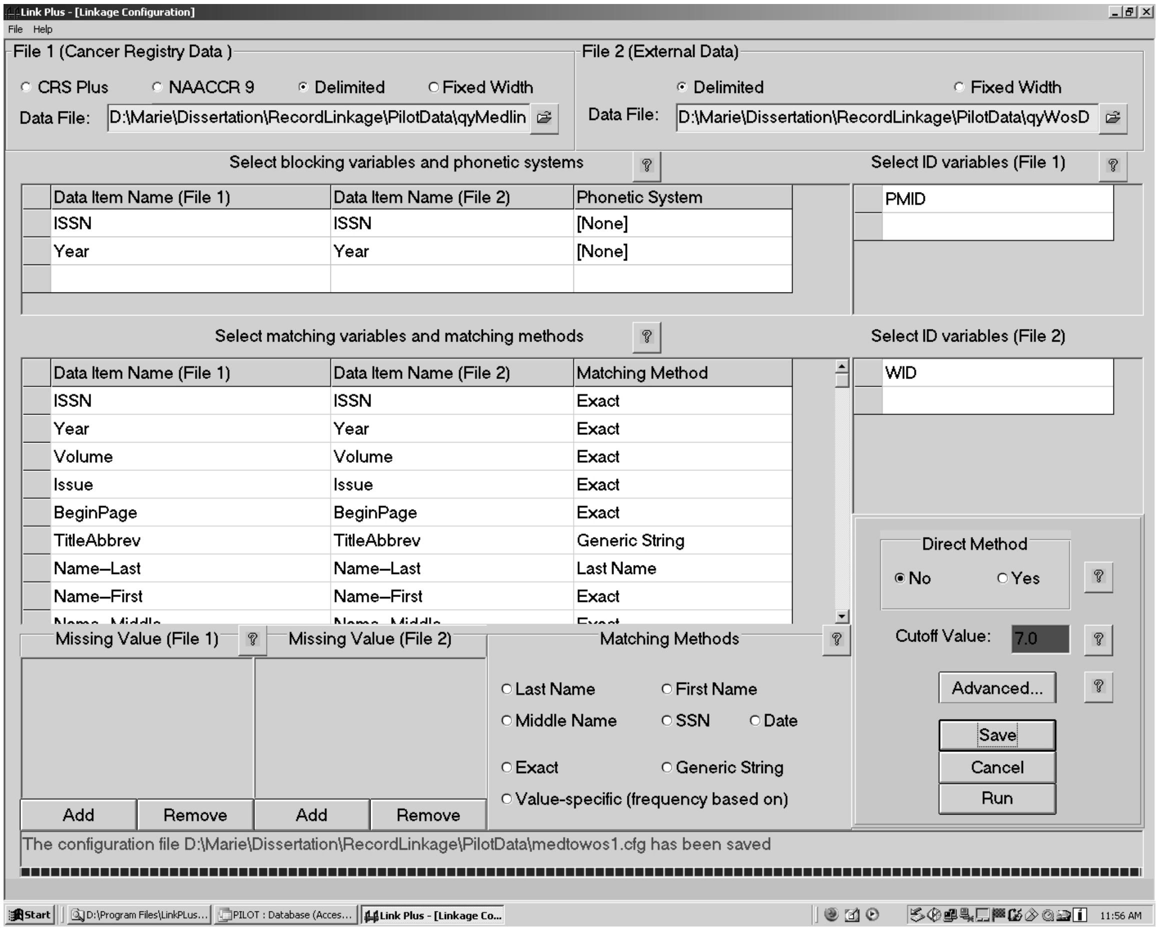
3.1.3 Measures (instrumentation and materials)

The deterministic modeling was performed using relational database queries in Microsoft Access. The probabilistic record linkage modeling has been developed with Link Plus (Figure 3.1), which is a record linkage program developed at the Centers for Disease Control and Prevention (CDC), Division of Cancer Prevention and Control (DCPC), in support of CDC's National Program of Cancer Registries (NCPR). Link Plus was written as a linkage tool for cancer registries. However, no theoretical or practical barriers exist to prevent using the program with data other than cancer registry data. Link Plus can be run in two modes: to detect duplicates in a database, or to link two files. The program computes probabilistic record linkage scores based on the theoretical frame work developed by Fellegi and Sunter. (1969), and facilitates a simple and efficient blocking mechanism by indexing the variables for blocking and comparing the pairs with the identical values on at least one of those variables. The option of computing the M-Probabilities using the EM algorithm for maximum likelihood estimation is available. Link Plus provides the following comparison methods that may be applicable to citation data:

Value-specific (frequency-based) comparison method that sets weights for matching values, based on the frequencies of values in the files being compared.

- Last name and first name comparison methods that incorporate both partial matching and value-specific matching to account for minor typographical errors, misspellings, and hyphenated names.

- Generic String method that incorporates partial matching to account for typographical errors. The string comparator used by LinkPLus is based on the methods of Jaro and Winker (Jaro 1989, Winkler 1990). The Jaro-Winkler string comparator is the comparator developed at the U.S. Census Bureau and used in the Census Bureau record linkage software, and commonly used in the record linkage field. The basis of the Jaro comparator is the count of common characters between the strings, where a character is counted as common if it occurs in the other string within a position distance that depends on the string length. The Jaro string comparator accounts for insertions, deletions and transpositions. The second enhancement due to Winkler (1990) gives increased value to agreement on the beginning characters of a string. This approach is based on findings of Pollock and Zamora (1984) that showed that the fewest errors typically occur at the beginning of a string and the error rates by character position increase monotonically as the position moves to the right. The Winkler enhancement adjusts the string comparator value upward by a fixed amount if the first four characters agreed; by lesser amounts if the first three, two, or one characters agreed. The Jaro-Winkler comparators have been found to be superior for matching of name and address data. Budzinsky (1991) concluded that the comparators due to Jaro and Winkler were the best among twenty in the computer science literature. Grannis (2004) compared and approximate string comparators in a study of name matching in deterministic record linkage. Approximate comparators included the modified Jaro-Winkler method, the longest common substring, and the Levenshtein edit distance. The Jaro-Winkler comparator achieved the highest linkage sensitivities of 97.4% and 97.7%.


Link Plus - [Linkage Configuration]
File
Help
Add
Remove
Add
Remove
Advanced...
Save
Cancel
Run
Start
D:\Program Files\LinkPLus...
PILOT : Database (Acces...
Link Plus - [Linkage Co...
11:56 AM

File 1 (Cancer Registry Data)

○ CRS Plus ○ NAACCR 9 ● Delimited ○ Fixed Width

Data File: D:\Marie\Dissertation\RecordLinkage\PilotData\qyMedlin

File 2 (External Data)

● Delimited ○ Fixed Width

Data File: D:\Marie\Dissertation\RecordLinkage\PilotData\qyWosD

Select blocking variables and phonetic systems

Data Item Name (File 1)	Data Item Name (File 2)	Phonetic System
ISSN	ISSN	[None]
Year	Year	[None]

Select ID variables (File 1)

PMID

Select matching variables and matching methods

Data Item Name (File 1)	Data Item Name (File 2)	Matching Method
ISSN	ISSN	Exact
Year	Year	Exact
Volume	Volume	Exact
Issue	Issue	Exact
BeginPage	BeginPage	Exact
TitleAbbrev	TitleAbbrev	Generic String
Name–Last	Name–Last	Last Name
Name–First	Name–First	Exact
Name–Middle	Name–Middle	Exact

Select ID variables (File 2)

WID

Direct Method

● No ○ Yes

Cutoff Value: 7.0

Missing Value (File 1)

Missing Value (File 2)

Matching Methods

○ Last Name ○ First Name

○ Middle Name ○ SSN ○ Date

○ Exact ○ Generic String

○ Value-specific (frequency based on)

The configuration file D:\Marie\Dissertation\RecordLinkage\PilotData\medtowos1.cfg has been saved

Figure 3.1 The Link Plus Linkage Configuration Interface

3.1.4 Model Selection

3.1.4.1 Deterministic Models

Five deterministic models were selected for evaluation based on review of the literature, committee recommendations, current bibliography management tools, and an alternate approach that did not rely on matching of author or title string fields.

Deterministic Model #0 (DMatch0): This model was evaluated to rule out the use of document titles as a single matching variable. Titles are nearly unique identifiers of documents as determined by frequency distributions. However due to inconsistencies in the recording of titles between Medline and WOS a model based on title as a single matching variable is not expected to perform well. The matching variable evaluated was Title (truncated at 50 characters).

Deterministic Model #1 (DMatch1): This model is based on the matchkey reported by Slach (1985). This model was selected for evaluation because it was developed for use with bibliographic citation data, has simplicity, was not based on complex rules or weighting schemes, and had a low reported rate of false positives. The matching variables evaluated were Year, First Author Last Name (first 4 characters), and Begin page.

Deterministic Model #2 (DMatch2): This model was recommended by the Committee as being a standard for current good practice and is very similar to DMatch1 with the exception of using the full last name of the first author. The variables evaluated were Year, First Author Last Name, and Begin page.

Deterministic Model #3 (DMatch3): This model is based on the matching criteria used by the RefWorks bibliography management tool to identify duplicates. The variables evaluated were First Author Last Name, Year, and Title (truncated at 50 characters).

Deterministic Model #4 (DMatch4): This model was designed to avoid the use of author and title text strings which may be difficult to match because of variations in wording, spelling and punctuation. The variables evaluated were ISSN, Year, Volume, Issue, and Begin Page.

3.1.4.2 Probabilistic Model

The development of the probabilistic model was an iterative process of experimentation, analysis of frequency distributions, and manual review of matched records for errors. There is an assumption of conditional independence in both the probabilistic scoring method and the EM algorithm (Winkler, 1999). The models assume that identifiers are independent, i.e. if there is a match on one variable there is not a second correlated variable that will have a very high probability of matching. For this reason both Journal ISSN and Journal Title elements were not combined in the list of candidate variables for probabilistic model development. Journal ISSN was selected over Journal Title because of less variability between the two databases (Table 3.9).

Table 3.9 Variability In ISSN And Journal Titles Between WOS And Medline

WOS ISSN	**Journal Title**	**Journal Abbrev (J9 - 29-character source abbreviation)**	**Journal Abbrev (JI - ISO abbreviation)**	**Medline ISSN**	**Journal Title**	**Journal Abbrev**
1538-2931	Cin-computers informatics nursing	Cin-Comput Inform Nurs	CIN-Comput. Inform. Nurs.	1538-2931	Computers, informatics, nursing : CIN	Comput Inform Nurs
0169-2607	Computer methods and programs in biomedicine	Comput Method Program Biomed	Comput. Meth. Programs Biomed.	0169-2607	Computer methods and programs in biomedicine	Comput Methods Programs Biomed
0010-468X	Computer programs in biomedicine	Comput Program Biomed	none	0010-468X	Computer programs in biomedicine	Comput Programs Biomed
0010-4809	Computers and biomedical research	Comput Biomed Res	Comput. Biomed. Res.	0010-4809	Computers and biomedical research, an international journal	Comput Biomed Res
1089-7771	IEEE transactions on information technology in biomedicine	Ieee Trans Inf Technol Biomed	IEEE T. Inf. Technol. Biomed.	1089-7771	IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society	IEEE Trans Inf Technol Biomed
0020-7101	International journal of bio-medical computing	Int J Bio-Med Comput	Int. J. Bio-Med. Comput.	0020-7101	International journal of bio-medical computing	Int J Biomed Comput
0266-4623	International journal of technology assessment in health care	Int J Technol Assess Health C	Int. J. Technol. Assess. Health Care	0266-4623	International journal of technology assessment in health care	Int J Technol Assess Health Care
1067-5027	Journal of the american medical informatics association	J Amer Med Inform Assoc	J. Am. Med. Inf. Assoc.	-	AMIA ... Annual Symposium proceedings [electronic resource] / AMIA Symposium. AMIA Symposium	AMIA Annu Symp Proc
1067-5027	Journal of the american medical informatics association	J Amer Med Inform Assoc	J. Am. Med. Inf. Assoc.	1067-5027	Journal of the American Medical Informatics Association : JAMIA	J Am Med Inform Assoc
1067-5027	Journal of the american medical informatics association	J Amer Med Inform Assoc	J. Am. Med. Inf. Assoc.	1091-8280	Proceedings : a conference of the American Medical Informatics Association / ... AMIA Annual Fall Symposium. AMIA Fall Symposium	Proc AMIA Annu Fall Symp
1067-5027	Journal of the american medical informatics association	J Amer Med Inform Assoc	J. Am. Med. Inf. Assoc.	1531-605X	Proceedings / AMIA ... Annual Symposium. AMIA Symposium	Proc AMIA Symp
1067-5027	Journal of the american medical informatics association	J Amer Med Inform Assoc	J. Am. Med. Inf. Assoc.	0195-4210	Proceedings / the ... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care	Proc Annu Symp Comput Appl Med Care
0724-6811	M D computing	M D Comput	M D Comput.	0724-6811	M.D. computing : computers in medical practice	MD Comput
0272-989X	Medical decision making	Med Decis Making	Med. Decis. Mak.	0272-989X	Medical decision making : an international journal of the Society for Medical Decision Making	Med Decis Making
0026-1270	Methods of information in medicine	Methods Inform Med	Methods Inf. Med.	0026-1270	Methods of information in medicine	Methods Inf Med

The variables evaluated for use in the probabilistic linkage model were:

- First Author Last Name
- First Author First Initial
- First Author Middle Initial
- Journal ISSN
- Year of Publication
- Volume
- Issue
- Begin Page
- EndPage Digit
- Supplement
- Document title (first 40 characters, first 50 characters, first 75 characters, TitleEnd)

The objective of the modeling experiments was to find a solution which minimized the “grey zone”, or the range of probabilistic scores in which true and false matches overlapped. Three sets of conditions that adversely impacted model performance were observed:

1) Inclusion of variables with systematic patterns of disagreement between datasets
2) Inclusion of variables with high rates of truly null data
3) Inadequate sampling of title strings

In the first condition, there is a systematic pattern of differences in ISSN’s between the WOS and Medline datasets. Conference Proceedings in WOS are indexed under a Journal ISSN, while the proceedings in Medline are indexed under multiple unique ISSN’s for the AMIA proceedings (Table 3.9). In the second condition, there are a high percentage of null values for the Middle Initial and Supplement variables in both the WOS and Medline dataset (Table 3.10 and 3.11. The Supplement variable was created with the intention that it might help distinguish conference proceeding from journal articles in the WOS database. However the inclusion of ISSN, Middle Initial and Supplement variables reduced the weight attributed to highly discriminate variables such as Title and Author (Table 3.10 and 3.11), and led to situations in which a pair of citations that matched on ISSN but not other more critical variables such as Title scored high enough to become a false match.

Table 3.10 Frequency Distributions Of WOS Variables, N= 11,752

Variable	% null	Unique Values	Frequency Dist
ISSN	0.00	16	0736-8593 (119) to 1067-5027 (1674)
Journal Abbreviation	0.00	16	COMPUT NURS (119) to J AMER MED INFORM ASSOC (1674)
Journal Title	0.00	16	COMPUTERS IN NURSING (119) to JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION (1674)
Year	0.00	41	1966 (26) to 1994 (825)
Volume	10.00	76	Volume 59 (18) to Volume 16 (370)
Issue	10.50	8	Issue 9 (8) to Issue 1 (2935)
Begin Page	0.01	1001	217 pages with count of 1 to Page 1 (222)
End Page Digit	4.70	14	EPD=1(1076) to EPD=4(1162), with exception of a few chars
Supplement	99.2	1	S (94)
First Author Last Name	1.80	5935	3924 Last Names with count of 1 to Miller (50)
First Author First Initial	1.80	26	Q (15) to J (1347)
First Author Middle Initial	46.30	26	X(6) to J(694)
Title50	0.0	11662	11594 unique titles to “UNTITLED” (13)
TitleEnd (Word)	0.0	2948	1693 unique words to “SYSTEM” (345)

Table 3.11 Frequency Distributions Of Medline Variables, N = 21,771

Variable	% null	Unique Values	Frequency Dist
ISSN	8.2	23	1538-2931 (101) to 1367-4803 (2198)
Journal Abbreviation	<0.001	26	Comput Inform Nurs (101) to Bioinformatics (2198)
Journal Title	0.3	26	Computers, informatics, nursing : CIN. (101) to Bioinformatics(Oxford, England) (2190)
Year	0.0	43	Year 1962 (18) to year 2003 (1810)
Volume	12.6	77	Vol 89 (8) to Vol 8 (1183)
Issue	20.6	18	Issue 15 (47) to Issue 1 (4240)
Begin Page	<0.001	1986	Pg 2429 (1) to pg 1 (336)
End Page Digit	<0.001	15	EPD=1 (2082) to EPD=8 (2253), with exception of a few chars
Supplement	95.4	1	S(990)
First Author Last Name	1.1	9670	6193 Last names with count of 1 to Miller (79)
First Author First Initial	1.1	26	Q(25) to D (2462)
First Author Middle Initial	49.0	26	X (7) to A (1202)
Title50	0.0	21485	21276 unique titles to “Law and Ethics” (14)
TitleEnd (word)	0.0		2418 unique words to “system” (628)

In the third condition, the document title was initially abbreviated to the first 40 characters as this was the shortest point at which titles broke across two lines in the raw citation data files. However it was found that this was a source of errors due to titles which are identical within the first 40 characters, such as studies published in multiple parts where the latter part of the title distinguishes the documents. An attempt was made to include up to the first 75 characters of titles, but this length string comparison crashed the LinkPlus software. A compromise solution found was to create two variables for title, one that sampled the first 50 characters, and one which sampled the last word from the title. The final set of variables was selected and the associated matching parameters obtained from the EM algorithm are shown in Figure 3.2.

```
                                             Linking Process
Indirect method Is employed
Field for Blocking
BeginPage

                                         Matching Parameters
           Matching Field   m-prob   u-prob      agree   disagree     matching method
                BeginPage  0.95000  0.00194    5.63580   -2.72506               exact
                   Volume  0.95000  0.03916    2.90250   -2.69047               exact
                     Year  0.95000  0.04783    2.72045   -2.68222               exact
                    Issue  0.95000  0.15732    1.63679   -2.57103               exact
             EndPageDigit  0.95000  0.09993    2.04986   -2.63100               exact
                 LastName  0.95000  0.00051    6.84846   -2.72637      generic string
             FirstInitial  0.95000  0.05884    2.53195   -2.67163               exact
                  Title50  0.95000  0.00005    8.93527   -2.72679      generic string
                 TitleEnd  0.95000  0.00339    5.12913   -2.72374               exact

m-prob: The probability that a matching variable agrees given that the comparison pair
being examined is a match
u-prob: The probability that a matching variable agrees given that comparison pair
being examined as a non-match
agree: The agreement weight assigned for an agreement on a given matching variable
disagree: The disagreement weight assigned for a disagreement on a given matching
variable
matching method: The method used for computing the weight on a given matching
variable.
```

Figure 3.2 Probabilistic Model Parameters.

3.1.5 Data Analysis

3.1.5.1 Truth Database and Model Test Environment

Model performance was assessed using a gold standard “truth” dataset. The truth dataset establishes the identity of the true matching citation in the Medline dataset for each citation in a sample taken from the WOS dataset. The first step in developing the truth dataset was to randomly split the WOS dataset into equal pools of potential case and control citations from which three ten-percent samples without replacement are drawn. The case citations are WOS citations for which there is a true matching citation in the Medline dataset. The control citations are WOS citations for which the identity of the true matching citation in Medline is known, but the citation is withheld from the Medline dataset during model testing. The WOS dataset was randomized into two approximately equal groups of 1600 cases and 1600 controls using a Visual Basic (VB) random number generator function. For any given initial seed supplied to the VB random function, the same number sequence is generated because each successive call to the Rnd() function uses the previous number as a seed for the next number in the sequence. By supplying the unique document i.d. number for each citation as the seed, new random numbers

Table 3.12 Distribution Of Records By Case/Control Status, Journal, And Decade After Randomization

	Cases						Controls				
	Decade						Decade				
Journal title	1960	1970	1980	1990	2000		1960	1970	1980	1990	2000
Artificial Intelligence In Medicine				104	151					97	97
Cin-Computers Informatics Nursing					55						66
Computer Methods And Programs In Biomedicine			189	431	189				169	440	191
Computer Programs In Biomedicine		93	127					92	125		
Computers And Biomedical Research	33	231	248	190	12		36	246	223	167	17
Computers In Nursing				41	12					52	14
Ieee Transactions On Information Technology In Biomedicine					117						93
International Journal Of Bio-Medical Computing		51	174	285				63	179	269	
International Journal Of Medical Informatics				167	218					153	198
International Journal Of Technology Assessment In Health Care				189	158					199	196
Journal Of Biomedical Informatics					72						80
Journal Of The American Medical Informatics Association				529	307					510	328
M D Computing			81	131	22				77	164	25
Medical Decision Making			76	226	122				100	225	122
Medical Informatics And The Internet In Medicine				9	50					16	61
Methods Of Information In Medicine	78	103	154	287	164		85	121	149	257	174
Total	111	478	1049	2589	1649		121	522	1022	2549	1662

sequences are generated. After the initial randomization and taking the top fifty percent of random numbers as cases the pools of cases and controls are similar in distribution by journal and year of publication as shown in Table 3.12. The pools of cases and controls were then each equally sampled three times for samples of size n =1,180 (or ten-percent of total dataset size of 11,752). Because there were citations in the WOS dataset for which there was not a true match in the Medline dataset, slightly oversize samples were required (10.3%) to identify a sufficient number of true cases in each sample. The three samples constitute the truth database, for which the true identity of matching citations in the Medline dataset was determined. The process used to locate the true matches consisted of a combination of detailed exact match relational database queries, manual review of citation records, manual review of journal archives and source documents where needed, and occasional use of the Babelfish website to translate document titles from German to English. The stepwise process followed is detailed in Table 3.13.

Table 3.13 Stepwise Process For Development Of Truth Database

Step	Process	Criteria	Decision
1	Query	Exact Match on 8 variables: Last Name, First Initial, Journal, Year, Volume, Issue, BeginPage, TitleAbbrev.	Accept citation pairs as True Match
2	Query	Replace hyphens with space in WOS Title, Exact Match on 8 variables	Accept citation pairs as True Match
3	Query	Replace hyphens with space in Medl Title, Exact Match on 8 variables	Accept citation pairs as True Match
4	Query	Add leading “a “ to WOS title, Exact Match on 8 variables	Accept citation pairs as True Match
5	Query	Replace colons in Medl Title with hyphens, Exact Match on 8 variables	Accept citation pairs as True Match
6	Query	Exact Match on 7 variables: Last Name, First Initial, Journal, Year, Volume, Issue, BeginPage	Manually review pairs to identify true matches
7	Query	Exact Match on 5 variables: Journal, Year, Volume, Issue, BeginPage	Manually review pairs to identify true matches
8	Query	Exact Match on 5 variables: Last Name, First Initial, Journal, Year, BeginPage	Manually review pairs to identify true matches
9	Query	Exact Match on 5 variables: Journal, Year, Volume, Issue, TitleAbbrev	Manually review pairs to identify true matches
10	Query	All Citations not yet matched	Manually search database with multiple single variable search strategies, including filters and wildcard searches

The truth database was further validated by investigation of all false positive and false negative matching errors found during the record linkage model testing to ensure they were true errors and not an error in the truth database. The model testing consisted of a record linkage of each sample of 1,180 WOS records to 20,001 Medline records from which the controls had been withheld (Table 3.14), for a total of three trials for each of five record linkage models.

Table 3.14 Record Linkage Test Environment

	WOS DS 10% sample N= 1180	Linked to -→	Medline DS N= 20,001
(Cases)	Original Journals n = 590		Original Journals n= 14149
(Controls)	Original Journals n = 590		(withheld, n= 0)
			Added Journals n = 5852

.

The output from the deterministic record linkage is a set of linked records (pairs of citations that matched on the model variables) and unlinked records (citations from the WOS sample for which no match was found. The document ID numbers of the links and non links are then compared to the truth database and the true match/non-match status scored as follows:

Linked (1): If comparison pair is actually a match (True Positive), Match = 1
If the comparison pair is not a match (False Positive), Match = 0
Unlinked (0): If the unlinked citation was a Control (True Negative), Match = 0
If the unlinked citation was a Case (False Negative), Match = 1

The output from the probabilistic record linkage is a set of comparison pairs of linked records that have received a total weight for probability of agreement that exceeds a threshold score. Unlinked records may be either a record for no likely match was found, or pairs of records for which the probability score was below the cut-point. The document ID numbers of the links and non links are then compared to the truth database and the true match/non-match status is again scored as follows:

Linked (1): If comparison pair is actually a match (True Positive), Match = 1
If the comparison pair is not a match (False Positive), Match = 0
Unlinked (0): If the unlinked citation was a Control (True Negative), Match = 0
If the unlinked citation was a Case (False Negative), Match = 1

3.1.5.2 ROC analysis

The performance of the record linkage models was evaluated through ROC curve comparison analysis that was performed using STATA statistical software. In recent years ROC curves have been increasingly adopted in the machine learning and data mining research communities. Receiver Operating Characteristics (ROC) curves are used as a metric for evaluating classification and prediction rules and visualizing their performance (Fawcett, 2004). The objective of record linkage is to classify pairs of records as matches or non-matches. Figure 3.3 shows a bimodal distribution of total weight scores for matches and non-matches in a

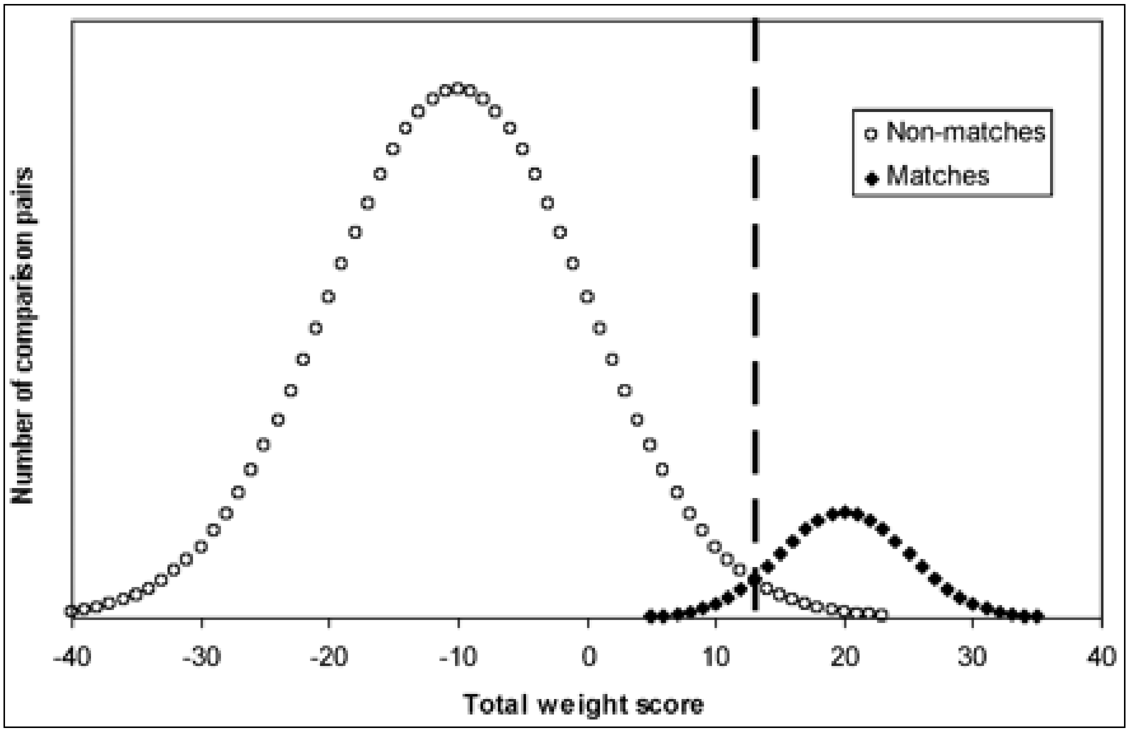


Figure 3.3 Number Of Comparison Pairs For Matches And Non-Matches By Total Weight Score In A Probabilistic Record Linkage Project (Blakeley, 2000)

hypothetical record linkage project. It is not usually possible to determine exactly which comparison pairs are matches and non-matches during the linkage process, just the observed number of comparison pairs (matches and non-matches) at any given total weight score are available. The task in record linkage is to set a cut-off weight (of the total weight) above which the majority of comparison pairs are true matches and below which the comparison pairs are categorized as true non-matches. The vertical dotted line in Figure 3.3 is a possible cut-off score.

A two-by-two table of link/non-link status by match/non-match status is shown below (Table 3.15), which is also referred to as a confusion matrix. A match can be considered to be equivalent to having the outcome of interest in an epidemiological study (e.g. death), and the performance of the record linkage in classifying the outcome can be quantified with the familiar terms:

Sensitivity (True positive Rate) = a/(a + c)
Specificity (True Negative Rate) = d/(b + d)
Positive predictive value = a/(a +b)
Negative predictive value = d/(c+d)

Table 3.15 Confusion Matrix

	Matches (1)	Non-matches (0)
Linked (1)	a (true positives)	b (false positives)
Unlinked (0)	C (false negatives)	d (true negatives)

In evaluating record linkage model performance, the terms are defined as:

* Sensitivity: How well the model detects matches
* Specificity: How well the model detects non-matches

These parameters will vary depending on the cut-off weight: moving it to the left in Figure 3.3 will increase the sensitivity, but also increase the number of false positives; moving it to the right will increase the specificity, but also increase the number of false negatives.

The Area under the ROC Curve (AUC) (Figure 3.4) is a single index of the ability of a test to classify true positive and true negative cases. ROC curves can be compared statistically (Hanley, 1983), and routines for comparison are available in Stata.

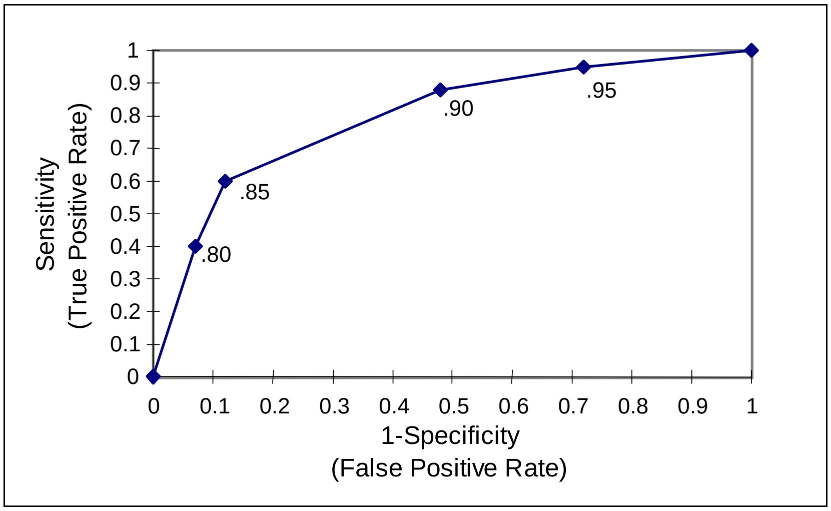


Figure 3.4 Area Under The ROC Curve (AUC)

The Sample Calculation for ROC curve comparison calculates the required sample size for the comparison of the areas beneath two ROC curves derived from the same cases. The sample size takes into account the required significance level and power of the test.

The required parameters of sample size calculation are:

1) Type I error - alpha: the probability of making a Type I error (a-level), i.e. the probability of rejecting the null hypothesis when in fact it is true.

2) Type II error - beta: the probability of making a Type II error (b-level), i.e. the probability of accepting the null hypothesis when in fact it is false.

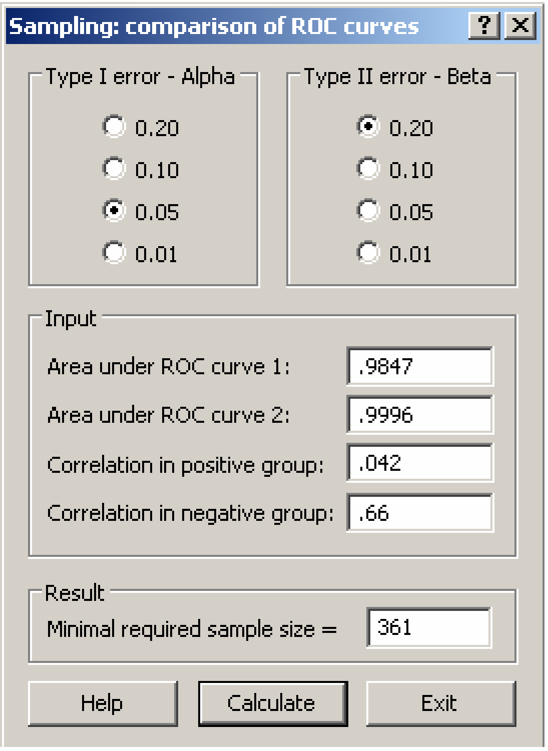
3) Area under ROC curve 1: hypothesized area for the first ROC curve.

4) Area under ROC curve 2: hypothesized area for the second ROC curve.

5) Correlation in positive group: the hypothesized rank correlation coefficient in the positive group (matched records)

6) Correlation in negative group: the hypothesized rank correlation coefficient in the negative group (non-matched records)

The minimum required sample for the model development was initially calculated using MedCalc software with estimates obtained from a pilot study, and re-calculated after model development. A minimum of 361 records is required in both the true match and non-match groups, for a total minimum sample size of approximately 720 records (Figure 3.5). A case/control design was used to sample records so adequate sample size was obtained for both the true match (case) and non-match (control) groups.



Sampling: comparison of ROC curves

Type I error - Alpha

○ 0.20
○ 0.10
● 0.05
○ 0.01

Type II error - Beta

● 0.20
○ 0.10
○ 0.05
○ 0.01

Input

Area under ROC curve 1: .9847
Area under ROC curve 2: .9996
Correlation in positive group: .042
Correlation in negative group: .66

Result

Minimal required sample size = 361

Figure 3.5 Sample Size Calculations For ROC Curve Comparison

3.2 Studies of the Effect of Record Linkage and Information Fusion

3.2.1 Sample Selection – Medical Informatics

The dataset for analysis of medical informatics that was developed for prior studies was the primary basis for this study (Synnestvedt et al, 2005). The dataset was defined by cross-referencing the Institute for Scientific Information's (ISI) Journal Citation Reports list of medical informatics journals for 2003 against a list of medical informatics journals from AMIA(AMIA, 2003). The twelve journals that both resources identified as important or relevant to medical informatics were selected for study. These twelve journals were also checked against the NCBI journals database for publication history, and the journals which were predecessors of some of the current journals were identified. The dataset covers the time period 1964-2004 (Table 3.1 in section 3.1.1 of Methods).

3.2.2 Sample Selection – HIV/AIDS

The methodology of record linkage with information fusion was validated with an alternate knowledge domain analysis. The first validation study used a sample of HIV/AIDS literature drawn from the AIDS subset of Medline and a sample of related journals drawn from WOS. Three infectious disease specialists in HIV/AIDS were polled for information on important journals in their field (Table 3.16), and all journals which received two or more votes were used to define the HIV/AIDS dataset. The challenge in analyzing the HIV/AIDS data is both the size of the literature and that it cannot be defined solely on the basis of Journals in WOS as most of the candidate journals cover either broad subject areas of medicine in general or infectious disease areas. HIV/AIDS specific subject terms are needed to select the data from WOS which may be problematical. An approach is taken that will demonstrate the benefit of the use of record linkage to define samples using an external standard, which in this case will be the AIDS subset of Medline. A second dataset was collected from WOS for the nine study journals with added subject terms (Figure 3.6), which became the baseline, or reference dataset.

Table 3.16 HIV/AIDS Journals And Coverage In WOS

WOS Journal Names	Votes	JCR 2005 Impact Factor	Years covered	In Study
AIDS CARE PSYCHOLOGICAL AND SOCIO MEDICAL ASPECTS OF AIDS HIV	1	Not avail	1992-	
AIDS	111	5.835	May 1987-	✓
AIDS PATIENT CARE AIDS PATIENT CARE AND STDS	11	1.944	Feb 1992 – Dec 1995 Feb 1996-	✓
AIDS RESEARCH AIDS RESEARCH AND HUMAN RETROVIRUSES	1	2.531	1986 1987 -	
ANNALS OF INTERNAL MEDICINE	11	13.254	1987-	✓
ARCHIVES OF INTERNAL MEDICINE	11	8.016	1983	✓
CLINICAL INFECTIOUS DISEASES	11	6.510	1992	✓
JOURNAL OF INFECTIOUS DISEASES	11	4.953	1983 -	✓
JOURNAL OF ACQUIRED IMMUNE DEFICIENCY SYNDROMES JAIDS JOURNAL OF ACQUIRED IMMUNE DEFICIENCY SYNDROMES	111	3.681	Oct 1992 – Aug 2002 Oct 2002 –	✓
JAMA JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION	11	21.455	1983 –	✓
JANAC JOURNAL OF THE ASSOCIATION OF NURSES IN AIDS CARE	1	Not avail	2004 -	
LANCET	1	18.316	1983 -	
NATURE	1	30.979	1983 -	
NATURE MEDICINE	1	30.550	1995 -	
NEW ENGLAND JOURNAL OF MEDICINE	11	34.833	1982 –	✓
SCIENCE	1	29.162	1983 –	

S=("acquired immune deficiency syndrome" OR "gay-related immune deficiency" OR "cellular immune deficiency" OR "acquired immunodeficiency syndrome" OR "human immunodeficiency virus" OR "human immune deficiency virus" OR HIV or AIDS) AND SO=(AIDS OR AIDS PATIENT CARE "AND" STDS OR ANNALS OF INTERNAL MEDICINE OR ARCHIVES OF INTERNAL MEDICINE OR CLINICAL INFECTIOUS DISEASES OR JOURNAL OF INFECTIOUS DISEASES OR JAIDS JOURNAL OF ACQUIRED IMMUNE DEFICIENCY SYNDROMES OR JAMA JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION OR NEW ENGLAND JOURNAL OF MEDICINE)
DocType=All document types; Language=English; Databases=SCI-EXPANDED, SSCI, A&HCI; Timespan=2003-2005

Figure 3.6 Query For Baseline HIV/AIDS Dataset, N= 4149

The Medline dataset consists of all citations for the study Journals with the added limit on the query of being in the AIDS subset of Medline (Figure 3.7)

```
("AIDS (London, England)"[Jour] OR "AIDS patient care and STDs"[Jour] OR "Annals of internal
medicine"[Jour] OR "Archives of internal medicine"[Jour] OR "Clinical infectious diseases : an official
publication of the Infectious Diseases Society of America"[Jour] OR "The Journal of infectious
diseases"[Jour] OR "Journal of acquired immune deficiency syndromes (1999)"[Jour] OR "JAMA : the
journal of the American Medical Association"[Jour] OR "The New England journal of medicine"[Jour])
AND AIDS[sb] AND ("2003/01/01"[PDAT] : "2005/12/31"[PDAT]
```

Figure 3.7 Query For Medline HIV/AIDS Dataset, N= 4,690

A comparison of the baseline WOS and Medline datasets showed over 10% fewer records retrieved from WOS, with fewer records returned from the majority of journals (Table 3.17). A second finding of this comparison was the systematic differences in ISSN's and Journal Titles, as found previously in the medical informatics dataset.

Table 3.17 Distribution Of Citations For WOS And Medline HIV/AIDS Datasets

WOS				Medline			
ISSN	**JournalAbbrev**	**#**	**%**	**ISSN**	**JournalAbbrev**	**#**	**%**
0269-9370	AIDS	1337	32.2	0269-9370	AIDS	1484	31.6
1087-2914	AIDS PATIENT CARE STDS	215	5.2	1087-2914	AIDS Patient Care STDS	501	10.7
0003-4819	ANN INTERN MED	41	1.0	1539-3704	Ann Intern Med	64	1.4
0003-9926	ARCH INTERN MED	39	0.9	0003-9926	Arch Intern Med	35	0.7
1058-4838	CLIN INFECT DIS	721	17.4	1537-6591	Clin Infect Dis	701	15.0
1525-4135	JAIDS	931	22.4	1525-4135	J Acquir Immune Defic Syndr	977	20.8
0022-1899	J INFEC DIS	589	14.2	0022-1899	J Infect Dis	629	13.4
0098-7484	JAMA-J AM MED ASSN	124	3.0	1538-3598	JAMA	84	1.8
		0	0	0098-7484	JAMA	19	0.4
0028-4793	N ENGL J MED	152	3.7	1533-4406	N Engl J Med	196	4.2
Total		4149	100.0			4690	100.0

A second dataset was then collected from WOS that consists of all citations for the study Journals from the same time period, i.e., no subject terms were added to the retrieval query (Figure 3.8). This dataset was then linked to the Medline dataset using probabilistic record linkage to define the comparison dataset for fusion of MeSH terms.

```
SO=(AIDS OR AIDS PATIENT CARE "AND" STDS OR ANNALS OF INTERNAL MEDICINE OR
ARCHIVES OF INTERNAL MEDICINE OR CLINICAL INFECTIOUS DISEASES OR JOURNAL OF
INFECTIOUS DISEASES OR JAIDS JOURNAL OF ACQUIRED IMMUNE DEFICIENCY
SYNDROMES OR JAMA JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION OR NEW
ENGLAND JOURNAL OF MEDICINE)
DocType=All document types; Language=English; Databases=SCI-EXPANDED, SSCI, A&HCI;
Timespan=2003-2005
```

Figure 3.8 Query for WOS Comparison HIV/AIDS Dataset, N= 20,314

After linkage, the WOS dataset contained 4,252 records (Table 3.18), which was the net result of dropping 340 records that were in the baseline WOS dataset, and adding 443 records identified from the Medline dataset.

Table 3.18 Distribution Of Records In HIV/AIDS Dataset, After Linkage Of WOS And Medline

Journal Abbrev	WOS	Medline	%
AIDS	1408	1408	33.1
AIDS PATIENT CARE STDS	212	212	5.0
ANN INTERN MED	42	42	0.9
ARCH INTERN MED	35	35	0.8
CLIN INFECT DIS	688	688	16.2
J INFEC DIS	629	629	14.8
JAIDS	959	959	22.6
JAMA-J AM MED ASSN	97	97	2.3
N ENGL J MED	182	182	4.3
Total	4252	4252	100

3.2.3 Sample Selection - Nursing Informatics

The methodology of record linkage with information fusion was also validated with an alternate database analysis. An additional sample of medical informatics citation data was collected from CINAHL, the Cumulative Index to Nursing & Allied Health Literature. Data was collected for four journals which overlap with the medical informatics dataset definition (Table 3.19). CINAHL's coverage is not as long-term as Medline and fewer journals are indexed, but the focus is on nursing and allied health literature. Indexing terms are based upon MeSH with addition of nursing / allied health specific terms called CINAHL Subject Headings (CINAHL, 2006).

Table 3.19 Informatics Journal Coverage In CINAHL

	Journal Title	**Years Indexed in WOS**	**Years Indexed In CINAHL**	**Records in CINAHL Dataset**
1538-2931	Cin-Computers Informatics Nursing	2002	2002 -	388
0736-8593	Computers In Nursing (1)	1992-2002	1983-2002	641
1067-5027	Journal Of The American Medical Informatics Association	1994 -	1994 -	546
0272-989X	Medical Decision Making	1983 –	2001 -	115
1463-9238	Medical Informatics And The Internet In Medicine	1999 -	2002 -	82
	Total CINAHL			1772

1: Continued by Cin-Computers Informatics Nursing

3.2.4 Data analysis

Data quality metrics were compared for datasets prepared without and with record linkage, and the effect on subsequent visualizations demonstrated. Data quality was compared by the percentage of non-null data in keywords and abstracts. Visualizations of baseline and prepared datasets were developed using CiteSpace version 2.1.R1, and metrics collected for changes in burst terms, nodes & links, ranking of key terms.

CHAPTER 4: PERFORMANCE OF RECORD LINKAGE MODELS

The competing objectives in developing a record linkage model are to create a model with a combination of variables that are sufficient to uniquely identify citations that is also not subject to missed matches because of differences within variables between two sets of data. If the variable set used in a record linkage model is insufficient to generate a unique key, the resulting linkage will contain false positive links, i.e. citations that match on key variables but are not the referring to the same publication. In addition the use of a non-unique key leads to a Cartesian product problem. In a database sense, a Cartesian product is the cross-product of all possible record pairs that match on the model variables. For each pair of records involved in a non-unique key, the resulting record-linkage prepared dataset will contain four records. The problem with Cartesian products in the context of citation data prepared with record linkage and information fusion is that there will be the insertion of key terms and abstracts into both correctly and incorrectly matched citations, resulting in a doubling or greater increase in the raw counts of key words and an incorrect association between key terms and cited documents.

If the variable set used in a record linkage model is sufficient for uniquely identifying citations but subject to missed matches because of differences in variables between data sets the result will be false negative links, or citations for which a match should have been found but was not. The problem with false negative links in the context of citation data prepared with record linkage and information fusion is that the missed matches may be correlated with a specific journal or period of time or type of article. An anomalous pattern of missed matches can lead to a skewed visualization in which an area of specialization or period of time within a knowledge domain is not well represented.

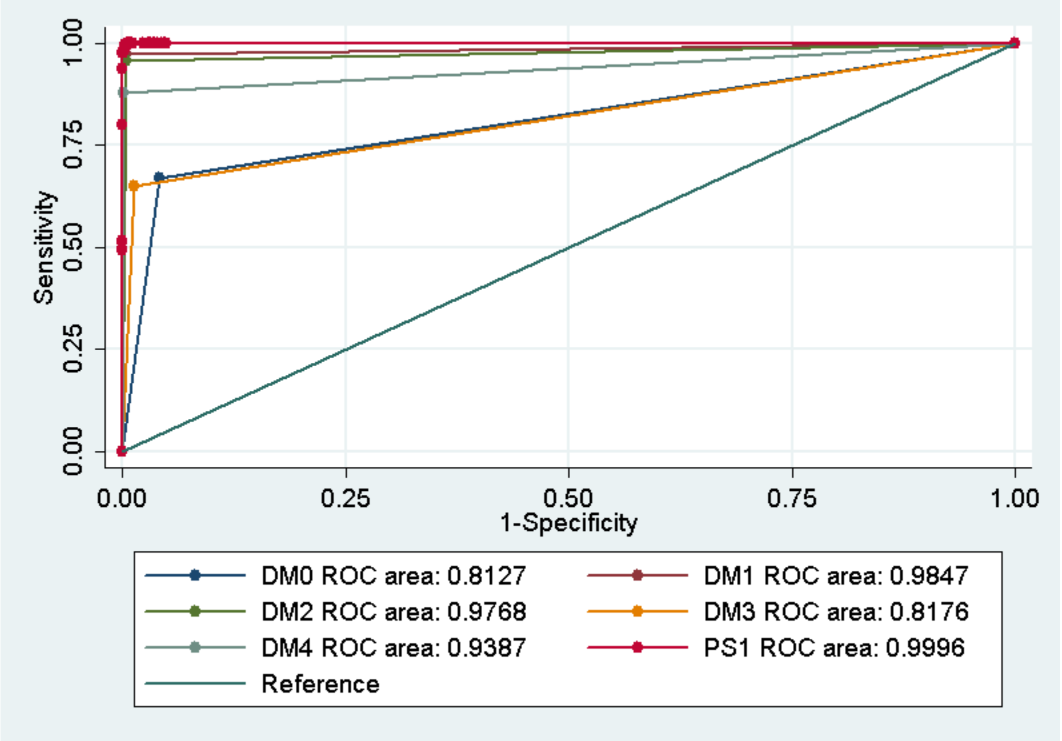
The five deterministic models and the probabilistic model are first compared by Receiver Operating Characteristic (ROC) curves for overall performance, followed by a ROC comparison of the highest performing deterministic model and the probabilistic model. The model errors are then examined for patterns of failure and the reasons for failure.

4.1 ROC Analysis – 6 Models

The results of the ROC analyses of three independent trials of linkage of samples of 1,180 WOS records to 20,000 Medline records are detailed in Table 4.1 and the AUC's (the area under the ROC curve) for the combined results of the trials are graphically summarized in Figure 4.1. The overall performance of models was stable across samples and there is a significant difference between models (P = 0.000).

Table 4.1 Detailed Statistics For Three Trials Of ROC Comparison Of Deterministic And Probabilistic Record Linkage Models

Ho: area(DM0) = area(DM1) = area(DM2) = area(DM3) = area(DM4) = area(PS1)							
Sample	Model	Obs	ROC Area	Std. Err.	[95% Conf.	Interval]	Prob>chi2
1	DM0	1198	0.8252	0.0103	0.80497	0.84534	0.0000
	DM1	1181	0.9831	0.0037	0.97572	0.99039	
	DM2	1181	0.9763	0.0044	0.96765	0.98490	
	DM3	1186	0.8289	0.0100	0.80932	0.84846	
	DM4	1181	0.9339	0.0070	0.92020	0.94760	
	PS1	1180	0.9990	0.0009	0.99726	1.00000	
2	DM0	1210	0.7915	0.0110	0.76989	0.81305	0.0000
	DM1	1181	0.9839	0.0036	0.97675	0.99105	
	DM2	1180	0.9763	0.0044	0.96767	0.98487	
	DM3	1189	0.8002	0.0104	0.77970	0.82066	
	DM4	1181	0.9424	0.0066	0.92943	0.95532	
	PS1	1180	1.0000	0.0000	0.99998	1.00000	
3	DM0	1186	0.8219	0.0102	0.80190	0.84188	0.0000
	DM1	1181	0.9873	0.0033	0.98091	0.99367	
	DM2	1181	0.9780	0.0042	0.96965	0.98630	
	DM3	1181	0.8237	0.0099	0.80434	0.84312	
	DM4	1180	0.9398	0.0067	0.92669	0.95297	
	PS1	1180	0.9999	0.0001	0.99980	1.00000	
	Key DM0= DMatch0 (Title50) DM1 = DMatch1 (LastName4,Year,BeginPage) DM2 = DMatch2 (LastName,Year,BeginPage) DM3 = DMatch3 (LastName,Year,Title50) DM4 = DMatch4 (ISSN,Year,Volume, Issue, BeginPage) PS1 = PScore (Probabilistic Model: Year, Volume, Issue, Begin page, EndPageDigit, Last Name, First Initial, Title50, TitleEnd)						


ROC Curves

Key:
DM0= DMatch0 (Title50)
DM1 = DMatch1 (LastName4,Year,BeginPage)
DM2 = DMatch2 (LastName,Year,BeginPage)
DM3 = DMatch3 (LastName,Year,Title50)
DM4 = DMatch4 (ISSN,Year,Volume, Issue, BeginPage)
PS1 = PScore (Probabilistic Model: Year, Volume, Issue, Begin page, EndPageDigit, Last Name, First Initial, Title50, TitleEnd)

Figure 4.1 ROC Comparison Of Deterministic And Probabilistic Record Models

The 6-way comparison of models does not indicate which pairs of models are significantly different from each other. The probabilistic model consistently has the highest AUC (.999 – 1.0), with the deterministic models of Slach (1985) and Committee based on Author Last Name, Year and Page having a slightly lower AUC between .97 8 and .987, and these models are selected for further comparison.

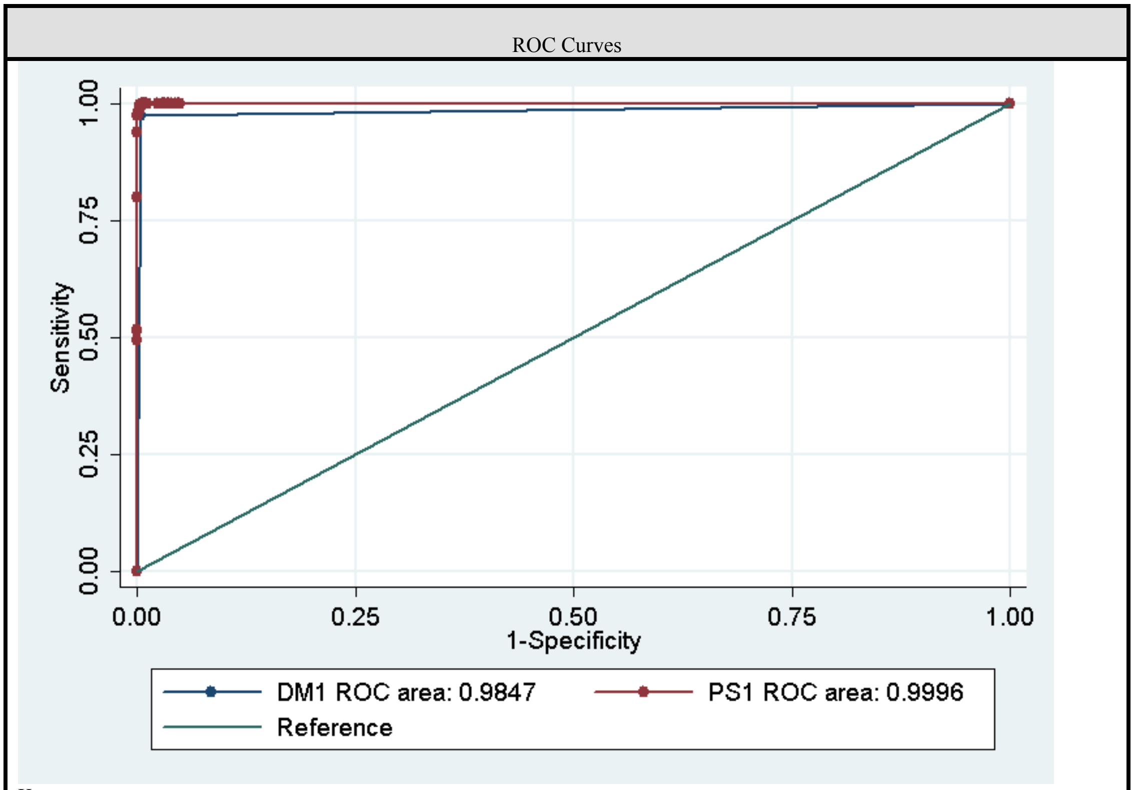
4.2 ROC Analysis – Probabilistic Model versus Deterministic Model 1

The overall performance of the DMatch1 and DMatch2 models are equivalent in terms of ROC area with overlapping confidence intervals. The deterministic model DMatch1 is chosen over DMatch2 for direct comparison to the probabilistic model as the match key of Slach has been recorded in the literature. The ROC curve comparison of the two models (DMatch1 and Pscore) was again stable across samples and significant (p=.0000) (Table 4.2 and Figure 4.2).

Table 4.2 Detailed Statistics For Three Trials Of ROC Comparison Of Best Deterministic Model To Probabilistic Model

Ho: area(DM1) = area(PS1)							
Sample	Model	Obs	ROC Area	Std. Err.	[95% Conf.	Interval]	Prob>chi2
1	DM1	1181	0.9831	0.0037	0.97572	0.99039	0.0000
	PS1	1180	0.9990	0.0009	0.99726	1.00000	
2	DM1	1181	0.9839	0.0036	0.97675	0.99105	0.0000
	PS1	1180	1.0000	0.0000	0.99998	1.00000	
3	DM1	1181	0.9873	0.0033	0.98091	0.99367	0.0001
	PS1	1180	0.9999	0.0001	0.99980	1.00000	

Key:
DM1 = DMatch1 (LastName4,Year,BeginPage)
PS1 = PScore (Probabilistic Model: Year, Volume, Issue, Begin page, EndPageDigit, Last Name, First Initial, Title50, TitleEnd)



Key
DM1 = DMatch1 (LastName4,Year,BeginPage)
PS1 = PScore (Probabilistic Model: Year, Volume, Issue, Begin page, EndPageDigit, Last Name, First Initial, Title50, TitleEnd)

Figure 4.2 Comparison Of Best Deterministic And Probabilistic Model

4.3 Analysis of Model Performance and Errors

We have said previously that the challenge in developing a record linkage model is to create a model with a combination of variables that are sufficient to uniquely identify citations, while also not being subject to missed matches because of differences within variables between two sets of data. Prior to discussion of model performance two sets of baseline data are presented for reference. First, Table 4.3 compares the models in terms of the performance of the variable sets in generating a unique key in the medical informatics datasets. The finding from this comparison is that the variable sets used in the deterministic models are unable to generate a completely unique key on either the WOS or Medline medical informatics datasets.

Table 4.3 Comparison Of Variable Sets In Generating Unique Key

Model	Variables	WOS Medical Informatics 1964-2004 (N=11,752) Uniquely Keyed Records n (%)	MEDLINE Medical Informatics 1962-2004 (N=21,771) Uniquely Keyed Records n (%)
DM0	Title50	11479 (97.7)	21277 (97.7)
DM1	LastName(4), Year, Begin Page	11712 (99.7)	21693 (99.6)
DM2	LastName, Year, Begin Page	11716 (99.7)	21703 (99.7)
DM3	LastName, Year, Title50	11651 (99.1)	21548 (99.0)
DM4	ISSN, Year, Volume, Issue, BeginPage	11710 (99.6)	21502 (98.8)
PS1	Year, Volume, Issue, BeginPage, EndPageDigit, LastName, FirstInitial, Title50, TitleEnd	11752 (100.0)	21771 (100.0)

The second table (Table 4.4) examines rates of agreement on single variables for the 3540 matched citations from the truth dataset used in model testing, and summarizes typical reasons observed for non-agreement. As expected from prior discussion of variability in ISSN and Journal Titles between WOS and Medline (Methods, Table 3.9), there are low rates of agreement for the Journal Abbreviation and Journal Title variables, and the ISSN agreement rate is not high. While Title was previously observed to have the highest discriminating value as a single variable based on frequency distributions within datasets (Table 3.10, Methods), it has a low rate of agreement between datasets that lowers it's usefulness as an variable in a deterministic model.

Table 4.4 Rates Of Agreement On Single Variables For 3540 Matched Citations

Variable	# in agreement	% agreement	Typical reasons for non-agreement
Year	3535	99.8	Indexing error in WOS
Begin Page	3531	99.7	1) Error in page number 2) Use of roman numerals vs. not
Issue	3525	99.6	1) 1964-1965 Methods of Information in Medicine 2) Indexing of conference proceedings under Journal Title (WOS)
Volume	3524	99.5	1) 1964-1965 Methods of Information in Medicine 2) Indexing of conference proceedings under Journal Title (WOS)
First Author First Initial	3523	99.5	Error in designation of first author
First Author Middle Initial	3445	97.3	1) Middle initial present (WOS) vs. null (Medline) 2) Middle initial present (Medline) vs. null (WOS) 3) Error in designation of first author
First Author Last Name	3402	96.1	1) Truncating last name (WOS) 2) Use of hyphens, apostrophes, spaces (Medline) vs. not (WOS) 3) Error in designation of first author 4) Author name null (Medline) vs. Author Name Present (WOS)
End Page Digit	3353	94.7	1-digit difference in end page
ISSN	3141	88.7	1) indexing of conference proceedings under Journal ISSN (WOS) 2) Use of ESSN (WOS) vs ISSN (Medline) 3) Conference proceedings without ISSN (Medline)
Title50	2363	66.7	Omitting article of speech (WOS) Omitting leading portion of title (WOS) Use of numeric characters (WOS) vs Text (e.g., 3 vs. three) Variable punctuation (e.g., hyphens vs. colons) Hyphenated terms (WOS) , e.g. "data-analysis" vs. "data analysis" Titles in German language (WOS) vs English language (Medline)
Journal Title	1479	41.8	1) Different spelling, wording, and punctuation of journal titles 2)) Indexing of conference proceedings under Journal Title (WOS)
Journal Abbreviation	1226	34.6	1) Different abbreviations 2)) Indexing of conference proceedings under Journal Title (WOS)

Because the performance of record linkage models was stable across samples and for purposes of discussion, the results of the individual trials have been combined for presentation of the confusion matrix data in Table 4.5, and a discussion of individual models follows.

Table 4.5 Combined Results Of Deterministic And Probabilistic Record Linkages, N=3540

Performance of Record Linkage Models (Combined results for all samples)	Confusion Matrix			Sensitivity a/(a+c)	Specificity d/(b+d)	AUC From Trials
		Matches (Cases)	Non-Matches (Controls)			
	Linked	**a** (true positives)	**b** (false positives)			
	Unlinked	**c** (false negatives)	**d** (true negatives			
DMatch0 (Title50)	N=3594	Matches	Non-Matches	.668	.958	.79 - .83
	Linked	1181	77			
	Unlinked	588	1748			
DMatch1 (LastName4,Year,BeginPage)	N=3543	Matches	Non-Matches	.974	.995	.987
	Linked	1724	8			
	Unlinked	46	1765			
DMatch2 LastName, Year, BeginPage)	N=3542	Matches	Non-Matches	.958	.996	.978
	Linked	1695	7			
	Unlinked	75	1765			
DMatch3 (LastName, Year, Title50)	N=3556	Matches	Non-Matches	.649	.986	.80 - .83
	Linked	1149	25			
	Unlinked	621	1761			
DMatch4 (ISSN, Year, Volume, Issue, BeginPage)	N=3542	Matches	Non-Matches	.879	.998	.93- .94
	Linked	1556	3			
	Unlinked	214	1769			
PScore (Probabilistic Model: Year, Volume, Issue, Begin page, EndPageDigit, Last Name, First Initial, Title50, TitleEnd)	N=3540	Matches	Non-Matches	.995	.997	.999 – 1.0
	Linked	1762	5			
	Unlinked	8	1765			

Deterministic Model #0 (DMatch0): The matching variable evaluated was Title (truncated at 50 characters). As expected, the sensitivity is relatively low due to the number of false negatives generated by difficulty of matching on Title. In addition, because Title is not a completely unique identifier, there are an excess number of linkages returned (3594 vs. 3540) due to false positive matches and the Cartesian product problem.

Deterministic Model #1 (DMatch1): The matching variables set was Year, First Author Last Name (first 4 characters), and Begin page, based on the matchkey reported by Slach (1985).

Deterministic Model #2 (DMatch2): The matching variables set was Year, First Author Last Name, and Begin page, based on recommendation by the committee as being a standard for current good practice and is very similar to DMatch1 with the exception of using the full last name of the first author. DMatch1 and DMatch2 are very similar models that have problems with false negatives when there are differences in spelling and punctuation of last names between datasets. Examples of variations in Author Names between datasets are listed in Table 4.6, and a complete listing can be found in the Appendix (Table A3). DMatch1 and DMatch2 also both return an excess number of links due to false positive matches and the Cartesian product problem. Two authors with same last name publishing at same time will be linked incorrectly – e.g., last names beginning with "Van ", or the multiple "C. Friedman" authors in the field of Medical informatics. A complete listing of linkage errors with the DMatch1 model can be found in the Appendix (Table A4).

Deterministic Model #3 (DMatch3): The variables evaluated were First Author Last Name, Year, and Title (truncated at 50 characters), based on the matching criteria used by the RefWorks bibliography management tool to identify duplicates. This model combines the difficulties and failures of the first 3 models, and ranks equally with DM0 as having the lowest AUC. Any differences in wording, spelling, or punctuation of Name or Title will result in false negatives, and there is a problem with false positives and an excess number of links due to the Cartesian product problem.

Table 4.6 Examples Of Variation In Author Names For Matched Citations

WOS FirstAuthor	Medline FirstAuthor		WOS FirstAuthor	Medline FirstAuthor
af Klercker, T	Klercker T		MALINDZA.GS	Malindzak GS Jr
ARZBAECH.RC	Arzbaecher RC		MCCONVILLE, KMV	Mc Conville KM
BARBOSA, MD	Barbosa M de Matos		MELO, MFV	Vidal Melo MF
BENJEBRIA, A	Ben Jebria A		MINAMIKAWATACHINO, R	Minamikawa-Tachino R
CAMPIONEPICCARDO, J	Campione-Piccardo J		MUSTAKAL.KK	Mustakallio KK
Cosp, XB	Bonfill Cosp X		NORDSCHO.CD	Nordschow CD
DAS, REG	Gaines RE		OCHOASANGRADOR, C	Ochoa-Sangrador C
DEBLIEK, R	de Bliek R		OQUIGLEY, J	O'Quigley J
DECARVALHO, LAV	de Carvalho LA		PATTISONGORDON, E	Pattison-Gordon E
DEMOOR, GJE	De Moor GJ		PIPBERGE.HV	Pipberger HV
DEPONTI, F	De Ponti F		POLIHRON.P	Polihroniadis P
deRoulet, D	de Roulet D		PRYER, DB	Pryor DB
DHOORE, W	D'Hoore W		REICHERT.PL	Reichertz PL
DOMBAL, FTD	de Dombal FT		Schoeffler, KM	Liu GC
EBENCHAIME, M	Eben-Chaime M		SHINOZAK.T	Shinozaki T
FAIRHURST, MC	Fairhust MC		Silveira, PSP	Panse Silveira PS
FEINSTEI.AR	Feinstein AR		SRINIVAS.R	Srinivasan R
FLATLEY, P	Brennan PF		STARTSMA.TS	Startsman TS
France, FHR	Roger France FH		Stoykova, B	Nixon J
GARFINKE.D	Garfinkel D		TAGLIACO.R	Tagliacozzo R
GONCEWINDER, C	Gonce-Winder C		Timothy, TYY	Lai TY
Gonzalez, JS	Solano Gonzalez J		VANALSTE, JA	van Alste JA
GUSTAFSO.DH	Gustafson DH		VANDAMME, M	van Damme M
Guvenir, HA	Altay Guvenir H		VANDENAKKER, TJ	van den Akker TJ
HENDERSO.C	Henderson C		VANDERLEIJE, BA	van der Leije BA
Houghton, J	Haughton J		VANGENNIP, EMSJ	van Gennip EM
JESDINSK.HJ	Jesdinsky HJ		VANKREEL, BK	van Kreel BK
KARBER, G	KAERBER G		vanOverbeeke, JJ	van Overbeeke JJ
Keravnou, ET	Eravnou ET		vanRoijen, L	van Roijen L
Kohl, P	Kokol P		VANZEE, GA	van Zee GA
LEAO, BD	Leao Bde F		VEGACATALAN, FJ	Vega-Catalan FJ
LLEWELLYNTHOMAS, HA	Llewellyn-Thomas HA		WHITINGOKEEFE, QE	Whiting-O'Keefe QE

Deterministic Model #4 (DMatch4): The variables evaluated were ISSN, Year, Volume, Issue, and Begin Page, to avoid matching on author and title text strings. This strategy did not perform as well as the Author-Year-Page models due to variability between the datasets. Primary sources of failure were:

-Differences in indexing of articles by journal ISSN. WOS indexes AMIA conference proceedings under JAMIA ISSN, Medline indexes under proceedings ISSN.

-Different use of print versus. electronic ISSN.

-Missing data in matching variables

DMatch4 is also a non-unique key, and excess links were returned.

Probabilistic Model (PScore): The variables selected for use in the probabilistic linkage model were:

- First Author Last Name
- First Author First Initial
- Year of Publication
- Volume
- Issue
- Begin Page
- EndPage Digit
- Title50, TitleEnd

The false positive errors in the probabilistic model (Table 4.7) are citations from the control group for which an alternate citation existed with an exact or highly similar match on at least 5 of the 9 model variables in a combination with a high enough weight to exceed the score threshold. The probabilistic model linkage selects the single best match, so in a non-experimental situation these errors are not as likely to occur as the true matching citation would be available for linkage.

Table 4.7 False Positive Errors In The Probabilistic Model

DS	First Author	Year	Vol	Issue	Pages	Title50	Title End	ISSN	Journal Abbrev
WOS	Kiel, JM	2000	17	1	27-28	Resolution 2000: Create an inviting e-practice	practice	0724-6811	M D COMPUT
Med	Kiel JM	2000	17	2	27-8	Positive outcomes, lower costs: using net-based IT	care	0724-6811	MD Comput
WOS	Goodman, KW	1999	16	3	17-+	Bioinformatics: Challenges revisited	revisited	0724-6811	M D COMPUT
Med	Goodman KW	1999	16	2	17-20	Health informatics and the Hospital Ethics Committ	Committee	0724-6811	MD Comput
WOS	Kiel, JM	1999	16	3	27-28	Going high tech: Size matters? Think again ...	..	0724-6811	M D COMPUT
Med	Kiel JM	1999	16	5	27-9	yourpractice.com: making the leap to the Internet	Internet	0724-6811	MD Comput
WOS	Sadegh-Zadeh, K	2000	20	3	227-241	Fundamentals of clinical methodology 4. Diagnosis	Diagnosis	0933-3657	ARTIF INTELL MED
Med	Sadegh-Zadeh K	1998	12	3	227-70	Fundamentals of clinical methodology: 2. Etiology	Etiology	0933-3657	Artif Intell Med
WOS	Aronson, AR	2001	-	-	17-21	Effective mapping of biomedical text to the UMLS m	Program	1067-5027	J AMER MED INFORM ASSOC
Med	Aronson AR	2000	-	-	17-21	The NLM Indexing Initiative	Initiative	1531-605X	Proc AMIA Symp

The false negative errors in the probabilistic model (Table 4.8) are primarily citations from the case group for which the correct match was found, but the probabilistic score did not meet the threshold cut-point due to insufficient matching.

Table 4.8 False Negative Errors In The Probabilistic Model

DS	First Author	Year	Vol	Issue	Pages	Title50	Title End	Journal Abbrev
WOS	YOUNG, DW	1972	11	1	15-&	EVALUATION OF A QUESTIONARY	QUESTIONARY	METHODS INFORM MED
Med	Young DW	1972	11	1	15-9	Evaluation of a questionnaire	questionnaire	Methods Inf Med
WOS	FINK, H	1966	5	1	19-&	VERGLEICH BIOLOGISCHER WIRKUNGEN MITTELS PROGRAMMI	PROBIT ANALYSE	METHODS INFORM MED
Med	Fink H	1966	5	1	19-25	[Comparison of biological effects by programmed pr	analysis	Methods Inf Med
WOS	JUHASZ, VP	1965	4	2	99-&	EIN EINFACHES VERSCHLUSSELUNGSSYSTEM FUR HANDLOCHK	HANDLOCH KARTEN	METHODS INFORM MED
Med	Juhasz VP	1965	4	2	99-101	[A simple coding system for edge-punched cards]	cards	Methods Inf Med
WOS	SACHS, L	1965	4	1	42-&	DER VERGLEICH ZWEIER PROZENTSATZE UND DIE ANALYSE	I	METHODS INFORM MED
Med	SACHS L	1965	45	-	42-5	[THE COMPARISON OF TWO PERCENTAGES AND THE ANALYSI	I.	Methods Inf Med
WOS	THURMAYR, R	1964	3	1	36-&	ERFAHRUNGEN BEI DER AUSWERTUNG DES ALLGEMEINEN KRA	KRANKEN BLATTKOPFES	METHODS INFORM MED
Med	THURMAYR R	1964	43	-	36-8	[EXPERIENCE IN THE EVALUATION OF "SUMMARY CHART SH	SHEETS".	Methods Inf Med
WOS	ARNAUD, P	1972	5	1	75-&	NEW METHOD FOR SPECTROPHOTOMETRIC ANALYSIS OF MIXT	.1	COMPUT BIOMED RES
Med	Arnaud P	1972	5	1	75-9	New method for the spectrophotometric analysis of	I	Comput Biomed Res
WOS	BLEICH, HL	1989	6	3	133-135	CLINICAL COMPUTING	COMPUTING	M D COMPUT
Med	Bleich HL	1989	6	3	132-5	Clinical computing	computing	MD Comput
WOS	PEARSON, WR	1985	2	5	45-&	PROGRAMMING-LANGUAGES .3.	.3	M D COMPUT
Med	Pearson WR	1985	2	5	45-9, 56	Programming languages III	III	MD Comput

Four of the eight records have titles in different languages, and seven of eight have an "&" character in the EndPageDigit position. The exception is Bleich (1989), which was not linked. There is a difference in the variable used for blocking (BeginPage) in this citation, which may indicate a need for "OR" blocking on multiple variables so a search for a match is done after a BeginPage match is not successful.

4.4 The Cartesian Product Problem

The citations in the medical informatics dataset used in model evaluation cover a period of 40 years from a relatively small field in the medical research literature. The variables used in the deterministic models did not generate completely unique keys on this dataset, but the percentage of records involved in non-unique keys was less than 1%. However a simple test of the stability of the models in domains other than medical informatics is to examine the performance of the variable sets in generating unique keys on alternate datasets (Table 4.9).

Table 4.9 Comparison Of Variable Sets In Generating Unique Key

Model	Variables	WOS Medical Informatics 1964-2004 (N=11,752) Uniquely Keyed Records n (%)	MEDLINE Medical Informatics 1962-2004 (N=21,771) Uniquely Keyed Records n (%)	WOS HIV/AIDS and General Medical JOURNALS 2003-2005 (N=20,314) Uniquely Keyed Records n (%)	MEDLINE HIV/AIDS SUBSET 2005 (N=17,005) Uniquely Keyed Records n (%)
DM0	Title50	11479 (97.7)	21277 (97.7)	15164 (74.6)	16620 (97.7)
DM1	LastName(4), Year, Begin Page	11712 (99.7)	21693 (99.6)	19737 (97.2)	16079 (94.6)
DM2	LastName, Year, Begin Page	11716 (99.7)	21703 (99.7)	19764 (97.3)	16147 (95.0)
DM3	LastName, Year, Title50	11651 (99.1)	21548 (99.0)	19766 (97.3)	16901 (99.4)
DM4	ISSN, Year, Volume, Issue, BeginPage	11710 (99.6)	21502 (98.8)	14209 (67.0)	16360 (96.2)
PS1	Year, Volume, Issue, BeginPage, EndPageDigit, LastName, FirstInitial, Title50, TitleEnd	11752 (100.0)	21771 (100.0)	20313 (99.9)*	17003 (99.9)

*There is a duplicate record in the dataset

As shown in Table 4.9, the performance of the deterministic variable sets decline as the datasets remain large, but cover shorter time periods. The percentage of records non-uniquely keyed by the best deterministic models (DM1 and DM2) increases from <1% to 5%. The citations from the 2005 HIV/AIDS subset of Medline have a 5% rate of null data for Author LastName, and a breakdown by the DM1 variable set shows up to 71 duplicates for a match key of LastName = null, Year = 2005, BeginPage = 1. As a result of the Cartesian product problem, record-linkage of the 2005 HIV/AIDS data with the DM1 variable set could generate 24,460 links from the records involved in duplicate keys, of which only 926 were correct (Table 4.10).

Table 4.10 Most Common Duplicate Keys In The 2005 Medline HIV/AIDS Data And The Cartesian Products

LastName	Year	BeginPage	Duplicates	CrossProduct
	2005	1	71	5041
	2005	3	63	3969
	2005	7	55	3025
	2005	6	54	2916
	2005	5	52	2704
	2005	8	50	2500
	2005	2	37	1369
	2005	4	29	841
	2005	9	26	676
	2005	10	7	49
	2005	11	6	36
Jame	2005	6	6	36
	2005	20	6	36
	2005	25	5	25
	2005	35	5	25
Jame	2005	2	5	25
	2005	32	5	25
	2005	24	5	25
Jame	2005	5	5	25
	2005	18	5	25
Jame	2005	3	5	25
	2005	127	5	25
	2005	54	5	25
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Total Records			926	24460

4.5 Summary of Findings

The ROC analyses of the five deterministic and one probabilistic model was stable over three trials and there was a significant difference between models (P = 0.000). The direct comparison between the probabilistic model (AUC = .999) and the best performing deterministic model (AUC = .98) was also significant (P = 0.000) for a difference in AUC. The AUC for both models are high, but analysis of model errors and the inability of the deterministic model variable set to generate unique keys shows that the deterministic model performance will decline in alternate datasets.

CHAPTER 5: THE EFFECTS OF RECORD LINKAGE AND FUSION

The evaluation of the effects of preparing citation data for visualization with a probabilistic record linkage and information fusion methodology (PRL-IF) consists of comparing pre-processing (baseline) WOS datasets and visualizations to post-processing (fusion) WOS datasets and visualizations. The baseline WOS data are first surveyed and compared to an alternate data source (Medline or CINAHL) for patterns of availability of key data elements (abstracts and keywords). A probabilistic record linkage approach is then used to link WOS records to the alternate data source, abstracts are added to the WOS file as needed and available, and MeSH or CINAHL terms are inserted in place of WOS keywords. The baseline and fusion WOS datasets are then compared for differences in data quality based on measures of availability of key words and abstracts. Visualizations of the baseline and fusion datasets are generated using identical parameters without adjustments for aesthetics for comparison purposes. The effects on visualization are described by measures of changes in burst terms, nodes/links, rankings of top terms, and rankings of highly cited documents. This evaluation is conducted on four sets of data in linkages of WOS to Medline and CINAHL in three knowledge domains: medical informatics Pre-1990, medical informatics Post-1990, HIV/AIDS 2003-2005, and nursing informatics 2002-2005.

5.1 Medical Informatics

The medical informatics data covers a time period of forty years from 1964-2004. As graphically presented in Figures 5.1 and 5.2, the WOS data do not begin including abstracts or keywords until 1990. For this reason the analysis is done separately for pre- and post- 1990 data. This step is taken so the effects of PRL-IF can be compared for data with and without abstracts or keywords in the baseline data, otherwise the effects seen would be averaged across two extremes of missing data.

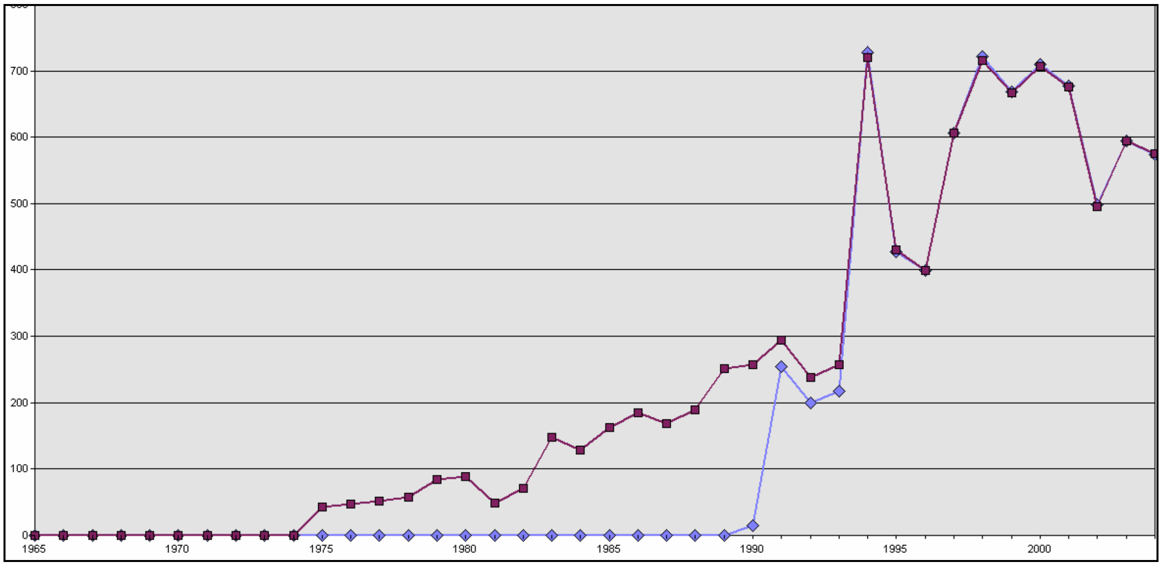


Figure 5.1 Documents With Abstracts By Year Of Publication, WOS (-▲-) Versus Medline (-■-), Showing Abstracts Available In WOS Starting In 1990

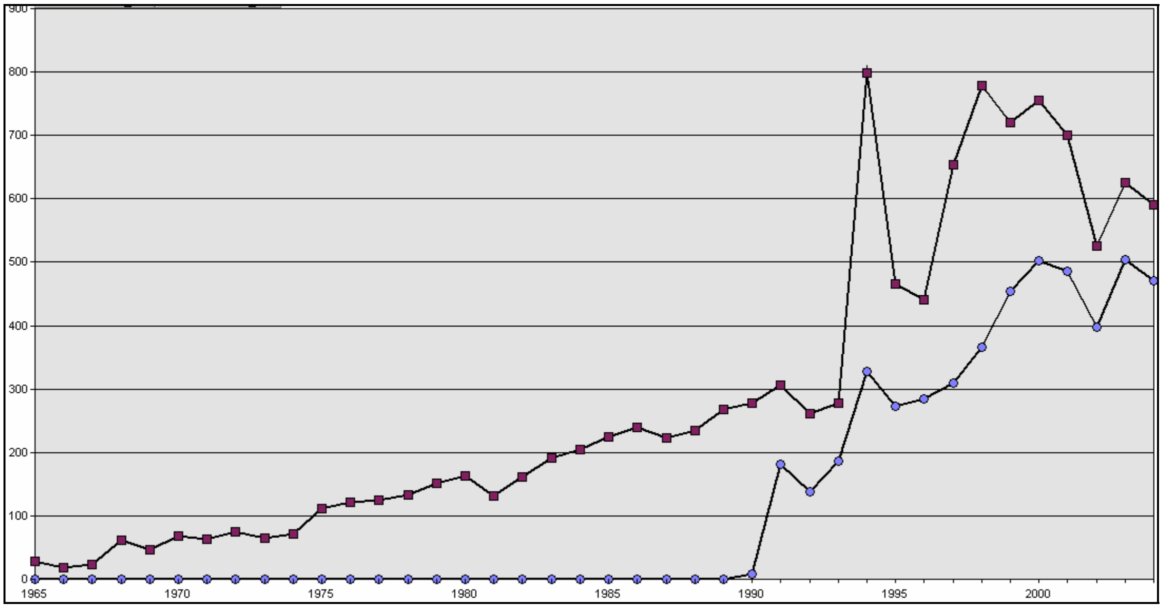


Figure 5.2 Documents With Key Words By Year Of Publication, WOS (●) Versus Medline (-■-), Showing Keywords Are Available In WOS Starting In 1990.

5.1.1 Medical Informatics Pre-1990

In addition to examining the data for patterns of availability of key variables by time, the data is surveyed for patterns of availability of abstracts and keywords by Journal (Table 5.1). If neither dataset has a high percentage of availability of key variables by a journal or journals which are representative of a sub-discipline within a domain, then that sub-discipline may not be well represented by keywords within the visualization and the issue should be documented. In the case of citations from prior to 1990 the potential increased availability of abstracts from Medline is less than 50% overall, but there is 100% availability of keywords from Medline. After fusion the overall record level data quality measures have increased but are only 56% complete because of the limited availability of abstracts (Table 5.2). However in addition to increasing the percentage of records with of keywords from <1% to >99%, the average number of keywords per record is now >10 due to the number of MeSH terms assigned to articles.

Table 5.1 Medical Informatics Pre-1990, Survey For Abstract And Keywords Data

WOS pre-1990 (N=3589)			Medline pre-1990 (N=4848)		
Journal	% Abstracts	% Keywords	Journal	% Abstracts	% Keywords
Comput Biomed Res	0%	0%	Comput Biomed Res	34.3%	100%
Comput Method Program Biomed	3.3%	1.8%	Comput Methods Programs Biomed	95.9%	100%
			Comput Nurs	21.3%	100%
Comput Program Biomed	0%	0%	Comput Programs Biomed	78.5%	100%
Int J Bio-Med Comput	0%	0%	Int J Biomed Comput	72.8%	100%
			Int J Technol Assess Health Care	54.0%	100%
M D Comput	0%	0%	MD Comput	12.7%	100%
Med Decis Making	0%	0%	Med Decis Making	72.0%	100%
Methods Inform Med	0%	0%	Methods Inf Med	10.8%	100%
Total	0.4%	0.2%	Total	47.8%	100%

Table 5.2 Medical Informatics Pre-1990, Pre And Post Fusion Data Quality Measures

Pre-1990	% Complete Pre-Linkage	% Complete With Fusion
Abstract	0.4%	56.8%
KeyWords/MeSH	0.2%	100%
Total Number of Keywords	32	38,381
Cited References	96.4%	96.4
Records with Abstract AND (Keywords/MeSH) AND Cited References	0.2%	56.2%

The effects of PRL-IF on visualization of the fusion data (Figure 5.4) compared to the baseline data (Figure 5.3) are a 35-fold increase in burst terms and doubling or greater increase in nodes and links between nodes (Table 5.3). The citation records included in the analysis are the same in both the baseline and fusion datasets, so there is no change in cited references, and consequently no change in the pattern of highly cited documents.

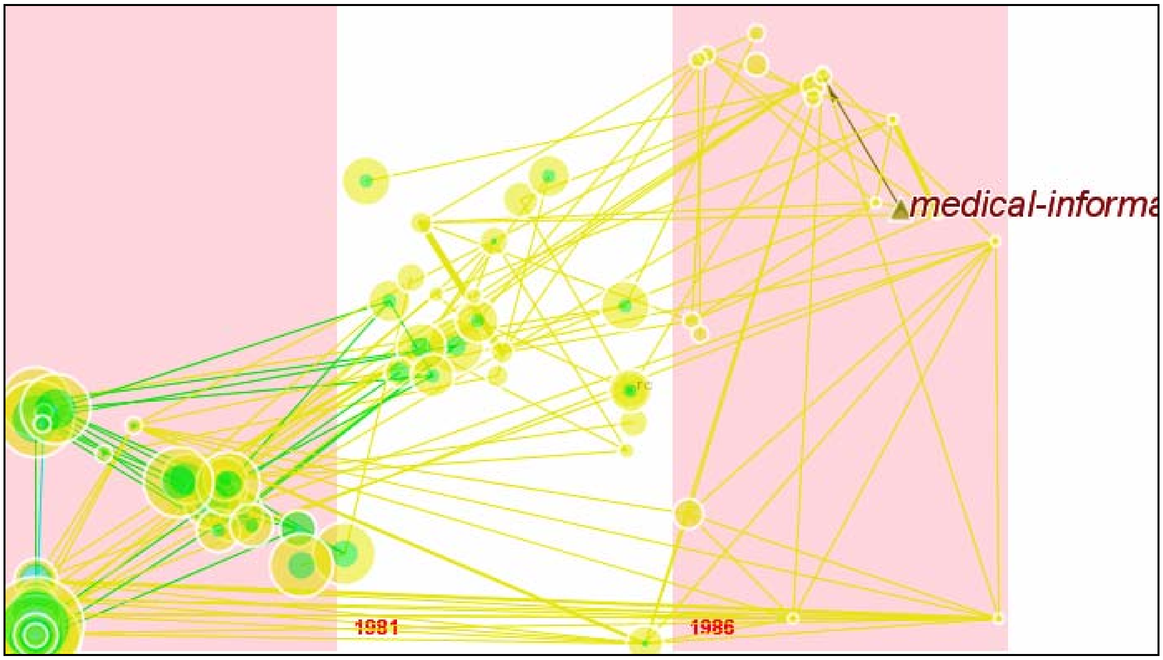


Figure 5.3 Medical Informatics 1976-1990, Pre-Linkage

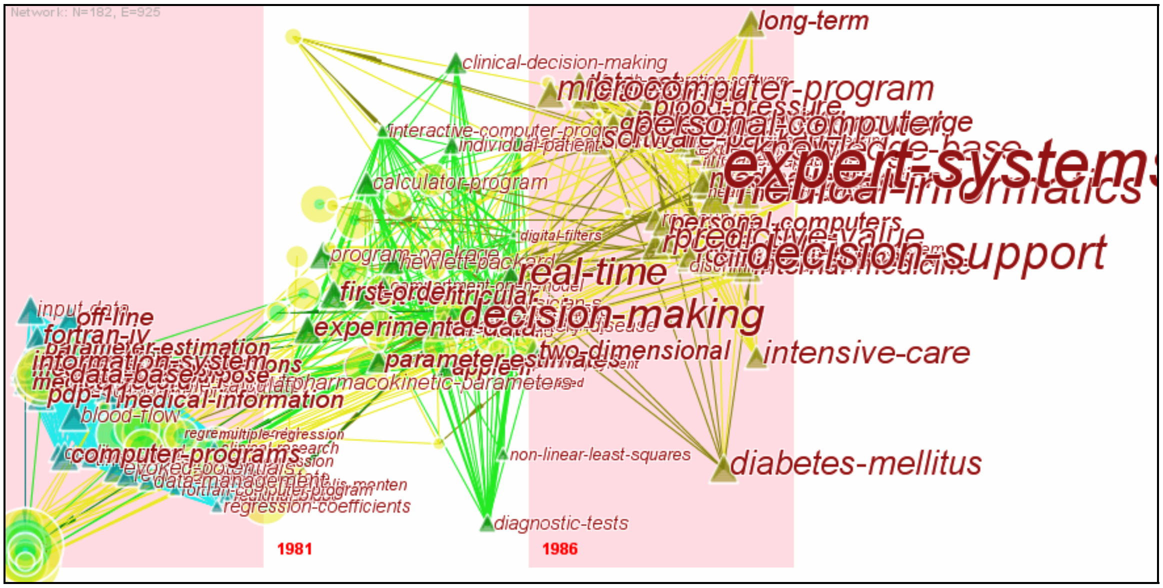


Figure 5.4 Medical Informatics 1976-1990, Post-Linkage. With Fusion There Is A 35-Fold Increase In Burst Terms And Doubling Of Nodes.

Table 5.3 Medical Informatics 1976-1990, Effects On Visualization Metrics

	Pre-Linkage (Figure 5.3)	With Fusion (Figure 5.4)
Analysis Type	Document-Term Co-citation	Document-Term Co-citation
Publication Years	1976-1990	1976-1990
Thresholding (c/cc/ccv)	4/2/20	4/2/20
Burst Terms In Range	19	653
Nodes & Links	80 & 192	191 & 1,598

In addition to increasing the number of keywords available to the visualization, the PRL-IF process has also changed the rankings (Table 5.4) and content of the top twenty burst terms. While the visualization may not be completely representative of medical informatics prior to 1990 due to incomplete availability of abstract data, there is a much more descriptive picture of the research fronts and what was initially thought to be a young domain with too few citations for assessment now appears to be actively "bursting" or changing field. Figure 5.5 shows the fusion visualization with the display of terms limited for clarity to the top 20 terms. Two research fronts not previously seen are expert systems (knowledge base, medical decision making) and personal computers in period between 1985 and 1990.

Table 5.4 Medical Informatics 1976-1990, Effect On Term Rankings

		Pre-Linkage			With Fusion	
Rank		Freq	Keyword		Freq	Keyword
1		20	medical-informatics		109	expert-systems
2	(No further terms found)				60	decision-support
3					58	medical-informatics
4					41	decision-making
5					35	real-time
6					33	personal-computer
7					32	knowledge-base
8					31	microcomputer-program
9					30	predictive-value
10					28	diabetes-mellitus
11					27	intensive-care
12					25	monte-carlo
13					24	decision-theory
14					23	software-package
15					21	blood-pressure
16					21	medical-knowledge
17					20	internal-medicine
18					20	medical-education
19					19	clinical-information
20					19	data-base
					19	experimental-data

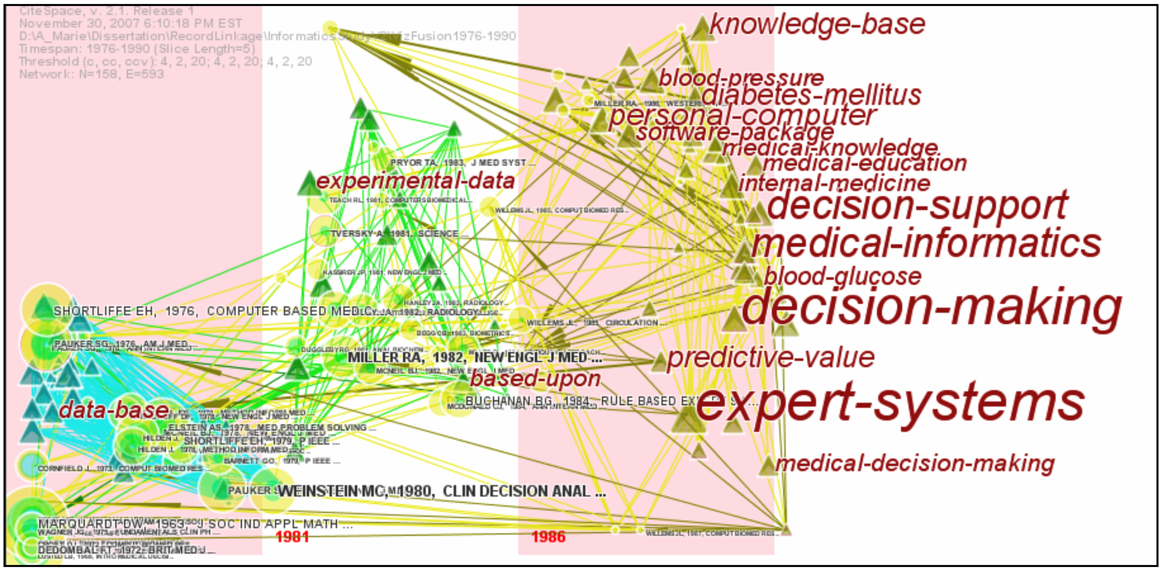

CiteSpace, v. 2.1. Release 1
November 30, 2007 6:10:18 PM EST
D:\A_Marie\Dissertation\RecordLinkage\InformaticsStudyV2\VizFusion1976-1990
Timespan: 1976-1990 (Slice Length=5)
Threshold (c, cc, ccv): 4, 2, 20; 4, 2, 20; 4, 2, 20
Network: N=158, E=593
knowledge-base
blood-pressure
diabetes-mellitus
personal-computer
software-package
medical-knowledge
medical-education
internal-medicine
decision-support
medical-informatics
blood-glucose
decision-making
predictive-value
expert-systems
medical-decision-making
experimental-data
based-upon
data-base
PRYOR TA, 1983, J MED SYST ...
TEACH RL, 1981, COMPUTERS BIOMEDICAL ...
WILLEMS JL, 1985, COMPUT BIOMED RES ...
TVERSKY A, 1981, SCIENCE ...
KASSIRER JP, 1981, NEW ENGL J MED ...
HANLEY JA, 1983, RADIOLOGY ...
HANLEY JA, 1982, RADIOLOGY ...
BEGG CB, 1983, BIOMETRICS ...
WILLEMS JL, 1985, CIRCULATION ...
SHORTLIFFE EH, 1976, COMPUTER BASED MED...
PAUKER SG, 1976, AM J MED ...
DUGGLEBY RG, 1981, ANAL BIOCHEM ...
MILLER RA, 1982, NEW ENGL J MED ...
MCNEIL BJ, 1982, NEW ENGL J MED
MILLER RA, 1986, WESTERN ...
BUCHANAN BG, 1984, RULE BASED EXPERT S ...
ELSTEIN AS, 1978, MED PROBLEM SOLVING ...
MCNEIL BJ, 1978, NEW ENGL J MED
SHORTLIFFE EH, 1979, P IEEE ...
HILDEN J, 1978, METHOD INFORM MED ...
BARNETT GO, 1979, P IEEE ...
CORNFIELD J, 1973, COMPUT BIOMED RES ...
WEINSTEIN MC, 1980, CLIN DECISION ANAL ...
MARQUARDT DW, 1963, J SOC IND APPL MATH ...
WAGNER JG, 1975, FUNDAMENTALS CLIN PH ...
DEDOMBAL FT, 1972, BRIT MED J ...
WILLEMS JL, 1987, COMPUT BIOMED RES ...
1981
1986

Figure 5.5 Medical Informatics 1976-1990, Fusion, With Display Limited To Top Terms

5.1.2 Medical Informatics Post-1990

The data survey for patterns of availability of abstracts (Table 5.5) and keywords (Table 5.6) by Journal shows that while >85% of records have abstracts overall in both datasets, there are several journals where there is the potential to enrich the WOS data with added abstracts such as Comput Biomed Res (99% vs. 64%). MeSH terms are consistently 99-100% available in the Medline data, while the WOS keyword data ranges from 34% to 87% complete. After fusion the addition of abstracts to a few journals and the insertion of MeSH terms in all records results in record level completeness increasing from 58% to 92% (Table5.7). There is a 7-fold increase in key terms, and the average number of keywords per document increases from 5 to 21.

Table 5.5 Medical Informatics Post-1990, Data Survey For Abstracts

WOS, Abstracts Available, post-1990 (N=8163)		Medline, Abstracts Available, post-1990 (N=11067)	
Journal	%	Journal	%
Artif Intell Med	92.4%	Artif Intell Med	94.3%
Comput Biomed Res	63.6%	Comput Biomed Res	99.1%
Cin-Comput Inform Nurs	66.1%	Comput Inform Nurs	80.2%
Comput Method Program Biomed	96.0%	Comput Methods Programs Biomed	97.1%
Comput Nurs	91.6%	Comput Nurs	76.3%
Ieee Trans Inf Technol Biomed	95.7%	IEEE Trans Inf Technol Biomed	96.4%
Int J Bio-Med Comput	90.9%	Int J Biomed Comput	96.4%
Int J Med Inform	91.0%	Int J Med Inform	93.4%
Int J Technol Assess Health C	80.7%	Int J Technol Assess Health Care	80.5%
J Amer Med Inform Assoc	95.6%	J Am Med Inform Assoc	81.1%
J Biomed Inform	92.8%	J Biomed Inform	96.2%
M D Comput	38.4%	MD Comput	24.1%
Med Decis Making	93.8%	Med Decis Making	78.6%
Med Inform Internet Med	98.5%	Med Inform Internet Med	100%
Methods Inform Med	99.3%	Methods Inf Med	91.6%
Total	89.6%	AMIA Annu Symp Proc Proc AMIA Annu Fall Symp Proc AMIA Symp Proc Annu Symp Comput Appl Med Care	100% 98.8% 99.5% 91.6%
		Total	87.7%

Table 5.6 Medical Informatics Post-1990, Data Survey For Keywords

WOS, Keywords Available, post-1990 (N=8163)		Medline, Keywords Available, post-1990 (N=11067)	
Journal	%	Journal	%
Artif Intell Med	77.3%	Artif Intell Med	99.8%
Comput Biomed Res	75.7%	Comput Biomed Res	100%
Cin-Comput Inform Nurs	51.2%	Comput Inform Nurs	100%
Comput Method Program Biomed	61.2%	Comput Methods Programs Biomed	99.8%
Comput Nurs	58.8%	Comput Nurs	99.5%
Ieee Trans Inf Technol Biomed	75.7%	IEEE Trans Inf Technol Biomed	99.7%
Int J Bio-Med Comput	41.9%	Int J Biomed Comput	100%
Int J Med Inform	52.6%	Int J Med Inform	100%
Int J Technol Assess Health C	65.6%	Int J Technol Assess Health Care	100%
J Amer Med Inform Assoc	45.0%	J Am Med Inform Assoc	99.6%
J Biomed Inform	83.6%	J Biomed Inform	100%
M D Comput	34.0%	MD Comput	100%
Med Decis Making	86.9%	Med Decis Making	100%
Med Inform Internet Med	71.3%	Med Inform Internet Med	100%
Methods Inform Med	65.7%	Methods Inf Med	100%
(AMIA Proceedings are indexed under JAMIA in WOS)		AMIA Annu Symp Proc Proc AMIA Annu Fall Symp Proc AMIA Symp Proc Annu Symp Comput Appl Med Care	100% 100% 98.6% 100%
Total	60.1%	Total	99.8%

Table 5.7 Medical Informatics Post-1990, Pre And Post Fusion Data Quality Measures

Post-1990	% Complete Pre-Linkage	% Complete With Fusion
Abstract	91.9%	93.8%
KeyWords/MeSH	61.6%	99.9%
Total Number of Keywords	26,752	179,021
Cited References	95.2%	95.2%
Records with Abstract AND (Keywords/MeSH) AND Cited References	57.8%	92.1%

The effects of PRL-IF on visualization of the fusion data (Figure 5.7) compared to the baseline data (Figure 5.6) are a doubling of burst terms and increase in nodes and links between nodes, resulting in increased information about the knowledge domain being available to a user (Table 5.8). The citation records included in the analysis are the same in both the baseline and fusion datasets, so again there is no change in cited references, and consequently no change in the pattern of highly cited documents.

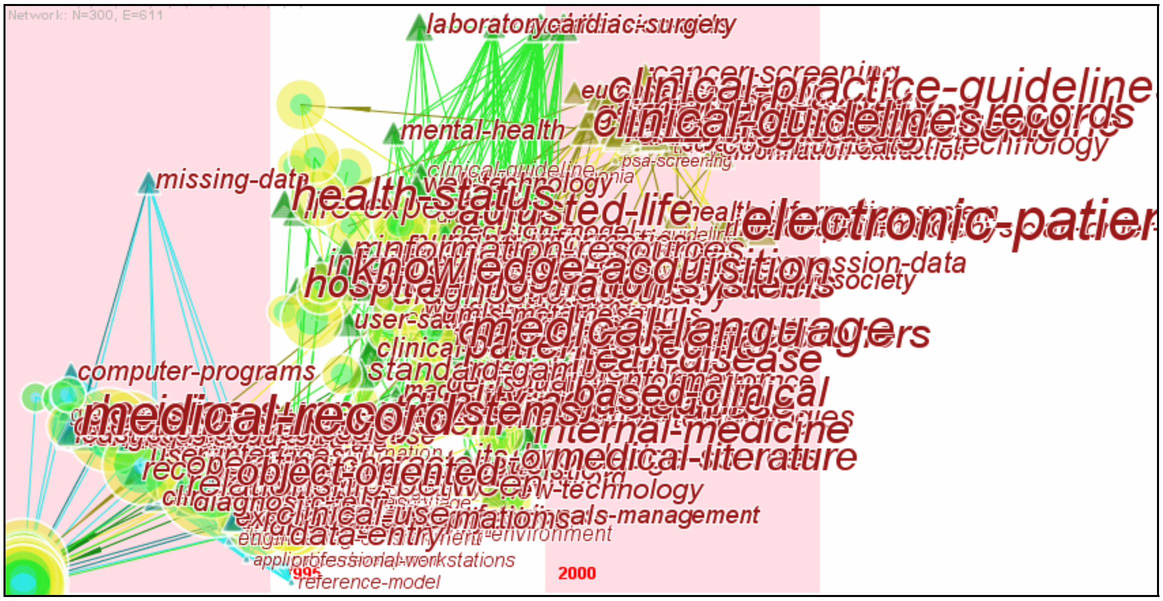


Figure 5.6 Medical Informatics 1990-2004, Pre-Linkage.

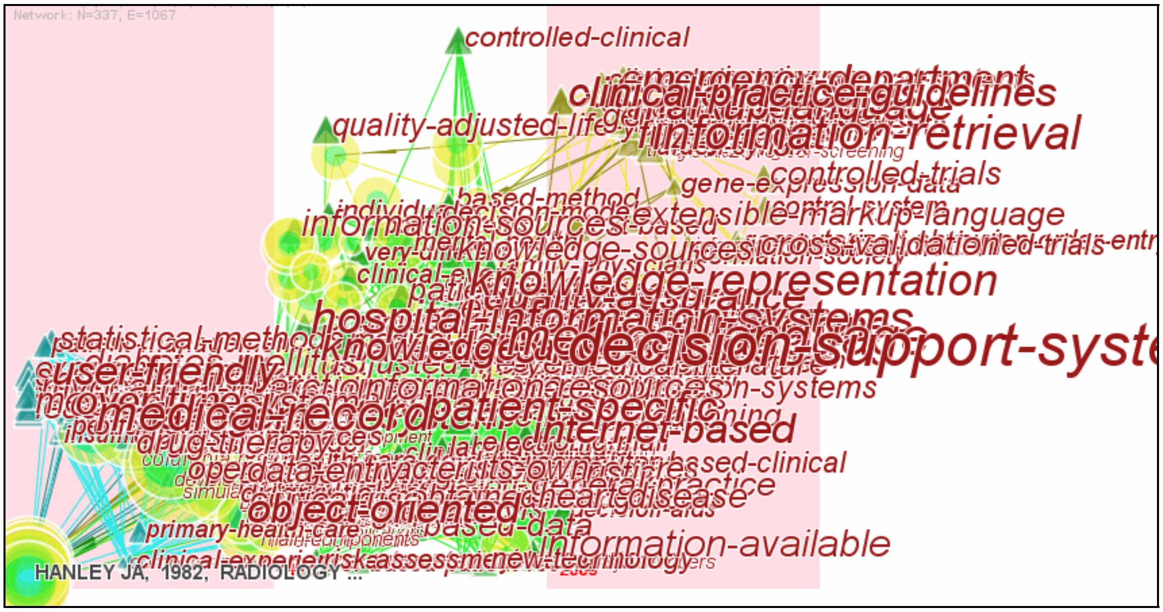


Figure 5.7 Medical Informatics 1990-2004, Post-Linkage. With Fusion There Is A Doubling In Burst Terms.

Table 5.8 Medical Informatics 1990-2004, Effects On Visualization Metrics

	Pre-Linkage (Figure 5.6)	With Fusion (Figure 5.7)
Analysis Type	Document-Term Co-citation	Document-Term Co-citation
Publication Years	1990-2004	1990-2004
Thresholding (c/cc/ccv)	7/3/30	7/3/30
Burst Terms In Range	1,759	3,436
Nodes & Links	437 & 8,369	521 & 13,950

In addition to increasing the number of keywords available to the visualization, the PRL-IF process has also changed the rankings (Table 5.9) and content of the top twenty key terms, which are now based on more complete, and less biased data .

Table 5.9 Medical Informatics 1990-2004, Effect On Term Rankings

		Pre-Linkage			With Fusion	
Rank		Freq	Keyword		Freq	Keyword
1		77	electronic-patient-record		81	decision-support-system
2		71	medical-record		63	medical-language
3		67	medical-language		60	medical-record
4		59	clinical-guidelines		52	information-retrieval
5		53	clinical-practice-guidelines		50	hospital-information-systems
6		52	knowledge-acquisition		47	knowledge-representation
7		50	patient-safety		42	patient-specific
8		47	health-status		40	clinical-practice-guidelines
9		46	adjusted-life		39	knowledge-acquisition
10		44	electronic-patient-records		38	quality-assurance
11		43	evidence-based-medicine		38	relational-database
12		42	fuzzy-logic		37	internet-based
13		41	decision-support-systems		37	management-system
14		41	patient-specific		37	object-oriented
15		41	quality-assurance		37	outcome-measures
16		40	adverse-drug-events		36	internal-medicine
17		40	based-clinical		36	markup-language
18		40	general-practitioners		36	user-friendly
19		40	quality-adjusted-life		35	over-time
20		39	hospital-information-systems		34	emergency-department
					34	fuzzy-logic
					34	general-practitioners

A comparison of pre-linkage and post-fusion visualizations with the display limited for clarity to top terms shows an overlap in terms, but with several key differences (Figure 5.8 and Figure 5.9). With fusion there is an absence of the previously high-ranking term "electronic patient record", there has been a shift in both the timing and ranking of "decision support system" and terms related to the internet now appear in 1995-20000.

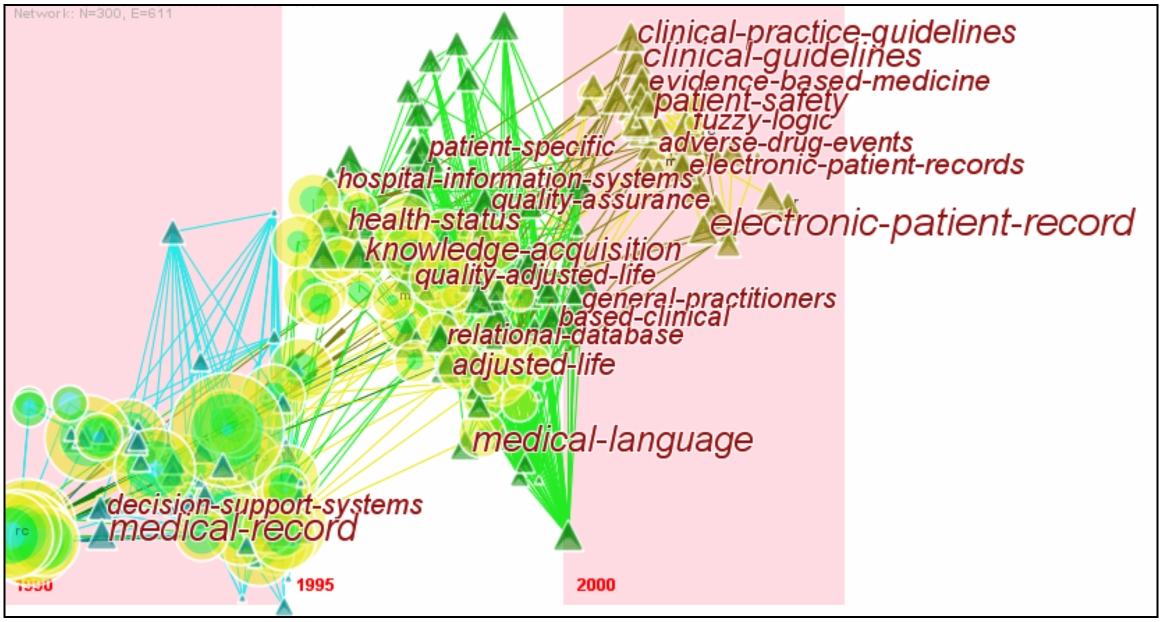


Figure 5.8 Medical Informatics 1990-2004, Pre-Linkage With Display Limited To Top Terms

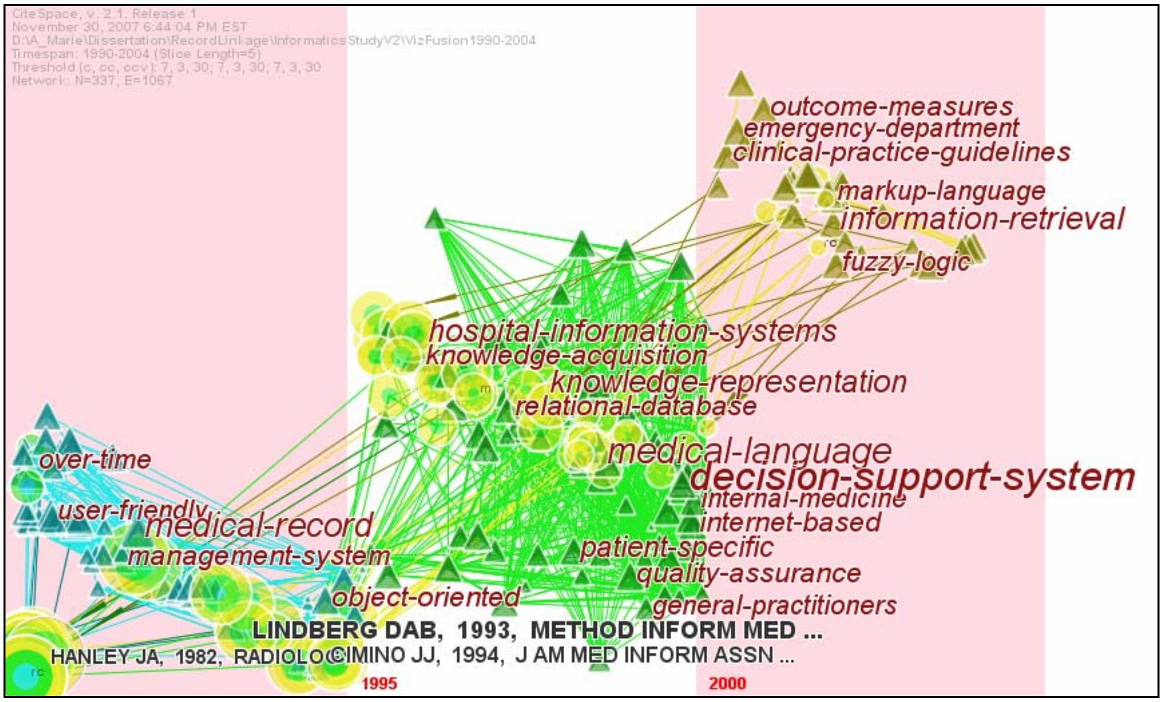


Figure 5.9 Medical Informatics 1990-2004, Fusion With Display Limited To Top Terms

In this study of medical informatics post-1990, the primary differences between the baseline and fusion datasets are the addition of abstracts to 30% of records and the change from 60% of records having "Keywords Plus" keywords and assigned descriptors to 100% of records having the terms from the MeSH hierarchical controlled vocabulary. The fusion of MeSH terms not only increases the percentage of records with key terms, it also increases the number of keywords assigned each record, eliminates duplication of terms between the WOS Keywords Plus and Author provided descriptors, and changes terminology such as "cost effectiveness-analysis" to the familiar MeSH term "cost-benefit analysis" (Table 5.10).

Table 5.10 Comparison Of The WOS Keywords And Mesh Terms Assigned To A Record

<table>
<tr><td colspan="3"><b>Health economic evaluations: The special case of end-stage renal disease treatment</b>
<b>This article synthesizes the evidence on the cost-effectiveness of renal replacement therapy and discusses the findings in light of the frequent practice of using the cost-effectiveness of hemodialysis as a benchmark of societal willingness to pay. The authors conducted a meta-analytic review of the medical and economic literature for economic evaluations of hemodialysis, peritoneal dialysis, and kidney transplantation. Cost effectiveness ratios were translated into 2000 U.S. dollars per life-year (LY) saved. Thirteen studies published between 1968 and 1998 provided such information. The cost-effectiveness of center hemodialysis remained within a narrow range of $55,000 to $80,000/LY in most studies despite considerable variation in methodology and imputed costs. The cost-effectiveness of home hemodialysis was found to be between $33,000 and $50,000/LY. Kidney transplantation, however, has become more cost-effective over time, approaching $10,000/LY Estimates of the cost per life-year gained from hemodialysis have been remarkably stable over the past 3 decades, after adjusting for price levels. Uses of the cost-effectiveness ratio of $55,000/LY for center hemodialysis as a lower boundary of society's willingness to pay for an additional life-year can be supported under certain assumptions.</b></td></tr>
<tr><td><b>WOS keywords</b></td><td><b>WOS descriptors</b></td><td><b>Medline MeSH</b></td></tr>
<tr><td>COST-EFFECTIVENESS ANALYSIS;
AMBULATORY PERITONEAL-DIALYSIS;</td><td>cost-effectiveness analysis;
dialysis;
kidney transplantation;</td><td>Attitude to Health
Cost-Benefit Analysis
Direct Service Costs/statistics & numerical data
Evidence-Based Medicine
Health Care Costs/*statistics & numerical data
Hemodialysis Units, Hospital/economics
Hemodialysis, Home/economics
Humans
Kidney Failure, Chronic/*economics/mortality/*therapy
Kidney Transplantation/*economics
Peritoneal Dialysis, Continuous Ambulatory/economics
Peritoneal Dialysis/*economics
Quality-Adjusted Life Years
Renal Dialysis/*economics
Social Values
Technology Assessment, Biomedical
Time Factors
Treatment Outcome
Value of Life/economics
Attitude to Health
Cost-Benefit Analysis
Direct Service Costs/statistics & numerical data
Evidence-Based Medicine
Health Care Costs/*statistics & numerical data
Hemodialysis Units, Hospital/economics
Hemodialysis, Home/economics
Humans
Kidney Failure, Chronic/*economics/mortality/*therapy
Kidney Transplantation/*economics
Peritoneal Dialysis, Continuous Ambulatory/economics
Peritoneal Dialysis/*economics
Quality-Adjusted Life Years
Renal Dialysis/*economics
Social Values
Technology Assessment, Biomedical
Time Factors
Treatment Outcome
Value of Life/economics</td></tr>
</table>

5.2 HIV/AIDS

The HIV/AIDS study demonstrates the use of record linkage to enrich data, as well as the use of record linkage to define or validate a sample based on an external “gold standard”. This is analogous to the use of record linkage to validate census data with the use of post-enumeration surveys (Jaro, 1989). In this study citation data from the AIDS subset of Medline is used to select the WOS dataset, as well as used to enhance abstract and keyword data in a WOS dataset. Due to the size of the HIV/AIDS literature, the study is limited to three recent years, 2003 to 2005. The baseline dataset is data selected from WOS based on nine journals and keywords for HIV/AIDS. The alternate dataset is all records from the AIDS subset of Medline for the same nine journals. The fusion dataset is selected by linking the AIDS Medline sample to a third sample consisting of all citations for the nine journals from WOS for study period 2003-2005 (N=20,314).

The data survey shows an overall similarity of total abstract availability, but several journals appear to have a higher availability of abstracts in the WOS baseline data (Table 5.11).

Table 5.11 HIV/AIDS 2003-2005, Data Survey For Abstracts

WOS, Abstracts Available (N=4149)			Medline, Abstracts Available (N=4692)			WOS, with linkage (pre-fusion, N=4252)		
Journal	%		Journal	%	#	Journal	%	#
Aids	76%	1337	AIDS	67%	1485	Aids	74%	1408
Aids Patient Care Stds	93%	215	AIDS Patient Care STDS	42%	501	Aids Patient Care Stds	95%	212
Ann Intern Med	66%	41	Ann Intern Med	39%	64	Ann Intern Med	62%	42
Arch Intern Med	82%	39	Arch Intern Med	77%	35	Arch Intern Med	80%	35
Clin Infect Dis	79%	721	Clin Infect Dis	80%	701	Clin Infect Dis	81%	688
Jaids	80%	931	J Acquir Immune Defic Syndr	79%	978	Jaids	79%	959
J Infec Dis	87%	589	J Infect Dis	89%	629	J Infec Dis	89%	629
J Jama-J Am Med Assn	41%	124	JAMA	36%	103	Jama-J Am Med Assn	45%	97
N Engl J Med	24%	152	N Engl J Med	21%	196	N Engl J Med	23%	182
Total	77%	4149	Total	69%	4692	Total	77%	4252

The explanation for this apparent difference is likely to be articles that are indexed in Medline but not in WOS. Of the 440 citations from Medline that do not link to WOS, 60% are classified as "news" publication type in Medline, and do not have abstracts. The most notable aspect of keyword availability is the lower rate of keywords in the WOS data for two very high impact journals (NEJM and JAMA) (Table 5.12).

Table 5.12 HIV/AIDS 2003-2005, Data Survey For Keywords

WOS, Keywords Available (N=4149)		Medline, Keywords Available (N=4692)		WOS, with linkage (pre-fusion)		
Journal	%	Journal	%	Journal	%	
Aids	92%	AIDS	99%	Aids	92%	1408
Aids Patient Care Stds	93%	AIDS Patient Care STDS	100%	Aids Patient Care Stds	96%	212
Ann Intern Med	90%	Ann Intern Med	100%	Ann Intern Med	86%	42
Arch Intern Med	87%	Arch Intern Med	100%	Arch Intern Med	83%	35
Clin Infect Dis	91%	Clin Infect Dis	99.9%	Clin Infect Dis	92%	688
Jaids	94%	J Acquir Immune Defic Syndr	99.5%	Jaids	94%	959
J Infec Dis	96%	J Infect Dis	100%	J Infec Dis	98%	629
Jama-J Am Med Assn	59%	JAMA	100%	Jama-J Am Med Assn	60%	97
N Engl J Med	61%	N Engl J Med	100%	N Engl J Med	61%	182
Total	91%	Total	99.5%	Grand Total	91%	4252

The use of record linkage to select the WOS sample based on the AIDS subset of Medline increases the total sample size only by 2.5%, but the actual change to the sample consists of both addition and deletion of records resulting in an 18% change in citations. In terms of proportion of records changed the Journals most affected are again high-impact journals NEJM and JAMA, with one-third of citations replaced (Table 5.13). The effects on data quality measures are primarily seen in the tripling of keywords after MeSH terms are inserted into the WOS data (Table 5.14). Despite minor differences in most data quality measures, the PRL-IF process results in an almost 7-fold increase in burst terms (Figures 5.10 & 5.11), and Table 5.15.

Table 5.13 HIV/AIDS 2003-2005, Journal Distribution Pre And Post-Linkage

Journal	Pre-Linkage	Removed	Added	Post-Linkage
AIDS	1337	-33	+104	1408
AIDS PATIENT CARE STDS	215	-9	+6	212
ANN INTERN MED	41	-10	+11	42
ARCH INTERN MED	39	-8	+4	35
CLIN INFECT DIS	721	-103	+70	688
J INFEC DIS	589	-62	+102	629
JAIDS	931	-25	+53	959
JAMA-J AM MED ASSN	124	-46	+19	97
N ENGL J MED	152	-44	+74	182
	4149	-340	443	4252

Table 5.14 HIV/AIDS 2003-2005, Pre And Post Fusion Data Quality Measures

	% Complete Pre-Linkage N=4149	% Complete With Linkage N=4252	% Complete With Linkage and Fusion N=4252
Abstract	77.0%	76.9%	76.9%
KeyWords/MeSH	90.5%	91.2%	99.8%
Total Number of Keywords	40,370	40,995	114,170
Cited References	98.1%	98.4%	98.4%
Records with Abstract AND (Keywords/MeSH) AND Cited References	74.9%	74.8%	76.6%

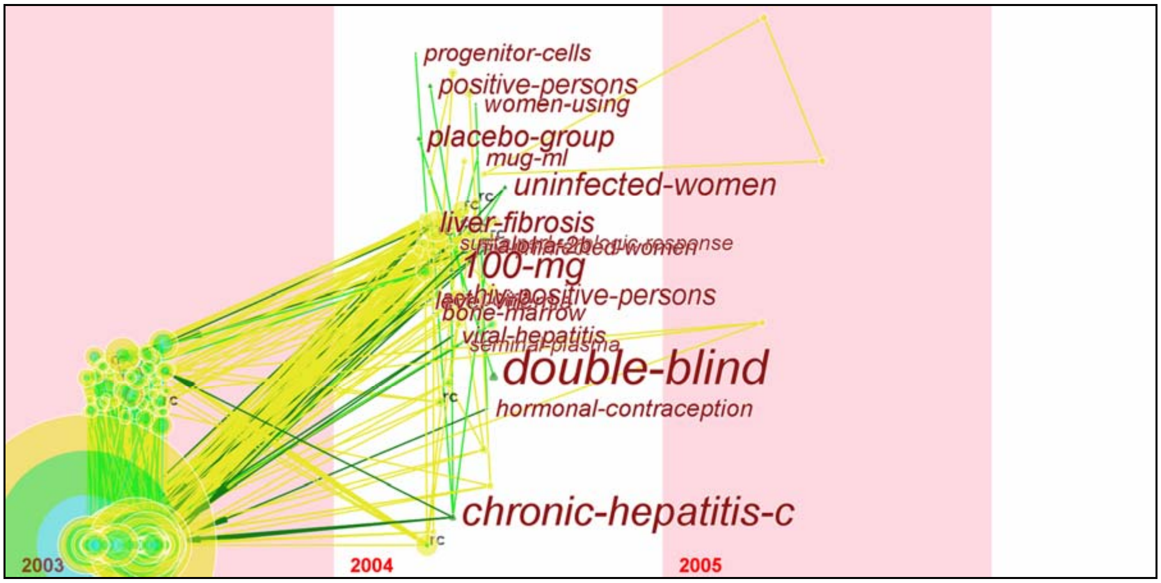


Figure 5.10 HIV/AIDS 2003-2005, Pre-Linkage

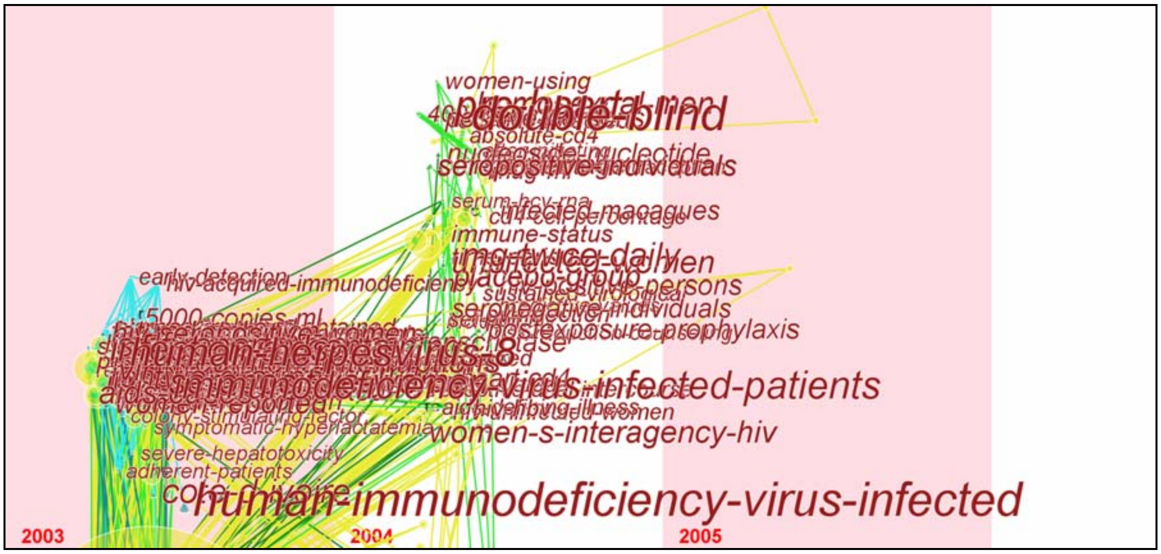


Figure 5.11 HIV/AIDS 2003-2005, With Linkage And Fusion

Table 5.15 HIV/AIDS 2003-2005, Effects On Visualization Metrics

	Pre-Linkage (Figure 5.10)	With Fusion (Figure 5.11)
Analysis Type	Document-Term Co-citation	Document-Term Co-citation
Publication Years	2003-2005	2003-2005
Thresholding (c/cc/ccv)	7/5/33	7/5/33
Burst Terms In Range	371	2,033
Nodes & Links	712 & 2,462	819 & 2,693

In addition to increasing the number of keywords available to the visualization, the PRL-IF process has also changed the rankings and content of the top twenty key terms (Table 5.16).

Table 5.16 HIV/AIDS 2003-2005, Effect On Term Rankings

		Pre-Linkage			With Fusion	
Rank		Freq	Keyword		Freq	Keyword
1		37	double-blind		34	double-blind
2		25	chronic-hepatitis-c		34	human-immunodeficiency-virus-infected
3		22	100-mg		25	human-herpesvirus-8
4		17	uninfected-women		25	pre-haart
5		13	liver-fibrosis		24	immunodeficiency-virus-infected-patients
6		13	placebo-group		22	infection-aids
7		12	hiv-positive-persons		21	seropositive-women
8		12	positive-persons		19	cote-d-ivoire
9		10	alpha-2b		19	mg-twice-daily
10		10	bone-marrow		19	self-report
11		10	progenitor-cells		19	two-groups
12		10	viral-hepatitis		18	regimen-containing
13		9	hormonal-contraception		17	aids-cases
14		9	level-viremia		17	treatment-interruptions
15		9	mug-ml		17	uninfected-women
16		9	women-using		16	containing-regimen
17		8	hiv-uninfected-women		16	one-patient
18		8	set-point		16	seronegative-women
19		7	sustained-virologic-response		16	virus-type-i
20		7	seminal-plasma		15	acquired-immune
					15	diabetes-mellitus
					15	homosexual-men
					15	older-adults
					15	women-s-interagency-hiv

A comparison of pre-linkage and post-fusion visualizations with the display limited for clarity to top terms shows a difference in research clusters (Figure 5.12 and Figure 5.13). With fusion the terms related to hepatitis-c and related issues such as liver fibrosis have been removed, and terms are added related to the longer survival of AIDS patients and the related complications of anti-retroviral therapy such as cardiac disease, diabetes, and lipodystrophy.

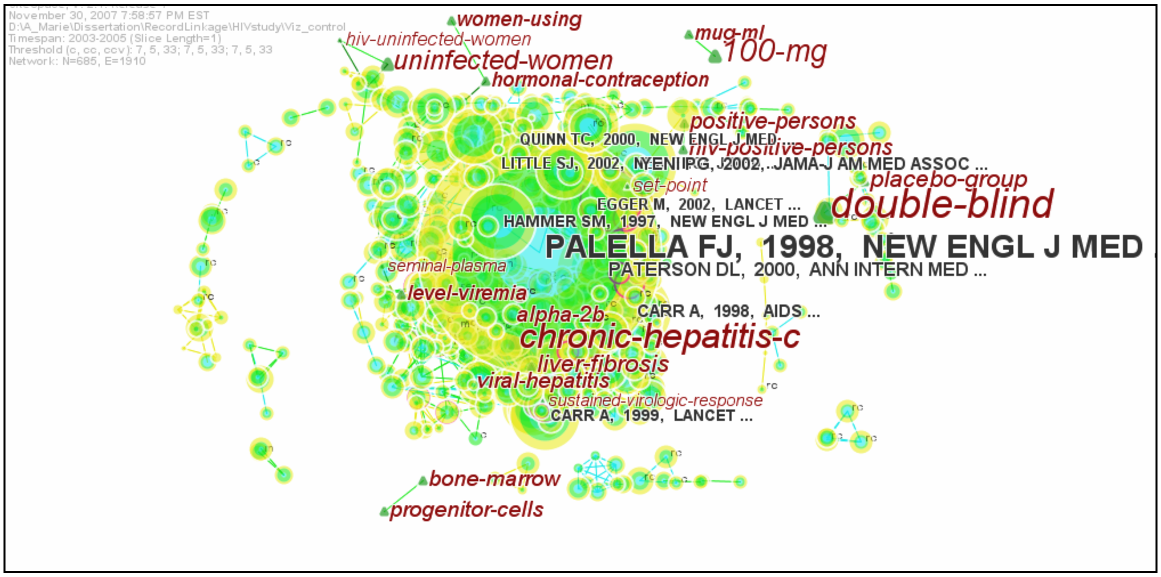

November 30, 2007 7:58:57 PM EST
D:\A_Marie\Dissertation\RecordLinkage\HIVstudy\Viz_control
Timespan: 2003-2005 (Slice Length=1)
Threshold (c, cc, ccv): 7, 5, 33; 7, 5, 33; 7, 5, 33
Network: N=685, E=1910
women-using
hiv-uninfected-women
uninfected-women
hormonal-contraception
mug-ml
100-mg
positive-persons
hiv-positive-persons
QUINN TC, 2000, NEW ENGL J MED ...
LITTLE SJ, 2002, NEW ENGL J MED ...
YENI PG, 2002, JAMA-J AM MED ASSOC ...
placebo-group
set-point
EGGER M, 2002, LANCET ...
double-blind
HAMMER SM, 1997, NEW ENGL J MED ...
PALELLA FJ, 1998, NEW ENGL J MED
PATERSON DL, 2000, ANN INTERN MED ...
seminal-plasma
level-viremia
alpha-2b
CARR A, 1998, AIDS ...
chronic-hepatitis-c
liver-fibrosis
viral-hepatitis
sustained-virologic-response
CARR A, 1999, LANCET ...
bone-marrow
progenitor-cells

Figure 5.12 HIV/AIDS 2003-2005, Pre-Linkage With Display Limited To Top Terms

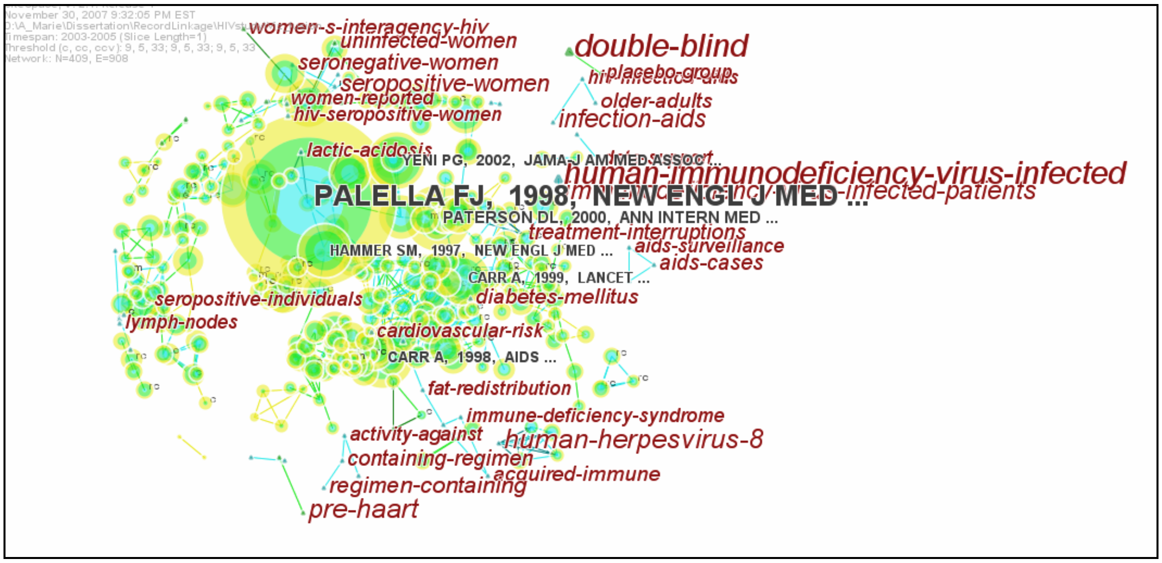


Figure 5.13 HIV/AIDS 2003-2005, Fusion With Display Limited To Top Terms

The use of record linkage to define the sample of citations in this study also changes the data related to cited references. As shown in Table 5.17, this has had little impact on the top 20 cited documents in this case. There have been minor changes in rankings and frequencies, but membership of the set has not changed.

Table 5.17. HIV/AIDS 2003-2005, Top 20 Citations

		Pre-Linkage				With Fusion		
Rank		Freq	Author	Year		Freq	Author	Year
1		394	PALELLA FJ	1998		392	PALELLA FJ	1998
2		140	PATERSON DL	2000		140	PATERSON DL	2000
3		126	YENI PG	2002		125	YENI PG	2002
4		121	CARR A	1998		122	CARR A	1998
5		108	HAMMER SM	1997		110	HAMMER SM	1997
6		103	CARR A	1999		105	CARR A	1999
7		96	QUINN TC	2000		95	QUINN TC	2000
8		94	EGGER M	2002		93	EGGER M	2002
9		85	LITTLE SJ	2002		85	LITTLE SJ	2002
10		81	HOGG RS	2001		82	STASZEWSKI S	1999
11		81	STASZEWSKI S	1999		81	HOGG RS	2001
12		79	MOCROFT A	1998		78	MOCROFT A	1998
13		77	MELLORS JW	1997		77	LEDERGERBER B	1999
14		76	GUAY LA	1999		76	GUAY LA	1999
15		76	LEDERGERBER B	1999		75	BICA I	2001
16		75	BICA I	2001		75	MELLORS JW	1997
17		70	GREUB G	2000		69	GREUB G	2000
18		69	DEEKS SG	2001		67	HIRSCH MS	2000
19		68	HIRSCH MS	2000		66	AUTRAN B	1997
20		66	AUTRAN B	1997		66	DEEKS SG	2001
		66	SULKOWSKI MS	2000				

5.3 Nursing Informatics

The nursing informatics study demonstrates again the use of record linkage to enrich data, as well as the use of record linkage to define a representative sample. The sample is defined by cross-referencing the medical informatics journal set from WOS against CINAHL and selecting the articles indexed by both databases from the four journals common to both databases. By taking this approach it is possible to select the subset of nursing specific informatics articles from the broader medical informatics journals, and to enrich the dataset with MeSH plus nursing specific keywords.

The data survey shows a difference between WOS and CINAHL in the years of indexing and the number of articles indexed per year where years overlap for the study journal set. The analysis is then limited to the four years (2002 – 2005) where all four journals are represented and a greater than 50% match can be obtained (Table 5.18 – 5.20). The most notable aspect of keyword availability is the lower rate of keywords in the WOS data for the nursing specific journal CIN (Table 5.22).

Table 5.18 The Baseline WOS Dataset.

WOS Journal	‘92	‘93	‘94	‘95	‘96	‘97	‘98	‘99	‘00	‘01	‘02	‘03	‘04	‘05	‘06	Total
COMPUT NURS CIN-COMPUT INFORM NURS	1	11	11	2	15	15	14	16	12	13	24	32	31	38	26	261
J AMER MED INFORM ASSOC			29	30	31	46	44	38	46	45	79	59	55	75	73	650
MED DECIS MAKING	31	38	44	39	50	54	60	53	44	48	53	48	51	53	49	715
MED INFORM INTERNET MED								24	20	23	24	22	21	28	19	181
Total	32	49	84	71	96	115	118	131	122	129	180	161	158	194	167	1807

Table 5.19 The Reference CINAHL Dataset.

CIN Journal	‘92	‘93	‘94	‘95	‘96	‘97	‘98	‘99	‘00	‘01	‘02	‘03	‘04	‘05	‘06	Total
COMPUT NURS CIN COMPUT INFORM NURS	35	49	66	80	67	81	83	64	48	63	47	78	85	93	90	1029
J AM MED INFORM ASSOC			19	24	18	32	31	29	47	43	100	65	64	74	0	546
MED DECIS MAKING										22	20	18	22	14	19	115
MED INFORM INTERNET MED											18	17	18	29	0	82
Total	35	49	85	104	85	113	114	93	95	128	185	178	189	210	109	1772

Table 5.20 The WOS Dataset Post-Linkage

Journal	‘92	‘93	‘94	‘95	‘96	‘97	‘98	‘99	‘00	‘01	‘02	‘03	‘04	‘05	‘06	Total
COMPUT NURS CIN-COMPUT INFORM NURS	1	10	11	2	15	15	14	16	12	13	23	32	30	38	24	256
J AMER MED INFORM ASSOC			10	17	11	25	27	23	37	34	74	56	49	68	0	431
MED DECIS MAKING										20	18	18	21	14	17	108
MED INFORM INTERNET MED											18	17	18	27	0	80
Total	1	10	21	19	26	40	41	39	49	67	133	123	118	147	41	875

Table 5.21 Nursing Informatics 2002-2005, Survey For Abstract Data

WOS, Abstracts Available (N=693)			CINAHL, Abstracts Available (N=757)	
Journal	%		Journal	%
CIN-COMPUT INFORM NURS	90.3%	100%	CIN COMPUT INFORM NURS	40.9%
J AMER MED INFORM ASSOC	95.2%	100%	J AM MED INFORM ASSOC	86.8%
MED DECIS MAKING	92.7%	100%	MED DECIS MAKING	98.7%
MED INFORM INTERNET MED	100.0%	100%	MED INFORM INTERNET MED	97.6%
Total	94.2%	100%	Total	71.1%

Table 5.22 Nursing Informatics 2002-2005, Survey For Keywords Data

WOS, Keywords Available (N=693)		CINAHL, Keywords Available (N=757)	
JournalAbbrev	%	Journal	
CIN-COMPUT INFORM NURS	66.9%	CIN COMPUT INFORM NURS	100.0%
J AMER MED INFORM ASSOC	82.5%	J AM MED INFORM ASSOC	100.0%
MED DECIS MAKING	93.2%	MED DECIS MAKING	100.0%
MED INFORM INTERNET MED	73.7%	MED INFORM INTERNET MED	100.0%
Total	81.7%	Total	100.0%

The use of record linkage to select the WOS sample based on the CINAHL subset of nursing informatics decreases the total sample size by 25%, and alters the distribution of documents by journal. The effect on data quality measures are primarily seen in the percentage of records with keywords and the doubling of keywords after CINAHL terms are inserted into the WOS data (Table 5.23). Despite a relatively small sample of 521 citations the PRL-IF process results in an almost 5-fold increase in burst terms (Figures 5.14 & 5.15, and Table 5.24).

Table 5.23 Nursing Informatics 2002-2005, Pre And Post Fusion Data Quality Measures

	% CompletePre-Linkage N=693	% Complete With Fusion N=521
Abstract	94.2%	96.3%
KeyWords/MeSH	81.7%	100%
Total Number of Keywords	4039	9099
Cited References	98.4%	98.8%
Records with Abstract AND (Keywords/MeSH) AND Cited References	79.7%	95.4%

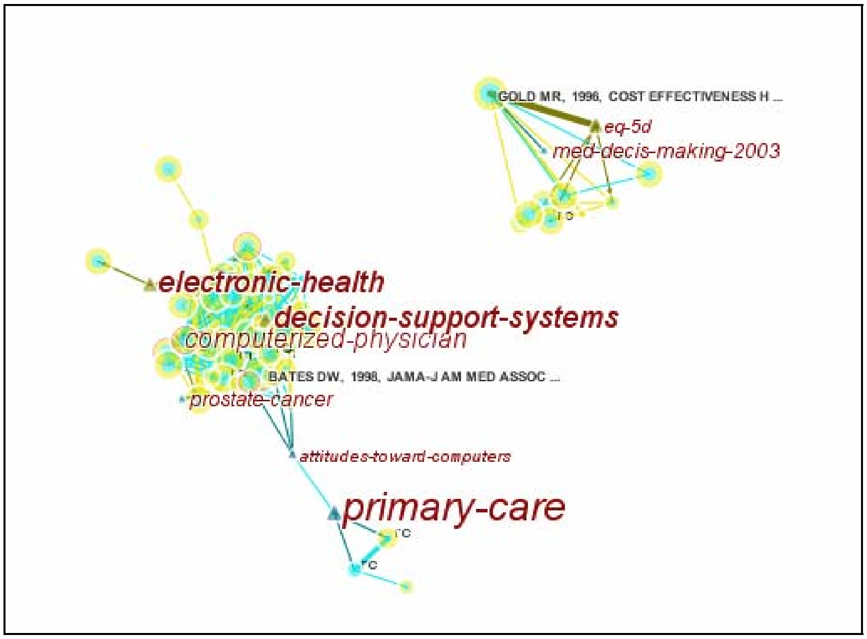


Figure 5.14 Nursing Informatics 2002-2005, Pre-Linkage

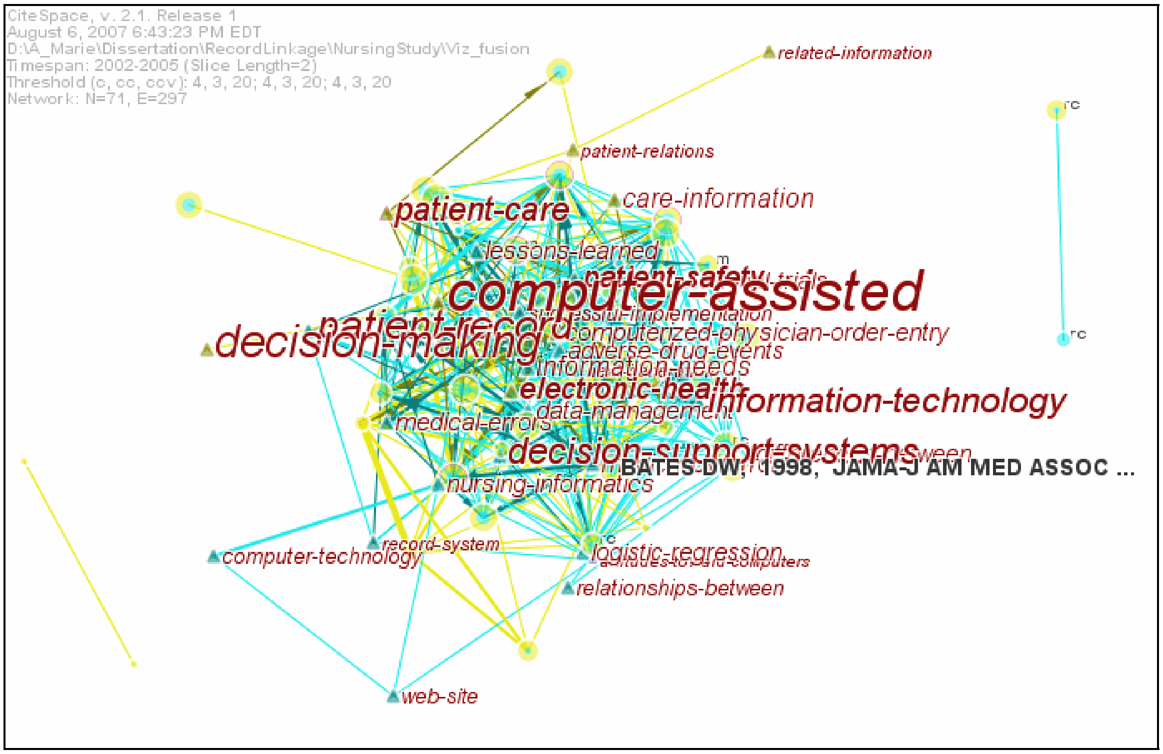


Figure 5.15 Nursing Informatics 2002-2005, With Fusion

Table 5.24 Nursing Informatics 2002-2005, Effects On Visualization Metrics

	Pre-Linkage N = 693 (Figure 5.14)	With Fusion N = 521 (Figure 5.15)
Analysis Type	Document-Term Co-citation	Document-Term Co-citation
Publication Years	2002-2005	2002-2005
Thresholding (c/cc/ccv)	4/3/20	4/3/20
Burst Terms In Range	71	342
Nodes & Links	70 & 266	72 & 311

In addition to increasing the number of keywords available to the visualization, the PRL-IF process has also changed the rankings and content of the top twenty key terms (Table 5.25).

Table 5.25 Nursing Informatics 2002-2005, Effect On Term Rankings

		Pre-Linkage			With Fusion	
Rank		Freq	Keyword		Freq	Keyword
1		22	primary-care		50	computer-assisted
2		14	decision-support-systems		31	Decision-making
3		14	electronic-health		26	patient-record
4		11	computerized-physician		21	Decision-support-systems
5		6	med-decis-making-2003		19	information-technology
6		6	prostate-cancer		17	patient-care
7		5	eq-5d		14	patient-safety
8		4	attitudes-toward-computers		13	electronic-health
9		(no further terms found)			11	care-information
10					11	information-needs
11					10	data-management
12					10	medication-errors
13					10	nursing-informatics
14					9	computerized-physician-order-entry
15					9	differences-between
16					9	lessons-learned
17					9	Medical-errors
18					8	adverse-drug-events
19					8	clinical-trials
20					8	logistic-regression

The use of record linkage to define the sample of citations in the nursing informatics study also changes the data on cited references. Citations have been both removed from and added to the membership of the 20 most highly cited references (Table 5.26).

Table 5.26 Nursing Informatics 2002-2005, Effect On Citation Rankings

		Pre-Linkage				With Fusion			Titles of Cites Added/Dropped with Linkage
Rank		Freq	Author	Year		Freq	Author	Year	
1		32	BATES DW	1998		31	BATES DW	1998	
2		29	~~GOLD MR~~	~~1996~~		22	HUNT DL	1998	
3		23	HUNT DL	1998		21	BATES DW	1999	
4		21	BATES DW	1999		18	*I MED	2001	
5		19	*I MED	2001		17	BATES DW	1995	
6		19	BATES DW	1995		17	LEAPE LL	1995	
7		18	LEAPE LL	1995		16	EVANS RS	1998	
8		17	EVANS RS	1998		**14**	**COVELL DG**	**1985**	**Information needs in office practice: are they being met?**
9		14	TEICH JM	2000		14	TEICH JM	2000	
10		12	CLASSEN DC	1997		12	OVERHAGE JM	1997	
11		12	MCDONALD CJ	1976		12	SITTIG DF	1994	
12		12	OVERHAGE JM	1997		11	CLASSEN DC	1997	
13		12	SITTIG DF	1994		11	MCDONALD CJ	1976	

Table 5.26 (continued)

		Pre-Linkage				With Fusion			Titles of Cites Added/Dropped with Linkage
Rank		Freq	Author	Year		Freq	Author	Year	
14		11	BATES DW	1997		11	TIERNEY WM	1993	
15		11	~~SACKETT DL~~	~~1978~~		10	BATES DW	1997	The utility of different health states as perceived by the general public.
16		11	TIERNEY WM	1993		10	DICK RS	1997	
17		11	~~TORRANCE GW~~	~~1986~~		10	MASSARO TA	1993	Measurement of health state utilities for economic appraisal.
18		10	DICK RS	1997		10	SHEA S	1996	
19		10	~~HANLEY JA~~	~~1982~~		9	ASH JS	1998	The meaning and use of the area under a receiver operating characteristic (ROC) curve.
20		10	JOHNSTON ME	1994		9	ASH JS	2003	
		10	MASSARO TA	1993		9	JOHNSTON ME	1994	
		10	SHEA S	1996		9	OVERHAGE JM	2001	
		9	ASH JS	1998		**9**	**RASCHKE RA**	**1998**	**A computer alert system to prevent injury from adverse drug events: development and evaluation in a community teaching hospital.**
		9	ASH JS	2003		8	*COMM QUAL HLTH CA	2001	
		9	GUSTAFSON DH	1999		8	ASH JS	2004	
		9	OVERHAGE JM	2001		8	BATES DW	1994	
						8	ELY JW	1999	
						8	GORMAN PN	1995	
						8	GUSTAFSON DH	1999	

A comparison of pre-linkage and post-fusion visualizations with the display limited for clarity to top terms shows added terms and a research cluster not previously seen (Figure 5.16 and Figure 5.17). With fusion there is a new group of terms identified related to patient safety (medical errors, medication errors, adverse drug events) with connections to the pivotal paper by Bates on “Effect of computerized physician order entry and a team intervention on prevention of serious medication errors”. The term “nursing informatics” also now appears in the visualization.

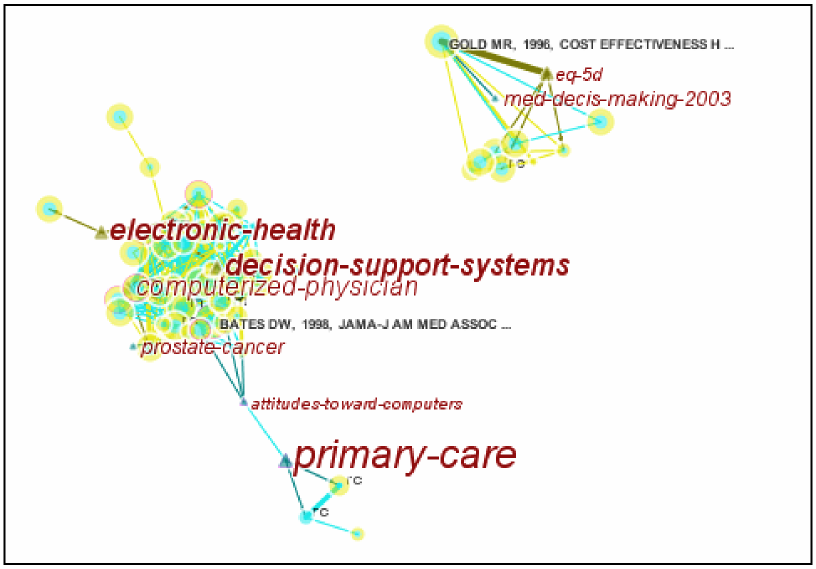


Figure 5.16 Nursing Informatics 2002-2005, Pre-Linkage, With Display Limited To Top Terms

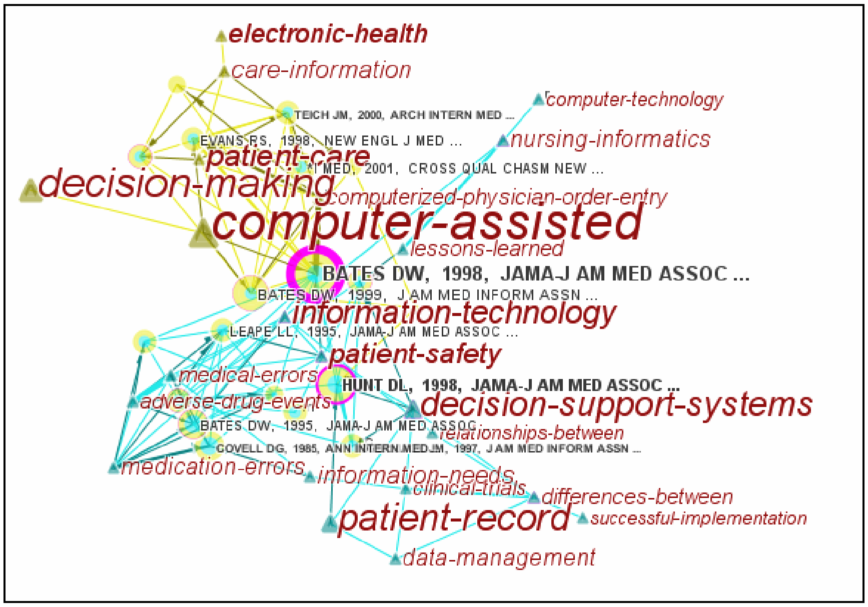

electronic-health
care-information
computer-technology
TEICH JM, 2000, ARCH INTERN MED ...
EVANS RS, 1998, NEW ENGL J MED ...
nursing-informatics
patient-care
I MED, 2001, CROSS QUAL CHASM NEW ...
decision-making
computerized-physician-order-entry
computer-assisted
lessons-learned
BATES DW, 1998, JAMA-J AM MED ASSOC ...
BATES DW, 1999, J AM MED INFORM ASSN ...
information-technology
LEAPE LL, 1995, JAMA-J AM MED ASSOC ...
patient-safety
medical-errors
HUNT DL, 1998, JAMA-J AM MED ASSOC ...
adverse-drug-events
decision-support-systems
BATES DW, 1995, JAMA-J AM MED ASSOC
relationships-between
COVELL DG, 1985, ANN INTERN MED
JM, 1997, J AM MED INFORM ASSN ...
medication-errors
information-needs
clinical-trials
differences-between
patient-record
successful-implementation
data-management

Figure 5.17 Nursing Informatics 2002-2005, Fusion, With Display Limited To Top Terms

5.3 Summary of Findings

The findings from four studies of three knowledge domains using two biomedical citation databases show the multiple points of improvement possible with use of a probabilistic record linkage and information fusion process. All studies showed increases in measures of data quality and increases in burst terms available for labeling visualizations. The fusion of MeSH terms increases the percentage of records with keywords data as well as increasing the number of keywords assigned each record, eliminates duplication of terms between the WOS Keywords Plus and Author provided descriptors, and changes terminology to familiar MeSH terms. In addition to multi-fold increases in burst terms, the reduction in missing data bias obtained through a probabilistic record linkage and information fusion process also improves the rankings and content of the top burst terms. In addition to enriching data, the use of record linkage to improve representative sampling also reduces bias resulting in improved cited reference data. The resulting knowledge domain visualizations are improved by a 1) a reduction in bias as a result of improved data quality and sample selection, and 2) a richer information space for user exploration.

CHAPTER 6: CONCLUSIONS AND FUTURE STUDIES

In this thesis a series of five studies has investigated record linkage models for biomedical citation data and explored the effects of using record linkage to prepare fused data sets for knowledge domain visualization.

6.1 Conclusion 1: Probabilistic versus deterministic models in the linkage of biomedical citation data

The analysis of the performance of record linkage models found the probabilistic model, the Slach deterministic model, and the "Good Practice" deterministic model to have AUC's between .98-1.0. The comparison of ROC curve analysis showed a statistically significant difference between the probabilistic model and the Slach deterministic model. But given that the AUC's were high for both approaches (probabilistic versus deterministic), the questions that might be asked are 1) is a high AUC sufficient to assess model performance and stability, and 2) is the probabilistic approach worth the effort? The answer from analysis of the Cartesian product problem is that the despite the high AUC, deterministic models have a weakness in the inability to generate a unique match key for citation records that is not sensitive to the differences between databases, and the deterministic model performance will not be stable in larger datasets. The Slach and "Good Practice" deterministic models have an underlying assumption that there is a low probability of multiple citations with the same author last name and same page occurring within the same year. This assumption will not hold true where there are highly prolific authors or multiple authors with the same last name publishing at the same time within a domain (C. Friedman is an example in medical informatics), or the citations records do not contain author information, or multiple citations occur within the same page of a journal (as in the conference abstracts from HIV/AIDS conferences). As shown in the case of HIV/AIDS data from year 2005, the use of an Author/Year/Page deterministic model has the potential to generate erroneous false positive links that would create a dataset with >50% incorrectly linked records. The assumption of low rates of Author/Year/Page duplicates in linkages of datasets that cross multiple knowledge domains will also generate erroneous links. The problem with deterministic models cannot be resolved by adding variables to a match key because the variability between datasets will then lead to an increase in false negative matches. The probabilistic model does not have this problem because the matching is based on a sum of weights across variables and includes highly discriminating title information. When reported in terms of accuracy the probabilistic model achieved > 99% accuracy. The most directly comparable performance of functions to map WOS and Medline citations was 70%-79% accuracy that has recently been reported by Bernstam et al (2006), who noted difficulties due to incompatible article representations between the two databases and inadequacy of simple string matching approaches. The Bernstam mapping functions are not described, but probabilistic record linkage approach would provide improved performance.

The analysis of record linkage models focused on one-to-one record linkages between two sets of data, but based on the findings inferences can be made about the utility of probabilistic record linkage for disambiguation of cited references in WOS. Disambiguation of cited references, or identification of variants in forms of a citation, is equivalent to a deduplication linkage in record linkage terms. WOS cited references contain only six possible elements of a citation record: First Author Last Name, First Author First Initial, Year, Journal Abbreviation, Volume, and Page. The extreme variability of the journal and conference proceedings abbreviations found in cited references make this element of the citation record a poor candidate for both exact or approximate string matching. A search of the medical informatics dataset found over 40 variants of citations to The Journal of the American Medical Informatics Association ranging from “J AM MED INFORM ASS” to “J JAMIA’. Variants of either of those will not be matched to the other in an exact match and can be matched to variants of other journal’s abbreviations in an approximate match. Cited references also contain a mixture of record types including citations to both journal articles and books. Due to this mixture, as well as missing data, the Volume and Page elements of the cited reference data are also not suitable for probabilistic linkage calculations based on frequency distributions. Thirty percent of cited references in the medical informatics dataset did not have Volume data and 25% did not have Page data. While 98% of cited references in the medical informatics data do have Author and Year data, this constitutes an insufficient match key to uniquely identify references, and a probabilistic record linkage approach based solely on the data available in WOS cited references will not disambiguate variants of citations.

6.2 Conclusion 2: A Probabilistic Record Linkage and Information Fusion Approach to Citation Data

The effects of preparing citation data for visualization with a probabilistic record linkage and information fusion methodology were initially assessed for two time periods in the domain of medical informatics in a linkage of data from WOS and Medline. The methodology was then further validated in two additional knowledge domains of HIV/AIDS and nursing informatics, and extended to an additional database (CINAHL). All four studies of three knowledge domains using two biomedical citation databases showed increases in measures of data quality, increases in burst terms in visualizations, and changes in rankings of top keywords. The resulting knowledge domain visualizations are improved by a 1) a reduction in bias and 2) a richer information space. These improvements are significant in at least two aspects. First, information visualization has the potential to be a powerful medium for finding causality, forming hypotheses, and assessing available evidence (Chen, 2005). Knowledge domain visualization is a form of information visualization that has the potential for use by a wide range of users, notably scientists, clinicians, science policy researchers, and medical librarians. However if there is a deficiency or anomaly in the data used to generate visualizations, knowledge domain visualization also has the potential to misinform. These visualizations are a form of data mining that can be skewed by sampling errors or systematic patterns of missing data. In the case of progressive knowledge domain visualization the identification of research fronts by burst analysis depends on data collected from titles, abstracts and keywords. A KDD approach to data preparation with probabilistic record linkage can be used to reduce deficiencies in the data, enrich the data, and improve the overall quality of patterns mined.

The second aspect of improvements to knowledge domain visualizations is in reducing barriers to end-user comprehension. Information visualization should be a visual exploration tool that enables the user to interact with the visualized content and comprehend its meaning. But users generally need two types of prior knowledge to understand the intended message in visualized information (Chen 2005):

- The knowledge of how to operate the device (or visualization)
- The domain knowledge of how to interpret the content.

With this methodology users who are not familiar with the literature of a given knowledge domain will have increased information available in the form of key terms to assist in interpreting the subject content of clusters and research fronts. And in the context of biomedical knowledge domains the enrichment with MeSH terminology that is familiar to the medical community may have implications for information retrieval.

6.3 Future Studies

The limitation of the probabilistic model in this study was the approximate string matching of the document Title element of the citation record. The LinkPlus record linkage software was unable to match on the full length of titles exceeding 50 characters without crashing, and situations were observed in some of the modeling trials where similar but semantically different titles received a scored high enough to create a false positive match. The Jaro-Winkler string methods may not be optimal for matching of document titles as Jaro seems designed for short strings, such as a last name (Cohen, 2003). Also the pattern of differences between titles from WOS and Medline does not correspond to the error assumptions of Jaro-Winkler. Jaro-Winkler expects character based errors such as insertions, deletion, and transpositions. However the differences observed between titles are largely of word or phrase level insertions and deletions. An area for further research in the development of probabilistic record linkage models for structured citation data would be the incorporation of token, n-gram or phrase level string comparators from the area of matching of unstructured free-form citations for use in title matching (Hylton 1996, Lawrence 1999, Pasula 2003, Cohen 2003, Wellner 2004). Improvements might also be found by adding a method for a containment operation ("is a contained in b?" rather than "is a equal to b?") in the matching of titles.

The limitations of this enriched information space obtained with data preparation are the poor aesthetics of dense visualizations and the increased difficulty for the user to operate and navigate the visualization. An area for future work would be zoomable user interfaces or ZUI's (Perlin, 1993) for dynamic information visualizations as have been developed for the static images of data visualization. A top problem in the data visualization field has been improving image quality in terms of information density. In systems that support interactive zoom, this means progressively adjusting detail as users zoom and maintaining fonts at a constant screen size (Hibbard, 2004). Google Maps is an example of this type of ZUI with interactive zoom and constant adjustment of labels for information density.

The potential benefits of a methodology for linking and fusing citation data from multiple sources with models that are highly specific and sensitive extends beyond the specific context of biomedical knowledge domain visualization. There are no theoretical reasons why this methodology would not be applicable to domains other than bio-medicine and citation databases other than those studied here. Citation analysis in any form may have a need to improve data quality through better sample selection or to enrich data with additional variables such as secondary authors and institutional affiliations. For the biomedical community that uses Medline there is a need to develop information retrieval strategies to identify articles that are important as well as relevant. Researchers have developed quality filters for Medline that return relevant articles that also conform to methodological quality standards. But even quality filters tuned for precision rather than recall retrieve thousands of articles about common conditions. A high precision query template for therapy returns over 3,800 results for ‘‘breast cancer’’ and the high recall version of the same query template returns over 40,000 results (Bernstam, 2006). Query templates can effectively retrieve high-quality articles, but results are generally not ordered by importance or quality. PubMed clinical query templates retrieve results in reverse chronological order (Bernstam 2006). Bernstam et al (2006) recently compared eight algorithms for identification of important articles: simple PubMed queries, clinical queries (sensitive and specific versions), vector cosine comparison, citation count, journal impact factor, PageRank, and machine learning based on polynomial support vector machines. Citation-based algorithms were found more effective than non-citation-based algorithms at identifying important articles. The most effective strategies were simple citation count and PageRank, which on average identified over six important articles in the first 100 results compared to 0.85 for the best non-citation-based algorithm.

As a preliminary test of the concept that probabilistic record linkage could be used to give Medline citations an “importance” variable by adding the “TC” (times cited) data element from WOS to Medline citations, a comparison was done of article ranking by cited reference citation counts versus the TC counts (Table 6.1). The WOS medical informatics dataset was used to obtain citation counts to JAMIA from all cited references that were then compared to the TC counts directly from the JAMIA citation records. The most challenging aspect of this analysis was first identifying variants of citations to JAMIA and related conference proceedings, with 691 variations found, which demonstrates the processing difficulty of using cited references to obtain counts. A comparison of the 25 most highly cited articles by each methods shows 72% agreement on articles included in the set, and for articles in common to both sets a change in ranking order with higher citation counts obtained form the TC variable.

Table 6.1 Comparison Of JAMIA Rankings By CR Versus TC

Top 25 Cited JAMIA articles from WOS CR						Top 25 Cited JAMIA articles from WOS TC				
Name	**Year**	**Vol**	**Page**	**CR_TC**	**Rank**	**FirstAuthor**	**Year**	**Page**	**WOS_TC**	**Rank**
						BATES, DW	1994	P404	54	21
BATES DW	1999	V6	P313	34	17	Bates, DW	1999	P313	146	2
						Bates, DW	2001	P299	61	20
CAMPBELL JR	1997	V4	P238	38	14	Campbell, JR	1997	P238	54	22
CAMPBELL KE	1994	V1	P218	54	8	CAMPBELL, KE	1994	P218	64	19
CHUTE CG	1996	V3	P224	56	7	Chute, CG	1996	P224	80	14
CIMINO JJ	1994	V1	P35	114	1	CIMINO, JJ	1994	P35	139	3
CIMINO JJ	1995	V2	P273	57	6	CIMINO, JJ	1995	P273	98	6
EVANS DA	1994	V1	P207	71	2	EVANS, DA	1994	P207	89	12
FRIEDMAN C	1994	V1	P161	61	5	FRIEDMAN, C	1994	P161	83	13
						HAYNES, RB	1994	P447	170	1
HUFF SM	1998	V5	P276	26	23					
HUMPHREYS BL	1998	V5	P1	51	10	Humphreys, BL	1998	P1	90	11
						Jha, AK	1998	P305	91	9
						Kane, B	1998	P104	132	5
LEE F	1996	V3	P42	26	24	Lee, F	1996	P42	50	23
MCCRAY AT	1994		P235	32	20					
MCDONALD CJ	1997	V4	P213	35	16	McDonald, CJ	1997	P213	73	16
MILLER RA	1994	V1	P8	39	13	MILLER, RA	1994	P8	90	10
MUSEN MA	1996	V3	P367	70	3	Musen, MA	1996	P367	97	7
OHNOMACHADO L	1998	V5	P357	65	4	Ohno-Machado, L	1998	P357	93	8
OVERHAGE JM	1997	V4	P364	31	21	Overhage, JM	1997	P364	73	15
RECTOR AL	1995	V2	P19	34	18					
ROSSE C	1998	V5	P17	26	26	Rosse, C	1998	P17	45	25
SAGER N	1994	V1	P142	50	11	SAGER, N	1994	P142	65	18
SHEA S	1996	V3	P399	37	15	Shea, S	1996	P399	132	4
SHIFFMAN RN	1999	V6	P104	27	22					
						Shojania, KG	1998	P554	48	24
SITTIG DF	1994	V1	P108	53	9					
SPACKMAN KA	1997		P640	33	19					
						Spitzer, V	1996	P118	70	17
TIERNEY WM	1995	V2	P316	46	12					

With the ability to link and rank articles, probabilistic record linkage has the potential for practical applications in biomedical library use. This methodology could be applied to merging of multi-database searches, thereby enabling querying by MeSH terms with results ranked by citations counts. Currently a visualization of MeSH term co-occurrence will give a view of the organization of articles by topic, but the visual cannot be used to identify relatively important papers within a topic. Addition of the TC data would offer another dimension for coding or filtering a topical visualization. The ability to combine the MeSH terms of Medline citation data with the Times Cited data of WOS also opens new possibilities for information visualizations of MeSH term co-occurrence.

List of References

AMIA. (2003). *American medical informatics association [homepage on the internet]. resource center - publications of interest - journals.* Retrieved Mar 16, 2005 from http://www.amia.org/resource/pubs/f3.html

Aphinyanaphongs, Y., Statnikov, A., & Aliferis, C. F. (2006). A comparison of citation metrics to machine learning filters for the identification of high quality MEDLINE documents. *Journal of the American Medical Informatics Association : JAMIA, 13*(4), 446-455.

Ayres, F. H., Huggill, J. A., & Yannakoudakis, E. J. (1988). The universal standard bibliographic code (USBC): Its use for clearing, merging and controlling large databases. *Program, 22*(2), 117-132.

Ayres, F. H., Nielsen, L. P. S., Ridley, M. J., & Torsun, I. S. (1996). USBC (universal standard bibliographic code) : Its origin and evolution. *Journal of Librarianship and Information Science, 28*(2), 839-91.

Babic, A. (1999). Knowledge discovery for advanced clinical data management and analysis. *Studies in Health Technology & Informatics, 68*, 409-413.

Baldwin, J. A. (1987). *Textbook of medical record linkage*. Oxford: Oxford University Press.

Belin, T. R., & Rubin, D. B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association, 90*(430), 694-707.

Bernstam, E. V., Herskovic, J. R., Aphinyanaphongs, Y., Aliferis, C. F., Sriram, M. G., & Hersh, W. R. (2006 Jan-Feb). Using citation data to improve retrieval from MEDLINE. *Journal of the American Medical Informatics Association, 13*(1), 96-105.

Bhattacharya, I., & Getoor, L. (2004). Iterative record linkage for cleaning and integration. *DMKD '04: Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery,* Paris, France. 11-18. from http://doi.acm.org/10.1145/1008694.1008697

Blakely, T., & Salmond, C. (2002 Dec). Probabilistic record linkage and a method to calculate the positive predictive value. *International Journal of Epidemiology, 31*(6), 1246-1252.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 57*(3), 359-377.

Christen, P. (2002). Probabilistic name and address cleaning and standardisation.

Christen, P., & Churches, T. (2005). *Febrl - freely extensible biomedical record linkage* No. Manual, release 0.3)

CINAHL (2006). How is PubMed / MEDLINE different from CINAHL? (retrieved from http://www.library.health.ufl.edu/help/CINAHL/Media/MEDLINE%20CINAHL%20comparison.pdf on 09/08/2006).

Clark, D. E. (2004). Practical introduction to record linkage for injury research. *Injury Prevention : Journal of the International Society for Child and Adolescent Injury Prevention, 10*(3), 186-191.

Cohen, W. W., Kautz, H., & McAllester, D. (2000). Hardening soft information sources. *KDD '00: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* Boston, Massachusetts, United States. 255-259. from http://doi.acm.org/10.1145/347090.347141

William W. Cohen, Pradeep Ravikumar & Stephen Fienberg (2003): A Comparison of String Metrics for Matching Names and Records in KDD Workshop on Data Cleaning and Object Consolidation 2003.

Contiero, P., Tittarelli, A., Tagliabue, G., Maghini, A., Fabiano, S., & Crosignani, P., et al. (2005). The EpiLink record linkage software - presentation and results of linkage test on cancer registry files. *Methods of Information in Medicine, 44*(1), 66-71.

Cook, L. J., Olson, L. M., & Dean, J. M. (2001 Jul). Probabilistic record linkage: Relationships between file sizes, identifiers and match weights. *Methods of Information in Medicine, 40*(3), 196-203.

Copas, J. B., & Hilton, F. J. (1990). Record linkage - statistical-models for matching computer records. *Journal of the Royal Statistical Society Series A-Statistics in Society, 153*, 287-320.

Cousins, S. A. (1998). Duplicate detection and record consolidation in large bibliographic databases: The COPAC database experience. *Journal of Information Science, 24*, 231-240.

Cousins, S. (1999). Virtual OPACs versus union database: Two models of union catalogue provision. *The Electronic Library, 17*(2), 97-103.

Coyle, K., Brown, L. G., & Brown, L. G. (1985). Record matching: An expert algorithm. *ASIS '85., Edited by Carol A. Parkhurst, White Plains, New York,* Knowledge Industry Publications Inc. for the American Society for Information Science.

Culotta, A., & McCallum, A. (2005). Joint deduplication of multiple record types in relational data. *Proceedings of the 14th ACM International Conference on Information and Knowledge Management. CIKM '05,* Bremen, Germany. 257.

Dunn, H. L. (1946). Record linkage. *American Journal of Public Health, 36*, 1412-1416.

Elfeky, M. G., Verykios, V. S., & Elmagarmid, A. K. (2002). TAILOR: A record linkage toolbox. *ICDE, San Jose, California, USA, February 2002,*

Fawcett, T. (2004). *ROC graphs: Notes and practical considerations for researchers* (Tech Report No. HPL-2003-4). Available: http://home.comcast.net/~tom.fawcett/public_html/papers/ROC101.pdf: HP Laboratories.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*(328), 1183-1210.

Garfield, E. (1972). Citation anlysis as a tool in journal evaluation. *Science, 178*, 471-479.

Gomatam, S., Carter, R., Ariet, M., & Mitchell, G. (2002 May 30). An empirical comparison of record linkage procedures. *Statistics in Medicine, 21*(10), 1485-1496.

Goyal, P. (1987). Duplicate record identification in bibliographic databases. *Information Systems, 12*(3), 239-242.

Grannis, S. J., Overhage, J. M., & McDonald, C. J. (2002). Analysis of identifier performance using a deterministic linkage algorithm.

Grannis, S. J., Overhage, J. M., Hui, S., & McDonald, C. J. (2003). Analysis of a probabilistic record linkage technique without human review. *AMIA.Annu.Symp.Proc.,* , 259-263.

Grannis, S. J., Overhage, J. M., & McDonald, C. J. (2004). Real world performance of approximate string comparators for use in patient matching. MEDINFO 2004 M. Fieschi et al. (Eds) Amsterdam: IOS Press © 2004 IMIA.

Han, J., & Kamber, J. (2001). *Data mining: Concepts and techniques*. San Francisco: Morgan Kaufmann Publishers.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29-36.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148*, 839-843.

Hernández, M. A., & Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the Merge/Purge Problem . *Data min. knowl. discov* (pp. 9-37)

Herskovic, J. R., & Bernstam, E. V. (2005). Using incomplete citation data for MEDLINE results ranking. Paper presented at the 316-320.

Jorge R. Herskovic, M. Sriram Iyengar, Elmer V. Bernst 2007 Using hit curves to compare search algorithm performance. Journal of Biomedical Informatics 40 (2007) 93–99

Hibbard B, (2004). The Top Five Problems that Motivated My Work IEEE Computer Graphics and Applications Volume 24 , Issue 6 (November 2004) Pages: 9 - 13

Hickey, T. B. (1980). *Development of a probabilistic author search and matching technique for retrieval and creation of bibliographic records* No. OCLC/OPR/RR-81-2). Dublin, Ohio: OCLC.

Hickey, T. B., & Rypka, D. J. (1979). Automatic detection of duplicate monographic records. *Journal of Library Automation, 12*(2), 125-142.

Hobbs, G. R. (2001). Data mining and healthcare informatics. *American Journal of Health Behavior, 25*(3), 285.

Hylton, J. A. (1996). *Identifying and merging related bibliographic records* No. TR-678.)Massachusetts Institute of Technology.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association, 84*(406), 414-420.

Jaro, M. A. (1995). Probabilistic linkage of large public-health data files. *Statistics in Medicine, 14*(5-7), 491-498.

Lawrence, S., Giles, C. L., & Bollacker, K. (1999). Autonomous citation matching. *Proceedings of the Third International Conference on Autonomous Agents,* Seattle, Washington.

Monge, A., & Elkan, C. (1997). An efficient domain-independent algorithm for detecting approximately duplicate database records. *The Proceedings of the SIGMOD 1997 Workshop on Data Mining and Knowledge Discovery,*

Newcombe, H. B. (1962). *Record linkage: Making maximum use of the discriminating power of identifying information*

Newcombe, H. B. (1988). *Handbook of record linkage: Methods for health and statistical studies, administration and business*

Newcombe, H. B., Kennedy, J. M., & Axford, A. P. (1959). Automatic linkage of vital records. *Science, 130*(3381), 954-959.

Newgard, C. D. (2006). Validation of probabilistic linkage for matching de-identified prehospital records to a state trauma registry. *Academic Emergency Medicine, 13*, 69-75.

Onorato, E. S., & Bianchi, G. (1981). Automatic identification of duplicates after multidatabase online searching. *Online Review, 5*, 445-451.

Pasula, H., Marthi, B., Milch, B., Russell, S., & Shpitser, I. (2003). Identity uncertainty and citation matching. *Advances In Neural Information Processing Systems,* (15), 1425-1432.

Porter E, Winkler W. Approximate string comparison and it's effect on an advanced record linkage system. In: Record Linkage Techniques: Proceedings of an International Workshop and Exposition; 1997; Arlington, VA: National Academy Press; 1997. chapter 6 p. 190-199.

Pyle, D. (1999). *Data preparation for data mining*. San Francisco: Morgan Kaufmann Publishers.

Ridley, M. J. (1992). An expert system for quality control and duplicate detection in bibliographic databases. *Program, 26*(1), 1-18.

Roos, L. L., & Wajda, A. (1991 Apr). Record linkage strategies. part I: Estimating information and evaluating approaches. *Methods of Information in Medicine, 30*(2), 117-123.

Shaw, W. M.,Jr. (1991). Subject and citation indexing. part I: The clustering structure of composite representations in the cystic fibrosis document collection. *Journal of the American Society for Information Science, 42*(9), 669-675.

Shaw, W. M.,Jr. (1991). Subject and citation indexing. part II: The optimal, cluster-based retrieval performance of composite representations. *Journal of the American Society for Information Science, 42*(9), 676-684.

Slach, J. E. (1985). Detection and elimination of duplicates from multidatabase searches. *Bulletin of the Medical Library Association, 73*(3), 235-237.

Synnestvedt, M. B., & Chen, C. (2003). Visualizing AMIA : A medical informatics knowledge domain analysis. Paper presented at the *Proceedings - AMIA Annual Symposium,* 1024.

Synnestvedt, M. B., & Chen, C. (2005). Design and evaluation of the tightly coupled perceptual-cognitive tasks in knowledge domain visualization. Paper presented at the *The 11th International Conference on Human-Computer Interaction (HCII 2005),* Las Vegas, Nevada.

Synnestvedt, M. B., Chen, C., & Holmes, J. H. (2005). CiteSpace II: Visualization and knowledge discovery in bibliographic databases. Paper presented at the *AMIA '05,* Washington, DC. 724-728.

Synnestvedt, M. B., Chen, C., & Holmes, J. H. (2005). Visual exploration of landmarks and trends in the medical informatics literature. Paper presented at the *AMIA '05,* Washington, DC. 1129.

Thornburg, G. (2005). Matching: Discrimination, misinformation and sudden death. *InSite 2005, , 2* 555-589.

Toney, S. R. (1992). Cleanup and deduplication of an international bibliographic database. *Information Technology and Libraries, 11*, 19-28.

Torra, V. (2003). Trends in information fusion in data mining. *Information fusion in data mining* (pp. 1-6). Berlin: Springer.

Torra, V., & Domingo-Ferrer, J. (2003). Record linkage methods for multidatbase data mining. ***Information fusion*** *in data mining* (pp. 101-132). Berlin : New York: Springer.

Torvik, V. I., Weeber, M., Swanson, D. R., & Smalheiser, N. R. (2003). A probabilistic similarity metric for medline records: A model for author name disambiguation. *AMIA.Annu.Symp.Proc.,* , 1033.

Trybula, W. J. (1998). Data mining and knowledge discovery. *Annual Review of Information Science and Technology, 32*, 197-229.

Verykios, V. S., Elmagarmid, A. K., & Houstis, E. N. (2000). Automating the approximate record-matching process. *Inf.Sci.Inf.Comput.Sci., 126*(1-4), 83-98.

Wagner, G., & Newcombe, H. B. (1970). Record linkage - its methodology and application in medical data processing. *Methods of Information in Medicine, 9*(2), 121-&.

Wajda, A., & Roos, L. L. (1987). Simplifying record linkage - software and strategy. *Computers in Biology and Medicine, 17*(4), 239-248.

Wajda, A., Roos, L. L., Layefsky, M., & Singleton, J. A. (1991 Aug). Record linkage strategies: Part II. portable software and deterministic matching. *Methods of Information in Medicine, 30*(3), 210-214.

Web of Science (2007) Thomson Scientific. http://scientific.thomson.com/products/wos/

Wellner, B., McCallum, A., Peng, F., & Hay, M. (2004). An integrated, conditional model of information extraction and coreference with application to citation matching. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence,* Banff, Canada. *, 70* 593-601.

Winkler, W.E. (1990), "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," Proceedings of the section on Survey Research Methods, American Statistical Association., 354-359

Winkler, W. E. (1999). *The state of record linkage and current research problems* from census.gov

Winkler, W. E. (2000). *Using the EM algorithm for weight computation in the fellegi-sunter model of record linkage* No. RR2000/05)

Winkler, W. E. (2003). Data cleaning methods. *Proceedings of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation,* Washington, DC.

Winkler, W. E. (2006). *Overview of record linkage and current research directions* No. Statistics 2006-2)from census.gov

Yannakoudakus, E. J., Ayres, F. H., & Huggill, J. A. W. (1990). Matching of citations between non-standardized databases. *Journal of the American Society for Information Science, 41*(8), 599-610.

Zhu, J. J., & Ungar, L. H. (2000). String edit analysis for merging databases. *KDD 2000,*

Zupan, B., Lavrac, N., & Keravnou, E. (1999). Data mining techniques and applications in medicine. *Artificial Intelligence in Medicine, 16*(1), 12.

APPENDIX

Table A1. WOS Dataset Survey

WOS Field Tags	Field Description	Records With data	Maximum Value (Text)
AB	Abstract	7327	z-tests for probabilistic output. Measures of performance of the expert
AU	Authors	11752	Zywietz, CW
BP	Beginning page	11750	V
C1	Author address	8536	Zonguldak Karaelmas Univ, Fac Engn, Dept Mech Engn, TR-67100 Zonguldak, Turkey.
CA	Group Authors	42	Telemed Adoption Study Grp
CR	Cited references	11072	ZYWIETZ C, UNPUB
DE	Author keywords	5171	Ziv-Lempel method
DT	Document type	11752	Software Review
EM	E-mail address	706	zywietz.christoph@biosigna.de
EP	Ending page	11750	VI
ER	End of record	11752	
GA	ISI document delivery number	11752	ZZ377
ID	Keywords Plus®	4913	ZIDOVUDINE; TRIAL
IS	Issue	10513	9-10
J9	29-character source abbreviation	11752	METHODS INFORM MED
JI	ISO source abbreviation	11315	Methods Inf. Med.
LA	Language	11752	English
NR	Cited reference count	11752	99
PA	Publisher address	11752	PO BOX 211, 1000 AE AMSTERDAM, NETHERLANDS
PD	Publication date	8003	WIN
PG	Page count	11752	9
PI	Publisher city	11752	THOUSAND OAKS
PT	Publication type (e.g., book, journal, book in series)	11752	S
PU	Publisher	11752	TAYLOR & FRANCIS LTD
PY	Publication year	11752	2005
RP	Reprint address	9341	Zywietz, CW, BIOSIGNA Inst, Feodor Lynen Str 21,Med Pk, D-30625
SC	Subject category	11196	THEORY & METHODS; ENGINEERING, BIOMEDICAL; MEDICAL INFORMATICS
SI	Special issue	98	Sp. Iss. SI
SN	ISSN	11752	1538-2931
SO	Full source title	11752	METHODS OF INFORMATION IN MEDICINE
SU	Supplement	1294	Suppl. S
TC	Times Cited	11752	98
TI	Document title	11752	ZUR VERWENDUNG DER FAKTORENANALYSE IN DER MEDIZINISCHEN DIAGNOSTIK
UT	ISI unique article identifier	11752	ISI:A1997YJ72800004
VL	Volume	10580	934

Table A2. Medline Dataset Survey

Medline Field Tags	Field Description	Records w/ data	Maximum Value (Text)	Equiv. WOS Field Tag
AB	Abstract	12011	ZX-81 to read data files from magnetic tape making the data analysis	AB
AD	Affiliation	10927	Zywietz.Christoph@MH-Hannover.de	
AID	Article Identifier	3934	YYV2KY15QYJ80TQW [pii]	
AU	Author	15696	Zywietz CW	AU
CI	Copyright Information	88	Copyright 2002 Elsevier Science Ireland Ltd.	
CN	Corporate Author	37	The UK MARIBS Breast Screening Study.	
DA	Date Created	15920	which outflow data are the only available.	
DCOM	Date Completed	15893	20060504	
DEP	Date of Electronic Publication	111	20051230	
DP	Date of Publication	15920	2006 May	PY
EDAT	Entrez Date	15920	2006/03/28 09:00	
FAU	Full Author	15666	Zywietz, C W	
FIR	Full Investigator Name	13	Webster, L	
FPS	Full Personal Name as Subject	67	Williams, G Z	
GN	General Note	134	treatment	
GR	Grant Number	2321	Z-T15-LM07037-04/LM/NLM	
GS	Gene Symbol	1	HUMCOL4A5	
IP	Issue	12966	Pt 2	IS
IR	Investigator Name	13	Webster L	
IRAD	Investigator Affiliation	13	Washington U, St Louis, MO	
IS	ISSN	15463	1559-4076 (Electronic)	SN
JID	NLM Unique ID	15897	9712259	
JT	Journal Title	15895	Symposium. AMIA Symposium.	SO
LA	Language	15920	ger	
LR	Date Last Revised	14872	20060510	
MH	MeSH Terms	15893	Zimeldine/adverse effects/toxicity	ID
MHDA	MeSH Date	15920	2006/05/05 09:00	
OAB	Other Abstract	7	vaccines is therefore called for. Ideal vaccines will be administered in	
OID	Other ID	194	POP: 00269942	
OT	Other Term	231	Youth	
OTO	Other Term Owner	231	PIP	
OWN	Owner	15920	NLM	
PG	Pagination	15918	V-VIII	BP
PHST	Publication History Status	320	2005/12/27 [aheadofprint]	
PL	Place of Publication	15912	United States	PI
PMID	PubMed Unique Identifier	15920	9988966	
PS	Personal Name as Subject	68	Williams GZ	
PST	Publication Status	15920	ppublish	
PT	Publication Type	15920	Validation Studies	DT
PUBM	Publishing Model	15920	Print-Electronic	
RF	Number of References	633	99	NR
RN	Registry Number/EC Number	1698	EC 6.1.1.1 (Tyrosine-tRNA Ligase)	
SB	Subset	15893	X	
SFM	Space Flight Mission	1	Soyuz TM22 Project	
SI	Secondary Source ID	1	GENBANK/AI111901	
SO	Source	15920	Proc Annu Symp Comput Appl Med Care. 1995;:96-100.	
STAT	Status	15920	Publisher	
TA	Journal Title Abbreviation	15920	Proc Annu Symp Comput Appl Med Care	J9
TI	Title	15920	Zora: a pilot virtual community in the pediatric dialysis unit.	TI
TT	Transliterated Title	259	Zusatzklassifikation zur Kennzeichung von Personen ohne akute Beschwerden	
VI	Volume	13176	Suppl	VL

Table A3. Full List Of Variation In Author Names Between WOS And Medline For Matched Citations

Wos FirstAuthor	Medline FirstAuthor		Wos FirstAuthor	Medline FirstAuthor
ABRAHAMS.S	Abrahamsson S		MACDONAL.LK	MacDonald LK
Af Klercker, T	Klercker T		MALINDZA.GS	Malindzak GS Jr
ALPEROVI.A	Alperovitch A		MCALISTE.NH	McAlister NH
ARZBAECH.RC	Arzbaecher RC		MCCONVILLE, KMV	Mc Conville KM
BARBOSA, MD	Barbosa M de Matos		MELO, MFV	Vidal Melo MF
BENBASSAT, M	Ben-Bassat M		MEYEREBRECHT, D	Meyer-Ebrecht D
BENJEBRIA, A	Ben Jebria A		MINAMIKAWATACHINO, R	Minamikawa-Tachino R
BLUMENFE.W	Blumenfeld W		MORI, AR	Rossi Mori A
CAMPIONEPICCARDO, J	Campione-Piccardo J		MUSTAKAL.KK	Mustakallio KK
CANTRAIN.FR	Cantraine FR		NIELSENKUDSK, F	Nielsen-Kudsk F
CHRISTENSENSZALANSKI, JJJ	Christensen-Szalanski JJ		NILANDWEISS, J	Niland-Weiss J
CORNFIEL.J	Cornfield J		Ning, OY	Ouyang N
Cosp, XB	Bonfill Cosp X		NORDSCHO.CD	Nordschow CD
DARGENIO, DZ	D'Argenio DZ		OBRIEN, KF	O'Brien KF
DAS, REG	Gaines RE		OCHOASANGRADOR, C	Ochoa-Sangrador C
DATRI, A	D'Atri A		OMARA, K	O'Mara K
DEBLIEK, R	de Bliek R		OQUIGLEY, J	O'Quigley J
DEBRUIJN, LM	De Bruijn LM		OSHAUGHNESSY, TJ	O'Shaughnessy TJ
DECARVALHO, LAV	de Carvalho LA		PATTISONGORDON, E	Pattison-Gordon E
DEMEDINACELI, L	de Medinaceli L		PEER, J	Pe'er J
DEMOOR, GJE	De Moor GJ		PIPBERGE.HV	Pipberger HV
DENICOLAO, G	De Nicolao G		PLUYTERWENTING, ESP	Pluyter-Wenting ES
DEPONTI, F	De Ponti F		POLIHRON.P	Polihroniadis P
DEROSIS, F	de Rosis F		PRADHAM, M	Pradhan M
deRoulet, D	de Roulet D		PRYER, DB	Pryor DB
DHOLLOSY, W	d'Hollosy W		Read, CY	Yetter Read C
DHOORE, W	D'Hoore W		REICHERT.PL	Reichertz PL
DIFELICE, P	Di Felice P		Riesco, AM	Manjarres Riesco A
DOMBAL, FTD	de Dombal FT		Schoeffler, KM	Liu GC
DUDDLESO.WG	Duddleson WG		SCHOEVAERTBROSSAULT, D	Schoevaert-Brossault D
EBENCHAIME, M	Eben-Chaime M		SHINOZAK.T	Shinozaki T
ELDHAHER, AHG	el-Dhaher AH		Siegel, JE	Hagen MD
FAIRHURST, MC	Fairhust MC		Silveira, PSP	Panse Silveira PS
FDEZVALDIVIA, J	Fernandez-Valdivia J		SMYTHSTARUCH, K	Smyth-Staruch K
FEINSTEI.AR	Feinstein AR		SRINIVAS.R	Srinivasan R
FLATLEY, P	Brennan PF		STARTSMA.TS	Startsman TS
France, FHR	Roger France FH		Stoykova, B	Nixon J
GARFINKE.D	Garfinkel D		TAGLIACO.R	Tagliacozzo R
GONCEWINDER, C	Gonce-Winder C		Timothy, TYY	Lai TY
Gonzalez, JS	Solano Gonzalez J		VANALSTE, JA	van Alste JA
GONZALEZHEYDRICH, J	Gonzalez-Heydrich J		VANBEMMEL, JH	van Bemmel JH
GOUVEIAOLIVEIRA, A	Gouveia-Oliveira A		VANBRUNT, EE	Van Brunt EE
GUSTAFSO.DH	Gustafson DH		VANDAMME, M	van Damme M
Guvenir, HA	Altay Guvenir H		VANDENAKKER, TJ	van den Akker TJ
HajianTilaki, KO	Hajian-Tilaki KO		VANDERLEER, OFC	van der Leer OF
HENDERSO.C	Henderson C		VANDERLEIJE, BA	van der Leije BA
HENDRICK.L	Hendrickson L		VANDORP, HD	van Dorp HD
Houghton, J	Haughton J		VANGENNIP, EMSJ	van Gennip EM
HUSSONVANVLIET, J	Husson-van Vliet J		VANHEIJST, G	van Heijst G
JESDINSK.HJ	Jesdinsky HJ		VANKREEL, BK	van Kreel BK
KARBER, G	KAERBER G		vanOverbeeke, JJ	van Overbeeke JJ
Keravnou, ET	Eravnou ET		vanRoijen, L	van Roijen L
Kohl, P	Kokol P		VANZEE, GA	van Zee GA
LEAO, BD	Leao Bde F		VEGACATALAN, FJ	Vega-Catalan FJ
LLEWELLYNTHOMAS, HA	Llewellyn-Thomas HA		WHITINGOKEEFE, QE	Whiting-O'Keefe QE
LONBERGHOLM, K	Lonberg-Holm K		WIJNAND, HP	Hauschke D
LOPEZCABRERA, A	Lopez-Cabrera A		ZWETSLOOTSCHONK, JHM	Zwetsloot-Schonk JH
LUECKE, RH	Leucke RH			

Table A4. Linkage Errors by Slach Deterministic Model (DM1), not made by Probabilistic Model

Error	DS	First Author	Year	ISSN	Journal Abbrev	Volume	Issue	Pages	Title+
False Positive	WOS	Kiel, JM	2000	0724-6811	M D COMPUT	17	1	27-28	Resolution 2000: Create an inviting e-practice
False Positive	Med	Kiel JM	2000	0724-6811	MD Comput	17	2	27-8	Positive outcomes, lower costs: using net-based IT to manage care.
False Positive	WOS	Kiel, JM	2000	0724-6811	M D COMPUT	17	1	27-28	Resolution 2000: Create an inviting e-practice
False Positive	Med	Kiel JM	2000	0724-6811	MD Comput	17	4	27-8	Buy software or "pay-per-view": the ASP option.
False Positive	WOS	Goodman, KW	1999	0724-6811	M D COMPUT	16	3	17-+	Bioinformatics: Challenges revisited
False Positive	Med	Goodman KW	1999	0724-6811	MD Comput	16	2	17-20	Health informatics and the Hospital Ethics Committee.
False Positive	WOS	Kiel, JM	1999	0724-6811	M D COMPUT	16	3	27-28	Going high tech: Size matters? Think again ...
False Positive	Med	Kiel JM	1999	0724-6811	MD Comput	16	5	27-9	yourpractice.com: making the leap to the Internet.
False Positive	WOS	van der Weijden, T	2003	0272-989X	MED DECIS MAKING	23	3	226-231	Unexplained complaints in general practice: Prevalence, patients' expectations, and professionals' test-ordering behavior
False Positive	Med	van Ginneken AM	2003	0026-1270	Methods Inf Med	42	3	226-35	Considerations for the representation of meta-data for the support of structured data entry.
False Positive	WOS	Haux, R	2002	0026-1270	METHODS INFORM MED	41	1	31-35	Health care in the information society: What should be the role of medical informatics?
False Positive	Med	Haux R	2002	1386-5056	Int J Med Inform	65	1	31-9	Master of science program in health information management at Heidelberg/Heilbronn: a health care oriented approach to medical informatics.
False Positive	WOS	Cai, YD	2000	1089-7771	IEEE TRANS INF TECHNOL BIOMED	4	2	152-158	Content-based retrieval of dynamic PET functional images
False Positive	Med	Cai D	2000	1367-4803	Bioinformatics	16	2	152-8	Modeling splice sites with Bayes networks.
False Positive	WOS	Friedman, C	2001	1067-5027	J AMER MED INFORM ASSOC	-	-	189-193	Evaluating the UMLS as a source of lexical knowledge for medical language processing
False Positive	Med	Friedman CP	2001	1067-5027	J Am Med Inform Assoc	8	2	189-91	Publication bias in medical informatics.
False Negative	WOS	DEROSIS, F	1979	0026-1270	METHODS INFORM MED	18	4	203-206	HEALTH-CARE REORGANIZATION AND INFORMATION-SYSTEM BUILDING IN ITALY
False Negative	Med	de Rosis F	1979	0026-1270	Methods Inf Med	18	4	203-6	Health care reorganization and information system building in Italy.
False Negative	WOS	DOMBAL, FTD	1972	0026-1270	METHODS INFORM MED	11	1	32-&	PATTERN-RECOGNITION - COMPARISON OF PERFORMANCE OF CLINICIANS AND NON-CLINICIANS - WITH A NOTE ON PERFORMANCE OF A COMPUTER-BASED SYSTEM
False Negative	Med	de Dombal FT	1972	0026-1270	Methods Inf Med	11	1	32-7	Pattern-recognition: a comparison of the performance of clinicians and non-clinicians--with a note on the performance of a computer-based system.
False Negative	WOS	VANZEE, GA	1978	0010-4809	COMPUT BIOMED RES	11	4	325-335	CONTRAST ENHANCING FILTER FOR BANDED CHROMOSOMES
False Negative	Med	van Zee GA	1978	0010-4809	Comput Biomed Res	11	4	325-35	A contrast enhancing filter for banded chromosomes.

Error	DS	First Author	Year	ISSN	Journal Abbrev	Volume	Issue	Pages	Title+
False Negative	WOS	VANBRUNT, EE	1970	0010-4809	COMPUT BIOMED RES	3	5	477-&	KAISER-PERMANENTE MEDICAL INFORMATION SYSTEM
False Negative	Med	Van Brunt EE	1970	0010-4809	Comput Biomed Res	3	5	477-87	The Kaiser-Permanente Medical Information System.
False Negative	WOS	VANDERLEIJE, BA	1983	0010-468X	COMPUT PROGRAM BIOMED	17	3	243-248	SIRAD - A PROGRAM FOR AUTOMATIC DOSIMETRY AND DATA TRANSFER FOR RADIOTHERAPY PLANNING
False Negative	Med	van der Leije BA	1983	0010-468X	Comput Programs Biomed	17	3	243-8	SIRAD: a program for automatic dosimetry and data transfer for radiotherapy planning.
False Negative	WOS	DAS, REG	1982	0010-468X	COMPUT PROGRAM BIOMED	15	1	13-21	ITERATIVE WEIGHTED REGRESSION-ANALYSIS OF LOGIT RESPONSES - A COMPUTER-PROGRAM FOR ANALYSIS OF BIOASSAYS AND IMMUNOASSAYS
False Negative	Med	Gaines RE	1982	0010-468X	Comput Programs Biomed	15	1	13-21	Iterative weighted regression analysis of logit responses: a computer program for analysis of bioassays and immunoassays.
False Negative	WOS	VANDAMME, M	1981	0010-468X	COMPUT PROGRAM BIOMED	13	3	239-250	THE PRESSURE AND FLOW DISTRIBUTION WITHIN A FILTERING CAPILLARY NETWORK
False Negative	Med	van Damme M	1981	0010-468X	Comput Programs Biomed	13	3	239-50	The pressure and flow distribution within a filtering capillary network.
False Negative	WOS	OQUIGLEY, J	1980	0010-468X	COMPUT PROGRAM BIOMED	12	1	14-18	WEIBULL - A REGRESSION-MODEL FOR SURVIVAL TIME STUDIES
False Negative	Med	O'Quigley J	1980	0010-468X	Comput Programs Biomed	12	1	14-8	Weibull: a regression model for survival time studies.
False Negative	WOS	DARGENIO, DZ	1979	0010-468X	COMPUT PROGRAM BIOMED	9	2	115-134	PROGRAM PACKAGE FOR SIMULATION AND PARAMETER-ESTIMATION IN PHARMACOKINETIC SYSTEMS
False Negative	Med	D'Argenio DZ	1979	0010-468X	Comput Programs Biomed	9	2	115-34	A program package for simulation and parameter estimation in pharmacokinetic systems.
False Negative	WOS	LUECKE, RH	1978	0010-468X	COMPUT PROGRAM BIOMED	8	1	35-43	PROGRAM TO SIMULATE DRUG ELIMINATION INTERACTIONS - WARFARIN AND BSP - ILLUSTRATIVE EXAMPLE
False Negative	Med	Leucke RH	1978	0010-468X	Comput Programs Biomed	8	1	35-43	A program to simulate drug elimination interactions: warfarin and BSP - an illustrative example.
False Negative	WOS	DEPONTI, F	1988	0020-7101	INT J BIO-MED COMPUT	22	1	51-64	QUANTITATIVE-ANALYSIS OF INTESTINAL ELECTRICAL SPIKE ACTIVITY BY A NEW COMPUTERIZED METHOD
False Negative	Med	De Ponti F	1988	0020-7101	Int J Biomed Comput	22	1	51-64	Quantitative analysis of intestinal electrical spike activity by a new computerized method.
False Negative	WOS	BENJEBRIA, A	1987	0020-7101	INT J BIO-MED COMPUT	21	2	137-151	EFFECT OF RESIDENT GAS-DENSITY ON CO2 ELIMINATION DURING HIGH-FREQUENCY OSCILLATION - A MODEL STUDY
False Negative	Med	Ben Jebria A	1987	0020-7101	Int J Biomed Comput	21	2	137-51	Effect of resident gas density on CO2 elimination during high-frequency oscillation: a model study.
False Negative	WOS	OMARA, K	1985	0020-7101	INT J BIO-MED COMPUT	17	1	31-48	ENHANCED X-RAY-IMAGING OF SPHEROIDS - AN O(N)

Error	DS	First Author	Year	ISSN	Journal Abbrev	Volume	Issue	Pages	Title+
									ALGORITHM FOR CHARACTERIZING CONVEX BLOBS
False Negative	Med	O'Mara K	1985	0020-7101	Int J Biomed Comput	17	1	31-48	Enhanced X-ray imaging of spheroids: an O(n) algorithm for characterizing convex blobs.
False Negative	WOS	VANALSTE, JA	1986	0010-4809	COMPUT BIOMED RES	19	5	417-427	ECG BASE-LINE WANDER REDUCTION USING LINEAR-PHASE FILTERS
False Negative	Med	van Alste JA	1986	0010-4809	Comput Biomed Res	19	5	417-27	ECG baseline wander reduction using linear phase filters.
False Negative	WOS	DEMEDINACELI, L	1984	0010-4809	COMPUT BIOMED RES	17	2	185-192	RAT SCIATIC FUNCTIONAL INDEX DATA MANAGEMENT-SYSTEM WITH DIGITIZED INPUT
False Negative	Med	de Medinaceli L	1984	0010-4809	Comput Biomed Res	17	2	185-92	Rat sciatic functional index data management system with digitized input.
False Negative	WOS	VANDENAKKER, TJ	1982	0010-4809	COMPUT BIOMED RES	15	5	405-417	AN ONLINE METHOD FOR RELIABLE DETECTION OF WAVEFORMS AND SUBSEQUENT ESTIMATION OF EVENTS IN PHYSIOLOGICAL SIGNALS
False Negative	Med	van den Akker TJ	1982	0010-4809	Comput Biomed Res	15	5	405-17	An on-line method for reliable detection of waveforms and subsequent estimation of events in physiological signals.
False Negative	WOS	BENJEBRIA, A	1981	0010-4809	COMPUT BIOMED RES	14	6	493-505	FINITE-ELEMENT SIMULATION OF GAS-TRANSPORT IN PROXIMAL RESPIRATORY AIRWAYS - COMPARISON WITH EXPERIMENTAL-DATA
False Negative	Med	Ben Jebria A	1981	0010-4809	Comput Biomed Res	14	6	493-505	Finite element simulation of gas transport in proximal respiratory airways: comparison with experimental data.
False Negative	WOS	DEROSIS, F	1988	0026-1270	METHODS INFORM MED	27	1	23-33	TREATMENT OF UNCERTAINTY IN AN ONCOLOGY PROTOCOL BY PROBABILISTIC AND ARTIFICIAL-INTELLIGENCE APPROACHES
False Negative	Med	de Rosis F	1988	0026-1270	Methods Inf Med	27	1	23-33	Treatment of uncertainty in an oncology protocol by probabilistic and artificial intelligence approaches.
False Negative	WOS	PEER, J	1982	0026-1270	METHODS INFORM MED	21	1	23-25	COMPUTER IMAGE-ANALYSIS OF THE OCULAR FUNDUS
False Negative	Med	Pe'er J	1982	0026-1270	Methods Inf Med	21	1	23-5	Computer image analysis of ocular fundus.
False Negative	WOS	BENBASSAT, M	1980	0026-1270	METHODS INFORM MED	19	2	93-98	A HIERARCHICAL MODULAR DESIGN FOR TREATMENT PROTOCOLS
False Negative	Med	Ben-Bassat M	1980	0026-1270	Methods Inf Med	19	2	93-8	A hierarchical modular design for treatment protocols.
False Negative	WOS	DIFELICE, P	1990	0169-2607	COMPUT METHOD PROGRAM BIOMED	31	2	125-137	FUNCTIONALITY OF THE ARPIA AMBULATORY INFORMATION-SYSTEM
False Negative	Med	Di Felice P	1990	0169-2607	Comput Methods Programs Biomed	31	2	125-37	Functionality of the ARPIA ambulatory information system.
False Negative	WOS	VANKREEL, BK	1989	0169-2607	COMPUT METHOD PROGRAM BIOMED	28	2	137-149	THERMODYNAMIC NETWORK MODELING OF TRANSFER ACROSS THE PERFUSED GUINEA-PIG PLACENTA, USING

Error	DS	First Author	Year	ISSN	Journal Abbrev	Volume	Issue	Pages	Title+
									SPICE
False Negative	Med	van Kreel BK	1989	0169-2607	Comput Methods Programs Biomed	28	2	137-49	Thermodynamic network modelling of transfer across the perfused guinea-pig placenta, using SPICE.
False Negative	WOS	ELDHAHER, AHG	1988	0169-2607	COMPUT METHOD PROGRAM BIOMED	26	1	63-70	MICROCOMPUTER-BASED SYSTEM TO MEASURE, RECORD AND PROCESS FLOW VOLUME CURVES, RESPIRATORY QUESTIONNAIRE DATA AND ENVIRONMENTAL EXPOSURE
False Negative	Med	el-Dhaher AH	1988	0169-2607	Comput Methods Programs Biomed	26	1	63-70	Microcomputer-based system to measure, record and process flow-volume curves, respiratory questionnaire data and environmental exposure.
False Negative	WOS	VANBEMMEL, JH	1987	0169-2607	COMPUT METHOD PROGRAM BIOMED	25	3	243-244	4TH-GENERATION MEDICAL INFORMATION-SYSTEMS - FOREWORD
False Negative	Med	van Bemmel JH	1987	0169-2607	Comput Methods Programs Biomed	25	3	243-4	Fourth-generation medical information systems.
False Negative	WOS	DHOORE, W	1993	0026-1270	METHODS INFORM MED	32	5	382-387	RISK ADJUSTMENT IN OUTCOME ASSESSMENT - THE CHARLSON COMORBIDITY INDEX
False Negative	Med	D'Hoore W	1993	0026-1270	Methods Inf Med	32	5	382-7	Risk adjustment in outcome assessment: the Charlson comorbidity index.
False Negative	WOS	VANBEMMEL, JH	1992	0026-1270	METHODS INFORM MED	31	4	235-246	ADVANCES IN AN INTERDISCIPLINARY SCIENCE
False Negative	Med	van Bemmel JH	1992	0026-1270	Methods Inf Med	31	4	235-46	Advances in an interdisciplinary science.
False Negative	WOS	VANBEMMEL, JH	1989	0026-1270	METHODS INFORM MED	28	4	227-233	EDUCATION IN MEDICAL INFORMATICS IN THE NETHERLANDS - A NATIONWIDE POLICY AND THE ERASMUS CURRICULUM
False Negative	Med	van Bemmel JH	1989	0026-1270	Methods Inf Med	28	4	227-33	Education in medical informatics in The Netherlands: a nationwide policy and the Erasmus curriculum.
False Negative	WOS	DEMOOR, GJE	1994	0020-7101	INT J BIO-MED COMPUT	35	1	1-12	STANDARDIZATION IN MEDICAL INFORMATICS IN EUROPE
False Negative	Med	De Moor GJ	1994	0020-7101	Int J Biomed Comput	35	1	1-12	Standardisation in medical informatics in Europe.
False Negative	WOS	VANDERLEER, OFC	1994	0020-7101	INT J BIO-MED COMPUT	35	-	87-95	THE USE OF PERSONAL DATA FOR MEDICAL-RESEARCH - HOW TO DEAL WITH NEW EUROPEAN PRIVACY STANDARDS
False Negative	Med	van der Leer OF	1994	0020-7101	Int J Biomed Comput	35	-	87-95	The use of personal data for medical research: how to deal with new European privacy standards.
False Negative	WOS	DEROULET, D	1994	0020-7101	INT J BIO-MED COMPUT	35	-	107-114	THE TECHNICAL CONDITIONS FOR AN OPEN-ARCHITECTURE
False Negative	Med	de Roulet D	1994	0020-7101	Int J Biomed Comput	35	-	107-14	The technical conditions for an open architecture.
False Negative	WOS	VANDORP, HD	1994	0020-7101	INT J BIO-MED COMPUT	35	-	179-186	THE AIM SEISMED GUIDELINES FOR SYSTEM-DEVELOPMENT AND DESIGN
False Negative	Med	van Dorp HD	1994	0020-7101	Int J Biomed Comput	35	-	179-86	The AIM SEISMED guidelines for system development and design.

Error	DS	First Author	Year	ISSN	Journal Abbrev	Volume	Issue	Pages	Title+
False Negative	WOS	DEMOOR, GJE	1994	0020-7101	INT J BIO-MED COMPUT	34	1	319-330	TOWARDS A META-SYNTAX FOR MEDICAL EDI
False Negative	Med	De Moor GJ	1994	0020-7101	Int J Biomed Comput	34	1	319-30	Towards a meta-syntax for medical edi.
False Negative	WOS	VANGENNIP, EMSJ	1992	0020-7101	INT J BIO-MED COMPUT	30	3	153-158	A VIEW OF THE WORKSHOP
False Negative	Med	van Gennip EM	1992	0020-7101	Int J Biomed Comput	30	3	153-8	A view of the workshop.
False Negative	WOS	MCCONVILLE, KMV	1991	0020-7101	INT J BIO-MED COMPUT	27	3	157-173	APPLICATION OF THE ENTROPY THEORY OF PERCEPTION TO AUDITORY INTENSITY DISCRIMINATION
False Negative	Med	Mc Conville KM	1991	0020-7101	Int J Biomed Comput	27	3	157-73	Application of the entropy theory of perception to auditory intensity discrimination.
False Negative	WOS	Silveira, PSP	1998	0010-4809	COMPUT BIOMED RES	31	1	1-17	Modeling and simulating morphological evolution in an artificial life environment
False Negative	Med	Panse Silveira PS	1998	0010-4809	Comput Biomed Res	31	1	1-17	Modeling and simulating morphological evolution in an artificial life environment.
False Negative	WOS	OBRIEN, KF	1994	0010-4809	COMPUT BIOMED RES	27	6	434-440	CONCERNING THE ANALYSIS OF 2X2-TABLES
False Negative	Med	O'Brien KF	1994	0010-4809	Comput Biomed Res	27	6	434-40	Concerning the analysis of 2 x 2 tables.
False Negative	WOS	WARNER, H	1993	0010-4809	COMPUT BIOMED RES	26	4	319-326	A VIEW OF MEDICAL INFORMATICS AS AN ACADEMIC DISCIPLINE
False Negative	Med		1993	0010-4809	Comput Biomed Res	26	4	319-26	A view of medical informatics as an academic discipline.
False Negative	WOS	MELO, MFV	1993	0010-4809	COMPUT BIOMED RES	26	2	103-120	ALVEOLAR VENTILATION TO PERFUSION HETEROGENEITY AND DIFFUSION IMPAIRMENT IN A MATHEMATICAL-MODEL OF GAS-EXCHANGE
False Negative	Med	Vidal Melo MF	1993	0010-4809	Comput Biomed Res	26	2	103-20	Alveolar ventilation to perfusion heterogeneity and diffusion impairment in a mathematical model of gas exchange.
False Negative	WOS	OBRIEN, B	1994	0272-989X	MED DECIS MAKING	14	3	289-297	WILLINGNESS-TO-PAY - A VALID AND RELIABLE MEASURE OF HEALTH STATE PREFERENCE
False Negative	Med	O'Brien B	1994	0272-989X	Med Decis Making	14	3	289-97	Willingness to pay: a valid and reliable measure of health state preference?
False Negative	WOS	Houghton, J	2000	0724-6811	M D COMPUT	17	4	34-38	A paradigm shift in healthcare - From disease management to patient-centered systems
False Negative	Med	Haughton J	2000	0724-6811	MD Comput	17	4	34-8	A paradigm shift in healthcare. From disease management to patient-centered systems.
False Negative	WOS	Ning, OY	1998	0724-6811	M D COMPUT	15	2	106-109	Using a neural network to diagnose the hypertrophic portions of hypertrophic cardiomyopathy
False Negative	Med	Ouyang N	1998	0724-6811	MD Comput	15	2	106-9	Using a neural network to diagnose the hypertrophic portions of hypertrophic cardiomyopathy.
False Negative	WOS	PAYNE, B	1993	0724-6811	M D COMPUT	10	4	231-267	THE 10TH ANNUAL DIRECTORY OF MEDICAL HARDWARE AND SOFTWARE COMPANIES
False Negative	Med		1993	0724-6811	MD Comput	10	4	231-67	The tenth annual directory of medical hardware and software companies.
False Negative	WOS	VANGENNIP, EMSJ	1992	0169-2607	COMPUT METHOD PROGRAM	37	4	265-271	DO THE BENEFITS OUTWEIGH THE COSTS OF PACS - THE RESULTS OF AN

Error	DS	First Author	Year	ISSN	Journal Abbrev	Volume	Issue	Pages	Title+
					BIOMED				INTERNATIONAL WORKSHOP ON TECHNOLOGY-ASSESSMENT OF PACS
False Negative	Med	van Gennip EM	1992	0169-2607	Comput Methods Programs Biomed	37	4	265-71	Do the benefits outweigh the costs of PACS? The results of an International Workshop on Technology Assessment of PACS.
False Negative	WOS	Gonzalez, JS	2000	0933-3657	ARTIF INTELL MED	19	1	75-89	Model-based spectral estimation of Doppler signals using parallel genetic algorithms
False Negative	Med	Solano Gonzalez J	2000	0933-3657	Artif Intell Med	19	1	75-89	Model-based spectral estimation of Doppler signals using parallel genetic algorithms.
False Negative	WOS	Riesco, AM	2000	0933-3657	ARTIF INTELL MED	18	1	57-82	A customisable framework for the assessment of therapies in the solution of therapy decision tasks
False Negative	Med	Manjarres Riesco A	2000	0933-3657	Artif Intell Med	18	1	57-82	A customisable framework for the assessment of therapies in the solution of therapy decision tasks.
False Negative	WOS	Keravnou, ET	1996	0933-3657	ARTIF INTELL MED	8	3	187-191	Temporal reasoning in medicine
False Negative	Med	Eravnou ET	1996	0933-3657	Artif Intell Med	8	3	187-91	Temporal reasoning in medicine.
False Negative	WOS	VANHEIJST, G	1995	0933-3657	ARTIF INTELL MED	7	3	227-255	A CASE-STUDY IN ONTOLOGY LIBRARY CONSTRUCTION
False Negative	Med	van Heijst G	1995	0933-3657	Artif Intell Med	7	3	227-55	A case study in ontology library construction.
False Negative	WOS	Read, CY	2004	1538-2931	CIN-COMPUT INFORM NURS	22	2	83-89	Conducting a client-focused survey using e-mail
False Negative	Med	Yetter Read C	2004	1538-2931	Comput Inform Nurs	22	2	83-9	Conducting a client-focused survey using e-mail.
False Negative	WOS	Elfrink, V	1999	0736-8593	COMPUT NURS	17	2	73-81	Designing an information technology application for use in community-focused nursing education
False Negative	Med		1999	0736-8593	Comput Nurs	17	2	73-81	Designing an information technology application for use in community-focused nursing education. Nightingale Tracker Field Test Nurse Team.
False Negative	WOS	FLATLEY, P	1994	1067-5027	J AMER MED INFORM ASSOC	-	-	1011-1011	ELDERS ATTITUDES AND BEHAVIOR REGARDING COMPUTERLINK
False Negative	Med	Brennan PF	1994	0195-4210	Proc Annu Symp Comput Appl Med Care	-	-	1011	Elders' attitudes and behavior regarding ComputerLink.
False Negative	WOS	DEBLIEK, R	1994	1067-5027	J AMER MED INFORM ASSOC	1	4	328-338	INFORMATION RETRIEVED FROM A DATABASE AND THE AUGMENTATION OF PERSONAL KNOWLEDGE
False Negative	Med	de Bliek R	1994	1067-5027	J Am Med Inform Assoc	1	4	328-38	Information retrieved from a database and the augmentation of personal knowledge.
False Negative	WOS	vanRoijen, L	1996	0266-4623	INT J TECHNOL ASSESS HEALTH C	12	3	405-415	Labor and health status in economic evaluation of health care - The health and labor questionnaire
False Negative	Med	van Roijen L	1996	0266-4623	Int J Technol Assess Health Care	12	3	405-15	Labor and health status in economic evaluation of health care. The Health and Labor Questionnaire.
False Negative	WOS	Cosp, XB	1996	0266-4623	INT J TECHNOL ASSESS HEALTH C	12	2	388-394	Evaluation of the regular practice of breast cancer screening in a health area
False Negative	Med	Bonfill Cosp X	1996	0266-4623	Int J Technol Assess Health Care	12	2	388-94	Evaluation of the regular practice of breast cancer screening in a health area.

Error	DS	First Author	Year	ISSN	Journal Abbrev	Volume	Issue	Pages	Title+
False Negative	WOS	Perry, S	1995	0266-4623	INT J TECHNOL ASSESS HEALTH C	11	4	795-796	Report from the Agencia D'Avaluacio de Tecnologia Medica (CAHTA)
False Negative	Med		1995	0266-4623	Int J Technol Assess Health Care	11	4	795-6	Report from the Agencia D'avaluacio de Tecnologia Medica (CAHTA).
False Negative	WOS	vanBemmel, JH	1996	0026-1270	METHODS INFORM MED	35	3	157-172	Medical informatics, art or science?
False Negative	Med	van Bemmel JH	1996	0026-1270	Methods Inf Med	35	3	157-72; 173-201	Medical informatics, art or science?
False Negative	WOS	DHOLLOSY, W	1995	0026-1270	METHODS INFORM MED	34	3	266-271	SEMIAUTOMATED DATABASE DESIGN BY THE END-USER
False Negative	Med	d'Hollosy W	1995	0026-1270	Methods Inf Med	34	3	266-71	Semi-automated database design by the end-user.
False Negative	WOS	vanGennip, EMSJ	1996	0020-7101	INT J BIO-MED COMPUT	43	3	161-178	Guidelines for cost-effective implementation of picture archiving and communication systems an approach building on practical experiences in three European hospitals
False Negative	Med	van Gennip EM	1996	0020-7101	Int J Biomed Comput	43	3	161-78	Guidelines for cost-effective implementation of Picture Archiving and Communication Systems. An approach building on practical experiences in three European hospitals.
False Negative	WOS	deRoulet, D	1996	0020-7101	INT J BIO-MED COMPUT	43	1	39-44	Technical means for securing health information
False Negative	Med	de Roulet D	1996	0020-7101	Int J Biomed Comput	43	1	39-44	Technical means for securing health information.
False Negative	WOS	vanOverbeeke, JJ	1996	0020-7101	INT J BIO-MED COMPUT	42	1	91-96	The Dutch 'Benefit-II' project: Do physicians benefit from using an electronic medical dossier?
False Negative	Med	van Overbeeke JJ	1996	0020-7101	Int J Biomed Comput	42	1	91-6	The Dutch 'Benefit-II' project: do physicians benefit from using an electronic medical dossier?
False Negative	WOS	DEMOOR, GJE	1995	0020-7101	INT J BIO-MED COMPUT	39	1	81-85	EUROPEAN STANDARDS DEVELOPMENT IN HEALTH-CARE INFORMATICS - ACTUAL AND FUTURE CHALLENGES
False Negative	Med	De Moor GJ	1995	0020-7101	Int J Biomed Comput	39	1	81-5	European standards development in healthcare informatics: actual and future challenges.
False Negative	WOS	MORI, AR	1995	0020-7101	INT J BIO-MED COMPUT	39	1	93-98	CODING SYSTEMS AND CONTROLLED VOCABULARIES FOR HOSPITAL INFORMATION-SYSTEMS
False Negative	Med	Rossi Mori A	1995	0020-7101	Int J Biomed Comput	39	1	93-8	Coding systems and controlled vocabularies for hospital information systems.
False Negative	WOS	PRYER, DB	1995	0020-7101	INT J BIO-MED COMPUT	39	1	105-109	MANAGING THE DELIVERY OF HEALTH-CARE - CARE-PLANS MANAGED CARE PRACTICE GUIDELINES
False Negative	Med	Pryor DB	1995	0020-7101	Int J Biomed Comput	39	1	105-9	Managing the delivery of health care: care-plans/managed care/practice guidelines.
False Negative	WOS	DECARVALHO, LAV	1995	0020-7101	INT J BIO-MED COMPUT	38	1	33-45	A COMPUTATIONAL MODEL FOR THE NEUROBIOLOGICAL SUBSTRATES OF VISUAL-ATTENTION
False Negative	Med	de Carvalho LA	1995	0020-7101	Int J Biomed Comput	38	1	33-45	A computational model for the neurobiological substrates of visual

Error	DS	First Author	Year	ISSN	Journal Abbrev	Volume	Issue	Pages	Title+
									attention.
False Negative	WOS	af Klercker, T	1998	0169-2607	COMPUT METHOD PROGRAM BIOMED	55	1	31-37	Decision support system for primary health care in an inter/intranet environment
False Negative	Med	Klercker T	1998	0169-2607	Comput Methods Programs Biomed	55	1	31-7	Decision support system for primary health care in an inter/intranet environment.
False Negative	WOS	DEBRUIJN, LM	1995	0169-2607	COMPUT METHOD PROGRAM BIOMED	48	1	151-156	SPEECH INTERFACING FOR DIAGNOSIS REPORTING SYSTEMS - AN OVERVIEW
False Negative	Med	De Bruijn LM	1995	0169-2607	Comput Methods Programs Biomed	48	1	151-6	Speech interfacing for diagnosis reporting systems: an overview.
False Negative	WOS	DENICOLAO, G	1995	0169-2607	COMPUT METHOD PROGRAM BIOMED	47	3	237-252	WENDEC - A DECONVOLUTION PROGRAM FOR PROCESSING HORMONE TIME-SERIES
False Negative	Med	De Nicolao G	1995	0169-2607	Comput Methods Programs Biomed	47	3	237-52	WENDEC: a deconvolution program for processing hormone time-series.
False Negative	WOS	FDEZVALDIVIA, J	1995	0169-2607	COMPUT METHOD PROGRAM BIOMED	46	3	187-205	A NEW METHODOLOGY TO SOLVE THE PROBLEM OF CHARACTERIZING 2-D BIOMEDICAL SHAPES
False Negative	Med	Fernandez-Valdivia J	1995	0169-2607	Comput Methods Programs Biomed	46	3	187-205	A new methodology to solve the problem of characterizing 2-D biomedical shapes.
False Negative	WOS	OSHAUGHNESSY, TJ	1995	0169-2607	COMPUT METHOD PROGRAM BIOMED	46	1	79-90	A COMPUTER-PROGRAM FOR THE STUDY OF SYNAPTIC TRANSMISSION AT THE NEUROMUSCULAR-JUNCTION
False Negative	Med	O'Shaughnessy TJ	1995	0169-2607	Comput Methods Programs Biomed	46	1	79-90	A computer program for the study of synaptic transmission at the neuromuscular junction.
False Negative	WOS	DATRI, A	1994	0169-2607	COMPUT METHOD PROGRAM BIOMED	45	1	123-125	MILORD - MULTIMEDIA INTERACTION WITH LARGE OBJECT-ORIENTED RADIOLOGICAL AND CLINICAL DATABASES
False Negative	Med	D'Atri A	1994	0169-2607	Comput Methods Programs Biomed	45	1	123-5	MILORD: Multi-media Interaction with Large Object-oriented Radiological and clinical Databases.
False Negative	WOS	Littlejohns, P	2000	0266-4623	INT J TECHNOL ASSESS HEALTH C	16	4	1039-1049	Guideline development in Europe - An international comparison
False Negative	Med		2000	0266-4623	Int J Technol Assess Health Care	16	4	1039-49	Guideline development in Europe. An international comparison.
False Negative	WOS	Stoykova, B	2000	0266-4623	INT J TECHNOL ASSESS HEALTH C	16	3	731-742	The UKNHS Economic Evaluation Database - Economic issues in evaluations of health technology
False Negative	Med	Nixon J	2000	0266-4623	Int J Technol Assess Health Care	16	3	731-42	The U.K. NHS economic evaluation database. Economic issues in evaluations of health technology.
False Negative	WOS	Siegel, JE	2001	0272-989X	MED DECIS MAKING	21	4	307-323	Does cost-effectiveness analysis make a difference? Lessons from pap

Error	DS	First Author	Year	ISSN	Journal Abbrev	Volume	Issue	Pages	Title+
									smears - Preface
False Negative	Med	Hagen MD	2001	0272-989X	Med Decis Making	21	4	307-23	Does cost-effectiveness analysis make a difference? Lessons from Pap smears. Symposium.
False Negative	WOS	Haux, R	2000	0026-1270	METHODS INFORM MED	39	3	267-277	Recommendations of the International Medical Informatics Association (IMIA) on education in health and medical informatics
False Negative	Med		2000	0026-1270	Methods Inf Med	39	3	267-77	Recommendations of the International Medical Informatics Association (IMIA) on education in health and medical informatics.
False Negative	WOS	Timothy, TYY	2004	1386-5056	INT J MED INFORM	73	5	415-431	Do doctors act on their self-reported intention to computerize? A follow-up population-based survey in Hong Kong
False Negative	Med	Lai TY	2004	1386-5056	Int J Med Inform	73	5	415-31	Do doctors act on their self-reported intention to computerize? A follow-up population-based survey in Hong Kong.
False Negative	WOS	France, FHR	2003	1386-5056	INT J MED INFORM	70	2	215-219	Case mix use in 25 countries: a migration success but international comparisons failure
False Negative	Med	Roger France FH	2003	1386-5056	Int J Med Inform	70	2	215-9	Case mix use in 25 countries: a migration success but international comparisons failure.
False Negative	WOS	Kohl, P	2001	1386-5056	INT J MED INFORM	63	1	1-4	Intelligent medical systems - preface
False Negative	Med	Kokol P	2001	1386-5056	Int J Med Inform	63	1	1-4	Intelligent medical systems - preface.
False Negative	WOS	Schoeffler, KM	2001	1067-5027	J AMER MED INFORM ASSOC	-	-	388-392	Standards for the Electronic Health Record emerging from health care's Tower of Babel
False Negative	Med	Liu GC	2001	1531-605X	Proc AMIA Symp	-	-	388-92	Standards for the electronic health record, emerging from health care's Tower of Babel.
False Negative	WOS	Guvenir, HA	2004	0933-3657	ARTIF INTELL MED	31	3	231-240	Diagnosis of gastric carcinoma by classification on feature projections
False Negative	Med	Altay Guvenir H	2004	0933-3657	Artif Intell Med	31	3	231-40	Diagnosis of gastric carcinoma by classification on feature projections.

VITA

Marie B. Synnestvedt

Home Address: 44 Whyte Drive
Voorhees, NJ 08043

Education:

1972-76	B.S.	Penn State University (Biology)
1987-89	M.S.Ed.	University of Pennsylvania (M.E.T.E.R.*)
2001-2007	Ph.D.	Drexel University (Information Science)

(*Measurement, Evaluation, and Techniques of Experimental Research)

Visa type and number: United States citizen

Hospital and Administrative Appointments:

1998 - 2006	Sr. Programmer/Analyst, Center for Clinical Epidemiology & Biostatistics University of Pennsylvania School of Medicine
2006 -	Sr. Data Analyst, Office of Human Research University of Pennsylvania School of Medicine

Specialty Certification:

Clinical Research Certificate Program, Center for Clinical Epidemiology & Biostatistics, University of Pennsylvania School of Medicine, In Progress

Awards, Honors and Membership in Honorary Societies:

2004	Drexel University IST Research Day Award
2004	Drexel University Evelyn Walker Armstrong Endowed Scholarship
2004	Drexel University Judith M. Feller ‘63 Endowed Scholarship for Students of L&IS

Bibliography: Selected Recent Research Publications, peer reviewed (print or other media):

Synnestvedt, M.B., Chen, C. and Holmes, J.H., CiteSpace II: Visualization and knowledge discovery in bibliographic databases. in AMIA '05, (Washington, DC. October, 2005). pp. 724-728.

Synnestvedt, M.B. and Chen, C., Design and evaluation of the tightly coupled perceptual-cognitive tasks in knowledge domain visualization. in The 11th International Conference on Human-Computer Interaction (HCII 2005), (Las Vegas, Nevada, 2005), Lawrence Erlbaum Associates.

Synnestvedt, M.B., Enriching Knowledge Domain Visualizations: Analysis of a Record Linkage and Information Fusion Approach to Citation Data. in AMIA '07, (Chicago, IL. November, 2007). pp. 711-716.