MEDICINFOSYS: AN ARCHITECTURE FOR AN EVIDENCE-BASED MEDICAL INFORMATION RESEARCH AND DELIVERY SYSTEM

by

Pif Edwards

Submitted in partial fulfillment of the
requirements for the degree of
Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
August 2010

DALHOUSIE UNIVERSITY

FACULTY OF COMPUTER SCIENCE

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "MEDICINFOSYS: AN ARCHITECTURE FOR AN EVIDENCE-BASED MEDICAL INFORMATION RESEARCH AND DELIVERY SYSTEM" by Pif Edwards in partial fulfillment of the requirements for the degree of Master of Computer Science.

Dated: August 3, 2010

Supervisor: _____
Dr. Vlado Kešelj

Readers: _____
Dr. Michael Shepherd

_____
Dr. Christian Blouin

DALHOUSIE UNIVERSITY

DATE: August 3, 2010

AUTHOR: Pif Edwards

TITLE: MEDICINFOSYS: AN ARCHITECTURE FOR AN EVIDENCE-BASED MEDICAL INFORMATION RESEARCH AND DELIVERY SYSTEM

DEPARTMENT OR SCHOOL: Faculty of Computer Science

DEGREE: M.C.Sc. CONVOCATION: October YEAR: 2010

Table of Contents

List of Tables

List of Figures

Abstract

Due to the complicated nature of medical information needs, the time constraints of clinicians, and the linguistic complexities and sheer volume of medical information, most medical questions go unanswered. It has been shown that nearly all of these questions can be answered with the presently available medical sources and that when these questions get answered, patient health benefits.

In this work, we design and describe a framework for Evidence-Based medical information research and delivery, MedicInfoSys. This system leverages the strengths of knowledge-based workers and of mature knowledge-based technologies within the medical domain. The most critical element of this framework, is a search interface, PifMed. PifMed uses gold-standard MeSH categorization (presently integrated into MEDLINE) as the basis of a navigational structure, which allows users to browse search results with an interactive tree of categories. Evaluation by user study shows it to be superior to PubMed, in terms of speed and usability.

Chapter 1

Introduction

> *"Medicine, in modern jargon, is a knowledge based business, and experienced doctors use about two million pieces of information to manage their patients. ...Clinical information can be defined as 'the commodity used to help make patient care decisions.' "* [78]

The above quote is very helpful in framing the modern medical situation in a way computer scientists can appreciate. Given a patient's situation, a physician is either certain or uncertain on how to proceed. We must provide a system that increases the level of medical certainty in patient care, which benefits both patients health and doctors confidence in a present, and similar future situations.

1.1 Motivation

There are three important factors at the center of the clinical context that motivate and mold this effort in health informatics: 1) the stakes are very high (for physicians as well for as patients); 2) time is in short supply and; 3) physicians have sophisticated and context-specific information needs which must be satisfied by an equally sophisticated and comprehensive knowledge base.

Physicians have fourteen years of post-secondary education and their level of diction reflects that education. This high level of diction makes the source material — medical documentation — often beyond the understanding of anyone outside the medical field, and its interpretation into medical practice requires years of experience. In order to plumb this highly sophisticated source material, equally sophisticated methods of information retrieval are required.

Physicians bring a huge volume of medical knowledge to bear on any reading and interpretation of a medical article. This fact makes seemingly straight-forward tasks in this domain, like the identification of similarities or differences in sentences, exceedingly difficult for computers to perform [23].

The conclusion that the systems which presently exist are insufficient is supported by the fact that the majority of medical questions go unanswered [92, 78, 28]. Physicians have much less time to pursue information needs (2–8 minutes) [73, 28], then it takes to satisfy all but the simplest of them (10–43 minutes on average) [73, 78, 49]. Several studies have shown that the majority of unanswered questions were answerable with present resources — between 77%–92% [73, 78] of the time — and would have changed patient management 40%–47% [73, 78] of the time. The answers are there, but within the present clinical context, physicians cannot find them due to a lack of time and the inadequacy of search systems.

It is important to note there are two distinct user-groups, those in the **research context** and those in the **clinical context**. The research information gathering tasks produce results which are meant to be generalized, the clinical information gathering tasks are meant to be interpreted into the context of a specific patient.

In the terms of information retrieval (IR) evaluation, the time constraints present in the clinical context support the weighting of precision over recall as the prudent evaluator of IR systems. That is, finding a small number of good articles is sufficient, perhaps even one if it has the precise answer. Contrast this with users in the research context which require a system that performs strongly in terms of recall [73]. That is, finding articles which cover many-to-all different perspectives on a topic.

A motivation and goal is to make best use of human efforts in the medical domain; to get medical research results into the hands of doctors. Medical findings which cannot be found, cannot be used to help patients. We need a better way to get answers from the laboratory to the physician.

1.2 Objectives

We aim to make a survey of current medical information needs and existing difficulties facing physicians with information needs: to reveal the roots of this problem, the extent and the affect on patient care. To do this, we must investigate the nature of medical information in terms of its linguistic complexity, ethical constraints, use and growth.

A second objective is to suggest a framework for getting physicians the information they need, in form they can use, in a way that reduces their workload — allows them

to focus on patient care —, that is affordable, timely, transparent and reliable.

A third objective is to demonstrate the maturity of the knowledge-based resources within the medical domain built to tackle the complex nature of medical information, and show how they can be used for information retrieval.

A forth objective is to implement a major part of the suggested framework: a complete IR prototype that demonstrates a paradigm shift away from ranked lists, into hierarchical categorization.

A fifth objective is to design a usability study, test this method of evaluation with a small pilot study and then execute a large-scale user study to show that a navigation structure based on hierarchical categorization is both better in terms of usability and in terms of effectiveness when compared to the leading medical domain search engine, PubMed.

1.2.1 Leverage Human Efforts

> *"Building on the generalization of human-computer optimization... we hypothesize that by including a human 'in-the-loop' we can leverage the intelligence of the human and the processing power of the computer to quickly solve the same problems with better solutions."* [75]

The above quote is a great overall description of this goal, specifically, we wish to leverage existing and on-going efforts of human-knowledge workers to aid in information retrieval. We identify four ways in which we do this: MeSH hierarchical categorization, MEDLINE Indexing, Hierarchical browsing and Informationist information collection.

1. **MeSH Hierarchical Categorization**: The MeSH taxonomy is a human effort, each category and relationship painstakingly constructed and maintained over a 50 year period. We wish to put this effort to use by bringing it to the forefront of our search UI.

2. **MEDLINE Indexing**: Each article in MEDLINE has been individually indexed by a human knowledge worker with an advanced life science degree. Our goal is to make maximal use of this indexing.

3. **Hierarchical Browsing**: The user no longer relies on a ranking algorithm to dictate an order of visitation of search results, instead the user chooses their own ordering by implicitly refining their query as they browse results. Instead of an objective function, we have a navigational structure which facilitates dynamic, subjective ordering, which both guides the user and instantly reacts to the user's evolving query.

4. **Informationist Information Collection**: The Informationist is inserted between the physician and the medical information, a specialist in information gathering and trained to understand complex medical questions and recognize potential answers.

In short, I wish to show the reader that a new division of labour is necessary within the medical field, to describe the task they need to do, the parameters they need to do it in, a framework for this division to function within, the tools needed to be efficient and effective, and to recommend a tested method of remuneration.

1.3 Thesis Statement

Information Retrieval within the Medical Domain needs to take advantage of specialized labour, knowledge-based methods and a new search interface paradigm to effectively satisfy information needs and constraints of physicians in the present medical information climate.

I will describe the design, execution and analysis of a user study that confirms this hypothesis. I will present statistically significant results of a collection of paired t-tests, which clearly indicate that in this domain, knowledge-based methods and my novel navigational structure are more effective, efficient and usable.

1.4 Contributions

There are two main contributions of this work: **MedicInfoSys** and **PifMed**.

1.4.1 MedicInfoSys

An architecture within which the physician and Informationist can divide up the task of information gathering, and the 'fruit of their labours' can be stored and reused by

other users. A framework in which the Informationist can use specialized tools, built especially for their task and their domain.

1.4.2 PifMed

Another contribution is PifMed: a hierarchical browsing system, with a collapsible tree-based navigational structure. This is a paradigm shift away from ranked lists and is proven in this thesis to be more effective, efficient and usable for browsing large result sets. This is a new method that puts the order of article visitation, the ranking of result sets, out of the black box of a ranking algorithm and into the head of the user. By leveraging the existing gold-standard categorization of human indexers, this method collects similar articles together, so articles can be categorically investigated or categorically ignored. This system allows users to focus queries as the browse, allows users to see a summarization of all articles down a given path (if you see the category name as a one-word summary) and shows the relationships between returned results.

Attention has been paid to aspects specific to this browsing model, such as default tree state (open-closed) and its impact on usability, tree-customization (deletions, session searching) and increasing and decreasing search sensitivity. We used the pilot study to refine this prototype for the user study. Based on the user study results, further refinements were implemented. These final changes are described and future improvements are outlined in the Future Work Section.

Chapter 2

Background and Related Work

2.1 Clinical Context

2.1.1 Clinical Information Climate

> *"US medical care is some 30% more expensive than that in Canada and Europe, where quality is comparable; and US medicine also has the most litigious malpractice climate in the world. Some have argued that this 30% surcharge on US medical care, about US$1000 per capita annually, is mostly medico-legal: either direct legal costs, or else the overhead of 'defensive medicine', i.e. unnecessary tests ordered by physicians to cover themselves in potential future lawsuits. In this tense climate, physicians and other medical data-producers are understandably reluctant to hand over their data to data miners."* [13]

With the stakes as high as they are in medicine, where daily decisions have life and death impacts, information must be accurate and timely, and sources must be reliable and trustworthy. If not, patients face death and injury and doctors face lawsuits and the loss of their livelihoods. With stakes this high questions of ethical responsibility must be addressed.

Data mining in the medical domain has three primary ethical issues: data ownership, fear of lawsuits and privacy [13]. First the question of data ownership; do patients own data about them, or do physicians own the data they collect, or do the insurance providers who paid for the tests own the data? Adding to the confusion are ethical questions surrounding the sale of human data and tissues. Second, there is the threat of lawsuits. Physicians and medical data-producers are wary that data provided could be used against them in a court of law, that accidental omissions and unrepresented context specific information may generate — or add leverage to — a case of malpractice. This is a situation where physicians have much to lose and little

to gain. Finally, there is the issue of privacy. Patient-physician confidentiality is a legal contract which patients and doctors both take very seriously. If there was any doubt in this confidence, patients may not be as forthcoming and the care of patients would suffer. These ethical question can be contextualized by the following four levels of identification:[13]

1. **Anonymous data**: No identification. (e.g. Tissue from a corpse.)
2. **Anonymized data**: Identification completely removed.
3. **De-identified data**: Patient ID encoded and encrypted.
4. **Identified data**: Patient given written informed consent.

These ethical questions are moot for level 1 and level 2; having an increasing impact on level 3; and these questions are vital and explicit for level 4 data. For many questions specific to patient diagnosis the context information available only in level 3 and 4 is critical and highly valued.

2.1.2 The Advantages of Medical Literature

For all experts, text is the primary channel for information exchange [81], the medical domain is no different. The medical literature is the predominant medium for researchers to make known their findings. Medical articles have well-structured conventions for the presentation of the material which provides multiple entry points into the information (Title, Abstract, Introduction, and specific sections headings to direct the readers' attention.) In 1987, The Ad Hoc Working Group for Critical Appraisal of the Medical Literature established guidelines for structuring headings within abstracts to reflect the content of publications in an effort to help people quickly assess content [23], increasing the usability of the literature for users. When present, this standardized structure of headings (Objective, Method, Results, Conclusion) can be used to the advantage of an IR system made sensitive to it.

Clinicians are not the only ones who have unmet medical information needs. Pharmaceutical companies in development of medications use the same resources and it is estimated that these companies derive 90% of drug targets from the literature [36]. Unfortunately, the amount of information is curbing their advancements, "surveys

suggest that about 50% of all potentially therapeutic compounds undergo attrition due to safety concerns and that about 50% of them had some indication in the literature already" [36].

Computerized medical records have been a suggested new source of research data. However, since medicine is primarily a patient-care activity and only secondarily acts as a research resource [13], it must be noted that there is a clear advantage of finding evidence in scientific literature since it is intended to be used as evidence, where data generated from medical records is not. When filling out patient medical records doctors are meant to focus on patient health not on the future needs of researchers. Though the use of these records is rife with pitfalls (privacy, legal-responsibility, their anecdotal and idiosyncratic nature, and habitual incompleteness [13]) they can be effectively used in a supporting role, for example in the assistance of automatic and interactive query formulation.

2.1.3 Amount of Information

The sheer volume of documents in this domain is its greatest blessing and ultimate obstacle. There is a vast array of information sources: commercial, governmental, academic, open-access, all of varying reliability. Over the last 20 years, primary sources (such as MEDLINE) are growing at a double-exponential pace [39]. In fact, MEDLINE has grown at a $\sim$4.2% compounded annual growth rate [39], and as of July 2010, MEDLINE has 18,182,098 [68] citations indexed and was increasing at a rate of more than 2300/day [68]. Medical research produces medical findings at a reported rate of publishing 55 clinical trials a day [50]. Each medical specialty has its own tale, but the same story; Epidemiologists "would need over 600 hours a month to read every new article published in their field" [21] and "the body of information on HIV doubles every 22 months, and, although half of that information is concentrated in 30 journals, the other half is spread through 593" [78]. In fact, according to research done in 1985, the biomedical knowledge-base doubles every 19 years, meaning that medical knowledge will quadruple during a professional lifetime [78]. If you take into account the double-exponential growth rate of medical information, the period it takes medical knowledge to double, is shortening, thereby worsening this problem.

This problem not only has the medical impact of lost opportunities for improved

patient care but financial impacts as well. "Studies by International Data Corporation estimate that an enterprise employing 1,000 knowledge workers wastes nearly $2.5 million per year due to an inability to locate and retrieve information" [36].

Medical professionals stand at the foot of an exponentially growing 'mountain' of information. For the proper functioning of the medical system patients must have confidence in their doctors' level of knowledge and for doctors to provide a state-of-the-art level of care they must have the state-of-the-art tools to navigate this 'mountain'.

In summary, the sheer volume of information and lack of an adequate way of searching it has the following consequences: (1) searching for answers to clinical questions is likely to fail; (2) keeping up to date in even one medical field requires an enormous effort and time — time and effort which doctors prefer to spend caring for patients; (3) advances in the field, medical breakthroughs, and all the "effort, creativity, and money that go into biomedical research is simply wasted" [19].

2.1.4 Information Needs of Clinicians

There are two types of information needs of clinicians: focused and general. The **focused** need is one where a specific question is formulated, specific situational factors are in play and the clinician requires an exact answer. The **general** information need is one where an overview is necessary and sufficient to satisfy the need, but from which a focused question may emerge [79].

Ely *et al.* [27] divided the process of asking and answering clinical questions into five steps: (1) recognizing an uncertainty, (2) formulating a question, (3) pursuing an answer, (4) finding an answer, and (5) applying the answer to patient care. Ely *et al.* also compiled a complete (and exhaustive) taxonomy of obstacles to clinical question answering [29]. Both the question process steps and obstacle taxonomy are useful in identifying where things go wrong, and to generate ideas on how we can help.

The satisfaction of an information need begins with recognizing one. The lack of recognition of a need, that is, the problem of not knowing that you don't know, is aggravated by rate which new clinical information is being generated. We could provide tools to help in this capacity, for example, article recommendations and new finding notifications may help recognize uncertainties by increasing awareness.

The primary reasons not to pursue questions were lack of time and lack of confidence that an answer could be found [30]. Thus any new information system must demonstrate what sort of questions can be answered and how quickly, especially when systems are faster and when questions that a system is capable of handling are unlike those in past systems.

Several studies show that when questions are answered, patients benefit. It has been shown "that conducting a MEDLINE search early in the hospitalization of a patient could significantly lower costs, charges, and lengths of stay" [43] and that "...answers to these questions came from MEDLINE and the information from the articles changed patient management 47% of the time" [73] and another study "reported that the use of an on-line information retrieval system improved the quality of clinicians' answers to clinical questions by 21%" [86].

2.1.5 Obstacles in Medical Question Answering

Doctors have obstacles in question answering which have nothing to do with the technology or resources available. These physician-related obstacles are not the problems we should be trying to solve. They include:

- the failure to recognize information needs,
- the decision to pursue questions only when answers are thought to exist,
- the preference for the most convenient rather than the most appropriate resource,
- and the formulation of questions in a way that is difficult to answer with general resources. [27]

The best use of our time is to focus on the development of a system which overcomes the resource-related obstacles:

- the excessive time and effort required to find answers in existing resources,
- the difficulty navigating the overwhelming body of literature to find the information needed,
- the lack of access to information resources,

- search technology that is unable to directly answer clinical questions,
- and the lack of evidence that addresses questions arising in practice. [27]

However, the side-effect of faster, easier to use systems, which provide precise answers to specific questions will be the redefinition of expectations. In this way, better technology will also help overcome physicians-related obstacles.

If a fast and reliable system was prevalent, a lower level of uncertainty may cause doctors to initiate a search. Furthermore, a routine "better safe then sorry" search before prescription, diagnosis or treatment may become commonplace. Such searches would reveal new findings and updated recommendations, exposing information needs which would have gone otherwise undetected. This is a circumstance where the mentioned physician-related obstacles are essentially solved. Due to obstacles created by the sheer volume of clinical information this circumstance necessitates better information systems than are presently available.

2.1.6 How Physicians Search

In a 2008 study [32] of American emergency rooms where a follow-up visit is unlikely and need is immediate, **Table 2.1** was developed. You can see in this study doctors favor IR systems 54% of the time, then print sources 28.5% and finally colleagues 14% of the time. The preference toward IR systems makes itself clear.

2.2 Evidence Based Medicine

> *"Evidence-based medicine (EBM) is a widely accepted paradigm for medical practice that involves the explicit use of current best evidence, that is, high-quality patient-centered clinical research such as reports from randomized controlled trials, in making decisions about patient care."* [23]

Within the field of EBM, the problem of question formulation "is the first and arguably the most important step in the EBM process. Without a well-focused question, it can be very difficult and time consuming to identify appropriate resources and search for relevant evidence" [73]. To solve this problem 'well-built question' methods, such as the PICO model have been suggested and are taught in EBM curriculum.

Source	Frequency	%
PDA-based drug information: Epocrates/Tarrascon	22	17.5
Micromedex	14	11
Pocket Pharmacopeia (print version)	11	8.5
Google	11	8.5
UpToDate.com	11	8.5
Consulted specialist	10	8
Miscellaneous texts	9	7
Consulted ED colleague	7	6
Tintinalli *et al*, 2003.	7	6
PubMed	5	5
Red Book:	4	3
Harrison's On-line	3	2
PDA-other (personal notes, 5-Minute Consult, PEPID)	3	2
eMedicine.com	3	2
Lange, EM On Call	2	1.5
Willis Eye Manual	2	1.5
Sanford Guide	2	1.5
Total	**126**	**98.5**

Table 2.1: Information sources Emergency Room doctors use to successfully satisfy information needs. [32]

For this problem of question formulation we could look at generic question templates, structured queries (implementing the PICO Model), and interactive query iteration may help in formulating questions.

2.2.1 PICO: The Well-Built EBM Question

The mnemonic PICO, stands for **P**atient/ Population/ Problem, **I**ntervention/ Exposure, **C**omparison and **O**utcome. This mnemonic is meant to be used by clinicians to aid in the formulation of an evidence-based question. This method, first suggested in 1995 [72], now pervasive, has generated a number of variants including PICOTT [73], PECODR [20] and PESICO [74]. As a companion to the PICO method questions were also divided into 6 types: **Clinical evidence**, concerning interpretation and gathering of evidence; **Diagnosis**, concerning selection and interpretation of diagnostic tests; **Prognosis**, concerning predicting complications and mapping a patient's progress; **Therapy**, concerning treatments; **Prevention**, reduce risk; and **Education**, how

to teach patients, families and oneself [72]. Many descriptions of the question types have simplified these six, to four types of questions: **Diagnosis**, **Therapy**, **Prognosis** and **Etiology**. In the literature, these four are regularly referred to as the 'Clinical Tasks.'

IR specialists are aware of the importance of query formulation, the users predisposition to 2–3 word queries, issues of lexical, syntactic and semantic ambiguity and the guess work commonly needed to predict: what the user means, is looking for, and their task. This PICO system implicitly gives keywords context and question type indicates search task. An IR expert can see the value of this and the improvement over standard 2–3 keyword queries.

PubMed has a clinical queries search mode [64] which qualifies the search with the selection of radio button to indicate the type of clinical query as seen in **Figure 2.1** [64]. This search tool, specialized for clinical queries, is based directly on the research of the Hedge Filter Group from McGill University [89]. A problem persists with the PICO method, not all questions can fit the PICO frame. Some drawbacks include inability to capture temporal information and anatomical qualifications [37]. It has also been noted that this model favors questions pertaining to treatment and interventions and is less conducive to well-built prognosis and etiology question formulation [37].

2.2.2 Strength of Evidence

One of the foundations of Evidence-Based Medicine is strong evidence, thus many models of evidence categorization have been created: SORT (Strength Of Recommendation Taxonomy) [26, 7], Oxford Centre Levels of Evidence [71], and GRADE [6, 90]. All systems are similar in that controlled randomized trials are the most highly rated form of evidence, followed by cohort studies in the middle and expert opinion ranking lowest.

2.2.3 EBM Informatics Infrastructure

If we look at the building blocks of an EBM informatics infrastructure: [8] 1) standardized terminologies and structures, 2) digital sources of evidence, 3) standards that facilitate health care data exchange among heterogeneous systems, 4) informatics processes that support the acquisition and application of evidence to a specific

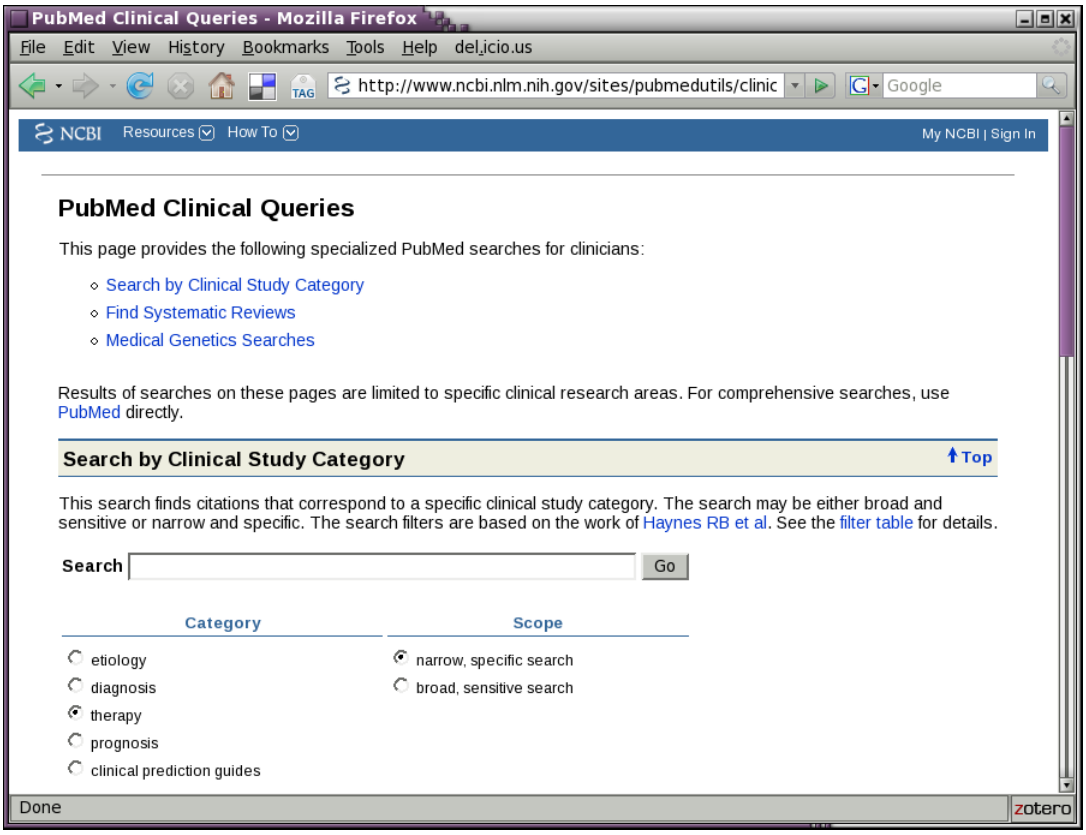


Figure 2.1: The PubMed Clinical Queries search UI.

clinical situation, and 5) informatics competencies; we can see our project fits in the fourth category and is a key element of the overall EBM informatics task.

2.3 Knowledge Sources

2.3.1 Ambiguity in the medical domain

Ambiguity is a central obstacle in all levels of language processing and information retrieval, the medical domain is no exception. In fact, ambiguity is arguably a bigger problem then in the general domain. Here are five key examples of medical domain specific ambiguity which are particularly problematic: *tokenization*, *acronyms*, *polysemy*, *synonymy* and *metonymy*.

Tokenization Identifying sentence boundaries is a problem, periods are used for sentences, abbreviations, decimals, and hierarchical delimiters, and it is not

uncommon to have sentences that begin with lowercase letters [39].

Acronym/ abbreviation With so many lengthy chemical compounds, anatomical terms and pathogen taxonomies it is easy to see the motivation to make regular use of shortened forms. (E.g. "PDA" can be "patent ducus arteriosus", "prosterior descending artery", "phorbol 12, 13 diacetate", "Parenteral Drug Association" [39], not to mention general uses like "personal digital assistant".) New acronyms are being introduced to the domain at an alarming rate of one in every five to ten abstracts [12]. Further complicating the problem, more than 8% of acronyms are ambiguous [39] and there are, on average, more than 15 possible interpretations for a given acronym [39].

Polysemy A single name can refer to more then one gene from a single species and from different organisms. (E.g. The Entrez Gene database contains more than 800 distinct genes that have been called P60) [39].

Synonymy The problem of many words having the same meaning may be particularly acute in this domain where for example, many trademark names refer to the same compound (e.g. ibuprofen is sold as Advil, Bufren, Motrin, Nuprin and Nurofen) [81]. This and other factors create the situation where six or seven synonyms for a single concept is common [81], resulting in a deeply problematic semantic ambiguity where "the probability of two experts using the same term to refer to the same concept is less than 20 per cent" [81].

Metonymy The use of a word for a concept or object which is associated with the concept/object originally denoted by the word. For example, in the phrase: *"The White House phoned..."*, the use of the word *White House* to mean *President* [42] is an example of the word "White House" used as a metonym. A string like *p53* could refer to the gene of that name, to the protein that it codes for, or to its mRNA [39].

2.3.2 NLM: PubMed, MEDLINE and ClinicalTrials.gov

Index Medicus, created in 1879, was a comprehensive index of medical journal articles which evolved into the US National Library of Medicine (NLM). This index was

supplanted by the PubMed (also a NLM project) and ceased publication in 2004. MEDLINE is the biggest database of medical journal abstracts indexed and searched by PubMed [39].

Citations in MEDLINE are collected from 5,455 (as of July 2010) [69] medical journals and it has 18,182,098 (as of July 2010) [68] total records from 1966 to the present with articles added to MEDLINE at the average rate of over 2300/day [68]. Each of these articles have been manually indexed by one of 100 human indexers with MeSH terminology, 712,675 were indexed in 2009 [63]. The MEDLINE database is one of the resources searched by PubMed, both are maintained by the NLM. Since PubMed searches MEDLINE and other resources, it is a little larger: it has 19,960,914 (as of July 2010) [68] total records from 1948 to the present. Other sources it searches are, for example: (1) the 438,252 [68] articles not yet indexed with MeSH terminology, but in the process of being processed (i.e. indexed with MeSH) into the MEDLINE system, and (2) the 471,316 [68] records from OLDMEDLINE which contains records from the years 1948 to 1965. 75% of the articles published in the last 25 years have abstracts in MEDLINE. It is free to search MEDLINE and PubMed, and they were directly searched 1.3 billion times in 2009 [63], an increase of 65% over the 776 million searches in 2008.

On April 11, 2003, in promotion of open access to scientific literature, a group drafted a statement known as THE BETHESDA STATEMENT, this was followed by THE BERLIN DECLARATION ON OPEN ACCESS TO KNOWLEDGE IN THE SCIENCES AND HUMANITIES, pushing for open access in reaction to rising subscription fees and decreasing library budgets [39]. In 2004, the NLM created PubMedCentral (PMC) [11], an on-line digital library of open-access journal articles, containing some or all the articles from about 154 journals and individual article submissions from many others [39]. Since 2005, all NIH funded researchers (in part or in full) were requested to submit manuscripts to PubMedCentral, adding 430,000 manuscripts (5TB compressed) to PMC [39]. In late 2007 that changed from voluntary submission, to a legally binding one, with the CONSOLIDATED APPROPRIATIONS ACT OF 2007 (H.R. 2764) [88]. As of 2007, 18% of recent and 12% [60] overall PubMed articles are available as full-text through open-access sites such as PubMedCentral, BioMedCentral [10] and the Public Library of Science [57].

ClinicalTrials.gov [58], maintained by the NLM, currently contains 61,557 trials in their database from 157 countries and receives over 40 million page views per month [61]. It is by far the largest repository of controlled randomized trials and observational studies [61]. This is a major directory of primary sources for anyone interested in biomedical research and in Evidence-Based Medicine.

2.3.3 The Cochrane Collaboration

> *"The Cochrane Collaboration is an international not-for-profit and independent organization, dedicated to making up-to-date, accurate information about the effects of health care readily available worldwide. It produces and disseminates systematic reviews of health care interventions and promotes the search for evidence in the form of clinical trials and other studies of interventions. The Cochrane Collaboration was founded in 1993 and named after the British epidemiologist, Archie Cochrane."* [15]

This collaboration, though originating in the UK, has branches in every continent for a total of 21 branches in 19 countries [14] (including Canada and the US). They produce a major EBM resource known as The Cochrane Database of Systematic Reviews. This collection is one of the sources (along with DARE, CENTRAL and others [41]) available as part of The Cochrane Library. Decisions regarding changes to its reviews are evaluated by committees of volunteers known as Cochrane Review Groups [83], which are made up of mostly medical professionals. Strictly organized and constantly updated, these reviews provide status flags which act as visual indicators of any content changes in the library. In **Figure 2.2**, you can see 3 (New Search, Conclusions Changed and Review) of the 9 flags (Review, Protocol, Methodology, New, New Search, Conclusions Changed, Major Change, Withdrawn, and Comment) used. Though freely available in Canada, UK and much of Europe, limited public access in the United States has prevented its universal adoption [33].

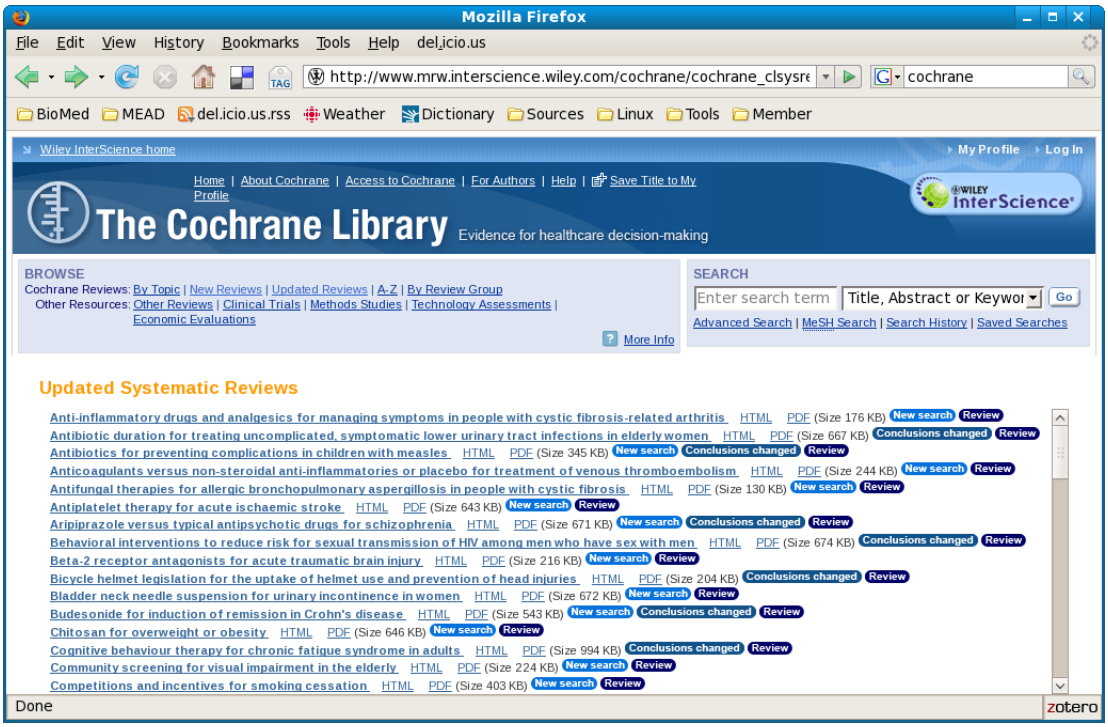


Figure 2.2: A screenshot taken from The Cochrane Library of flagged search results.

2.3.4 Up-to-date, DynaMed, Google and Wikipedia

UptoDate.com

UptoDate.com [46] is a commercial, Internet-based service which provides medical information directed at primary care medical practitioners. As the name indicates, the published monographs from this source are regularly updated by its 3,800 authors, editors, or clinical experts all of whom are listed on the website. The service is available off-line for $1500 (with quarterly updates for a year), or on-line and on PDA for $500/ year. UptoDate.com practices many aspects of EBM including structured queries (like PICO) and uses the GRADE [90] system to indicate Strength of Recommendation.

In a 2008 observational study of 424 hospitals increased usage of UpToDate.com (measured in hits per week) were "significantly associated with a shortened severity-adjusted length of stay and lower risk-adjusted patient safety adverse outcome rates" [9]. This study also showed that the 424 hospitals with UpToDate compared against

the 3091 hospitals without UpToDate "were associated with significantly lower risk-adjusted complication rates and patient safety adverse outcome rates" [9]. This second point loses its potency when you notice that this is an observational study, that is to say, confounding factors must be considered. For example, there is the possibility that hospitals with UpToDate.com subscriptions were better hospitals on the whole and therefore provided better care and lower complication rates and a generally shorter length of stay.

DynaMed

DynaMed is a similar regularly updated subscription-based service available on-line and on PDAs (Palm, PocketPC, Windows Smartphone, BlackBerry, and iPhone). This site EBM based service uses the Strength of Recommendation Taxonomy (SORT) [26] to delineate the strength of evidence. The source material for the reviews within DynaMed are searched using PubMed Clinical Queries [64], the Cochrane Database of Systematic Reviews, and the National Guideline Clearinghouse is the source for medical guidelines. A complete list of primary and secondary sources is available on the DynaMed website [25].

Google and Wikipedia

Though patients may get nervous of the idea of doctors googling their symptoms on the Internet, there is mention of its use in the literature. Google scholar is used by doctors [16], sometimes preferred [91], and there is some evidence that it does provide decent results [91].

Wikipedia's quality is steadily increasing as is its reputation. Though still frowned upon in a court of law [76] and the medical office, some improvements and recent developments must be noted. The combination of concretely referenced articles which hyperlink to reputable sources, the integration of clinical taxonomies such as MeSH and the familiarity of its interface make it a useful starting point for many clinical queries and in some cases adequate for general information needs. Recent studies [16, 34] find that while only 10% of doctors edit or contribute to Wikipedia's content [34], nearly 50% use it for clinical queries [16] and that it is nearly error-free on the topic of drugs [34].

In both of these cases, the use of these web services are inevitable due to the off-duty habits of clinicians and the reality that these resources are both easy to use and familiar, two qualities that are potent and desperately needed elements of a clinical information service. These popular services set the expectations, and for better or worse any other service must contend with them as competitors.

2.4 Knowledge-Based Sub-systems

2.4.1 Ontologies

A multitude of definitions and theory surround the concept of an ontology. Here, we will define what it is and what it is not in the context of this paper, based on the research in the subject. An ontology primarily serves as a tool to solve the problem of semantic ambiguity.

If I were to say "speaking in the language of a statistician... the result was *statistically significant*" I intend the listener to have a sense of the word 'significant' according to the domain of mathematical statistics. That is, according to the distribution and the experimental design we have a result which could lead to causation. A very specific meaning in which "statistics study" is the context for interpretation. Where the common use of "...the result was significant" would not carry those specific mathematical connotations. The ontology is the conceptualization of that domain knowledge, a tool to specify a context to frame meaning.

The ontology is a framework of communication. For two agents to agree on a subset of meanings (or senses) of a body of terminology is to agree to 'commit' to a specific ontology. So, an ontology needs only to define the terms of communication; two agents that 'commit' to an ontology are agreeing on a shared vocabulary. The deeper needs of answering arbitrary queries and solving problems is the concern of the knowledge base [35]. The ontology makes as few constraints about the world it is modeling as possible to maintain a consistent terminology and maximize the freedom of the ontology committal agents to instantiate as needed [35].

2.4.2 WordNet

The creators of WordNet [84] would not consider it an ontology, but rather, an on-line lexical database, a dictionary designed for a computer to read. Where human dictionaries define words with a list of descriptive sentences, WordNet is more like a thesaurus, defining words in relation to other words which share its meaning — more specifically — share the same sense. Each of these 'word-senses' is a collection of synonyms called 'synsets' and are meant to represent one distinct concept. Presently, WordNet contains 155,287 words, 117,659 synsets and 206,941 word-sense pairs [85].

Defining words in a way computers can use to interpret human communication means addressing the ambiguous ways humans use language: polysemy, the same word form may belong to more than one set; synonymy, different word-forms belong to the same synset; hyper/hyponymy, noun synsets must be organized hierarchically to represent ISA relations; meronymy, synsets which are conceptually related components of each other must indicate HASA relations; to name a few. Thus, the demarcation of these synsets and the definition of their relationships was not a trivial task. A task that was performed painstakingly by George Miller and his team of linguists at Princeton from 1985–1995, and is on-going, with the most recent version released in 2006. With this difficult groundwork laid, WordNet (free to download and use) has become a central resource for computational linguists, so much so that 434 papers have been published on WordNet [18] and the conference dedicated to its study and use is now in its 5^{th} year [4].

2.4.3 UMLS

> *"The objective of this program... is to solve what is the most fundamental barrier to the application of computers in medicine; namely, the lack of a standard language in medicine. We will attempt to build that vocabulary, a language that will cross between the biomedical literature and the observations on the patient, as well as the educational applications in the school, a language which allows those areas to be interrelated."*
>
> –Donald A. B. Lindberg, M.D., March 19, 1985 [38]

The UMLS is not strictly a formal ontology as described in the first sub-section

of this section. The UMLS is more similar to WordNet, that is, organized like a very precise thesaurus with several distinct frameworks of hierarchical and semantic relations added to its structure [92].

Before I begin describing the details of the UMLS, I would like to make some clarifications. The UMLS is a project, an acronym, which stands for (U)nified (M)edical (L)anguage (S)ystem. Under the umbrella of this project are several components. First, there are 3 knowledge sources the UMLS Metathesaurus; UMLS Semantic Network and the UMLS SPECIALIST Lexicon and Lexical tools [62]. Second, there is the UMLS Knowledge Server. Thirdly, the MetaMap program and finally the RRF Browser. In general, when people refer to the UMLS they are referring to the UMLS Metathesaurus, UMLS Semantic Network and UMLS SPECIALIST Lexicon in combination.

The UMLS Metathesaurus attempts to integrate all of the disparate and specialized medical terminologies, categorizations and thesauri into one unified super-set hence the name 'Metathesaurus'. It includes more than 100 source vocabularies from the entire domain of medicine, including such varied sources as:

- Diagnostic and Statistical Manual of Mental Disorders (DSM-IV),
- HCPCS Version of Current Dental Terminology,
- WHO Adverse Drug Reaction Terminology (WHOART),
- Standard Product Nomenclature (USFDA).

A complete list is available from [67].

UMLS Metathesaurus

The Metathesaurus attempts to tackle the problem of synonymy — different lexical forms (words) with the same meaning — by linking synonymous words to distinct, unique (and numbered) concepts it has defined. This way all synonymous concepts from all the source materials can be equated, allowing a framework for the exchange of knowledge between these vocabularies. The 2009 release of the UMLS Metathesaurus contains information on 2,181,676 concepts [66], has over 9,840,386 million concept

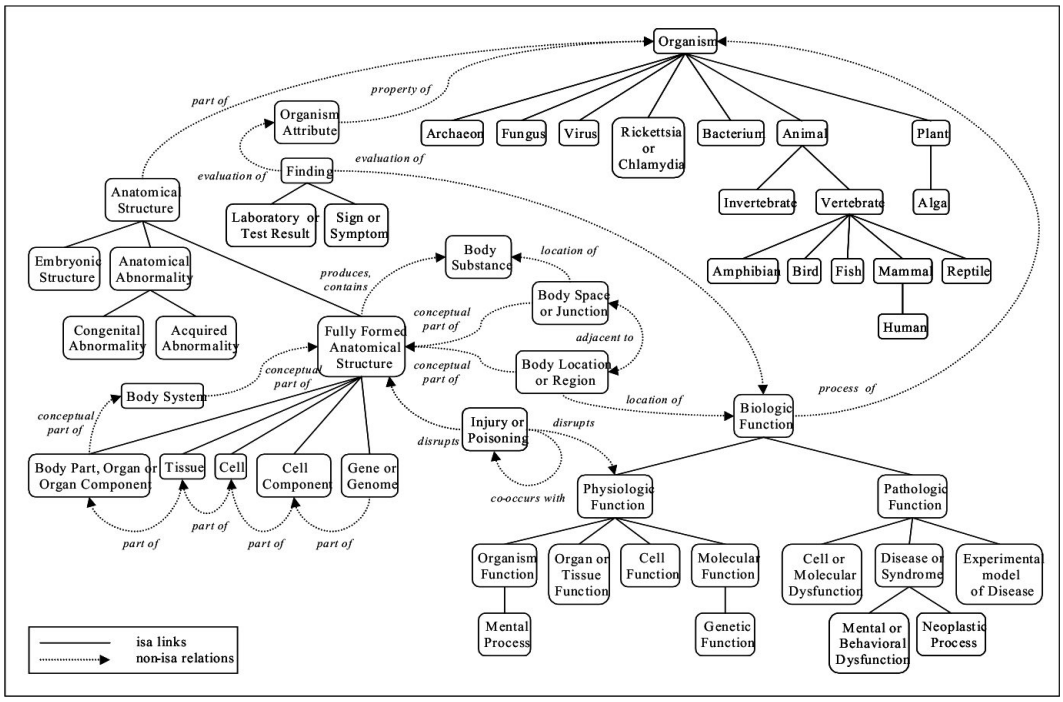


Figure 2.3: UMLS Semantic Network **associative** and **hierarchical** relationships. [65]

names [66] from 129 controlled vocabularies, and is available (at least in part) in 19 different languages [66].

UMLS Semantic Network

The Semantic Network provides categorization for all the concepts represented in the UMLS Metathesaurus and adds a hierarchical semantic structure to the Metathesaurus through a set of semantic types and relations between these types [48]. This is done in the attempt to tackle the problem of hyponymy, for example, Ibuprofen is a subclass of Anti-Inflammatory, and both are is a subclasses of Drug. All terms from every source vocabulary is linked to at least one concept in the Metathesaurus. All concepts in the Metathesaurus are linked to at least one of the 135 semantic types in the current Semantic Network. These semantic types are related to each other by at least one of the 54 relationships currently in use by the Semantic Network [65].

The primary relationship in the Semantic Network is the ISA relationship (**Figure**

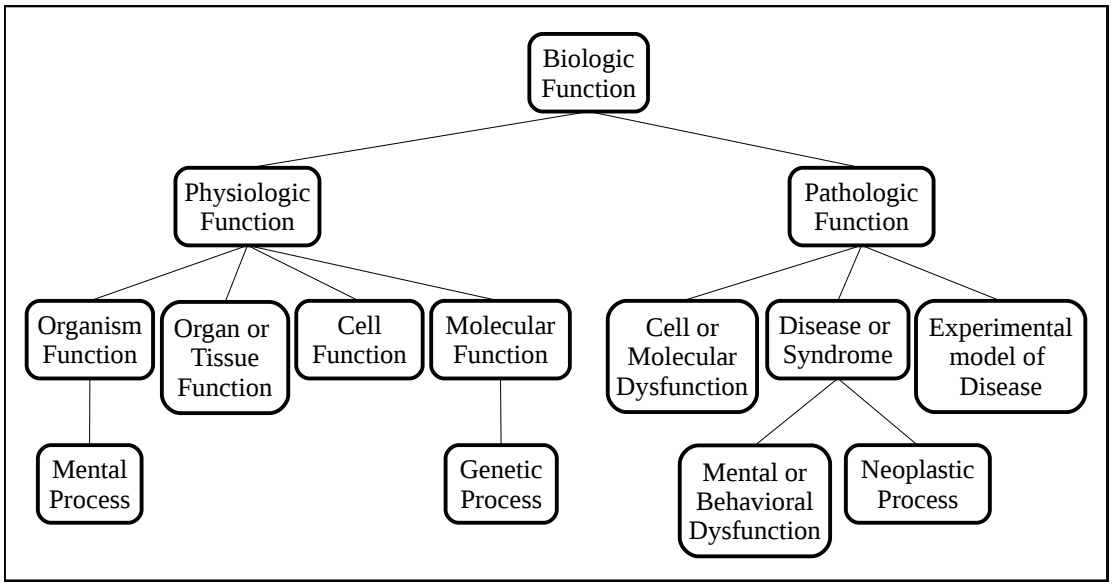


Figure 2.4: This sample from the UMLS Semantic Network shows the Biologic Function hierarchy which illustrates ISA relationships.[65]

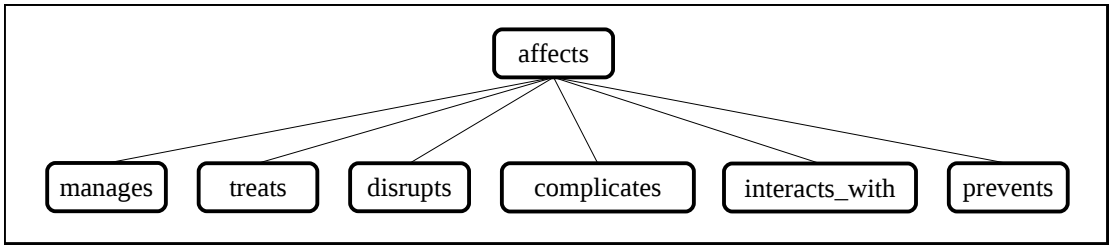


Figure 2.5: UMLS Semantic Network Affects hierarchy. An example hierarchy for associative relationships; the relationships used in the Semantic Network are they themselves hierarchically related with ISA relationships [65]

2.4), this is used to create the hierarchy of concepts necessary to solve semantic issues arising from hyponymy. In addition, five major categories of associative relationships are defined which are themselves relations: PHYSICALLY RELATED, SPATIALLY RELATED, TEMPORALLY RELATED, FUNCTIONALLY RELATED, and CONCEPTUALLY RELATED [65]. **Figure 2.5** shows an example hierarchy of an associative relationship and **Figure 2.3** shows examples of associative relationships and hierarchical relationships in a single graph representation.

UMLS SPECIALIST Lexicon and NLP Tools

The need was recognized for a bridge between the UMLS Metathesaurus and free text applications. These components of the UMLS were developed to foster development of — and for use in — natural language processing and information retrieval systems. The UMLS SPECIALIST lexicon "is a syntactic lexicon of biomedical and general English words, providing orthographic, morphological and syntactic information," [48] has 297K records (over 482K inflectional forms) [2]. The 20,000 word lexicon was generated from a variety of sources including MEDLINE articles, the UMLS, medical dictionaries and general-use dictionaries [55]. There are 6 tools in the 2008 UMLS SPECIALIST NLP Toolkit, they are open source, freely available and each is developed specifically for a standard NLP task.

Tokenization Wordind — Wordind is a tokenizer and word index generator.

Normalization Norm — Normalizes strings and words into the a form preferred by the UMLS Metathesaurus that is ignoring alphabetic case, inflection, spelling variants, punctuation, genitive markers, stop words, diacritics, symbols, ligatures, and word order [1].

Part-of-speech tagging dTagger — a Part of Speech (POS) tagger specifically built for use in the medical domain. It includes a trained model, one trained on a set of annotated MEDLINE abstracts from MedPost corpus (genomics) [1].

Spell Checking GSpell — a spell checker, but it treats a space as a letter allowing the correction of errors in word compounding [1].

LexAccess2008/2009 To allow easy access to the UMLS SPECIALIST Lexicon, LexAccess2008 is provided. It is written in Java and provides Java APIs for use as a component in other applications or can be used as an end-user tool. **Table 2.2** and continued in **Table 2.3** shows example output from this tool.

In this example you can see sensitivity to spelling variants such as 'CrT', 'Crt' and 'cRT' demonstrating potential pitfalls due to the ambiguous nature of the domain. Not only are there 14 possible acronyms of 'CRT' could be referring to, but 1 abbreviation (an adjective) and 3 spelling variants. This demonstrates not only how

```
$> CRT
{base=CRT
entry=E0420176
cat=noun
variants=uncount
variants=groupuncount
variants=plur
variants=metareg
acronym_of=Certified Record Techniques
acronym_of=cardiac resuscitation team|E0420190
acronym_of=cathode-ray tube|E0420189
acronym_of=choice reaction time|E0420188
acronym_of=chromium release test|E0420187
acronym_of=complex reaction time|E0420186
acronym_of=computerized renal tomography|E0420185
acronym_of=copper reduction test|E0420184
acronym_of=corrected retention time|E0420183
acronym_of=cortisone resistant thymocyte|E0420182
acronym_of=cranial radiation therapy|E0420181
acronym_of=capillary refilling time|E0420180
acronym_of=chemoradiation therapy|E0420179
acronym_of=conformal radiation therapy|E0420178
}
```

Table 2.2: Shown here is sample output from LexAccess2008.

simple it is to find ambiguity in this domain, but also points to the use of this tool as a possible piece of the solution.

MetamorphoSys

The third component, MetamorphoSys, 'the UMLS installation wizard and Metathesaurus customization tool, installs one or more of the UMLS Knowledge Sources and enables us to create customized Metathesaurus subsets' [3]. Part of the MetamorphoSys package is the RRF browser, which allows us to browse your customized installation of the Metathesaurus, as shown in **Figure 2.6**.

MetaMap

The last tool from the NLM to be discussed is MetaMap. MetaMap, also known as MMTx, was developed to map biomedical text to the Metathesaurus. MetaMap is

```
 {base=CRT
entry=E0420177
cat=adj
variants=inv
position=attrib(3)
position=pred
stative
abbreviation_of=certified|E0220630
abbreviation_of=corrected
}
{base=Crt
spelling_variant=CRT
entry=E0420191
cat=noun
variants=uncount
acronym_of=calreticulin|E0304049
}
{base=cRT
entry=E0420192
cat=noun
variants=uncount
acronym_of=competitive reverse transcriptase|E0420193
}
{base=CrT
entry=E0420194
cat=noun
variants=metareg
acronym_of=crista terminalis input site|E0420195
}
```

Table 2.3: Shown here is the continued output from LexAccess2008.

also used to semi-automatically relate MeSH terminology to MEDLINE papers [1]. It is semi-automatic in that human indexers approve MetaMap's choices, by selecting the specific MeSH terms on which both MetaMap and the indexers agree and removing the others [3]. Two of the Q&A systems discussed later, CQA-1.0 and Essie, make use of MetaMap.

As a word of warning, installation the UMLS Knowledge sources takes about 2-12 hours and requires about 20GB to 45GB of storage (our installation took 5 hours and totals 42.5GB). MetaMap must be installed separately.

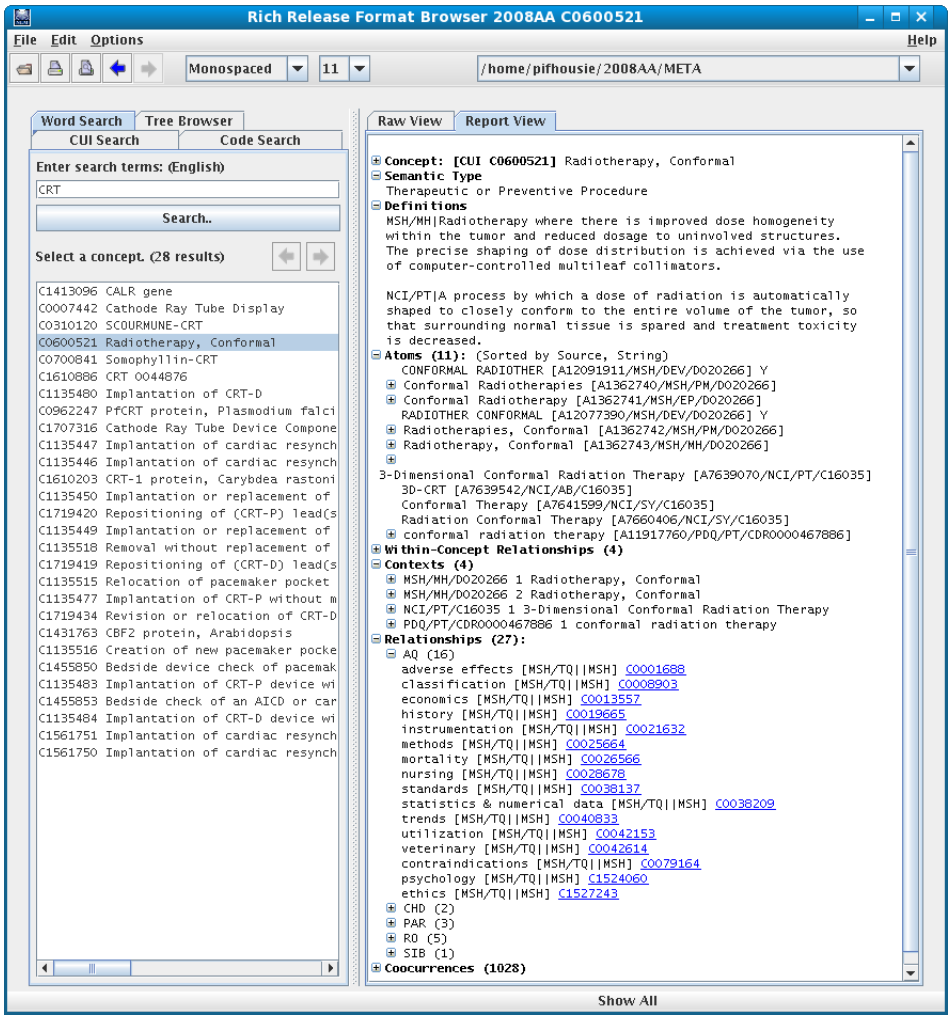


Figure 2.6: UMLS Metathesaurus search results for `CRT` as shown in the RRF browser. On the left side we see the search term and results. 'Conformal Radiotherapy' is selected. In the REPORT VIEW on the right hand side we see: the unique concept ID (CUI); the semantic type (taken from the Semantic Network); a short definition; variants; contexts (showing in which taxonomies the term is represented); and relationships shows connections to other concepts.

2.4.4 MeSH

MeSH, which stands for Medical Subject Headings, was developed and is maintained by the National Library of Medicine (NLM), an agency within the National Institute of Health (NIH). First published in 1960 [53], the NLM staff regularly updates this vocabulary, now releasing a new edition of MeSH each year. The 2010 version of MeSH contains 25,588 descriptors [53].

MeSH is a taxonomy, thus descriptors are arranged alphabetically and hierarchically. At the root there are 16 broadly defined main categories, which are further divided into alphabetically-ordered sub-categories. As we follow a path from the root category, down through the 11-level hierarchy, the concepts change from very general, near the root, to very specific, close to the leaves.

Unlike many other controlled vocabularies, MeSH was explicitly designed by medical librarians to organize medical document collections.

The top level of the MeSH Hierarchy is shown in **Table 2.4** and the children of Anatomy are shown in **Table 2.5**.

1. Anatomy [A]
2. Organisms [B]
3. Diseases [C]
4. Chemicals and Drugs [D]
5. Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. Psychiatry and Psychology [F]
7. Biological Sciences [G]
8. Natural Sciences [H]
9. Anthropology, Education, Sociology and Social Phenomena [I]
10. Technology, Industry, Agriculture [J]
11. Humanities [K]
12. Information Science [L]
13. Named Groups [M]
14. Health Care [N]
15. Publication Characteristics [V]
16. Geographicals [Z]

Table 2.4: This figure shows the top level concepts of the MeSH hierarchy.

Besides these descriptors there are 172,000 entry terms, synonymous with descriptors, to assist entry to the MeSH system, for example Heart Attack is an entry term for Myocardial Infarction. As well as being the key indexing and categorization paradigm for the NLM, this MeSH terminology is one of the source vocabularies in the UMLS Metathesaurus.

As point of interest, the MeSH terminology has begun to infiltrate the main stream. If you were to type in a disease name into Wikipedia most times a call-out box in the upper right-hand corner displays links to the MeSH terminology (see **Figure 2.7**). Clicking on word 'MeSH' will take you to the Wikipedia entry for

Anatomy [A]
- Body Regions [A01]
- Musculoskeletal System [A02]
- Digestive System [A03]
- Respiratory System [A04]
- Urogenital System [A05]
- Endocrine System [A06]
- Cardiovascular System [A07]
- Nervous System [A08]
- Sense Organs [A09]
- Tissues [A10]
- Cells [A11]
- Fluids and Secretions [A12]
- Animal Structures [A13]
- Stomatognathic System [A14]
- Hemic and Immune Systems [A15]
- Embryonic Structures [A16]
- Integumentary System [A17]

Table 2.5: This figure shows the children of the Anatomy concept in the MeSH hierarchy. For further exploration visit the on-line interactive MeSH Browser, http://www.nlm.nih.gov/mesh/MBrowser.html

'Medical Subject Headings' and clicking on the number beside it will take you to the entry in the MeSH browser on the NLM website.

2.5 Current Solutions to Health Information Needs

2.5.1 Informationist

> *"We believe it's time to face up to the fact that physicians can't, and shouldn't, try to do all or even most medical information retrieval themselves. ...Better they should focus their scarce discretionary professional time on reading, discussing, and reflecting in ways that truly deepen their conceptual and practical understanding of medicine than on the mechanics of finding, extracting, and synthesizing information from the published literature."* [19]

The idea behind this solution is to create a position akin to a medical librarian. A person whose primary responsibility is to answer doctors clinical questions, present during rounds, available after out-patient visits, to be seen as an important member

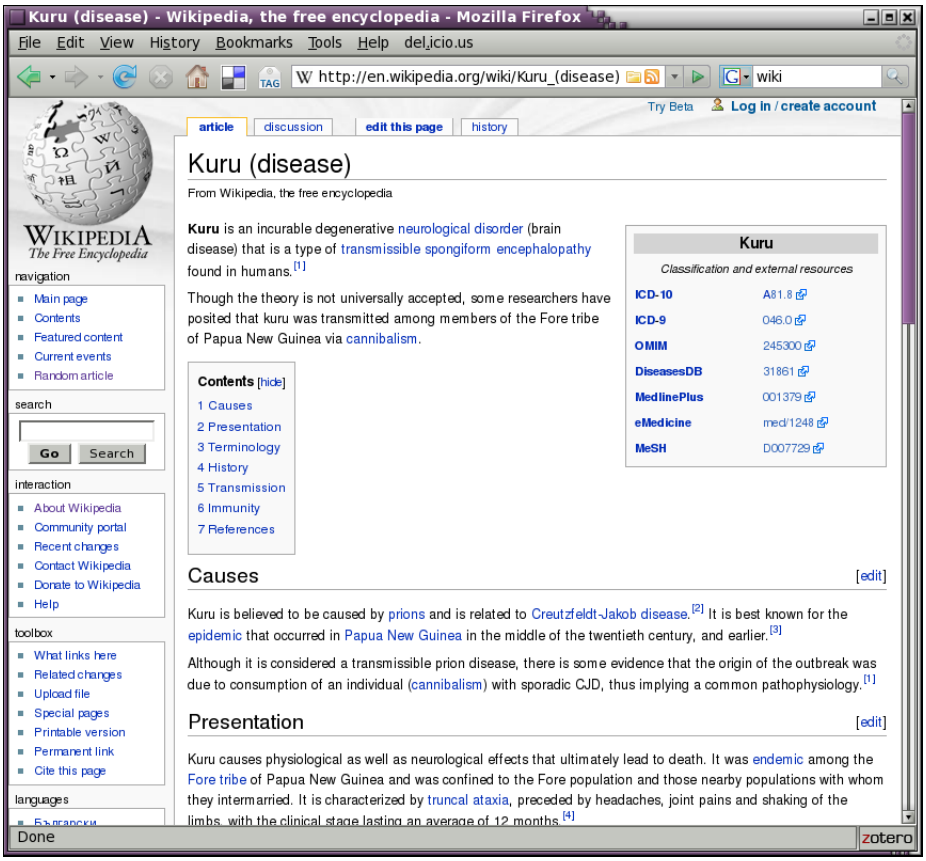


Figure 2.7: This figure shows Wikipedia's search results for the query: `Kuru disease`. You can see on the right side to box containing links to several resources including MedlinePlus and MeSH.

of the medical team. A person trained in equal parts clinical work and information science [19]. It has been shown that they often help clinicians formulate their questions [19], which is one of the major obstacles in clinical question answering [27, 29].

Unfortunately, on-site medical librarians and 'Informationists' are uncommon outside academic centers [19]. As an alternative, off-site clinical question answer services have be suggested, established and studied.

A system in the UK called ATTRACT, started in 1997, would deliver (via Fax) an Evidence-Based medicine summary created by an information manager within 6 hours. This service was rated 'useful' by 31% and 'very useful' by 69% of the 40 doctors participating in the study. Over half said the summaries changed their practice [17]. The average cost per question was $27.30. In an similar study in Australia, questions were answered for a fee of $27.50 per question and questions

were answered within 1 to 12 days [17]. Following the study all 9 doctors said they were willing to pay for the service and 50% said they would use it at least twice per month. The time to respond to questions was seen as an important factor is the perceived usefulness, by the participating doctors [17].

These services go a long way to help solve obstacles in answering clinical questions, such as "lack of time", "difficulty formulating questions", "selection of resources", and "difficulty finding optimal search strategies" [17, 27]. If this service is seen as an alternative (in some cases) to blood tests or scans, one could quickly alleviate a portion of the workload on expensive equipment and services with a price tag of $27.50 per question which is well below the price of even the most inexpensive test.

2.5.2 Essie

> ...O, be some other name!
> What's in a name? that which we call <concept-name='a rose'>
> By any other name would smell as [valid];
>
> –William Shakespeare, Romeo and Juliet, 1594.

Essie (formerly referred to as SE) [40] is a concept-based search engine developed in 2000 at the NLM for its site ClinicalTrials.gov [49, 58]. Essie's ranking algorithm can be best described as "all the right pieces in all the right places" [40]. Since the search engine is phrase based, (as opposed to single word tokenized), 'the right pieces' are these phrases from the query. Since the engine heuristically ranks locations in the structure of the document differently, such as the title which is ranked high and footnotes which are ranked low, 'the right places' are where in the document these pieces are found [40].

Essie uses the UMLS to identify concepts and as the basis for synonymous phrase-based query expansion. Once a phrase token is identified the UMLS is used to identify the concept it references, Essie then 'expands' the query by adding other synonymous phrases, phrases which also reference the same concept in the UMLS, thus searching for matches of all synonymous phrases, and thus searching for the concept 'by any other name' not just the queried phrase.

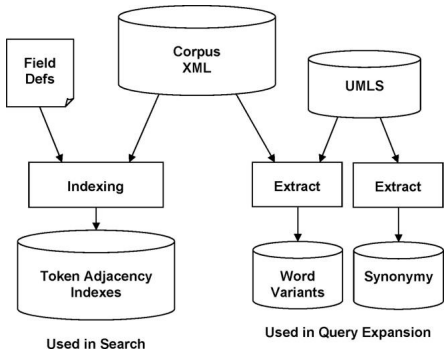


Figure 2.8: The Essie index architecture. [40]

This concept based query expansion necessitates unusual documents scoring, not on the usual how many occurrences (frequency) but on where in the document (location) a concept is placed. This is due to the fact that phrase proliferation generated by query expansion equates many very different phrases [40].

Essie competed in the 2003 and 2006 TREC Genomic track. In 2003 it was the best performing search-engine and in 2006 it "achieved results comparable to those of the highest-ranking systems" [40].

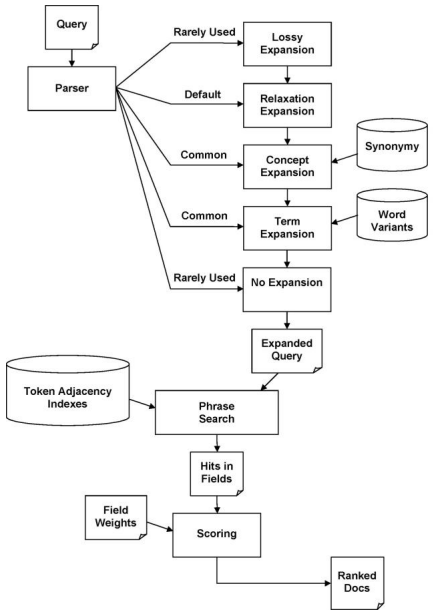


Figure 2.9: The Essie query architecture. [40]

One advantage of the concept-based system is the utilization of the 97,000 (at the

time, now 172,000) entry terms from MeSH (as part of the UMLS). For example the common term for 'ascorbic acid' is 'vitamin C', Essie would relate the query term `vitamin C` to the concept Ascorbic Acid (via the UMLS) and search both, performing innate translation of many common usage words into clinical terminology.

Essie has two main phases. First it indexes the search corpus, by tokenizing and recording the position of every token occurrence in the corpus. Position information is important for ranking results, and determining token adjacency, which is key information for phrase matching (i.e. words in a phrase are adjacent). This indexing results in a look-up table shown in **Figure 2.8**.

During this phase two other tables are generated, the synonymy data-set and the variant data-set. The former using concept expansion, the latter using term expansion. Concept expansion uses the UMLS in the manner described above and term expansion, uses the UMLS SPECIALIST Lexicon to include term variants such as plurals, possessives, hyphenation, compound words and alternate spellings but not non-noun inflectional variants.

The second phase is the search phase. In this phase the query is entered, parsed, broken into fragments (called relaxation expansion in **Figure 2.8**) and then expanded through concept expansion then term expansion. All of the phrases generated are searched, scored and returned according to rank, as shown in **Figure 2.9**.

There are a few limitations to this system. A relatively static corpus is a must to allow time required for the extensive indexing [40]. Query expansion has its dangers as "the explosive nature of the expansions makes the implementation vulnerable to failure when given a very long query" [40]. Finally, it is important to note the hardware requirements for this system are heavy: 64GB RAM and an index ten times the size of the document set [40]. These are very heavy requirements when compared to the MedQA system — discussed next — which runs on a personal computer.

2.5.3 MedQA

Since 40% of medical questions are *What is...* questions [28], the MedQA question answering system focused on these definitional questions. It uses both the World Web Web via Google and MEDLINE (indexed with Lucene) as resources for answering questions. MedQA is available for use on-line at: http://askhermes.org/MedQA/.

The MedQA team plan on adding other question types as they development of the system.

In MEDLINE, normally articles that report original research use a document structure known by the mnemonic: IMRAD (Introduction, Methods, Results, and Discussion) [91]. The authors took advantage of this structure in two distinct ways. First, to determine relevancy of an given article to the query, the 'Results' section was the focus since a recent user study had shown that physicians prefer this section when determining the relevancy [91]. Second, in identifying definitional sentences they found these sentences were more likely to be found in the Introduction and Background sections [91]. The authors made a training set of sentences classified by section, then used machine learning techniques (naïve Bayes) to classify unknown sentences from MEDLINE into the classes Introduction, Background, Methods, Results, Conclusion and Other with 78.6% accuracy.

To initiate a search, the user first types in a definitional question. Next, MedQA identifies noun phrases, and forms a query with only these terms. The query is then used to retrieve relevant documents which are tokenized into sentences and clustered. To generate an answer, centroid-based summarization is applied twice. First, to remove redundancies, MedQA selects one sentence based on TF*IDF weighted cosine similarity to be the most representative of its cluster. Then again to the collection of selected representative sentences to generate a final coherent summary. The user receives a result separated in two sections; Web and MEDLINE, (see Figure 2.10).

Web search has its pitfalls. On-line definitions can often be irrelevant to the medical domain. "For example, 'heart' was defined as both '. . . one of the most successful female fronted bands in the annals of hard rock' and 'a hollow, muscular organ that pumps blood through the blood vessels by repeated, rhythmic contractions;' " [91] To deal with this problem on-line medical dictionaries are also queried, the TF*IDF scores are then compared, if this similarity measure fall below a given threshold the web result is discarded.

MedQA was evaluated by four physicians in comparison to three other on-line systems Google, One-Look and PubMed. We assume reader's familiarity with Google, and PubMed was described in detail previously. OneLook, however, requires a brief description. OneLook is a federated search engine which has indexed over 900 other

Figure 2.10: MedQA search results for the query: `What is Kuru?`. In the SUMMARY section, each extracted sentence is followed by a link to the source and each source in the SUMMARY FROM MEDLINE subsection is hyperlinked. The second OTHER RELEVANT SENTENCES section provides highly ranked non-definitional extractions all of which are liked to primary sources through MEDLINE.

dictionaries [70]. A search on OneLook returns relevant results from any dictionary it has indexed. Results appear in the form of a list of hyperlinks to the source site, broken into categories such as General, Art, Science and most importantly Medicine.

Evaluation results indicated PubMed and OneLook were bettered in most evaluation criteria, quality of answer, ease of use, time taken, and actions taken, by Google and MedQA. Qualitative test scores, gathered by questionnaire showed Google was the preferred system in terms of ease of use and quality of answer overall. Quantitatively, Google provided ranked results in less then a second and MedQA generated its summary in an average of 16 seconds. However, since information is spread among

sites, the evaluation of Google results (identifying definitions) was more time consuming. MedQA was the highest rated system in terms of time spent and number of actions.

An interesting and encouraging point is one of hardware requirements, MedQA was written in Perl and runs on a Macintosh PowerPC with dual 2 GHz CPUs and 2GB of memory [91].

2.5.4 CQA-1.0

This prototype Q&A system was developed by the NLM around the fundamentals of EBM: PICO built questions, strength of evidence and clinical task type. This system views the clinical answering task "...as 'semantic unification' between information needs expressed in a PICO-based frame and corresponding structures automatically extracted from MEDLINE citations" [23]. The idea is that these three fundamentals of EBM taken together create the perfect structure for codifying the knowledge needed to answer clinical questions [23]. These facets are:

(1) the four main clinical tasks:

- **Therapy**: Selecting treatments to offer a patient, taking into account effectiveness, risk, cost, and other relevant factors (includes Preventionselecting actions to reduce the chance of a disease by identifying and modifying risk factors).
- **Diagnosis**: This encompasses two primary types: Differential diagnosis: Identifying and ranking by likelihood potential diseases based on findings observed in a patient. Diagnostic test: Selecting and interpreting diagnostic tests for a patient, considering their precision, accuracy, acceptability, cost, and safety.
- **Etiology/Harm**: Identifying factors that cause a disease or condition in a patient.
- **Prognosis**: Estimating a patients likely course over time and anticipating likely complications. [23]

(2) a well-built clinical question (PICO):

- What is the primary *problem* or disease? What are the characteristics of the patient (e.g., age, gender, or co-existing conditions)?
- What is the main *intervention* (e.g., a diagnostic test, medication, or therapeutic procedure)?
- What is the main intervention *compared to* (e.g., no intervention, another drug, another therapeutic procedure, or a placebo)?
- What is the *desired effect* of the intervention (e.g., cure a disease, relieve or eliminate symptoms, reduce side effects, or lower cost)? [23]

(3) strength of evidence (SORT):

1. A-level evidence is based on consistent, good-quality patient outcome-oriented evidence presented in systematic reviews, randomized controlled clinical trials, cohort studies, and meta-analyses.
2. B-level evidence is inconsistent, limited-quality, patient-oriented evidence in the same types of studies.
3. C-level evidence is based on disease-oriented evidence or studies less rigorous than randomized controlled clinical trials, cohort studies, systematic reviews, and meta-analyses. [23]

Term Re-Ranker

This term-based re-ranking algorithm, relies on matching terms in a natural language query phrase to sentences identified as outcome sentences. It is important to note this method makes little use of the EBM facets described above and is most useful in combination with the EBM re-ranker, described in the following section and the only method for use of natural language question as input.

Outcomes are full sentences unlike the other elements. Each sentence in the abstract is given a score to determine the likelihood that it is an outcome sentence, then the system returns all those rated above a certain threshold. The score is a

combination of elements captured in the following formula:

$$\begin{aligned} S\_outcome = \lambda_1 S\_cues \\ + \lambda_2 S\_unigram \\ + \lambda_3 S\_ngram \\ + \lambda_4 S\_position \\ + \lambda_5 S\_length \\ + \lambda_6 S\_semanticType \end{aligned}$$

Table 2.6 is a brief description of the components of the above formula.

S_cues	uses cue phrases heuristically developed by the team.
S_unigram	uses a 'bag-of-words' classifier.
S_n-gram	developed on corpus of positive outcome predictors using odds ratio.
S_position	closer to the end of the abstract is better.
S_length	a probability based on the length of the abstract that it contains an outcome statement.
S_semantic type	contains UMLS concepts related to outcome statements.

Table 2.6: The explanation of the components of the scoring formula. [23]

EBM Re-Ranker

The authors do not believe that free-form natural language queries are well-suited to question-answering systems. Instead, their system structures the query according to the familiar PICO framework, a standard of the EBM curriculum. The benefit is the physician — instead of the system — translates their information need into a frame-based representation [23], a problematic interpretation for a computer system. This interface also "...force[s] physicians to 'think through' their questions" [23] which leads to better 'thought-out' queries.

Here taken from [23] are some example questions and their PICO query frames:

1. Does quinine reduce leg cramps for young athletes? (Therapy)

`search task`: therapy selection

`primary problem`: leg cramps

`co-occurring problems`: muscle cramps, cramps
`population`: young adult
`intervention`: quinine

2. How often is coughing the presenting complaint in patients with gastroesophageal reflux disease? (Diagnosis)
`search task`: differential diagnosis
`primary problem`: gastroesophageal reflux disease
`co-occurring problems`: cough

3. What's the prognosis of lupoid sclerosis? (Prognosis)
`search task`: patient outcome prediction
`primary problem`: lupus erythematosus
`co-occurring problems`: multiple sclerosis

4. What are the causes of hypomagnesemia? (Etiology)
`search task`: cause determination
`primary problem`: hypomagnesemia

The EBM re-ranking scheme uses the follow formulas:

$$S\_EBM = S\_PICO + S\_StrengthOfEvidence + S\_Task \tag{2.1}$$

$$S\_PICO = S\_problem + S\_population + S\_intervention + S\_outcome \tag{2.2}$$

$$S\_StrengthOfEvidence = S\_journal + S\_study + S\_date \tag{2.3}$$

$$S\_Task = S\_PositiveIndicators - S\_NegativeIndicators \tag{2.4}$$

Sample Output from PICO Extractors

The PICO extractors parse the abstract, tagging phrases and sentences as Problem, Population, Intervention, or Outcome. It was noted that outcomes are usually complete sentences and tagged as such, while interventions, population, and problems are noun phrases [23]. In this sample output from the PICO extractors, the *italic text* is the extracted text and the Sans Serif Text immediately following is the tag. This sample is in response to the question "*In children with an acute febrile illness, what is the efficacy of single-medication therapy with acetaminophen or ibuprofen in reducing fever?*" [23]:

Title: Antipyretic efficacy of ibuprofen vs acetaminophen

Author: Kauffman RE, Sawyer LA, Scheinbaum ML

Journal: Am J Dis Child. 1992 May;146(5):622-5

Abstract: OBJECTIVE– To compare the antipyretic efficacy of ibuprofen, placebo, and acetaminophen. DESIGN– Double-dummy, double-blind, randomized, placebo-controlled trial. SETTING– Emergency department and inpatient units of a large, metropolitan, university-based, childrens hospital in Michigan. PARTICIPANTS– *37 otherwise healthy children aged 2 to 12 years* Population with *acute, intercurrent, febrile illness* Problem. INTERVENTIONS– Each child was randomly assigned to receive a single dose of *acetaminophen* Intervention (10 mg/kg), *ibuprofen* Intervention (10 mg/kg) (7.5 or 10 mg/kg), or *placebo* Intervention (10 mg/kg). MEASUREMENTS/MAIN RESULTS– Oral temperature was measured before dosing, 30 minutes after dosing, and hourly thereafter for 8 hours after the dose. Patients were monitored for adverse effects during the study and 24 hours after administration of the assigned drug. *All three active treatments produced significant antipyresis compared with placebo* Outcome. *Ibuprofen provided greater temperature decrement and longer duration of antipyresis than acetaminophen when the two drugs were administered in approximately equal doses* Outcome. No adverse effects were observed in any treatment group. CONCLUSION– *Ibuprofen is a potent antipyretic agent and is a safe alternative for the selected febrile child who may benefit from antipyretic medication but who either cannot take or does not achieve satisfactory antipyresis with acetaminophen* Outcome.

Publication Type: Clinical Trial, Randomized Controlled Trial

PMID: 1621668
Strength of Evidence: grade A [23]

This system uses MetaMap to map noun phrases into the concepts in the UMLS. This mapping was applied to each of the elements of the PICO frame. Population and problem were separated due to there conceptual differences and the fact that often they are not presented together in abstracts. Intervention and comparison were merged as they are conceptually similar and difficult for the system to distinguish. Though the system looks exclusively at abstracts, structured abstracts are common, though varied in naming structure. This system takes advantage this structure when present.

This extractor has two functions: (1) for use in the EBM re-ranking algorithm; and (2) for output for the user in the Information extraction UI. In the UI, the user could be shown any of the extracted passages they prefer, the default is to only return the outcome sentences (with the title, and bibliographic information) as studies show they are key sentences in abstracts for determining relevancy [23]. For the PICO component (**Equation 2.2**) of the EBM re-ranker (**Equation 2.1**), each tagged section was matched against the PICO frame query, a full match scored a 1.0, a partial match 0.5.

Strength of evidence evaluation is based on three components (as shown in **Equation 2.3**). First, the most recent articles are given greater weight. Second, highly trusted sources are given greater weight. The third weight depends on the where the type of study sits in the SORT taxonomy. The type of determined by is determined by the meta-data associated with the article. **Table 2.7** shows the MeSH tags which accompany and to what evidence level each tag is associated.

Evidence	MeSH/Publication type
Level A	Meta-analysis, randomized controlled trials, cohort study, follow-up study
Level B	Case-control study, case series
Level C	Case report, in vitro, animal and animal testing, alternatives studies

Table 2.7: The MeSH/Publication indicators for each of the three evidence levels (shown on the left). [23]

To determine the clinical task for use in **Equation 2.4**, MeSH terms (and their

children) are categorized into indicators of the four task types, shown in Table 2.8.

Task	Positive Indicators	Negative Indicators
Therapy	MeSH Terms: Clinical Trials, Random Allocation and Therapeutic Use	
Diagnosis	MeSH Terms: Diagnosis	Positive Therapy Indicators
Prognosis	MeSH Terms: Survival Analysis, Disease-Free Survival, Treatment Outcome, Health Status, Prevalence, Risk Factors, Disability Evaluation, Quality Of Life, and Recovery Of Function.	
Etiology	MeSH Terms: Population At Risk, Risk Factors, Etiology, Causality, and Physiopathology.	Positive Therapy Indicators

Table 2.8: The MeSH term indicators for the four task types (shown on the left). [23]

If any of these terms are marked as a `Major Topic` (indicated by a '*' next to the MeSH term in PubMed and by `<DescriptorName MajorTopicYN=Y>` in the MEDLINE Citation XML) that terms weight is increased.

Results

The baseline for comparison are expertly generated Boolean PubMed queries. Each of these queries took an average of 40 minutes for the first author (a Medical Librarian and Medical Doctor) to generate. These go far beyond the ability of your average PubMed user, but definitely demonstrate the system against the most expert of PubMed users. For example, the question, *What is the best treatment for analgesic rebound headaches?* resulted in the PubMed query:

```
((('analgesics'[TIAB] NOT Medline[SB]) OR 'analgesics'[MeSH Terms]
OR 'analgesics'[Pharmacological Action] OR analgesic[Text Word])
AND (('headache'[TIAB] NOT Medline[SB]) OR 'headache'[MeSH Terms]
OR headaches[Text Word]) AND ('adverse effects'[Subheading] OR
side effects[Text Word])) AND hasabstract[text] AND English[Lang]
AND 'humans'[MeSH Terms] [23]
```

The relevancy was evaluated based on following criteria:

- **Precision at ten retrieved documents** (P10) the precision of the top ten results.

- **Mean Reciprocal Rank** (MRR) is a measure of the rank of the first relevant result.

- **Total Document Reciprocal Rank** (TDRR) is the sum of the MRR of the relevant documents.

For the purposes of testing, any results that are helpful or contain the answer count as a successful result. **Table 2.9** shows a summary of testing with this system.

	P10	MRR	TDRR
PubMed (baseline)	0.281	0.526	1.353
Term	0.481 (+29%)*	0.513 (+44%)	1.945 (+44%)*
EBM	0.677 (+141%)*	0.936 (+78%)*	2.671 (+98%)*
Combo	0.688 (+145%)*	0.962 (+83%)*	2.680 (+98%)*

Table 2.9: A summary of one of the tests performed with CQA-1.0. The Term re-ranker and EBM re-ranker are described above, the Combo ranks based on equal weight given to Term and EBM re-rankers then combined. All results with '*' following the score are results which are statistically significant at the 1% level.

Limitations

This system performs very well and does accomplish the task it sets out to do, but it is very slow, prohibitively slow. I contacted the authors and asked for on-line access to the system, which they granted. On my few informal tests, the response time was between 18–40 minutes. Also the interface was extremely basic, clearly not developed for any user who was not involved in its development, though at no point did they state this system was ready for general use.

2.5.5 Semantic Clustering

Semantic clustering is an attempt to improve the way we return results. Instead of returning a ranked list of results, semantic clustering returns the results grouped by concept. To indicate the content of each cluster, UMLS concepts are used to label clusters, this was so the user gets an overview of the results inside.

Semantically-related results are organized into clusters. Cluster names are presented to the user a cluster can be selected and the contents of the cluster are displayed. Inside the cluster, each article is displayed as a short extractive summary of three parts: the title, the main intervention, and the top-scoring outcome sentence [22]. The extracted outcome sentence is the automatically identified using CQA-1.0 [23]. The outcome sentence serves as an entry point into the article, which the reader can use to judge relevance. The clusters are ordered by size (number of articles), the articles inside each cluster are sorted chronologically (newest first).

The idea is to 'drill-down' into the information, to view your results at deepening levels of granularity, unlike Google which presents pages of results which hyperlink out of the Google interface. This would, through series of turn-down switches, show more information about an article at each deeper level. For example, "Top-level answers to 'What is the best drug treatment for X?' consist of categories of drugs that may be of interest to the physician. Each category is associated with a cluster of abstracts from MEDLINE about that particular treatment option. Drilling down into a cluster, the physician is presented with extractive summaries of abstracts that outline the clinical findings. To obtain more detail, the physician can pull up the complete abstract text, and finally the electronic version of the entire article (if available)" [22].

This clustering method was compared to lexical clustering, which is clustering based solely on keywords. Not only did it not improve the PubMed baseline, but the cluster names were incoherent and therefore unhelpful in organizing results [44, 22].

There are 4 advantages we would like to mention. First is redundancy management; that is, redundant information is gathered together, since all interventions in a particular cluster are conceptually related [22]. Secondly the concepts labels give an information overview. In this form of presentation the clustered results provide a feel for information landscape, this is something difficult to get a sense of from browsing a ranked list. Third, obviously irrelevant articles are bundled together and can be categorically ignored, saving the precious resources of time, patience and 'screen real-estate'. And finally, it provides an opportunity for easy semantic-based relevance feedback. Clusters can be selected and deselected to indicate preference and focus iterative searches.

Chapter 3

MedicInfoSys – An Architecture for Medical Information Delivery

3.1 Research Problem

Neither software nor users operate in a vacuum. A mistake often made by software developers is to create software, without strictly defining the parameters within which that software was designed to operate. For the Informationist to be effective, they must clearly understand the overall and specific needs of the End-User, as well as the capabilities implemented into the system layer. A description of the overall framework in which each class of user (End-User and Informationist) and software component operates, must be defined. Most importantly, it is essential to the success of this system that we illuminate each user to the context in which each role is meant to operate, if we are to redefine expectations, roles and habits in medical question answering.

In this chapter we present the overall solution, the vision. Here we present an architecture to improve delivery of information to those working in the medical domain. **Figure 3.1** is a diagram of this architecture and **Appendix B** is a larger annotated version of this diagram. The following sections describe its aspects: Section 3.2 describes the End-User Layer, Section 3.3 details the Informationist Layer and Section 3.4 describes the System Layer and each component in a sub-section thereof.

3.2 End-User Layer

The highly valued time of medical professionals is limited and overburdened. This coupled with their varying degrees of comfort, competency and fluency with information technology points to the need for delegation to a specialist. In this layer, the end user's view of the system is represented. Once the end user identifies an information

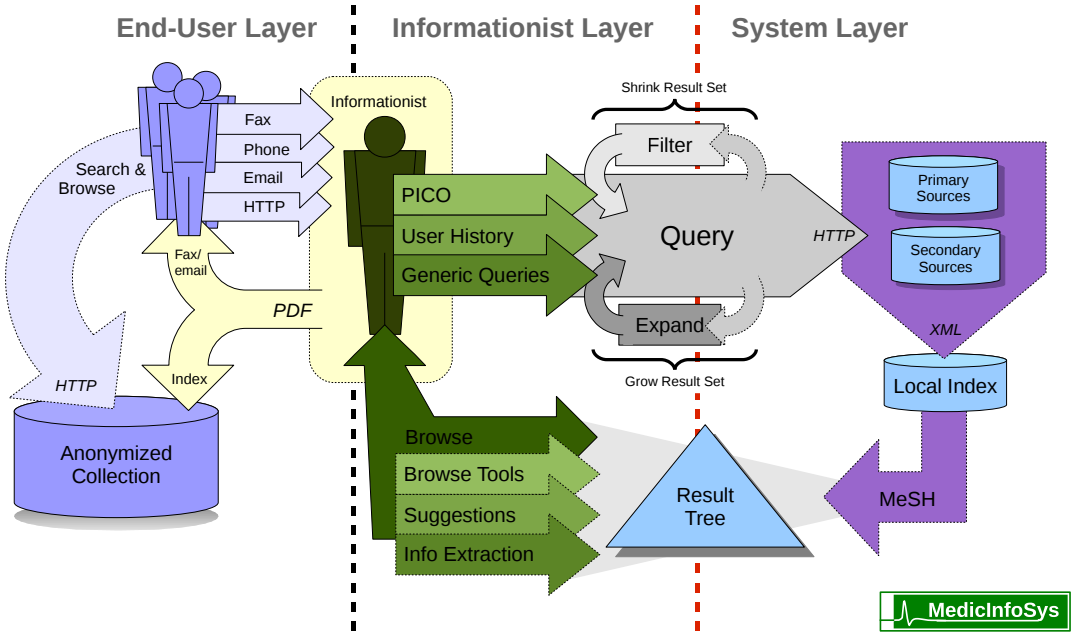


Figure 3.1: A proposed architecture for this MedicInfoSys (pronounced: 'medicine-fo-sys') medical information system is divided into 3 layers: 1) the **End-user layer**; 2) the **Informationist layer**; and 3) the **System layer**. The dotted lines represent these boundaries and indicate the interface of each user to the rest of the system to the right.

need, they have two avenues to find answers: search the collection of previous adequately answered medical queries for their query-answer pair, or communicate their information need to an Informationist. In some cases the request will be a one-way communication of a well-defined or generic sort, other times the user will benefit from interaction with the Informationist to help the user clarify their need and focus it into a precise query.

3.2.1 End Users

The intended core users of this system would be those directly involved with patient care (such as general practitioners, researchers, surgeons, specialists, nurses, psychiatrists, rehabilitaters, therapists). However, the information needs of health administrators, social scientists, health advisers and public servants may be well served by this system. The primary sources in collections such as PubMed, contain the newest

research and statistics on all health related fields including economics, ethics, social trends, prevention, law, and technology; which are valuable sources of information for making informed budgetary decisions, in support of litigation, grant applications and strengthening policy positions. [5]

3.2.2 PDF Report

The keys to this system are brevity, accessibility and transparency (of its evidence-based sources). The time constraints placed on medical professionals are severe, therefore customized reports must be timely and brief. In addition, with the stakes as high as they are in the medical field, all reports must be aggregate in nature, gathered from trusted Evidence-Based Medicine (EBM) sources, and explicitly referenced from sources accessible to the end-user preferable immediately via the Internet. We use a similar experimental phone/fax system in the UK as an example for MedicInfoSys, they reported: [17] a mean time to answer clinical questions of 45 minutes, a maximum length of two pages (one-sided), a maximum turn-around time of 8 hours and an average cost of $27.50 per answer. We expect the same approximate timings, costs and length of report with our proposed system.

The PDF is a format of choice for medical articles, compact and secure. Many freely available readers exist for any platform, it prevents manipulation, capable of embedding high-resolution images for medical diagrams, hyperlinks to web-resources, and has security layers to password protect printing and viewing, which is a priority if handling private patient data.

This PDF is faxed/emailed to the user (users preference) and sent to searchable collection for later reference. If the user approves the content, the report is anonymized, indexed and made available to all authorized users of the system.

3.2.3 Interface with Informationist Layer

Bad assumptions and ambiguity create extra-work and wasted effort in every workplace in any domain where one person gives a task to another. In a domain as sophisticated and time-sensitive as this is, it is imperative to make an extra effort to ensure that all parties are 'on the same page'. The structured query is a way to mitigate time wasted because of the delegation of search tasks; to promote a mutual

understanding and limit time-wasting irrelevant results.

We need to structure the query to draw out the users' information need. The user usually does not instantly know the best words or how to phrase an uncertainty. A structured model like PICO helps the user through the uncertain process of developing a query. Along the way a structured query contextualizes the components (i.e. query keywords) for the system implicitly, thus reducing the ambiguity inherent in common unqualified keyword searches. The qualifications (such as Population) help the system/informationist to narrow query expansion, prioritize the 'where' a word of phrase was found according to rhetorical structure (e.g. introduction, methods...) for the purposes of ranking, summarization and information extraction. Other qualifications clarify what task (such as diagnosis) the doctor is engaged in, which will help the system direct the search, better rank what is relevant, and best frame the answer. To put it a different way, help them (end-users) give us (Informationists/system) the pieces we need to solve their question by helping them formulate the query through a process we define, and define that process with terminology they are familiar with and we can compartmentalize.

Finally, further consultation, by phone or email, maybe necessary for some clarification. Both parties should have contact information available should the need arise.

3.3 Informationist Layer

This layer concerns the use of the system from the Informationist's point of view. This position as defined in the literature [19] historically goes by many names: medical librarian, medical researcher or medical knowledge worker. The qualified user is one with a medical background strong enough to research most detailed technical medical questions that health professionals have to offer and to provide answers or direct the end-user to the source of the answer to their information need. This service could be provided by a single government agency, a single private company, a selection of regional or specialist providers, or by accredited public, private and individual 'freelance' service providers. Remuneration could be based on an hourly fee, on a per query basis, a monthly retainer, or yearly contracts. Quality Assurance could be guaranteed by the particular end-user using the service in combination with a

regulatory body. Since Canada's public health care system would pay for this service the resultant research can be collected free of copyright restrictions, indexed and made available as web-based public medical information resource.

3.3.1 Input

The ways the Informationist has to input queries into the system have been categorized into three avenues: **PICO**, **User Profile** and **Generic Questions**.

PICO

In the PICO query input category the user enters the appropriate information into specialized fields (Population, Intervention, Comparison, Outcome). This information is used to formulate and query the system. PICO is the most common EBM 'well-built question' method, but several others exist including PICOTT and PEDCOR. The user may select the method which is most appropriate for their information need.

User Profile

All queries in the User Profile input category depend on some user specific information. For example, user search history is needed to revisit past query result sets; and private information access is necessary to pursue automatic query generation based on specific electronic patient records. A set of articles selected from the results of previous queries can be used to automatically generate queries-by-example. A user profile is needed to collect, store and retrieve this set.

Generic Queries

Certain questions are common enough that effective heuristics have been developed to answer them specifically. This input category helps users make use of these customized search methods. The user selects from a short list of generic questions, fills in the appropriate fields, then the system employs the optimized search method developed for that generic query type. Examples of these generic questions include: *What is the cause of symptom X?*; *What is the recommended dosage of X?*; and *How should*

I manage disease or finding X?. These three generic questions taken together make up the majority of all generic clinical queries [28].

Filter

This filtering module is meant to narrow results sets and increase precision. The three input categories above influence the filtering of the query, but the user may customize these filters arbitrarily. These filters include: the Clinical Query Filters developed at McGill University, specifying the Strength of Evidence by various models, the use of standard Boolean Operators, and a set of filters like those used in PubMed's Advanced Search to limit publication type, specify database, journal, author, etc. Also, an automatic disambiguation function will detect ambiguous word candidates and interact with the user to disambiguate candidate terms (e.g. *Did you mean... X or Y?*)

Expand

This Query Expansion module is meant to broaden results sets and increase recall. Like the filter module this module is only partly in the domain of the system layer, since the user has influence and can make customizations. This module uses the MeSH Entry Terms and the UMLS as its primary means of query expansion. The UMLS Metathesaurus can be used for the identification medical concepts in the query, allowing for the addition of synonyms for those terms into the query and for queries based on concept ids for databases equipped to accept them. The UMLS Semantic Network can be utilized to add related concepts appropriate to the query. Abbreviations, metonymy, hypo/hypernymy can also be addressed and/or exploited by this module, increasing recall and affecting better re-ranking.

3.3.2 Output

The output of the system is displayed in a browsable hierarchical tree structure based-on the MeSH controlled vocabulary. Results are organized into categories based on the conceptual attributes that best represent them, and ordered within each category by their degree of lexical similarity to the query. This structure retains its state during and after browsing, providing a predictably behaving structure for interaction, easing

a user's re-visitation of previous query result sets and maintaining progress from complete or interrupted browsing sessions. Nodes, branches and sub-trees may be deleted at the users discretion, for the purpose of trimming dead ends and focusing the result set. These result sets may be saved for future reference and used as the basis for a query-by-example type automation.

Browse Tools

This category represents tools meant to aid the user in navigating the result set and include: keyword highlighting, FIND IN MESH and secondary search. Keyword highlighting is fairly straight-forward, the user enters terms they would like to be highlighted in the results to draw attention to them as they scan the results. FIND IN MESH is a tool that locates MeSH concepts by name in the result tree hierarchy and centers the view over them for the users convenience. Secondary search is meant to focus a given result set by searching within that result set using a keyword search.

Suggestions

These navigation aids are meant to use the information available on the user and the query to direct the user to likely starting points (MeSH categories which match query terms) for browsing, to rank results within each category in the tree and to recommend articles based on the results so-far selected, based on the user profile and based on information task.

Information Extraction

As needed and appropriate, automatic extractive summarization will be utilized. Some queries may be sufficiently specific, adequately detailed or heuristically predisposed to precise answers or customized extractions. For example, well-structured clinical queries and certain structural features of many medical articles can enable outcome extraction which is very useful for determination of relevance for many clinical users. Also identification of salient sentences is likely for some query types and all properly defined queries will produce candidates sentences for extraction and those ranked highly enough to surpass a heuristically based threshold would be presented to the user using this tool.

3.3.3 Update PDF Files

Quotes sources could be checked against Cochrane Library by any user viewing the PDF after the creation date for updates and corrections. If an update to an existing PDF is requested from the user layer, the references in the existing PDF would be used to search for new papers which use them as references, as a starting point for the Informationist.

3.4 System Layer

3.4.1 Sources

Since all sources recommended are available on-line they will be queried via HTTP. Many sources make their results available via XML, but further research is needed to investigate XML availability for each on-line system. The results from all sources will be collected and indexed into one local database for every individual query.

Primary Sources

Primary Sources are the unfiltered root sources of medical information, medical journals and Clinical trials. The primary source in this category in MEDLINE. PubMed, MEDLINE and all NLM collections can be searched via HTTP with a publicly available tool developed by the NLM known as *efetch*. Results can be returned as XML formatted data. NLM sources are freely available and well-respected.

Secondary Sources

Secondary sources are collected, edited aggregates of the primary sources. These include clinical guidelines, Cochrane reviews, UpToDate.com, Dynamed, Micromedex, Harrison's On-line and Wikipedia. All these sources listed are known to be used by physicians and have varying licensing agreement, levels of quality and utility. Some heuristics for generic queries use specific collections in this set.

3.4.2 Local Index

The local index is the collection point for all these disparate sources of on-line information. Since indexing is done during run-time and results are to be displayed to the user as categorized a Boolean indexing system is recommended. Stop-words would be removed and a Porter stemmer would be applied to the set. This local index would be used primarily for building the MeSH tree, and searched for secondary searching and keyword highlighting.

3.4.3 MeSH-Based Browse Tree

Each article is categorized into a MeSH conceptual heading. These articles are then placed in a tree where each leaf is an article and the parent nodes between the leaf and the root are the concepts names in the MeSH hierarchy. These categories are very general near the root and increase in specificity the further from the root in the hierarchy you venture. The tree starts with an upper bound of 16 main MeSH categories and is at maximum 11 levels deep. The user navigates this hierarchy in the same way as a file hierarchy where folders are concepts and files are results.

By bringing MeSH categories to the forefront we add another tool to the toolbox of the researcher. Just as users learn and memorize authors, journals and keywords which consistently return interesting results, so now they can keep key categories in mind when searching for articles.

3.5 Conclusion

Adam Smith [77] points out the importance of matching skills with tools, of creating specific tools for specific tasks, for specialized training for each worker and to efficiently connect the product of each sub-task to complete the final product. The product here is knowledge; knowledge that when interpreted by the end user satisfies a medical information need. The end-user thoughtfully interprets their need into the fields of a structured query, the Informationist uses that information from the above layer to locate and extract pieces of the answer from the available medical information, collect them in structured format and pass it back to the end user in a short 2 page form — which is transparent to its sources — allowing the end user to either

satisfy their need, or use the extracts as expert advice on where to continue their search.

The next section describes an implementation and specific method of evaluation of the Systems Layer.

Chapter 4

PifMed – A Hierarchical Information Navigation System

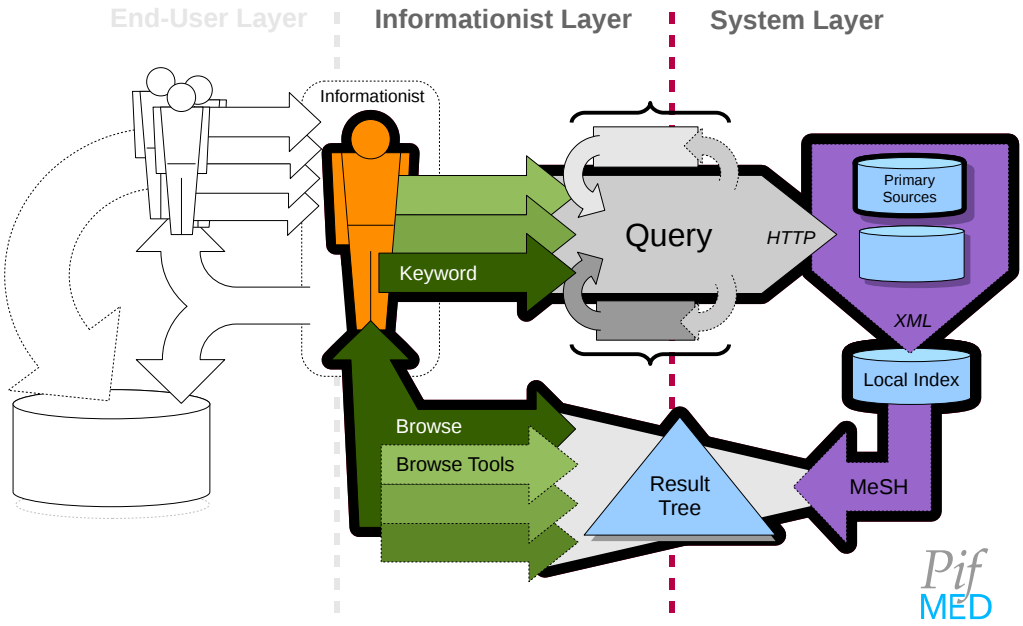


Figure 4.1: A focused view of the MedicInfoSys architecture described in the previous chapter. The thick dark outline draws the readers attention to the components implemented in this prototype of PifMed and all labels have been removed with the exception of those specific components implemented. The **End-User Layer** is ghosted-out since this implementation is between the **Informationist Layer** and **System Layer** only.

4.1 Introduction

Due to the maturity of MEDLINE and UMLS, we have an opportunity, within the medical domain, to begin exploration of ideas which may in the near future be applied

to the web in general.

With the application of knowledge-based methods we can categorize the results of search, but then also use these hierarchical categories as the basis for a browsing tool. These methods can organize the large results sets we receive in the present information climate, in a potentially more usable way for users.

In this chapter we describe the implementation and pilot testing of an IR system which presents MEDLINE results to the user, categorized by the MeSH terminology and hierarchically browsable.

This chapter describes a partial implementation of the MedicInfoSys architecture described in the previous chapter and depicted in **Figure 3.1**. **Figure 4.1** shows the specific components that have been implemented.

Section 4.2 defines and discusses the specific problem this implementation is focused to solve. Section 4.3 introduces a new model of IR User Interface, compares it to ranked lists (4.3.2), and discuss its theoretical advantages (4.3.3, 4.3.4, 4.3.5). In Section 4.4 we describe our implementation of this model, its dependencies (4.4.1), its structure and design choices (4.4.2), its features and use (4.4.3). Nearing the end of the chapter, we describe the method of evaluation (4.3.5) and the results of the pilot study (**??**) meant to test this method of evaluation. Finally we discuss its limitations (4.5) and lessons learned (4.6) to conclude the chapter.

4.2 Research Problem

With the overall growth in medical knowledge comes an increase in relevant documents for any given medical query. Given these circumstances the question is: *When queries match thousands of documents is it adequate to simply rank and list these large result sets?* That is: *Is there a more helpful method of organization of large result sets?*

4.2.1 The Problem of Ranked Lists and Large Result Sets

Conventional presentation of large results sets leave lower-ranked results generally ignored and unexplored by users. It is futile to continue to list items with the knowledge that most will never be browsed and may even be detrimental to the user's task.

Our browsing models need: to keep pace with this volume of information, to produce better ways to direct users to pertinent articles, and to construct ways to help users navigate as they browse.

A list provides few options for exploration. We can skim, periodically sampling results; or plod through every result. Since these results are listed in a linear fashion, when browsing them, we are constrained to visit them more or less linearly. But our understanding is not so constrained. We can think from general to specific. This type of thinking is reflected in a hierarchical browsing model and this model is the basis of our system.

4.3 The Categorical Exploration Solution

We propose a new browsing model: a hierarchical category tree model, in place of a ranked list model. In this model, the user is presented the results in a interactive browsable tree where the nodes are category names and the leaves are articles. The parent-child relationships between nodes are the hypo/hypernymic relationships predefined in the hierarchical category system (such as the MeSH Taxonomy). The user browses the system in the same way they would a file hierarchy in Windows, Linux or Mac OS.

4.3.1 Exploration of the Information Landscape

With a hierarchical model, uninteresting results are often categorized together, these sub-trees or leaves can be minimized, ignored or deleted. Likewise interesting results are frequently categorized within the same category node or sub-tree, enabling fruitful focused searching similar articles. In this way the user can get a feel for the information landscape — for what is 'out there' — the hierarchical structure acts as an informative guide for exploration, each node as a signpost indicating what lies down its path, rather than a long uninformative ranked list where each result is independent of the one before and the one after.

From Browsing Structure to Navigational Structure

By shifting from a ranked list to a browsable tree, we move from a **browsing structure**, to a **navigation structure**. Navigation requires a 'geography': stars, street signs, landmarks or maps; a way for persistent features to guide the user. A list algorithm creates a substantively different, wholly disposable ordering for each different query phrase, however, a hierarchy persists despite the query. Thus the hierarchical taxonomy acts as a persistent 'geography' to be remembered. Each search session shows paths through the tree which can be used to aid future travels. Here is a tool with a knowable, predictable, sensible underpinning. Only with this system can we navigate a result set, MeSH provides the map. Lists, on the other hand, are constrained to one path, browsed and skimmed in only one direction, down.

In a list each item must be examined in linear order, whereas like any tree the order in this tree is not set and thus the search path is not dictated by the structure but up to the user to decide. In fact, like any tree, each article, by comparison, is very close to 'the top of the list'; each article is less than 10 steps from the root. And since each article is in multiple categories, there is more than one path to each article. In a list any duplication of articles would be frowned on as needless redundancy, here we can see it as beneficial. Unlike a list, this non-linear structure makes no judgement on 'the best order', but allows each user to find their own order of visitation, and only with this system are different paths to the same result possible.

The user is in fact looking at all the results not just the top ten. That is, this category tree is a representation of all of the results. Every article is reachable from the beginning state, the root, in less than 12 clicks, since the MeSH hierarchy is 11 levels at its deepest. Furthermore, relevant results are likely to be found in 'bunches' that is, a very good paper will be in a category with similar papers.

This system adds a step in the information finding task, first they form a query, they then select a category of interest. Instead of looking for a paper directly after querying the system, they search for a relevant category. This searching for a relevant category is an act of focusing the query. Instead of looking at a list of articles, they are looking at a map of articles.

4.3.2 Tree Structure Versus List Structure

Instead of a list browsing structure, we use a tree browsing structure. A parallel can be seen between computer search and human search. Think of the list data structure vs the tree data structure. Each item in a list must be viewed sequentially, the order of visitation is dictated by the structure. In a tree, the order of visitation is non-linear, a choice is made at each node, starting at the root about which node (and thus sub-tree) should be visited. For tree search we have a well-defined objective ordering (alphabetical), objective comparator (greater/less than), and an objective end state (exact match). How can we use tree search for IR? We have no objective end-state or comparator, since relevancy is in the eye of the beholder, so to speak, it is entirely subjective.

If we see the human as part of the algorithm, and the users' subjective judgment as the comparator, then we can see the category tree structure as enabling tree search for information retrieval. At each node in the tree (category), the user chooses to visit the node (i.e. see its children) or chooses to move to a sibling, just as the comparator in tree search makes a choice at each node. Each leaf in the tree (article) is a possible match for the end state, but this matching is complex and subjective to each user. Since only the user can judge relevancy at each node and leaf, it is beneficial to let the user into the problem solving mechanism, into the algorithm. When the human and computer co-operate this way, strengths of each benefit the task. If the human is taken as part of the algorithm, we have an arguably objective ordering (hypo/hypernymic hierarchical categorization), a subjective comparator (more/less specifically relevant judgement of category) and an subjective end-state (satisfaction of information need).

We see the user as part of the ranking algorithm, the users subjectivity is the objective function (perhaps better named the subjective function) which judges the best order of visitation. Furthermore, since articles are placed in multiple categories, there are multiple paths to the same article, so the user can decide which path is best.

In this way we can include the user in a way not possible with ranking algorithms. Since, in the end, only the human makes the valid choices on what is relevant, this system enables the user to subjectively, dynamically and interactively set the order, and makes no assumptions about the degree to which our users' sense of relevancy matches that of an 'average user'.

4.3.3 Subjectivity as Guide Instead of Obstacle

Objectivity is still has a major role in this system, but rather than judging relevancy in relation to the query, it is judging relevancy of hypo/hypernymic relationships (known as ISA relationships) in the categorization hierarchy. Determining if two categories have a hyponymic or hypernymic relationship is much less subjective than if Order X is the best ordering of results given Query Phrase Y. Consensus is possible for the majority of categories and near consensus for all but a few. This can be supported by the popularity of WordNet and intensive work on ontologies.

There is an advantage of the tree structure that mitigates any subjective disagreement with the objective categorization hierarchy. The category system persists and is knowable, that is, the category relationships endure from query phrase to query phrase and these relationships can be learned and remembered. Though, ranking algorithms also persist their rankings can not be anticipated. That is, knowing how the ranking algorithm works will not help you locate articles in a list ranked by that algorithm. Even the creator of a ranking algorithm can not predict its specific behaviour on a given query phrase, for example, predict where on a list to find articles about Urban Health or South Africa for any given query. But since the categories persist between uses, once a user knows where a category is, they typically can quickly find it and any articles within. So if the users' subjective judgement disagrees with the placement of a category, they can use each experience to learn the objective categorization system.

4.3.4 Implicit Query Refinement

Each category over which the user browses in an opportunity to focus the formulation of the query. By blurring the lines of distinction between query formulation and browsing we can make the search process more responsive to human thought. There is an opportunity in a hierarchical categorization tree for the user to merge query refinement and article selection into the same action, browsing.

Even experienced users of computer search engines have used a query phrase which was 'the best they could do at the moment' rather than 'the best that could be done', that is, used a query that quickly came to mind, just to get started, because the right words were not coming to mind. It is these queries in particular that would benefit from this type of query refinement. An important point is that new query phrases

need not be entered, the system needs no feedback from the user to produce a new set for the user to browse. The refinement is an implicit function of the browsing process.

Another way to view this model is as one that presents the results of the query phrase conjoined with each category name. So in a sense, each result set is actually thousands of queries, deterministically created, sensibly organized and intuitively browsable. For example, by using the query `Artificial Intelligence`, the query `Artificial Intelligence AND Computing Methodologies` is automatically generated and the results of which are placed into the Computing Methodologies category in the tree, likewise, the `Artificial Intelligence AND Quality of Life` query is implicitly generated and the results are placed in the Quality of Life category, and so on for each MeSH descriptor. Thus, as many query refinements as there are MeSH descriptors — ∼24,000 — are created and hierarchically organized for browsing.

It is important to note that this system is uniquely able to provide the user with the knowledge that NO articles exist for a given category and a given query with large result sets, and quickly. The entire ranked list must be scanned sequentially to verify this fact.

Since it is a simpler task to recognize 'the right keyword' (in a category name), than to generate 'the right keyword' from thin air, the user can pick a category name similar to the category they have in mind, then look at the more specific sub-categories within to try to trigger a recognition of 'the right keyword'.

4.3.5 Relationships between Results

There is a dimension of query search results which is ignored by ranked lists. A ranked list, ranks solely on the dimension of relevancy to the query, that is, only the relationship to the query informs the ordering. There is another important dimension which PifMed bases its organization, the relationship of each result to each other result. All results are related to the query, but they are presented to the user in how they are related to each other, in categories. These categories are then related to each other to form a tree. A ranking system assumes independence in this dimension for simplicity sake, focusing instead on guessing the subjective needs of user based on the query.

The query is a means to an end, it is the article that is 'the thing'. The article is the goal — the knowledge contained within it. The query is a tool, ephemeral and disposable, so why use this dimension as the key to organize articles. Should we not use the articles themselves as the basis for organization? Shouldn't two different but similar queries produce the same results? Then why rank them differently? The fact that two semantically similar, but lexically different queries will generated a very different ranking is an inevitability with ranked lists, however, with a tree the categorization remains the same, and so will the order of visitation.

Two users may use identical queries for completely different information needs. The query is a moving target, yet when we rank, we rank based on query. That means every identical query produces an identically ranked list. The subjective nature of human perception causes problems for these objective computational models; no matter how good they are, no one objectively ordered list will likely satisfy the subjective information need of all users. Furthermore, when we rank based on Query alone, our ranking is only as good as the query. Many users create poor queries because they don't really know what they are looking for, and the query is not a good match for the information need. All the assumptions of ranking systems fail when this is the case. Since a hierarchy does not rank the results, it makes no assumptions of this kind, and the user can first browse categories (not articles) for relevancy, and upon finding one, can then continue searching from a 'more relevant footing'.

The sensitivity of ranking systems is essentially flawed, due to the reliance on objectively judging a subjective aspect, relevance. The query is much too indeterminate to act as a foundation for an objective ordering. Objectivity is a better fit in categorization and let the inscrutable subjectivity stay with its source, the user, and steer them in navigation. Finally since articles are related to each other, an assumption the user can make is that all papers in a sub category of a category of interest will be more specifically interesting or less specifically interesting. This is has been somewhat of a revelation for a few test users.

4.4 MeSHLINE aka PifMed

The prototype system that we present is called MeSHLINE, taken from a combination of the words MeSH and MEDLINE, which is apt since this project is essentially bringing

the MeSH categorization in MEDLINE to the forefront and viewing its use as browsing mechanism with an experimental eye. The original prototype was called 'PifMed', and throughout the testing, development and documentation we used this name so both names appear.

4.4.1 Dependencies

In this section we give a brief description of the services, modules and standards on which MeSHLINE depends.

MeSH

MeSH is used as the basis for the navigation structure in MeSHLINE. Unlike other controlled vocabularies, MeSH was explicitly designed by medical librarians to organize medical literature, thus ideally suited to this project's categorization task and that is why it was our choice among so many others.

MeSHLINE does not use all MeSH has to offer. First, the entry terms are not used. Inclusion of the over 172,000 entry terms (common synonyms for the MeSH category titles) would create many duplicate categories, impeding usability. In PifMed, each category is named after the `Preferred Term` of its `Preferred Concept`. For example, in this MeSH Descriptor definition for Aspirin:

```
Aspirin [Descriptor]
    Aspirin                          [Concept, Preferred]
        Aspirin                          [Term, Preferred]
        Acetylsalicylic Acid             [Term]
        2-(Acetyloxy)benzoic Acid        [Term]
    Solprin                          [Concept, Narrower]
        Solprin                          [Term, Preferred]
    Ecotrin                          [Concept, Narrower]
        Ecotrin                          [Term, Preferred]
```

we can see that Solprin, Ecotrin and Acetylsalicylic Acid results would all be placed in the category Aspirin, since Aspirin is the `Preferred Concept` for the whole definition and 'Aspirin' is the `Preferred Term` for this concept. Second, MESHLINE categorizes using descriptor terms only, qualifiers are not used (see **Appendix C** for

complete list of qualifiers terms). Qualifiers are only used in widening categorization of results, which is later described in detail.

MEDLINE

MEDLINE is the largest database searched by PubMed. MeSHLINE depends on one attribute unique to MEDLINE: all articles in MEDLINE have been indexed with MeSH terms by one of the 100 knowledge workers at the NLM. Each MeSH term can optionally be marked as `Major Topic` of a given article by the indexer. A second important feature of MEDLINE is that 79% of all articles in MEDLINE have abstracts. [52] MeSHLINE requires abstracts for tree constructions, but it is important to note that MeSHLINE misses out on ∼20% of the articles within MEDLINE due to this dependency. However, often articles without abstracts are generally less relevant to many medical queries. Many are old but some classes of articles such as letters to the editor, corrections and opinion pieces often do not have abstracts regardless of when they were published. These two categories makes up most of the citations excluded from MeSHLINE's search results.

efetch

To query and retrieve results from MEDLINE, MeSHLINE uses an NLM developed tool, named *efetch*, which is available in several programming languages from the NLM website [51]. This public domain tool allows MeSHLINE to query MEDLINE (though many other NLM databases accessible with this tool) directly via HTTP and receive results in XML (though a number of other formats are available).

Perl/Tk

MeSHLINE is written completely in Perl and the interface is written in the Perl port of Tk known as Perl/Tk. The version of *efetch* incorporated into MeSHLINE was available in Perl on the NLM website [59] and to the best of our knowledge still is.

4.4.2 System Information Flow

A high-level overview of the information flow of the system during a search session is given in **Figure 4.2**.

To begin a session the user must first, choose a pre-existing index, or enter a query into the entry box and click QUERY MEDLINE. The existing indices are selectable under the INDEX tab of the menubar. These large indexes of five and ten thousand articles (e.g. Informatics-10000, ADHD-5000) are presets provided primarily for testing purposes but also serve as a place holder for future user profile-based functionality.

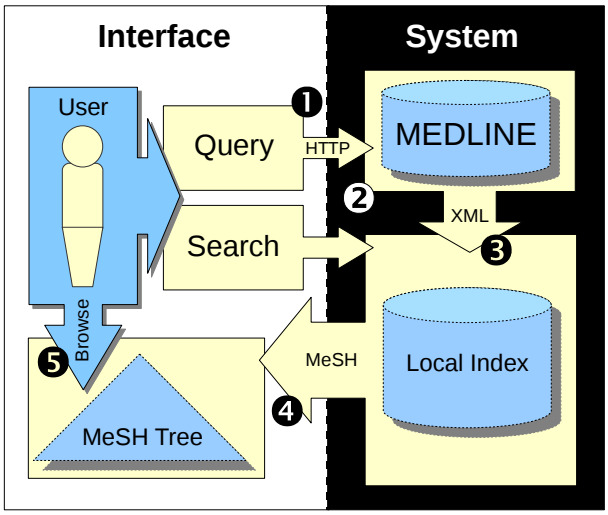


Figure 4.2: A high-level depiction of the MeSHLINE information flow: 1. user queries MEDLINE; 2. results downloaded in XML via *efetch*; 3. XML indexed locally, for iterative searching; 4. results displayed with a MeSH-based tree structure; 5. MeSH tree is browsed, modified and iteratively searched by user as needed.

When the user clicks QUERY MEDLINE, the system creates a conjunction of (ANDs) what the user has typed to the user specified preset limits (located under the MEDLINE QUERY tab in the menubar). The minimally required limits are `(hasabstract[text] AND medline[sb])` and a limit to the number of results returned (in the range 10–10000). Other optional limits include: `Clinical Trial[ptyp]`, `review[ptyp]`, `humans[mesh]`, `free full text[sb]` and `english[lang]`. ANDs and ORs are accepted as valid Boolean search operators in the user specified portion of the query.

The query is sent to MEDLINE via HTTP using the *efetch* utility. These session

results are quickly locally indexed so the user may iteratively search this local set with the SEARCH function (though most test users did not search the local set and prefer to re-query MEDLINE when the need did arise).

The MEDLINE results, now indexed, are used to build a tree from the MeSH terms attributed to each article. Each MeSH term in an article points to a node on the browsable tree where that article is placed. The tree is custom-built to the query, so only the nodes necessary to reach each article are added to the tree.

Indexing

The local index is based on the Boolean model. The words are stemmed, the stop words are removed, as well as the low-frequency terms. We chose the Boolean model for three major reasons; first, since we index while the users waits, we needed it to be fast; second, the ranking is not of great importance as we are categorizing the results; and finally, most categories return less than five articles and often only one, thus the overhead of using the vector space model for indexing and ranking would not be justified.

As a final point on the topic, presently if more than one article is mapped to the same node, the articles are presented in the order they are received from MEDLINE, therefore we benefit from the MEDLINE ranking system.

4.4.3 Interface Design

The interface is designed to be simple and familiar, intentionally focusing the attention of the user on the functionality instead of a logo or colour scheme. In pursuit of familiarity, the choice of colour for the text and graphics is inspired by Google's choice of colors. No unnecessary graphics or icons are added to sway users preference or to add confounding factors to influence test results.

4.4.4 Navigation

This method of presentation is familiar to most computer users due to its use in file hierarchies. Presented with the system no users needed any direction on its use. However, early use of the system identified two clear problems frustrating browsing.

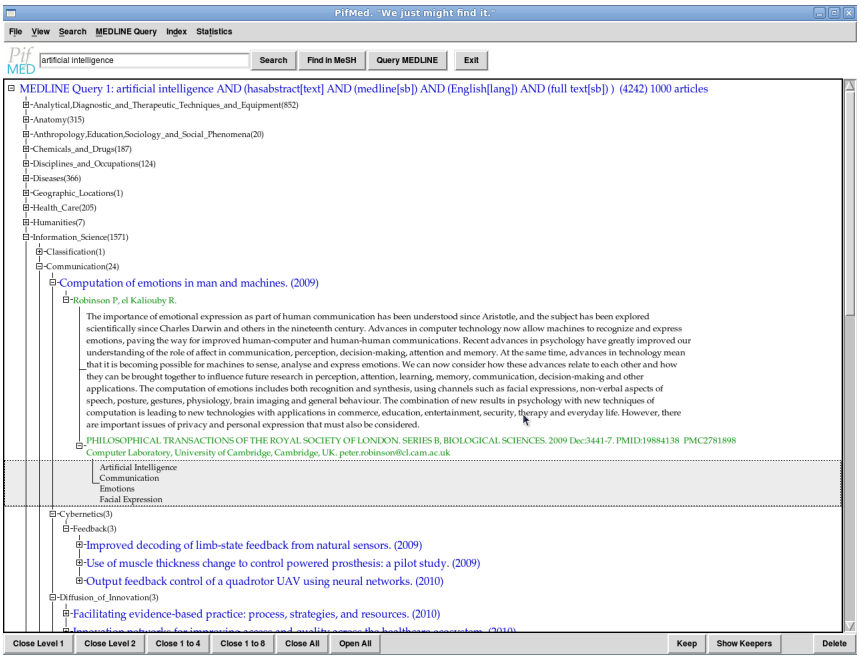


Figure 4.3: A screenshot of the MeSHLINE user interface. The root of the tree is the MEDLINE query in full. In brackets following the root is the total number of title nodes in the tree, followed by the total number of articles. These are not equal because most articles have more than one MeSH term attributed to them. The black text are the MeSH terms organized into a hierarchy. The blue nodes within the tree are title nodes. Inside a title node is a green author node. Inside the author node is an abstract in black, followed by the bibliographical information. The child node of the green bibliographical node (selected in this image) is the list of the MeSH descriptors attributed to this article.

We identified these as: 'lost in MeSH' and 'too many clicks'. We will discuss each of these presently.

Closed and Open Tree

There are three obvious ideas on how to present the starting state of the MeSH tree: all nodes closed, all nodes open, or partly open/closed.

'All closed' has two drawbacks. First, the user quickly become frustrated with opening nodes 5—10 levels deep before seeing any search results. Second, this leads the user to forget the query and focus on the MeSH terminology.

'All open' maybe seen as better, but has two drawbacks of its own. The first problem can be best described in the words of users: '*Too much clicking*', i.e. a lot

of time is spent closing nodes. This is less frustrating since the short-cut of closing nodes up the hierarchy—closer to the root—on long paths saves clicking. The second drawback is that it defeats the purpose of the categorization, since an 'all open' tree reads like a long list. When the results are presented in this way, the user begins at the top and scrolls a great deal, which was precisely the browse behavior we were trying to avoid, the categorization (i.e. MeSH terminology) falls to the background and is ignored almost completely.

These initial tree-states maybe selected in the view menu as the default state if they are preferred. Also in the application window, buttons to set the tree into the Open All, Close All states are located along the bottom, below the results (see **Figure 4.3**).

3-Click Tree

Our solution was to decide how many clicks were too many clicks; when did the user begin to get frustrated? To us, the answer was four. We designed a tree state which requires only three clicks to get the information you need to decide if an article is interesting or not.

Once a search is completed, the major MeSH headings are shown. These categories are both sufficiently distinct from each other and broad enough in scope to indicate to the user what each likely contain and what they likely do not. When one of these nodes are selected and opened (click 1) all sub-categories with these main headings are also closed, forcing the user to further focus their expectations on what results are possibly contained within. Once one of these nodes is selected and opened (click 2) all paths within that sub-tree are fully opened down to the article title node. When the user identifies a title of interest, opening this title node (click 3) will reveal the authors, abstract, bibliographical information and the MeSH terms this article has been indexed with, in an open state.

This method seems to have the best of both worlds and since its implementation no user has mentioned '*too much clicking*' commentary.

Find in MeSH

The FIND IN MESH function is put in place to address the problem of the complexity of the MeSH terminology. A user may know of a category but be unclear on the path from the root to that known category. If the user enters the category name into the entry box and clicks the FIND IN MESH button, MeSHLINE will open all nodes between it and the root, then center the screen over the found category. If the category exists in MeSH but not in the tree generated by the search results, all nodes between the root and where the node should be are opened. If the category does not exist the FIND IN MESH function leaves the tree unchanged.

This function is particularly useful when the user finds an article of interest and while browsing the other MeSH keywords used to categorize this article spots a category of interest. The FIND IN MESH feature can quickly center the screen on the contents of these newly identified categories for browsing.

At the moment, no MeSH Entry terms are implemented in this feature, only preferred MeSH Descriptor terms. The inclusion of these entry terms will be the concern of future work and will likely increase the usefulness of this feature.

Once this feature was implemented the question of the type '*Where is this (pointing to a category in a list of terms on-screen) category*' disappeared.

4.4.5 Session Search

Search

To allow users to focus their searches, a 'search within results' functionality was added (see **Figure 4.4**). Once the user has executed a MEDLINE Search, they can enter a keyword or several keywords into the search box and hit SEARCH. The system then queries the local index and returns a sub-set of articles, all of which contain the keyword (or keywords) in their title, abstract, author, or bibliographic information. This subset is then organized into the same MeSH tree structure and the root of this new tree is then rooted to the main MEDLINE Query from which it was created.

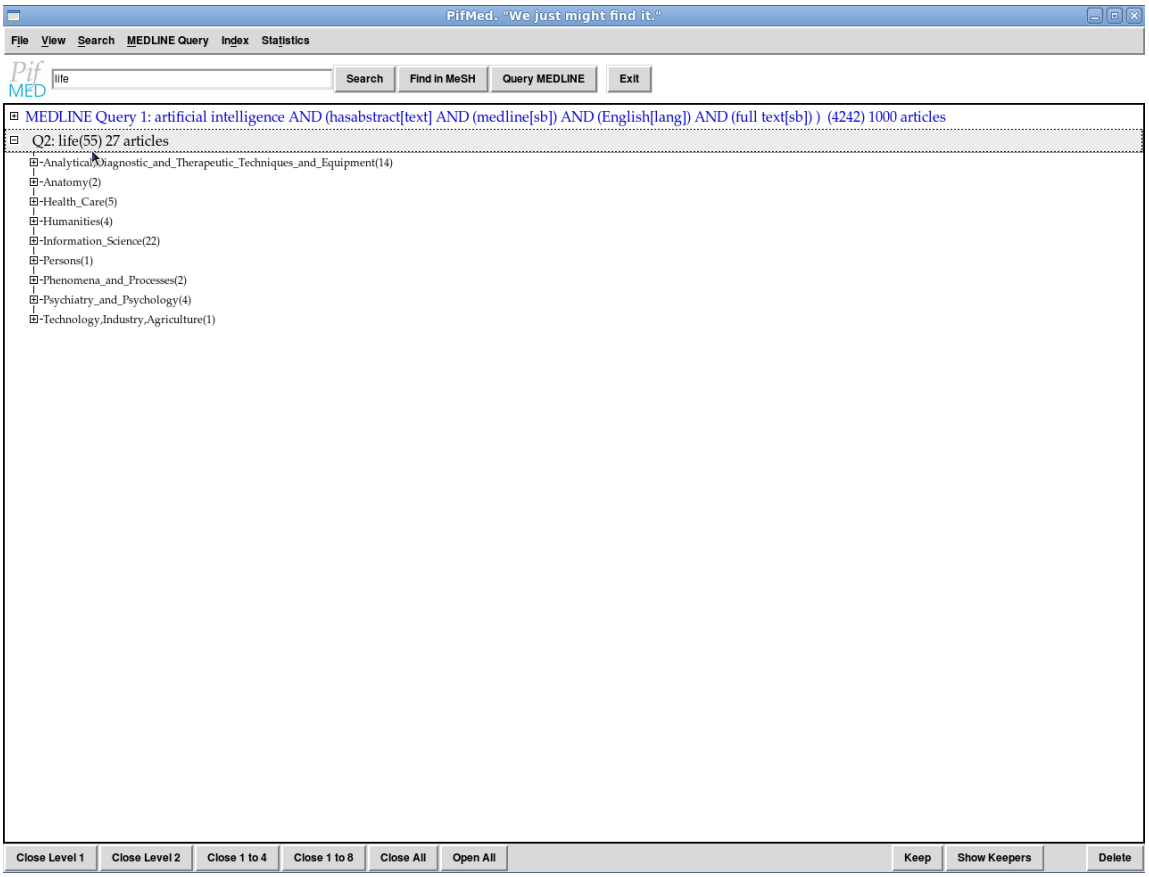


Figure 4.4: A screen-shot of the PifMed user interface. You can see the main MEDLINE Query result 'Artificial Intelligence' as a closed node and beneath it a search node 'life' open.

Keepers

When the user finds an article they like, they may click the KEEP button on the bottom right of the screen (see **Figure 4.3**). This collects the author, title, bibliographical information and abstract into a text file for the user to print, email or use however they see fit. Each time they click KEEP, the selected article is added to the bottom of the file. This file maybe viewed by clicking the SHOW KEEPERS button. Each session opens a new 'keepers' file.

Delete

Articles, branches, sub-trees even entire queries may be deleted by selecting any of the above and clicking the DELETE button. To clear clutter and prune away material

the user knows to be of no interest (or relevance) helps many users to focus on paths which show more promise. As you can see from **Figure 4.3** the DELETE button is in the lower right hand side.

Widening and Narrowing Result Sets

The search function is one way to focus a result set. A MeSH-based method to widen and narrow the results was also implemented. **Figure 4.8** shows the SEARCH menubar available within PifMed. There are three options on the bottom of the tab 'Narrowest...', 'Narrow...' and 'Wide...'. These functions take advantage of the `Major Topic` field in MEDLINE. If the 'Narrowest...' option is selected, PifMed adds a child `Article node` to a `MeSH node` only if that category has been attributed to that article and has been marked as a `Major Topic` of the article in question. If the 'Wide...' option is selected the PifMed adds a child `Article node` to a `MeSH node` only if that category has been attributed to that article and whether or not it has been marked as a `Major Topic` of the article in question. The default setting of 'Narrow...' adds all articles marked as a `Major Topic` and those who have a MeSH Qualifier of a MeSH Descriptor marked as a `Major Topic`. This is the only time I use the MeSH Qualifiers in PifMed. **Figure 4.5** shows three sub-searches '`robot`' on a MEDLINE query of '`Artificial Intelligence`'. The number in brackets following each query (and each category) show and increasing number of nodes with each search. This is due to the fact first search is set to 'Narrowest...', the second to 'Narrow...' and the third to 'Wide...'. That is, 'Narrowest...' searches have the fewest nodes associated with each category, thus resulting tree is smallest overall.

The theory is that with experience with PifMed and MeSH users will find favorite MeSH categories, or that in a particular search result a category may be particularly fruitful and may exhaust all articles in that category and would like to expand articles to include articles which are not as strongly associated with since the user's interest in the category is so high.

4.4.6 Menu Bar Tools

This section describes the functions and options available in the PifMed menu bar. There are four headings: FILE, VIEW, SEARCH, MEDLINE SEARCH, INDEX and

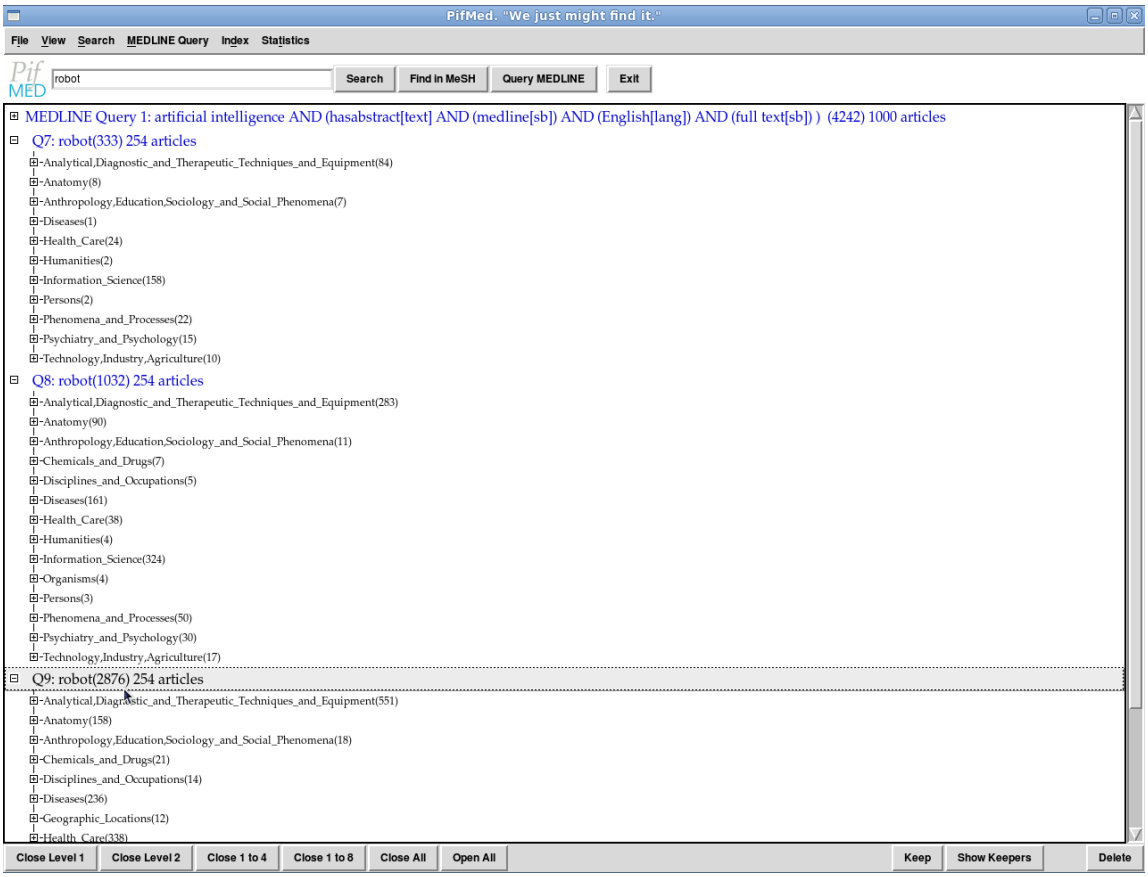


Figure 4.5: A screen-shot of the PifMed user interface. You can see the main MEDLINE Query result 'Artificial Intelligence' as a closed node and beneath it three open search nodes named 'robot'. The first search node was set to 'Narrowest...', the second search node was set to 'Narrow...', and the third search node was set to 'Widest...'. Notice the bracketed number of nodes in each tree increase with each passing query.

STATISTICS.

File Menu

The FILE menu tab as you can see in **Figure 4.6** has 5 options: `About MeSH`, `About MeSH History`, `About MeSH Structure`, `Do User Evaluation`, `Show Code` and `Exit`. The first three describe MeSH for interested users, `Do User Evaluation` is for user testing, `Show Code` shows the Perl code and is for development use specifically but any interested user can select to see the source code in a scrollable window, and `Exit` is self-explanatory.

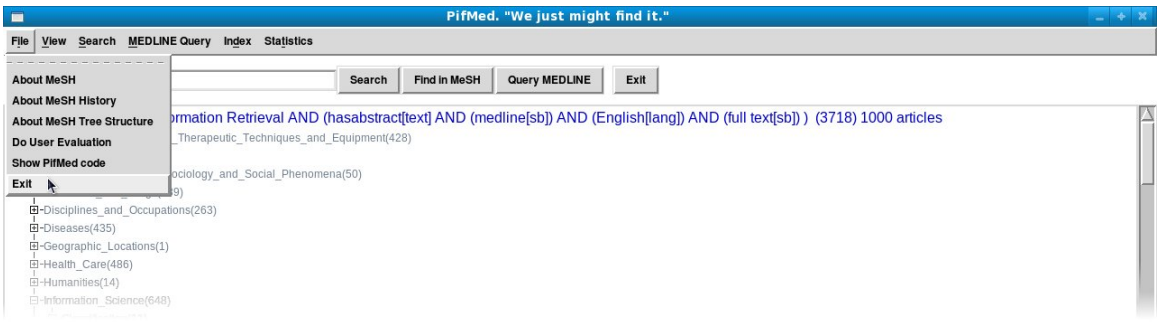


Figure 4.6: A screen-shot of the PifMed user interface. You can see the `File` tab of the menubar open. Shown are the default settings.

View Menu

The VIEW menu tab as you can see in **Figure 4.7** has 11 options, divided into 3 sections: Section 1, `Close Tree`, `Open Tree`, `3-Click Tree`, `Close Level 1`, `Close Level 2`, `Close Levels 1 to 4` and `Close Levels 1 to 8`; Section 2: `Google Look` and `PubMed Look`; Section 3: `Close default` and `3-Click default`. The commands in section 1 of this menu tab resets the tree in the fashion indicated by the command name. For example, `Close Level 2` sets children of the main category nodes to 'Closed' and leaves the rest of the nodes as they were. The options after the first divider set the default look to the Google Color scheme or the PubMed Color scheme. The last set of options set the default tree state to 'All Closed' or '3-Click Tree'. The first set of commands affect the tree immediately, the other two sections are settings which affect all future trees created.

Search Menu

The SEARCH menu tab as you can see in **Figure 4.8** has 8 options, divided into 3 sections: Section 1, `Search`, `Return All Documents`, `Show Collected References`; Section 2, `'AND' all query terms by default` and `'OR' all query terms by default`; Section 3, `Narrowest Search by default`, `Narrow Search by default`, and `Wide Search by default`. The 2 of the commands in section 1 are duplicated by on-screen buttons (SEARCH and SHOW KEEPERS the other command `Return All Documents` makes a new full tree with all documents and categories (repairs all DELETE actions). The third section is the only way for the user to set the level of sensitivity of the search

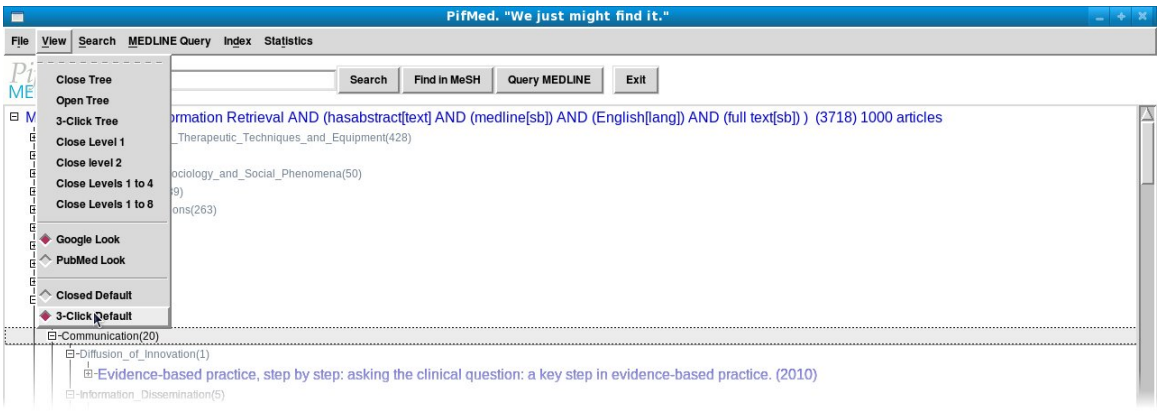


Figure 4.7: A screen-shot of the PifMed user interface. You can see the `View` tab of the menubar open. Shown are the default settings.

tree as described in the Widen and Narrowing Result Sets section.

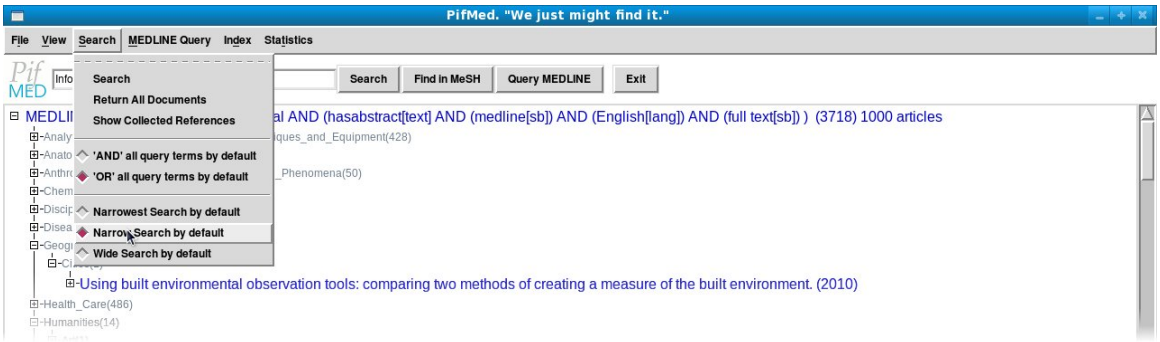


Figure 4.8: A screen-shot of the PifMed user interface. You can see the `Search` tab of the menubar open. Shown are the default settings.

MEDLINE Query Menu

The MEDLINE Query menu tab as you can see in **Figure 4.9** has 20 options, divided into 4 sections. The first command executes a MEDLINE Query in the same manner as the button. The second section sets the maximum number of query results the user would like to return. Useful in reducing wait time for quick searches or alternatively for returning thousands of results for an in-depth result set which — for example — would be necessary for beginning, or continuing, a literature review. The third section shows the setting for the default search parameters added

by PifMed to every MEDLINE Query (and appears on-screen directly after the Query Phrase). You can see that `hasabstract[text] AND (medline[sb])` are the minimum requirements for PifMed. The forth section gives a selection of `Article Type Limits` requested by users through the Pilot and User studies. These last two sections are some (but not all) of the options available in `Advanced Search` in PubMed.

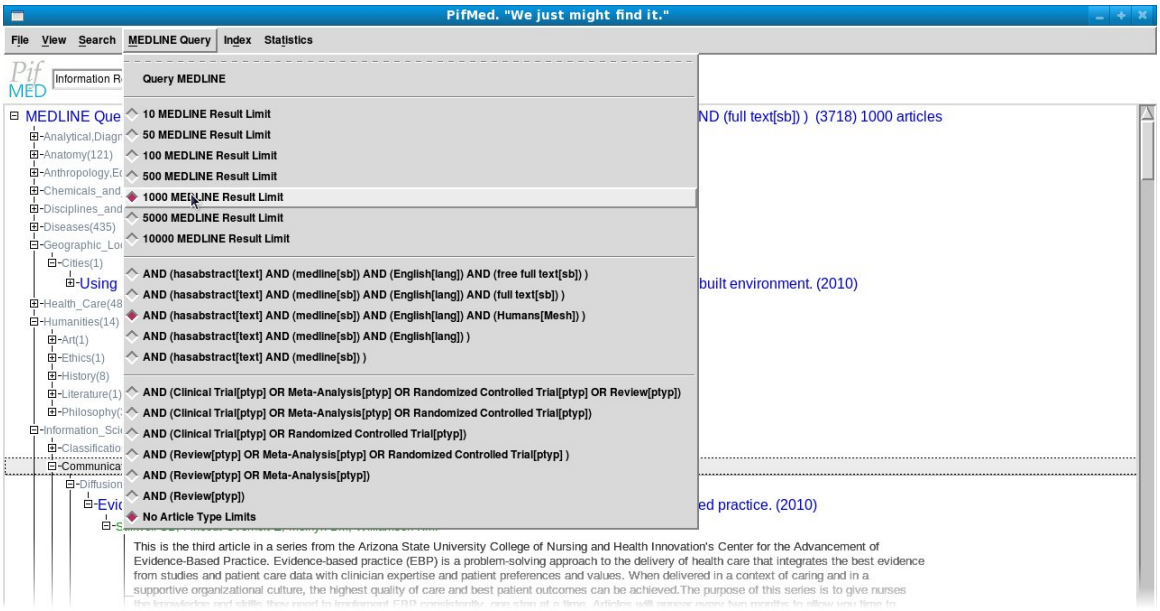


Figure 4.9: A screen-shot of the PifMed user interface. You can see the `MEDLINE Query` tab of the menubar open. Shown are the default settings.

Index Menu

The INDEX menu tab as you can see in **Figure 4.10** has 13 options, divided into 2 sections. The second section are user created indexes. The user creates them by, first, executing a MEDLINE Query and, second, selecting the `Create Index` command from this menu. The index is then added to this menu. `Armageddon` destroys the presently active index. `Select MEDLINE Query` is the default index and overwritten with every MEDLINE Query.

Statistics Menu

The STATISTICS menu tab as you can see in **Figure 4.11** has 3 options: `Show Stopword List` shows the list of words used for indexing in a separate window; `Show`

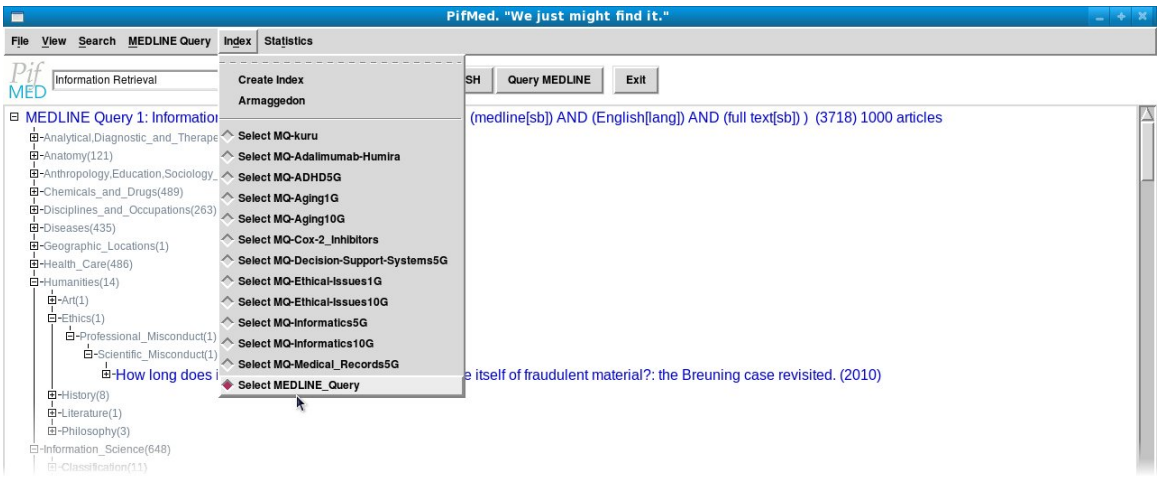


Figure 4.10: A screen-shot of the PifMed user interface. You can see the `Index` tab of the menubar open. Shown are the default settings.

`Stopword Statistics` shows the user the frequency counts for each stopword (and total stop words removed) in a separate window; `Show Index Word Statistics` shows the user all words used in the index sorted by frequency count (plus total words and total singletons removed) in a separate window.

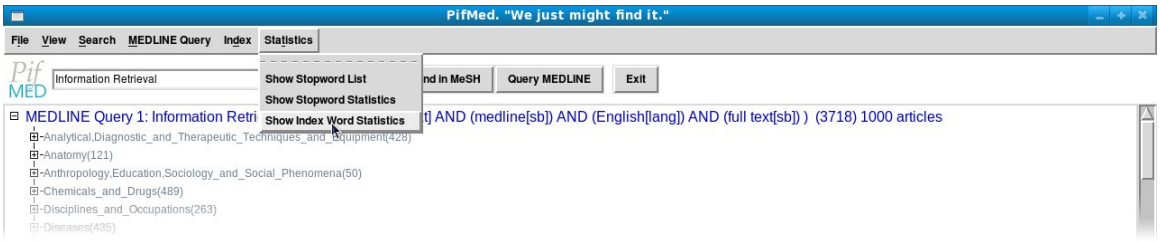


Figure 4.11: A screen-shot of the PifMed user interface. You can see the `Statistics` tab of the menubar open. Shown are the default settings.

4.5 Limitations

Very large query result sets (10,000+) are too slow for general use as shown in **Table 4.1**.

To what degree previous knowledge of MeSH maybe necessary to best navigate this system is still unknown. We believe use promotes understanding of the MeSH terminology, but willingness to learn MeSH, what rate users learn MeSH and what

Result Limit	MEDLINE Download	Index XML	Build Tree	Display Tree	Total
100	1s	1s	1s	1s	4s
500	6s	3s	6s	3s	18s
1000	10s	6s	12s	11s	39s
10000	138s	63s	118s	425s	744s

Table 4.1: MEDLINE search results for `neoplasms` returning limits of 100, 1000, and 10,000 of 716,517 possible results. 100 is quite fast, 1000 is still acceptable, but ∼12.5 minutes to manage 10,000 results is much to slow for all but the most dedicated user. The default limit for results is 500 (average 16–22s total wait time) for this reason.

depth of knowledge of MeSH is needed to best make use of this categorization are all unknowns. At the moment users did not seem uncomfortable with the system, but comfort levels have not been measured and we anticipate this may be an impediment to adoption as the system of choice.

4.6 Lessons Learned

One benefit of this model lies in the persistence of state of the users browsing. We would like to extend state persistence past one session by implementing user profiles. These user profiles would log past queries and maintain previous results trees in the state they were in when they were last modified. This would be particularly useful for an information monitoring task, where search trees could be re-queried, new entries added automatically marked as 'unseen' or 'new' and this marking reflected on parent nodes up the hierarchy.

User profiles and persistence make tree pruning a more worthwhile effort. Efforts deleting branches and articles would not go wasted. These 'pruned' search results may be valuable enough to share with other users, or be used as the basis for automatic query generation for new searches.

In conclusion, we have shown knowledge-based methods can be used to speed browsing of large result sets. The pilot study indicates users may prefer this browsing model to the existing conventional model. The encouraging results of our pilot study, mixed with the small sample size, justify a larger user-study to reinforce these findings.

Chapter 5

Results and Evaluation

5.1 Introduction

A user study was designed to test and measure the usability of the prototype system. To test the user study, a pilot study was completed. Participants completed tasks with both systems, in an informal environment (at my desk in the CS Playgrounds) and answered a questionnaire about their experience and thoughts on the system. Use-time for each system was compared and ratings and comments from the questionnaire were analyzed.

5.2 Pilot Study

First, the user was given a brief 5–6 minute power-point presentation which demonstrated the system. Then the user was asked to use the MeSHLINE system to find an article of interest. The user was made aware before they began that they would later use the same query on PubMed to find the same article they found on MeSHLINE. This process was repeated once. After these searches were completed the user was asked to fill out a short questionnaire. There was no monetary remuneration for these pilot users, nor did they the see the MeSHLINE system before the test took place.

Qualitative and quantitative data was collected during the user study. For quantitative data collection MeSHLINE reported all user actions to a log. This data was analyzed to measure how long it took to complete queries and where the user (and system) spent most of their time.

The qualitative data was collected primarily through a questionnaire which immediately followed the system and was administered through MeSHLINE itself.

5.2.1 Questionnaire

A three-part questionnaire was issued after the test was completed. The first part was meant to determine their similarity to the intended user. These questions ask them to rate their familiarity with: Computers, Computer Search, PubMed, MeSH and Medicine. An ideal candidate would score 5–7 out of seven in each of these categories.

The second part of the questionnaire was meant to judge the usability of the system. 11 questions in total, each question asks the user to rate MeSHLINE in terms of an aspect of usability (Effectiveness, Efficiency, Easy-to-Learn, Error Tolerance, Engagement) [82] and then directly compare MeSHLINE to PubMed on this aspect. The final question in this section asked for an overall rating of MeSHLINE and an overall comparison of the two systems. All ratings were from 1 (low/poor) to 7 (high/great).

The third-part of the questionnaire was in the form of short answer and asked the user to state favorite and least favorite feature of MeSHLINE and finally left room for comments, suggestions and to recommend improvements.

5.2.2 System and Hardware

The test system was Desktop PC with a dual 3.2 GHz Intel Pentium 4, 2GB of RAM, 100GB HD, running Linux-based Fedora 8 OS and a 17" LCD screen.

5.2.3 Pilot Study Participants

Three health informatics graduate students participated in this pilot study. They were contacted informally via email at the suggestion of professors within the Faculty of Computer Science here at Dalhousie University.

The intended users of this system are regular users of PubMed. Therefore we primarily focused on the needs of medical researchers, medical students and medical librarians. The familiarity of our ideal user would score a 5–7 out of 7 in each category: Computers, Computer Search, MeSH, PubMed and Medicine. According to the results shown in **Table 5.1** you can see that knowledge of MeSH is the low point, but overall, our pilot group is satisfactorily close to our target user. We expect

1	2	3	Min	Median	Max	Average	Familiarity with...
7	7	7	7	7	7	7	Computers
7	7	6	6	7	7	6.7	Computer Search
5	4	2	2	4	5	3.7	MeSH
5	6	6	5	6	6	5.7	PubMed
4	7	7	4	7	7	6	Medicine
			4.8	6.2	6.4	5.8	Average Participant

Table 5.1: In Part I of our questionnaire each user is asked to rate their familiarity in each of these areas. Our ideal user would score a 5–7 in each of these categories.

familiarity with MeSH to be the low point for the majority of users, these results reflect this assumption.

5.2.4 Quantitative Results

The quantitative data in **Table 5.2** presented some unexpected results. The fact that MeSHLINE was ∼80% faster than PubMed was a surprise. To explain the difference my instinct was to find fault with the study design. It may the case that hidden in the order of use is a confounding factor that is affecting the outcome. Perhaps the fact that the user found the paper with one tool means that tool lead them to that result, a result that would not so easily be repeated using the other tool. This may be the case, however, each of these users had a very specific paper they were looking for in at least one of their two queries, which throws doubt on this as an explanation.

To remove possible influence of confounding factors in this regard, the order of the systems was randomized in future studies.

5.2.5 Qualitative Results

On the whole, the users in our pilot study found MeSHLINE preferable to PubMed, though not resoundingly so. Users gave MeSHLINE the edge in terms of Easy-to-Learn (5.1), Effectiveness (4.8) and Efficiency (4.8) and lower grades for Error Tolerance (4.4) and Engagement (4.1). The overall usability averages to 4.7 out of 7. See **Table 5.3** for the full results.

User2	User3	Query
156	103	Q1: MeSHLINE (secs)
245	212	Q1: PubMed (secs)
89	**109**	**Q1: Difference (secs)**
57.05%	**105.83%**	**MeSHLINE Faster by X%**
213	96	Q2: MeSHLINE (secs)
341	220	Q2: PubMed (secs)
128	**124**	**Q2: Difference (secs)**
60.09%	**129.17%**	**MeSHLINE Faster by X%**
108.5	116.5	Average Difference in seconds
58.57%	117.50%	Average Result
	142	Total Average search time: MeSHLINE
	254.5	Total Average search time: PubMed
	112.2	Total Average difference
	79.23%	Total Average Result

Table 5.2: The results of our quantitative data collection. The quantitative data from user one was unfortunately corrupted and not usable for analysis.

min	median	max	average	measure
2	5	7	4.8	Effective
3	5	7	4.8	Efficient
3	4	6	4.1	Engaging
2	4	7	4.4	Error Tolerant
4	5	7	5.1	Easy to Learn
2	**5**	**7**	**4.7**	**Total Average Usability**

Table 5.3: The final analysis of the qualitative results. Averages of each of the measures of usability along the right are used to calculate the final score, which is on the bottom line.

5.3 User Study

5.3.1 Motivation

The results of the pilot study were encouraging, but the sample size was much too small to draw conclusions. Participation was much improved for the user study. 27 participants versus the three for the pilot study means populations were sufficient to perform statistical paired t-tests to rigorously compare the means, and justify inferred conclusions with statistical significance and show degrees of confidence in those conclusions.

5.3.2 Pilot Study versus User Study

There are four major differences between the design of the user and the design of the pilot study: the user study had three queries per participant while pilot study only two; in the user study Part I of the questionnaire was expanded to better determine the users' familiarities; the order of the search engines was randomized for each task iteration; and users were compensated 5 dollars for their participation.

5.3.3 Terminology

In this section I define some terminology to keep the analysis clear.

An **Outcome Pair** is a keyword phrase used by the user identically on each search engine and the resulting time it takes on each; e.g., Query(129,33) where '129' is the number of seconds this user has taken to find and article of interest on PubMed and '33' is the number of seconds to find the same or as interesting article on PifMed.

The **Query Phrase** is the phrase used for any given Outcome Pair, since the phrase is identically used on each search engine it could be written ' 'cancer'(129,33)' and since each user has three different query phrases to complete the study, it could be written 'Q1(X,Y)','Q2(X,Y)' or 'Q3(X,Y)'.

A **Task Iteration** is one complete task:

1. Producing a Query Phrase,
2. Executing this Query Phrase in Search Engine A,
3. Browsing to a result of Interest,

Min	Median	Max	Mean	Familiarity with...
5	7	7	6.9	Computers
5	7	7	6.8	Computer Search
1	2	6	2.5	MeSH
1	3	7	3.5	PubMed
2	5	7	4.9	Max(PubMed, Medicine, Biology, Psychology, Health/Bioinformatics)

Table 5.4: The profile of our General User population. Our General User must score at least a 5 in both Computers and Computer Search to be in this population.

4. Telling the Tester '*Done*' or '*Nothing Interesting*',

5. Entering the identical Query Phrase on Search Engine B,

6. Browsing to the same or equally interesting result,

7. Telling the Tester '*Done*' or '*Nothing Interesting*'.

For complete participation, each user completes 3 Task Iterations, the result is 3 Outcome Pairs (one for each task iteration).

5.4 Population and Data Partitions

5.4.1 Populations

Each user rates (from 1 to 7) their own familiarity with 9 aspects: Computers, Computer Search, MeSH, PubMed, Medicine, Biology, Psychology, Health Informatics and Bioinformatics. We use these aspects to identify populations within the pool of users. We have two sets: General Users and Target Users. A General User (see **Table 5.4**) is any user scoring at least a 5 in both Computers and Computers Search which we feel are the key competencies. For this study all 27 participants tested fit this profile. A Target User (see **Table 5.5**) is a user scoring at least a 6 in Computers and Computer Search and at least a 4 in one of the following PubMed, Medicine, Biology, Psychology, Health Informatics and Bioinformatics. For this study 22 participants tested fit this profile.

Min	Median	Max	Mean	Familiarity with...
7	7	7	7.0	Computers
6	7	7	6.8	Computer Search
1	2.5	7	3.0	MeSH
1	4.5	7	4.2	PubMed
4	5	7	5.5	Max(PubMed, Medicine, Biology, Psychology, Health/Bioinformatics)

Table 5.5: The profile of our TARGET USER population. Our TARGET USER must score a 6–7 in each of the first two categories, and 4–7 in at least one of the following fields: PubMed, Medicine, Biology, Psychology, Health Informatics, Bioinformatics to be part of this population

5.4.2 Query Result Set Size

Each query returns a number of results. Should that number of results be less than or equal to 20, it is considered a short result set. The reason for choosing '20' is 20 results completely fit on one page of the PubMed interface. As a result we have two sets to analyze: **All Queries** and **Long Queries**. Though I see that the reader may immediately think 'long queries' to mean many keywords, however the word 'results' is used far often in this section to be unambiguous, though it would be more accurately called: 'All queries with short search result sets removed.'

5.4.3 Independence Assumptions

There is another way to partition these results which depends on whether the reader agrees with an assumption of independence in the queries. That is, we can assume that each outcome pair is independent of each other outcome pair, or all outcome pairs are dependant on the user. If we assume each outcome pair is independent we have 27 x 3 = 81 results. If we do not make this assumption then we have 27 results. In the later case the result is the total time spent on each search engine over the course of the test. For example, Q1(129,33):Q2(14,91):Q3(61,64) would give a result of UserX(PubMed(129+14+61), PifMed(33+91+64)) = UserX(204,188).

It reasonable to assume that times produced by users may depend on that user, but it is not a certainty. It is for this reason that I have completed the analysis with and without this assumption. These two assumptions are called **Independence of**

Outcome Pairs, and **Dependence of Outcome Pairs**.

5.4.4 Groupings Used for Analysis

As a result of these 3 data partitions, this analysis focuses on $2^3 = 8$ groupings of the data:

1. General Users, All Queries, Independence of Outcome Pairs.
2. General Users, All Queries, Dependence of Outcome Pairs.
3. General Users, Long Queries, Independence of Outcome Pairs.
4. General Users, Long Queries, Dependence of Outcome Pairs.
5. Target Users, All Queries, Independence of Outcome Pairs.
6. Target Users, All Queries, Dependence of Outcome Pairs.
7. Target Users, Long Queries, Independence of Outcome Pairs.
8. Target Users, Long Queries, Dependence of Outcome Pairs.

5.4.5 Other Groupings and Results

There are a few data partitions which were tested and have no statistical significance. Their descriptions in this section, serve to rule them out as confounding factors of the final analysis.

Order in Outcome Pair

The order of which search engine is used has been randomized, this was done to rule it out as a possible confounding factor. Overall, users took ∼4.5 seconds longer with the first engine than the second as shown in **Table 5.6**.

This result shows that removing it as a possible confounding factor (by randomizing order) was prudent but not essential since a paired t-test shows its level of statistical significance is low, with a two-tailed p-value of 0.576 as shown in **Table 5.7**.

Min	Median	Max	Mean	Variance	Search Engine Order
General Users					
22.0	84.0	240.7	93.4	2725.1	Average First
16.3	76.3	257.0	88.9	3340.2	Average Second
Target Users					
22.0	76.3	240.7	90.9	3107.9	Average First
16.3	75.0	255.7	86.2	3450.8	Average Second

Table 5.6: Comparison of order of use for each of the populations.

Paired t-Test (by Query)	1^{st} vs 2^{nd}
Observations	81
Hypothesized Mean Difference	0.0000
Observed Mean Difference	4.4815
Variance of the Differences	5160.3433
df	80
t-Statistic	0.5615
$P(T \leq t)$ one-tail	0.2880
t Critical one-tail	1.9901
$P(T \leq t)$ two-tail	0.5760
t Critical two-tail	2.2844

Table 5.7: The results of a paired t-test comparing of order of search engine use for General User population.

Fails

For a given query if a user reported: *'I can't find anything relevant'* or *'I'm bored of looking'* the engine used was marked as a `FAIL` in that Outcome Pair, but the time taken for the user to come to this conclusion was recorded and used in the time comparison study. For the General User, five fails were reported for PifMed and six fails for PubMed, and Target Users reported five fails for each of the search engines. This result provides no insight, it is mentioned here for the sake of completeness.

Discretization of Use-Time Pairings

If we ignore to what degree PifMed was faster than PubMed, but instead if, for each Outcome Pair, we record a `TRUE` if PifMed was faster than PubMed and `FALSE` if not, we would have a Boolean result for each Outcome Pair. **Table 5.8** shows the average results for each population. These results are encouraging, but if PifMed was only 1 second faster 62% of the time and slower by 20 seconds 38% of the time these results would be misleading. I include these results in support of the full use time tests which follow.

	General Users	Target Users
All Queries	62.96%	59.09%
Long Queries	67.90%	63.92%

Table 5.8: The mean results of the Boolean `PifMed is Faster` data. The percentage is how often PifMed was faster then PubMed for each population and each search result set size partition.

Inter-Task Iteration Comparison

Each user performed three pairs of queries. Assuming the user was familiar with PubMed, as user gained familiarity with the PifMed, it is possible that they may perform faster with each iteration. There is another possibility, that with each iteration the user got better at the task. Finally, there is the a third possibility that the user — in an attempt to 'do well' or 'please the tester' — choose queries which suit PifMed rather than PubMed or choose articles faster/less critically with PifMed.

Iteration	PubMed	PifMed	Mean Difference	Total Mean time
General User				
Q1	103.78	91.81	+11.96	195.59
Q2	92.11	85.56	+6.56	177.67
Q3	104.11	69.15	+34.96	173.26
Target User				
Q1	93.45	105.77	-12.32	199.26
Q2	79.41	81.36	-1.95	160.77
Q3	101.73	69.27	+32.45	171.00

Table 5.9: The mean times for each Task Iteration, independent of user.

Though I would like to report any speed increase that comes with use is a result of the usability of PifMed, that more experience with PifMed is directly attributable faster times, these other two confounding factors must be ruled out.

1.A) Overall Task proficiency If with each iteration the user got better at the task, then incremental improvement with both engines would be seen between each iteration. A general trend can be seen, that overall mean use time for each iteration generally declines with use. However, as shown in **Table 5.9** neither the General User nor the Target User population average faster speeds with each iteration of the task with PubMed but with PifMed both user groups complete the task faster with each iteration. This is most pronounced in the Target User group where PubMed has a faster mean use time than PifMed and with each iteration its mean use time decreases until PifMed averages a faster use mean time to task completion. If 4 or more task iterations were completed we could see if this trend would continue, However, given the data we have, viewing the overall downward trend in use time as a result of a the possible confounding factor of increasing task proficiency, is not founded by the analysis of the results. This overall downward trend through iterations is directly attributable to PifMed's pronounced decreasing use time, why PifMed's use time decreases is still under scrutiny, but this factor can be ruled as a minor influence at best.

1.B) Query Choice Since the user was free to choose their own query, it is possible that with each iteration of the task, the user — putting motivation aside for the

Iteration	General User	Target User
Q1	342.4	349.9
Q2	336.9	319.0
Q3	329.5	303.5

Table 5.10: The average size of the search result set with each task iteration.

moment — selected queries which were better suited to PifMed than to PubMed. If so inclined, the user may try queries which return larger or smaller result sets thinking that a bigger or smaller set may be easier to browse with PifMed. Either way, the average size of the search results did not significantly change as shown in **Table 5.10**. There is a general decrease in search result set size but it would have no impact since searching with either search engine through 350 results in under five minutes (no user took longer than 5 minutes for any given task) would not be significantly different than searching through 300 results in the same amount of time.

I can see no pattern in the choice of query keywords or number of keywords to indicate any user 'gaming the system'. The **Appendix D.2.4** includes all queries by iteration, should the reader be interested in challenging or confirming this conclusion.

The Hawthorne Effect The Hawthorne Effect is based on a longitudinal study [80] on workers at The Hawthorn Works in Chicago during the 1920s. The popular understanding of this theory is: since users know they are being tested they perform more efficiently in response. In fact the theory as postulated by Elton Mayo [47] is: the participants were motivated to improve their performance since they were flattered by the attention paid by the researchers. This is a significant difference, we will take the theory as originally postulated.

There are some very important differences between the Hawthorne Study and this usability study (adapted from [45]),

1. The Hawthorne Studies were *Longitudinal studies*; this study was a *one-off test.*

2. The participation was mandated at the factory, but voluntary in my study.

3. Studies were strange and interesting to Hawthorne participants; studies are common-place in the university setting.

4. Hawthorne participants were at work, where poor performance may impact employment; not a factor in my test.

5. Hawthorne participants were experts using familiar tools; in this study, at least one of the tools tested was novel to every user.

If the reader believes this effect impacts this study, it must be noted that it would impact both interfaces, PifMed and PubMed. Furthermore, the effect would have the greater impact on PubMed since it is the more familiar tool, the one these expert participants would most able to improve their performance.

In conclusion I can find no evidence of it, nor can I rule it out. We can see on an intuitive level that when users are presented with the system they have used it zero times, and by the time they reach their last query they will have a better idea of how to use it and what to expect, resulting in faster times. Furthermore, in observation of participants I noticed users spending more time looking through the categories in the first two queries, by the third query users seemed to do less exploring of categories and more exploring of the literature. This may explain the trend.

One reason this is important is: to determine the usefulness PifMed, we need to determine the usefulness of MeSH as a categorization system. Given most users are unfamiliar with MeSH, it could easily have presented a serious problem in terms of usability. The results do not make that case, in fact, it seems that users — as expected — would adapt quickly to this (or any other) sensible categorization system given the opportunity. Some users questioned its sensibility:

> *The MeSH taxonomy has an unintuitive top-level for people interested in molecular biology. Entering through chemical and drugs, to the amino-acids to get to 'proteins' was unexpected. Finding 'organelle' as a category eluded me. I guess that a user with a specific background, unfamiliar with*

> *MESH, may have a hard time to see what the lower levels contain just from a node name.*

but overall users made little complaint about it as a categorization system. In fact the most repeated comment (∼18 different users) was of the ilk: "*...loved the categorization...*".

In conclusion, I do believe users became more proficient with PifMed with each task iteration, that users quickly adapted to the MeSH categorization and thus it became more useful to users with experience, but due to confounding factors, a confidence level and p-value would not be well-founded.

5.5 Questionnaire

5.5.1 Questionnaire: Part I

This part of the questionnaire was used to partition the participants into useful populations, discussed above in Section 5.4.1.

5.5.2 Questionnaire: Part II

After the search tasks were complete the user was asked to fill out a questionnaire (see **Appendix D.1** for full questionnaire). Part II of that questionnaire is the basis for our determination of the usability of PifMed alone and the usability of PifMed in comparison to PubMed. **Table 5.12** shows the results of this part of the questionnaire for the General User population and likewise **Table 5.14** for the Target Users. Each question in **Table 5.12** has two rows. The first row displays the results for PifMed alone and the second row displays the results for the comparison of PifMed and PubMed. To better understand the comparison results, **Table 5.11** is a reproduction of the guide users were given to answer comparison sub-question.

The reader will notice, each question in **Table 5.12** and **Table 5.14** has an aspect in bold before each question. These aspects are known as the '5 Es' of usability: Effectiveness, Efficiency, Engagement, Error Tolerance, and Easy-to-Learn. [82] Each instance asks a question to determine one of those aspects. There are 2 questions for each aspect. The final question asks for an overall rating.

```
 Guide for comparison of PubMed vs.  PifMed
-----------------------------------------------------------
1 = Always would use PubMed for this reason
2 = Usually would use PubMed for this reason
3 = Liked PubMed a little more for this reason
4 = No Preference.  Each performed equally in this regard
5 = Liked PifMed a little more for this reason
6 = Usually would use PifMed for this reason
7 = Always would use PifMed for this reason
-----------------------------------------------------------
```

Table 5.11: The guide given to participants to rate their preferences in the search engines after the search tasks were completed.

min	med	max	mean	Questions
3	6	7	5.4	**Effectiveness**: Did PifMed give you relevant results?
2	5	7	5.2	
1	4	7	4.2	**Efficiency**: Did PifMed respond quickly?
1	4	7	3.7	
1	6	7	5.9	**Engaging**: Did PifMed encourage you to explore the results?
1	6	7	5.6	
4	7	7	6.6	**Error Tolerance**: Did you notice any errors in PifMed?
4	4	7	4.7	
3	7	7	6.1	**Easy to Learn**: Was PifMed easy to learn and understand?
2	5	7	4.9	
1	6	7	5.5	**Effectiveness**: Did PifMed help you make up your mind on
2	6	7	5.5	what you were looking for?
2	6	7	5.3	**Error Tolerance**: How would you rank your confidence
2	5	7	5.0	in the results?
3	6	7	5.7	**Easy to Learn**: Do you feel like you have a good
2	4	7	4.7	understanding of the capabilities of PifMed?
2	6	7	5.6	**Efficiency**: Rate the ease (or difficulty) in retracing
1	5	7	5.1	your steps.
2	6	7	5.5	**Engaging:** Did PifMed help you browse to interesting
3	6	7	5.6	papers you did not expect?
3	6	7	5.8	**Overall**: Please rate the ease of using PifMed overall.
2	5	7	5.1	
2	6	7	5.6	Average for **PifMed**
2	5	7	5.0	Average for **Comparison**
2	6	7	5.3	Average **Overall Usability**

Table 5.12: The results of Part II of the questionnaire for the General User population. In bold before each question is the aspect of usability which the following question is evaluating. There are two questions from each aspect. Each question has two rows: the top row asks to rate PifMed in relation to the question, the second row is a comparison between PifMed and PubMed, that is to say it asks the user to rate strength of preference to PifMed(5–7) or PubMed(3–1) in relation to the question.

min	median	max	average	measure
1	6	7	5.4	Effective
1	4.5	7	4.7	Efficient
1	6	7	5.5	Engaging
2	6	7	5.4	Error Tolerant
2	5.5	7	5.4	Easy to Learn
2	**6**	**7**	**5.3**	**Total Average Usability**

Table 5.13: The final analysis of the qualitative results for the General User population. Averages of each of the measures of usability (on the right) are used to calculate the final score, which is on the bottom line.

Each of these aspects can be measured as the mean score of all ratings of questions of a that aspect. For example, the 4 ratings for "**Effectiveness**: *Did PifMed give you relevant results?*" and "**Effectiveness**: *Did PifMed help you make up your mind on what you were looking for?*" from each user can be averaged to give PifMed an overall rating for Effectiveness, **Table 5.13** gives these results for the General User population and **Table 5.15** gives these results for the Target Users.

5.5.3 Questionnaire: Part III

Part III of the Questionnaire was of a short answer format, where users are given the opportunity to say in their own words, what they thought of PifMed.

Question: What was your least favourite feature of PifMed?

Rethinking of the names of the buttons Query MEDLINE, Find in MeSH and Search. Many users didn't know what MeSH was, so the source of this button's name was a bit of a puzzle. In hindsight 'Find Category' would be a better name. Users consistently hit Search to Query MEDLINE or the `enter` key, neither of which initiated the desired behavior from the system. Binding the `enter` key to the Query MEDLINE function, renaming it 'Search MEDLINE' and swapping position with the Search button would take care of most of the complaints about this UI issue. Also the search button should be renamed 'Search Subset'. The user may not know what that means, but at least they won't think it means 'Query MEDLINE'.

The wait time between pressing Query MEDLINE and seeing results was a

min	med	max	mean	Questions
3	6	7	5.3	**Effectiveness**: Did PifMed give you relevant results?
2	5	7	5.0	
1	4	7	4.3	**Efficiency**: Did PifMed respond quickly?
1	4	7	3.8	
1	6	7	5.8	**Engaging**: Did PifMed encourage you to explore the results?
1	6	7	5.5	
4	7	7	6.5	**Error Tolerance**: Did you notice any errors in PifMed?
4	4	7	4.6	
3	7	7	6.2	**Easy to Learn**: Was PifMed easy to learn and understand?
3	5	7	4.9	
1	5	7	5.2	**Effectiveness**: Did PifMed help you make up your mind on
2	5	7	5.3	what you were looking for?
2	6	6	5.1	**Error Tolerance**: How would you rank your confidence
2	4	7	4.8	in the results?
3	6	7	5.7	**Easy to Learn**: Do you feel like you have a good
2	4	7	4.6	understanding of the capabilities of PifMed?
2	6	7	5.9	**Efficiency**: Rate the ease (or difficulty) in retracing
3	5	7	5.2	your steps.
2	5.5	7	5.3	**Engaging**: Did PifMed help you browse to interesting
3	5	7	5.4	papers you did not expect?
4	6	7	5.8	**Overall**: Please rate the ease of using PifMed overall.
2	5	7	5.0	
2	6	7	5.6	Average for **PifMed**
2	5	7	4.9	Average for **Comparison**
2	5	7	5.2	Average **Overall Usability**

Table 5.14: The results of Part II of the questionnaire for the Target User population.

min	median	max	average	measure
1	5	7	5.2	Effective
1	5	7	4.9	Efficient
1	6	7	5.5	Engaging
2	5.5	7	5.3	Error Tolerant
2	6	7	5.4	Easy to Learn
2	**5**	**7**	**5.2**	**Total Average Usability**

Table 5.15: The final analysis of the qualitative results for the Target User population. Averages of each of the measures of usability (on the right) are used to calculate the final score, which is on the bottom line.

major complaint among users, both verbally and written in this section. This wait time makes the lack of a status bar, or hourglass — a visual cue that PifMed was working and not frozen — to be a glaring omission on my part. This should be added.

Four Users complained about MeSH: categories were too broad, overlapped in ways they did not like, unfamiliar, not representative of what was contained within them.

Three Users wanted to see the full article, this functionality would be built into a next version should there be one.

Three Users complained that a 'Highlight Keywords' functionality should be implemented, and the ability to dynamically enter them like 'Find' in a web-browser is needed.

Those were the primary complaints, **Appendix D.2.2** has a complete listing.

Question: What was your favourite feature of PifMed?

Overwhelmingly, the favourite feature was the categorization of results, at least 18 users made a comment of this kind. Many users liked the tree-structure and ability to open and close whole sections of results, allowing navigation to bunches of interesting results as well as ones that were not expected.

Many users like the KEEP button and SHOW KEEPERS which displays the list of collected results. Many said PifMed was faster to reach interesting results.

Question: Suggestions, Improvements, Comments?

This section repeated much of what was said in the first section with a few exceptions. One users suggested ranking categories by relevancy to the query; 'Forward' and 'Back', web browser functionality was suggested to aid backtracking; three users did not like the 3-CLICK TREE default set up. There are also some comments about functionality which was built-in at the time of the test, but I did not describe to the user in fear of overwhelming them. One of these functionalities is for narrowing and broadening search results, another was to specify 'advanced search' settings like in PubMed.

5.5.4 Usability Conclusions

Overall, PifMed is preferred to PubMed. **Table 5.16** and **Table 5.17** show a mean and a confidence interval that strongly support the determination that on each aspect

of usability PifMed is preferred to PubMed.This is slightly stronger with the General Users than with the Target Users. I expect this is due to the fact that researchers are unwilling to give up on PubMed with only ∼20 minutes of experience with the system. That being said, the overall preference for PifMed — though it is minor — is a significant achievement. Since PifMed is a prototype system, responsiveness is not optimized, I believe this accounts for the lower Efficiency scores. I expected Engagement to be a strong point and the test results indeed show this to be the case.

Usability would increase as users became more familiar with the system. A few user complaints were handled by implemented features, but the users were unaware of them since my brief initial demonstration was constrained to the main functionality, thus too brief to introduce all implemented features. For example, overriding the default tree-state of `3-Click Tree` to `All Closed`; after the test was completed I demonstrated this setting for a user who made this complaint, then asked if it satisfied their complaint and they said it did. Another user complained of an inability to set 'Limits': "*...limit by clinical trials...*" as in PubMed, a few of these limits have been implemented in the MEDLINE Query menu tab (described in Section 4.4.6 and shown in **Figure 4.9**). Limit to clinical trial is one of the available options, so it is fair to say knowledge of this feature would partially satisfy this complaint, however, a selection of limits as broad as that available in PubMed is not available, so I cannot say if this feature as implemented was sufficient to fully satisfy the complaint.

	Effective	Efficient	Engaging	Error Tolerant	Easy to Learn
Mean	5.398	4.667	5.667	5.407	5.568
Std Deviation	1.088	1.171	1.179	0.861	1.016
Sample Variance	1.184	1.881	1.389	0.741	1.033
95%CI from	4.968	4.204	5.200	5.067	5.166
to	5.829	5.130	6.133	5.748	5.970

Table 5.16: Summarized descriptive statistics of the General User population in terms of usability (presented in full in **APPENDIX D.3.1**). The mean and confidence interval for each aspect show PifMed to be slightly preferred to PubMed. Strongest in Engagement and weakest in Efficiency. The overall preference for PifMed is slightly higher here than for the Target Users

	Effective	Efficient	Engaging	Error Tolerant	Easy to Learn
Mean	5.205	4.784	5.477	5.250	5.352
Std Deviation	1.031	0.961	1.220	0.820	0.969
Sample Variance	1.063	0.924	1.488	0.673	0.938
95% CI from	4.747	4.358	4.937	4.886	4.923
to	5.662	5.210	6.018	5.614	5.782

Table 5.17: Summarized descriptive statistics of the TARGET USER population in terms of usability (presented in full in **APPENDIX D.3.1**). The mean and confidence interval for each aspect show PifMed to be slightly preferred to PubMed. Strongest in Engagement and weakest in Efficiency.

The confidence intervals for the means of the aspects of usability are encouraging. This result means that we have statistical evidence that we can expect users to prefer PifMed after only one trial use of the system. If the user suggestions from Part III of the questionnaire were implemented, these ratings would likely increase. These conclusions give further motivation to continue to develop this prototype after completion of this thesis.

5.6 Paired *t*-Tests

We measure the use-time of each system in pairs to afford evaluation using paired t-tests. Paired t-tests have several statistical attributes which strengthen inferences drawn through their use. Since all factors (time of day, individual participant, query, etc...) are the same, except our factor of interest, we can have a great deal of confidence that any difference between the pairs is due to our factor of interest. That is, each participant is their own control, so we can assume any confounding factors (i.e. sources of experimental error) will have equal influence on both recorded values in the pair, except the one we specifically change to evaluate the effect, so this way we can marginalize the influence of confounding factors [24, 87, 56]. Furthermore, this experimental set-up strengthens the assumption of equal variance between the two samples, and any change in the variance (as well as the mean) can be attributed to our factor of interest.

5.6.1 Research Question

Research Question: *In general, is the proposed navigation structure more effective than ranked lists for* MEDLINE*?*

5.6.2 Test Series A: General Users

Test #1: General Users, All Queries, Independence of Outcome Pairs

Null Hypothesis: General Users find articles of interest in MEDLINE with PifMed in an equal or greater amount of time as with PubMed.

Alternate Hypothesis: General Users find articles of interest in MEDLINE with PifMed in significantly less time than with PubMed.

Additional Assumptions: Independence of Outcome Pairs.

Paired t-Test #1	
PubMed Mean	100.00
PifMed Mean	82.17
PubMed Variance	4099.1
PifMed Variance	1830.5
Observations	81
Hypothesized Mean Difference	0
Observed Mean Difference	**17.8272**
Variance of the Differences	5227.2448
df	80
t-Statistic	2.2192
$P(T \leq t)$ one-tail	**0.0147**
$P(T \leq t)$ two-tail	0.02930

Table 5.18: The results of the paired t-test on the GENERAL USER population on all Queries and assuming Independence of queries. We can reject the Null Hypothesis at a p-value of less than 0.015.

We can see in **Table 5.18** that we must reject the Null Hypothesis and accept the Alternate Hypothesis at a p-value of 0.0147.

Test #2: General Users, All Queries, Dependence of Outcome Pairs

Null Hypothesis: General Users find articles of interest in MEDLINE with PifMed in an equal or greater amount of time as with PubMed.

Alternate Hypothesis: General Users find articles of interest in MEDLINE with PifMed in significantly less time than with PubMed.

Additional Assumptions: Dependence of Outcome Pairs.

Paired t-Test #2	
PubMed Mean	100.00
PifMed Mean	82.17
PubMed Variance	1974.5
PifMed Variance	810.0
Observations	27
Hypothesized Mean Difference	0
Observed Mean Difference	**17.83**
Variance of the Differences	2196.3
df	26
t-Statistic	1.9766
$P(T \leq t)$ one-tail	**0.0294**
$P(T \leq t)$ two-tail	0.05878

Table 5.19: The results of the paired t-test on the GENERAL USER population on all Queries and not assuming Independence between Outcome Pairs. We can reject the Null Hypothesis at a p-value of less than 0.03.

We can see in **Table 5.19** that we must reject the Null Hypothesis and accept the Alternate Hypothesis at a p-value of 0.0294.

Test #3: General Users, Long Queries, Independence of Outcome Pairs

Null Hypothesis: General Users find articles of interest in 'large' MEDLINE result sets ($N > 20$) with PifMed in approximately the same amount of time as with PubMed.

Alternate Hypothesis: General Users find articles of interest in 'large' MEDLINE result sets ($N > 20$) with PifMed in significantly less time than with PubMed.

Additional Assumptions: Independence of Outcome Pairs.

Paired t-Test #3	
PubMed Mean	106.62
PifMed Mean	80.18
PubMed Variance	4113.1
PifMed Variance	1729.8
Observations	71
Hypothesized Mean Difference	0
Observed Mean Difference	**26.44**
Variance of the Differences	5143.4
df	70
t-Statistic	3.1061
$P(T \leq t)$ one-tail	**0.00137**
$P(T \leq t)$ two-tail	0.00274

Table 5.20: The results of the paired t-test on the GENERAL USER population on 'Long Queries' and assuming Independence of Outcome Pairs. We can reject the Null Hypothesis at a p-value of less than 0.0015.

We can see in **Table 5.20** that we must reject the Null Hypothesis and accept the Alternate Hypothesis at a p-value of 0.0014.

Test #4: General Users, Long Queries, Dependence of Outcome Pairs

Null Hypothesis: General Users find articles of interest in 'large' MEDLINE result sets ($N > 20$) with PifMed in an equal or greater amount of time as with PubMed.

Alternate Hypothesis: General Users find articles of interest in 'large' MEDLINE result sets ($N > 20$) with PifMed in significantly less time than with PubMed.

Additional Assumptions: Dependence of Outcome Pairs.

Paired t-Test #4	
PubMed Mean	109.5
PifMed Mean	81.12
PubMed Variance	2487.0
PifMed Variance	841.0
Observations	27
Hypothesized Mean Difference	0
Observed Mean Difference	**28.38**
Variance of the Differences	2796.8
df	26
t-Statistic	2.7887
$P(T \leq t)$ one-tail	**0.00488**
$P(T \leq t)$ two-tail	0.00976

Table 5.21: The results of the paired t-test on the GENERAL USER population on 'Long Queries' and not assuming Independence of Outcome Pairs. We can reject the Null Hypothesis at a p-value of less than 0.005.

We can see in **Table 5.21** that we must reject the Null Hypothesis and accept the Alternate Hypothesis at a p-value of 0.0049.

5.6.3 Test Series B: Target Users

Test #5: Target Users, All Queries, Independence of Outcome Pairs

Null Hypothesis: Target Users find articles of interest in MEDLINE with PifMed in an equal or greater amount of time as with PubMed.

Alternate Hypothesis: Target Users find articles of interest in MEDLINE with PifMed in significantly less time than with PubMed.

Additional Assumptions: Independence of Outcome Pairs.

Paired t-Test #5	
PubMed Mean	95.79
PifMed Mean	83.56
PubMed Variance	4754.2
PifMed Variance	1880.2
Observations	66
Hypothesized Mean Difference	0
Observed Mean Difference	**14.58**
Variance of the Differences	5913.3
df	65
t-Statistic	1.5399
$P(T \leq t)$ one-tail	**0.06422**
$P(T \leq t)$ two-tail	0.12844

Table 5.22: The results of the paired t-test on the TARGET USER population on all Queries and assuming Independence of queries. We can reject the Null Hypothesis at a p-value of less than 0.075.

We can see in **Table 5.22** that we could only reject the Null Hypothesis and accept the Alternate Hypothesis at a p-value of 0.0642. This test result does not strongly support this finding, since the p-value is greater than 0.05, which is the conventional threshold of statistical significance.

Test #6: Target Users, All Queries, Dependence of Outcome Pairs

Null Hypothesis: Target Users find articles of interest in MEDLINE with PifMed in an equal or greater amount of time as with PubMed.

Alternate Hypothesis: Target Users find articles of interest in MEDLINE with PifMed in significantly less time than with PubMed.

Additional Assumptions: Dependence of Outcome Pairs.

Paired t-Test #6	
PubMed Mean	95.79
PifMed Mean	81.21
PubMed Variance	2316.9
PifMed Variance	924.9
Observations	22
Hypothesized Mean Difference	0
Observed Mean Difference	**14.58**
Variance of the Differences	2548.8
df	21
t-Statistic	1.3542
$P(T \leq t)$ one-tail	**0.09504**
$P(T \leq t)$ two-tail	0.19007

Table 5.23: The results of the paired t-test on the GENERAL USER population on all Queries and not assuming Independence between Outcome Pairs. We can reject the Null Hypothesis at a p-value of less than 0.1.

We can see in **Table 5.23** that we can only reject the Null Hypothesis and accept the Alternate Hypothesis at a p-value of 0.095. This test result does not strongly support this finding, since the p-value is greater than 0.05, which is the conventional threshold of statistical significance.

Test #7: Target Users, Long Queries, Independence of Outcome Pairs

Null Hypothesis: Target Users find articles of interest in 'large' MEDLINE result sets ($N > 20$) with PifMed in an equal or greater amount of time as with PubMed.

Alternate Hypothesis: Target Users find articles of interest in 'large' MEDLINE result sets ($N > 20$) with PifMed in significantly less time than with PubMed.

Additional Assumptions: Independence of Outcome Pairs.

Paired t-Test #7	
PubMed Mean	102.2
PifMed Mean	78.3
PubMed Variance	4771.1
PifMed Variance	1724.7
Observations	57
Hypothesized Mean Difference	0
Observed Mean Difference	**23.91**
Variance of the Differences	5938.5
df	56
t-Statistic	2.3427
$P(T \leq t)$ one-tail	**0.01136**
$P(T \leq t)$ two-tail	0.02272

Table 5.24: The results of the paired t-test on the TARGET USER population on 'Long Queries' and assuming Independence of Outcome Pairs. We can reject the Null Hypothesis at a p-value of less than 0.015.

We can see in **Table 5.24** that we must reject the Null Hypothesis and accept the Alternate Hypothesis at a p-value of 0.0114.

Test #8: Target Users, Long Queries, Dependence of Outcome Pairs

Null Hypothesis: Target Users find articles of interest in 'large' MEDLINE result sets ($N > 20$) with PifMed in an equal or greater amount of time as with PubMed.

Alternate Hypothesis: Target Users find articles of interest in 'large' MEDLINE result sets ($N > 20$) with PifMed in significantly less time than with PubMed.

Additional Assumptions: Dependence of Outcome Pairs.

Paired t-Test #8	
PubMed Mean	105.73
PifMed Mean	78.92
PubMed Variance	2946.9
PifMed Variance	871.1
Observations	22
Hypothesized Mean Difference	0
Observed Mean Difference	**26.82**
Variance of the Differences	3377.8
df	21
t-Statistic	2.1643
$P(T \leq t)$ one-tail	**0.02105**
$P(T \leq t)$ two-tail	0.04211

Table 5.25: The results of the paired t-test on the GENERAL USER population on 'Long Queries' and not assuming independence of queries. We can reject the Null Hypothesis at a p-value of less than 0.025.

We can see in **Table 5.25** that we must reject the Null Hypothesis and accept the Alternate Hypothesis at a p-value of 0.021.

5.6.4 Statistical Conclusions

The mean use-time for PifMed is always faster in the final analysis, no matter how the data has been partitioned. PifMed is strongest when short search result sets are removed, for both populations ($p = 0.0013$ & $p = 0.0113$). If the reader does not accept the independence of queries, there is still strong statistical significance ($p = 0.0049$ & $p = 0.0211$) to the finding that users can find articles in large result sets approximately 27 seconds faster with a PifMed than with PubMed.

	General Users		Target Users	
Query Independence	All Query	Long Query	All Query	Long Query
Mean Difference	17.8272	26.4366	14.5758	23.9123
$P(T \leq t)$ one-tail	**0.0147**	**0.0014**	0.0642	**0.0114**
No Query Independence				
Mean Difference	17.8272	28.3827	14.5758	26.8182
$P(T \leq t)$ one-tail	**0.0294**	**0.0049**	0.0950	**0.0211**

Table 5.26: A brief summary of the statistical analysis. Statistically significant p-values (p <0.05) are in bold.

	Test #1	Test #2	Test #3	Test #4	Test #5	Test #6	Test #7	Test #8
PubMed Mean	100.00	100.00	106.62	109.50	95.79	95.79	102.23	105.73
PifMed Mean	82.17	82.17	80.18	81.12	83.56	81.21	78.32	78.92
PubMed Var.	4099.1	1974.5	4113.1	2487.0	4754.2	2316.9	4771.1	2946.9
PifMed Var.	1830.5	810.0	1729.8	841.0	1880.2	924.9	1724.7	871.1
Observations	81	27	71	27	66	22	57	22
Mean Diff.	**17.83**	**17.83**	**26.44**	**28.38**	**14.58**	**14.58**	**23.91**	**26.82**
Var. of Diffs	5227.2	2196.3	5143.4	2796.8	5913.3	2548.8	5938.5	3377.8
df	80	26	70	26	65	21	56	21
t-Stat	2.2192	1.9766	3.1061	2.7887	1.5399	1.3542	2.3427	2.1643
p-value	**0.0146**	**0.0294**	**0.0013**	**0.0049**	0.0642	0.0950	**0.0113**	**0.0211**

Table 5.27: A detailed summary of the statistical analysis. The Mean Differences and statistically significant p-values (p <0.05) are in bold.

Furthermore, if speed was negligible, PifMed is still the more usable, since all users rated PifMed as more usable than PubMed. The analysis of the results show PifMed to be faster and a better user experience for this task.

5.7 Limitations

This system is meant for navigating large result sets, thus very short result sets can be seen as a limitation. There is no doubt ranked lists are a more effective means of displaying 20 or less results, these results can be seen to support this. This is an acceptable limitation since this is not the task PifMed is meant to tackle.

One obstacle that need be tackled for the adoption of a novel system like PifMed over a conventional system like PubMed is: human habit and conventional thinking. In general, people like familiarity, and like to use tools that are familiar. Thus its novelty is a deterrent for many users. This limitation is encountered by many new technologies and is mitigated with experience with the system, wider use by colleagues and where possible, any degree of formal training with the system.

This system relies on the efforts of the NLM Indexers. One could see this dependency as a limitation. Should these indexers stop indexing, no new articles could be added to the tree. This is not a certainty. Many publishers (and authors) presently include suggested MeSH categorization meta-data with citation submission to the NLM. Furthermore should the task of indexing be resumed by users, Wikipedia shows high-quality content can be publicly generated.

This system will only function within the confines of MEDLINE citations. This limitation to the medical domain is only a constraint of this implementation, not of the navigational structure itself. Should any other digital library, such as the ACM Portal [31], provide a deep and rigorous hierarchical categorization and make the citations and categorization system publicly available, this system could be quickly adapted to another domain. In fact, preliminary tests were done on the ACM category hierarchy. The category names were found to be too general, and the hierarchy too shallow, to provide usable results.

Many of the limitations of this system can be addressed in a second implementation:

- Lack of Status/Progress bar.

- No Web Version.
- Confusing names and order of buttons.
- No full-text retrieval.
- Categories in citations are not hyper-linked to main category.
- Button to toggle open/close node to small (allow node title to toggle node state).
- Lack of further subdivision of the MeSH categorization (use of MeSH Qualifiers).

Some of the known limitations require more effort:

- Long initial wait time after MEDLINE query.
- No web browser style forward/back buttons.
- Lack of keyword highlighting in the search tree.

A few of the known limitations are systemic and will remain:

- '*I don't like the MeSH categorization*'.
- Reduce initial wait time to equal that of PubMed.

Since PubMed loads only 20 search results to screen at a time, it is unlikely for PifMed to equal PubMed initial load time. If retrieval time from NLM was eliminated (duplicate MEDLINE locally), and if a multi-threaded version was created to eliminate time spent in the initial screen draw (by dynamically adding nodes as they were needed), the subset of articles returned by any given search query would still need to be parsed to create a custom tree. Increased efficiency and optimization is possible, even likely, but the an initial search time will persist for PifMed.

5.8 Conclusion

We have shown users prefer PifMed to PubMed in terms of usability, demonstrated by the analysis of Part II of the questionnaire. It has also been shown that users browsed to interesting results faster with PifMed than with PubMed. Paired t-tests have shown this statistically significant result to have improved significance when small results sets are partitioned out of the data set. This finding strengthens the evidence supporting my hypothesis that use of the hierarchical navigation structure is more effective and efficient then the conventional method for browsing large result sets.

Its success with large result sets is a key finding for 3 reasons:

1. **Query Expansion**. Query expansion returns increases the size of result sets. For query expansion to be viable we need a way for users to adequately navigate those results, PifMed is such a way, thus query expansion can be used in future versions to grow small and medium-sized result sets.

2. **Growth of Corpus**. There is every indication MEDLINE will continue to grow, thus result sets for any given query will continue to grow. However, the rate of growth within Categories growth will be slower, that is, ∼50,000 new articles/week will be categorized across 24,000 categories. The hierarchical categorization mitigates the rate of growth by localizing growth into categories.

3. **Longevity of Queries** Changes in result sets will be focused into categories instead of drastically reorganizing ranked lists. Since no matter how many articles are added, old articles will always remain in their original category: PifMed behaves predictably over time. The hierarchical categorization mitigates the rate of growth by maintaining the legacy of each query over time. The disposable nature of ranking makes PubMed's ranking highly sensitive to growth in the corpus, resulting in a very different top 20 each month for many queries. So re-finding an article from an identical month-old (or year old) query might be difficult; PifMed will be less sensitive to this rate of change.

The reliance on the MeSH Taxonomy was a major concern. Since it is the foundation of this implementation of this novel browsing model, if users rejected it as

navigational structure, the whole project would fail. However, these results show that users accepted and quickly adapted to this largely unfamiliar taxonomy. Evidence of this is shown by the very few complaints about MeSH in the written comments or verbal comments. But the strongest support for the choice to base the navigation system on MeSH comes in the plain fact that users used it quickly and effectively to navigate to interesting results using only MeSH as their guide.

Many of the complaints and suggestions were of missing features in this specific implementation. This shows that the underlying model was satisfactory for both novices and domain experts. Furthermore, the majority of these comments are specific and constructive enough to be easily solved with a second implementation.

Chapter 6

Conclusion

6.1 Research Problem

The problems physicians and researchers face in medical question answering are well-documented, well-studied and detrimental for patients' health. It has been shown nearly all of these questions are answerable with the present resources and when they get answered, it benefits the patient's quality of care. Simply put, physicians do not have the time, nor the expertise, to properly search the enormous amount of medical information that presently holds the answers to their information needs. The rate of growth of information within the medical field will continue to worsen this problem. A position, the Informationist, has been suggested and outlined by those within the EBM community as a solution. The Informationist is an expert at understanding the information needs of clinicians, and at searching the medical corpus for answers.

6.2 Solution

I have suggested a three layer framework, MedicInfoSys, for delegating the task of medical question answering to the Informationist.

We can think of our solution as a puzzle: The Informationist is a piece of the puzzle; PifMed is a piece of the puzzle; Structured Queries are a piece of the puzzle. To solve the obstacles in medical question answering, MedicInfoSys, puts the pieces of the puzzle together. The Structured Query draws out the information need from the end-user and contextualizes it to resolve ambiguities. The Informationist is a medical specialist in information technology and retrieval, trained and tasked with providing medical answers, transparent to their sources. PifMed is a domain specific, expert tool, utilizing mature knowledge-based technologies to enable users to learn and navigate a persistent information landscape. This tool turns a query-and-list website, into a Digital Library where familiar sections are located, browsed and persist

over time.

The interfaces between the layers are as important as the layers themselves, and are also studied in this thesis. The Structured Query, acts as the interface between the End User Layer and the Informationist Layer, drawing out the need and reducing ambiguity. The hierarchical navigation structure, bridges the Informationist Layer and the literature, acting as a map of the literature, akin to sections 'in the stacks' of a physical library, which are browsed, a geography which the user can learn and over time sharpen skills to an expert level.

To pursue the metaphor a little further, every use of the system is like a trip to the library, whose layout is MeSH. With time, the layout becomes known to the patron/user. And with extended use and an effort to understand the layout, users can become experts of the collection, allowing speedy navigation and highly competent referencing. However, by analogy, a list is akin to a list of directions, useless outside its customized context, disposable and momentary.

PifMed does not solve the problem. It is a piece of the solution. PifMed is meant to be used by Informationists. These Informationists are meant to take on the information finding role from clinicians. Clinicians are meant to formulate a query within a structure, PICO, which has been explicitly designed within the EBM community, to draw out the information needs of the physician and put it into a form useful for formulating queries. Informationist take this Structured query information, and with their expertise draw the answers out of the Primary and Secondary sources, then return that information to the physician in a timely manner, and package an efficient, transparently summarized 2 page PDF so the doctor can make informed decisions on the basis of that PDF or use it as a head start in their own research of the literature.

There are two outside perspectives on this service that accurately position it in the mind of the end-user and give a good idea of how it is intended to be used. First, the paradigm in which this service should fit, to be used and properly thought of by physicians is of the same category as a blood test or a biopsy, one which aids medical understanding and assists in the tasks of diagnosis, prognosis, therapy or etiology. The second perspective is to view this as the consultation of a specialist, an information specialist. This is a colleague — an expert in the ways of information

— to listen and help answer questions for physicians, to allow physicians to focus on what they do best: diagnose, treat, and heal patients.

6.3 Implementation

One unique aspect of the knowledge-based resources within the medical domain, is their maturity and accessibility makes their implementation in a system, practical. We have focused on one of these resources (MeSH) to show that they can be effectively used to help people navigate the literature. Only due to the deep integration of this knowledge-based resource into MEDLINE, do we have the ground-work for our search tool prototype, PifMed, which presents search results to the user in a browsable, collapsible tree, based on the MeSH Taxonomy.

Within the MedicInfoSys framework, we have identified the most critical and software-dependant part, then designed and implemented a prototype, PifMed. Along with the implementation we also designed a method of evaluation (a user study), and tested this method with a pilot study. Participant comments and results of the pilot study were used to improve the design of our prototype and strengthen our confidence in our method of evaluation.

6.4 Evaluation

The implemented system was evaluated by a user study, which directly compared PifMed to PubMed. The results of user ratings through a questionnaire and through use-time comparison showed PifMed to be preferable to PubMed, exhibit higher usability in all its aspects (Effective, Efficient, Easy-to-learn, Error Tolerance and Engagement) and perform the task in significantly less time (27 seconds less on average). The performance gap is widened on large result sets.

6.5 Future Work

6.5.1 PifMed Web Version

Some of the limitations found in the user study have been addressed with the implementation of a second version of PifMed. **Figure 6.1** shows the a prototype

web version of PifMed, which was built with a combination of HTML, XML, XSLT, Javascript, Perl and PHP, to address some of the limitations of the original Perl/Tk version. Specifically, lack of web version, status bar (built-in to web browser), interactive hyperlinked category names, click node title to toggle node and full text retrieval. Being web-based, the foundation was coded in HTML, the browsable tree was implemented in Javascript, the back-end search engine remains in Perl, the articles are returned in XML and rendered with XSLT and the whole system is glued together with PHP.

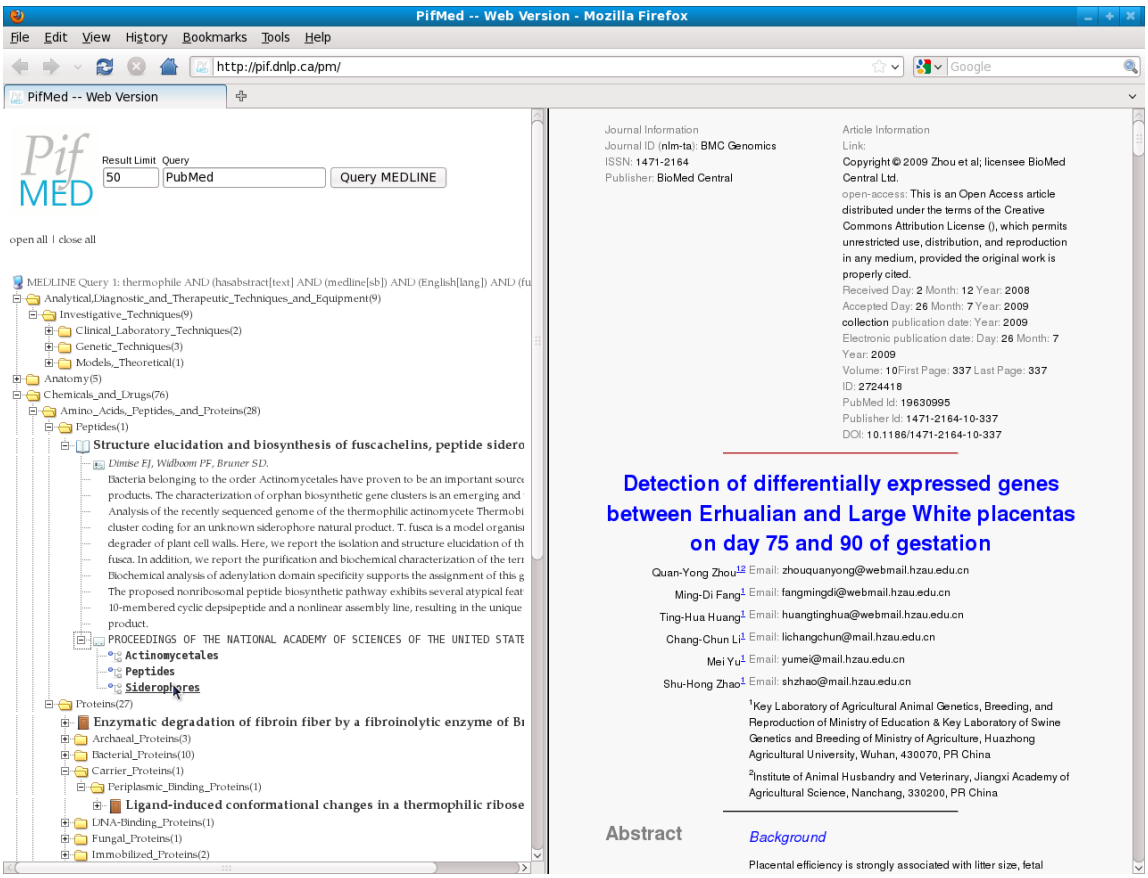


Figure 6.1: A screenshot of the web version of PifMed. The frame on the left shows the browsable tree, the frame on the right shows an article, rendered into HTML from XML retrieved from the NLM using *efetch*.

In the Perl/Tk version a small box with a '+' or '-' must be clicked to toggle open the node. On the advice of users, the whole node title can be clicked to open or close a node. Though only a few users wrote this down as a complaint during the

user study, I witnessed many users clicking title to open them and a few users seems consistently frustrated by this limitation. Since node highlighting seemed of little use, clicked nodes are toggled open or closed and are no longer highlightable.

All the categories in which each article has been placed — available at bottom of each fully open article node — are now hyper-linked to the main category node in the tree. This greatly enhances the usefulness of this information and changes the overall browsing experience of the user. This was not formally tested, but informally, the few users who tried it specifically praised this functionally. However, these links make it more difficult to retrace your steps and so a forward/back (by node) tool is required.

Full PubMedCentral articles are made available in XML format by the NLM. The HTTP tool *efetch* is used to return these articles by PubMedID. The NLM also provides XSL files to optionally render the XML into HTML, Text or PDF formats. The HTML rendered versions (used here) are hyperlinked internally as well as to external sources like charts, figures and other articles. These files are rendered with a PHP version of XSLT. The frame which divides the article and the search results is draggable, so the user may dynamically increase the screen real-estate of the article or the search results as they see fit.

In addition, icons where added before each node to indicate node type: open/close folder (category), open/closed book (article), ID card (author), library catalogue card (bibliographic information) and tree node (other categories). I believe it makes the interface more attractive and interesting looking, whether these icons are actually helpful is unknown.

Work on this prototype is on-going.

6.5.2 Future PifMed Revisions

The implementation and testing of the features outlined in Chapter three: PICO Frame Interface, Generic Query Interface, Query Filtering (Hedge Filters), Query Expansion (Entry Terms and UMLS), User Profiles; additional Primary Sources (other than MEDLINE) and Secondary Sources; Information Extraction fields (PICO), Browse Tools (e.g. Keyword Highlighting) and Suggestions (related articles based on KEEPERS)

6.5.3 Categorization

Though MeSH has been proven to be an effective starting point, a system of categorization better suited to this task may be developed. However, if problems with MeSH are identified, changes can be recommended to the NLM via the suggestion interface on their website [54], so users (i.e. Informationists) may solve problems as they are recognized, with the present evolving system of categorization.

One could ask: *Are the choices made by the human indexers at the NLM good enough?*: Our tests indicate they are sensible and hold up to public scrutiny. With time, as Informationists become more familiar with MeSH and the articles within, they may see problems, or have suggestions for categorization. The NLM has no known, specific mechanism for processing indexing suggestions, but in my correspondence with the NLM, they said:

> "If you believe that your article was indexed incorrectly, i.e. certain MeSH headings are missing and/or certain MeSH headings are used inappropriately, you should contact custserv@nlm.nih.gov, indicating PMID and specific MeSH headings, which, in your opinion, are incorrect. We will then re-examine this specific article, and either correct indexing or explain to you why we think the article was indexed properly."

which it is clear that if specific problems with categorization are found, they will be handled on a case-by-case basis by the NLM. (see **Appendix E.1** for full correspondence)

6.5.4 MedicInfoSys Implementation

The End-User Layer / Informationist Layer interface needs to build implemented and tested. This would be a major undertaking. It would require: the recruitment of test populations of Informationists and End Users; the implementation of the anonymized database, servers, accounts, and search interface (PifMed); the standardization of the documents meant for the anonymized collection and the development of a tool to create them; the development of the query communication interface between Informationist and end-user and methods of feedback for clarification of end-user information needs.

The PICO model does have its limitations which are discussed in Section 2.2.1. Though the Structured Query can act as a starting point for any dialogue to negotiate an understanding of the information need, there will likely be incremental improvements that can be made to the Structured Query over time. Certainly not all questions can be foreseen, but as problems arise, Informationists can be polled for suggestions to this interface.

A system to make the best use of the human knowledge worker in support of physicians is at the heart of this work. One could postulate, that some day these knowledge workers could be replaced by an AI agent, presently and for the foreseeable future we still need trained and knowledgeable people to perform these tasks. Still this system is needed to meet the needs of medical professionals which are short on time, but in the future, the Informationist Layer could be phased-out.

Bibliography

[1] Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 17–21, 2001. PMID: 11825149.

[2] Alan R Aronson, Allen Browne, Anantha Bangalore, Betsy Humphreys, Brian Carlsen, Chris Lu, David Sheretz, Guy Divita, Karen Thorn, Kin Wah Fung, Laura Roth, Mark Tuttle, Olivier Bodenreider, Stephanie Lipow, Steve Emrick, Stuart Nelson, Suresh Srinivasan, Tammy Powell, Tom Rindflesch, Vivian Auld, and William Hole. Umls basics, 2008. http://www.nlm.nih.gov/research/umls/pdf/UMLS_Basics.pdf. Last access July 2010.

[3] Alan R Aronson, James G Mork, Clifford W Gay, Susanne M Humphrey, and Willie J Rogers. The nlm indexing initiative's medical text indexer. *Studies in Health Technology and Informatics*, 107:268–72, 2004. PMID: 15360816.

[4] Global WordNet Association. The global wordnet assocation (gwa), 2009. http://www.globalwordnet.org/. Last access July 2010.

[5] Medical Library Association. Role of expert searching in health sciences libraries. *Journal of the Medical Library Association*, 93(1):42–44, January 2005. PMID: 15685273, PMCID: 545120.

[6] David Atkins, Dana Best, Peter A Briss, Martin Eccles, Yngve Falck-Ytter, Signe Flottorp, Gordon H Guyatt, Robin T Harbour, Margaret C Haugh, David Henry, Suzanne Hill, Roman Jaeschke, Gillian Leng, Alessandro Liberati, Nicola Magrini, James Mason, Philippa Middleton, Jacek Mrukowicz, Dianne O'Connell, Andrew D Oxman, Bob Phillips, Holger J Schnemann, Tessa Tan-Torres Edejer, Helena Varonen, Gunn E Vist, John W Williams, and Stephanie Zaza. Grading quality of evidence and strength of recommendations. *BMJ (Clinical Research Ed.)*, 328:1490, June 2004. PMID: 15205295.

[7] [No authors listed]. Sort: the strength-of-recommendation taxonomy. *American Family Physician*, 71(1):19–20, 2005. PMID: 15666558.

[8] Suzanne Bakken. An informatics infrastructure is essential for evidence-based practice. *Journal of the American Medical Informatics Association : JAMIA*, 8:199–201, 2001. PMID: 11320064.

[9] Peter A. Bonis, Gary T. Pickens, David M. Rind, and David A. Foster. Association of a clinical knowledge support system with improved patient safety, reduced complications and shorter length of stay among medicare beneficiaries

in acute care hospitals in the united states. *International Journal of Medical Informatics*, In Press, Corrected Proof.

[10] BioMed Central. Biomed central | the open access publisher, September 2008. http://www.biomedcentral.com/. Last access July 2010.

[11] PubMed Central. Pubmed central homepage, April 2009. http://www.ncbi.nlm.nih.gov/pmc/. Last access July 2010.

[12] Jeffrey T Chang, Hinrich Schtze, and Russ B Altman. Creating an online dictionary of abbreviations from medline. *Journal of the American Medical Informatics Association: JAMIA*, 9:612–20, December 2002. PMID: 12386112.

[13] Krzysztof J. Cios and G. William Moore. Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, 26:1–24, 2002.

[14] The Cochrane Collaboration. Cochrane centres and branches, December 2008. http://www.cochrane.org/contact/centresmap_new0507.htm. Last access July 2010.

[15] The Cochrane Collaboration. The cochrane collaboration - about the cochrane collaboration, 2010. http://www.cochrane.org/docs/descrip.htm. Last access July 2010.

[16] Ben Comer. Docs look to wikipedia for condition info: Manhattan research, April 2009. http://www.mmm-online.com/docs-look-to-wikipedia-for-condition-info-manhattan-research/article/131038/. Last Visited 2010.

[17] Herma C H Coumou and Frans J Meijman. How do primary care physicians seek answers to clinical questions? a literature review. *Journal of the Medical Library Association: JMLA*, 94:55–60, 2006. PMID: 16404470.

[18] Andras Csomai. Wordnet bibliography. http://lit.csci.unt.edu/%7Ewordnet/. Last access July 2010.

[19] F Davidoff and V Florance. The informationist: a new health profession? *Annals of Internal Medicine*, 132:996–8, June 2000. PMID: 10858185.

[20] Martin Dawes, Pierre Pluye, Laura Shea, Roland Grad, Arlene Greenberg, and Jian-Yun Nie. The identification of clinically important elements within medical journal abstracts: Patient-population-problem, exposure-intervention, comparison, outcome, duration and results (pecodr). *Informatics in Primary Care*, 15:9–16, 2007.

[21] Dina Demner-Fushman, Susan Hauser, and George Thoma. The role of title, metadata and abstract in identifying clinically relevant journal articles. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, pages 191–5, 2005. PMID: 16779028.

[22] Dina Demner-Fushman and Jimmy Lin. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. pages 841–848, Sydney, Australia, 2006. Association for Computational Linguistics.

[23] Dina Demner-Fushman and Jimmy Lin. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33:63–103, 2007.

[24] Jay L. Devore. *Probability and Statistics for Engineering and the Sciences*. Duxbury Press, 6th edition, 2004.

[25] DynaMed. Content sources, 2009. http://www.ebscohost.com/dynamed/sources.php. Last Visited 2010.

[26] Mark H Ebell, Jay Siwek, Barry D Weiss, Steven H Woolf, Jeffrey Susman, Bernard Ewigman, and Marjorie Bowman. Strength of recommendation taxonomy (sort): a patient-centered approach to grading evidence in the medical literature. *American Family Physician*, 69:548–56, February 2004. PMID: 14971837.

[27] John W Ely, Jerome A Osheroff, M Lee Chambliss, Mark H Ebell, and Marcy E Rosenbaum. Answering physicians' clinical questions: obstacles and potential solutions. *Journal of the American Medical Informatics Association: JAMIA*, 12:217–24, 2005. PMID: 15561792.

[28] John W Ely, Jerome A Osheroff, Mark H Ebell, G R Bergus, B T Levy, M Lee Chambliss, and E R Evans. Analysis of questions asked by family doctors regarding patient care. *BMJ (Clinical Research Ed.)*, 319:358–61, August 1999. PMID: 10435959.

[29] John W Ely, Jerome A Osheroff, Mark H Ebell, M Lee Chambliss, Daniel C Vinson, James J Stevermer, and Eric A Pifer. Obstacles to answering doctors' questions about patient care with evidence: qualitative study. *BMJ (Clinical Research Ed.)*, 324:710, March 2002. PMID: 11909789.

[30] John W Ely, Jerome A Osheroff, P N Gorman, Mark H Ebell, M Lee Chambliss, Eric A Pifer, and P Z Stavri. A taxonomy of generic clinical questions: classification study. *BMJ (Clinical Research Ed.)*, 321:429–32, August 2000. PMID: 10938054.

[31] Association for Computing Machinery. The ACM Portal, 2007. http://portal.acm.org/portal.cfm. Last visited July 2010.

[32] Mark A. Graber, Bradley D. Randles, John W. Ely, and Jay Monnahan. Answering clinical questions in the ed. *The American Journal of Emergency Medicine*, 26:144–147, February 2008.

[33] David A. Grimes, Melody Y. Hou, Laureen M. Lopez, and Kavita Nanda. Do clinical experts rely on the cochrane library? *Obstet Gynecol*, 111:420–422, February 2008.

[34] Lisa Grossman. Should you trust health advice from the web? *New Scientist*, July 2009. http://www.newscientist.com/article/mg20327185.500-should-you-trust-health-advice-from-the-web.html?full=true. Last Visited 2010.

[35] Thomas R. Gruber. Towards principles for the design of ontologies used for knowledge sharing. Kluwer Academic Publishers, Deventer, The Netherlands, 1993.

[36] Roger Hale. Text mining: getting more value from literature resources. *Drug Discovery Today*, 10:377–379, March 2005.

[37] Xiaoli Huang, Jimmy Lin, and Dina Demner-Fushman. Evaluation of pico as a knowledge representation for clinical questions. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, pages 359–63, 2006. PMID: 17238363.

[38] B L Humphreys, D A Lindberg, H M Schoolman, and G O Barnett. The unified medical language system: an informatics research collaboration. *Journal of the American Medical Informatics Association: JAMIA*, 5:1–11, 1998. PMID: 9452981.

[39] Lawrence Hunter and K. Bretonnel Cohen. Biomedical language processing: What's beyond pubmed? *Molecular Cell*, 21:589–594, March 2006.

[40] Nicholas C Ide, Russell F Loane, and Dina Demner-Fushman. Essie: a concept-based search engine for structured biomedical text. *Journal of the American Medical Informatics Association: JAMIA*, 14(3):253–63, June 2007. PMID: 17329729.

[41] Wiley InterScience. Wiley InterScience: Reference Work: The Cochrane Library 2008, Issue 3, 2008. http://www3.interscience.wiley.com/cgi-bin/mrwhome/106568753/ProductDescriptions.html?CRETRY=1&SRETRY=0. Last access July 2010.

[42] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice Hall Inc, New Jersey, NJ, USA, 1 edition, 2000.

[43] M S Klein, F V Ross, D L Adams, and C M Gilbert. Effect of online literature searching on length of stay and patient care costs. *Academic Medicine: Journal of the Association of American Medical Colleges*, 69:489–95, June 1994. PMID: 8003169.

[44] Jimmy Lin and Dina Demner-Fushman. Semantic clustering of answers to clinical questions. In *AMIA Annual Symposuim*, pages 458–62, Chicago, October 2007. PMID: 18693878.

[45] Ritch Macefield. Usability studies and the hawthorne effect. *Journal of Usability Studies*, 2(3):145–154, 2007.

[46] UpToDate Marketing. Uptodate inc., 2009. http://www.uptodate.com/ home-/about/index.html. Last access July 2010.

[47] Elton Mayo. *The human problems of an industrial civilisation*. The Macmillan Company, New York, 1933.

[48] Alex T McCray, Alan R Aronson, A C Browne, T C Rindflesch, A Razi, and S Srinivasan. Umls knowledge for biomedical language processing. *Bulletin of the Medical Library Association*, 81:184–94, April 1993. PMID: 8472004.

[49] Alex T McCray and N C Ide. Design and implementation of a national clinical trials registry. *Journal of the American Medical Informatics Association: JAMIA*, 7(3):313–23, June 2000. PMID: 10833169.

[50] Emma Meats, Jon Brassey, Carl Heneghan, and Paul Glasziou. Using the turning research into practice (trip) database: how do clinicians really search? *Journal of the Medical Library Association*, 95, April 2007. PMC1852632.

[51] National Library of Medicine (US). Data news and update information. http://www.nlm.nih.gov/bsd/revup/revup_pub.html. Last access 2010.

[52] National Library of Medicine (US). Fact Sheet – MEDLINE. http:// www.nlm.nih.gov/pubs/factsheets/medline.html. Last access 2010.

[53] National Library of Medicine (US). Fact Sheet: Medical Subject Headings – MeSH. http://www.nlm.nih.gov/pubs/factsheets/mesh.html. Last access 2010.

[54] National Library of Medicine (US). Suggestion for medical subject headings change, November 2001. http://www.nlm.nih.gov/mesh/meshsugg.html. Last access July 2010.

[55] National Library of Medicine (US). Fact Sheet: SPECIALIST Lexicon, March 2006. http://www.nlm.nih.gov/pubs/factsheets/umlslex.html. Last access July 2010.

[56] Government of Alberta. Guide to field experimentation: Experimental design, January 2010. http://www1.agric.gov.ab.ca/ $department/deptdocs.nsf /all/sag3024. Last visited July 2010.

[57] PLoS Board of Directors. Public library of science. http://www.plos.org/. Last Visited 2010.

[58] National Institute of Health (US). Clinicaltrials.gov, August 2008. http://clinicaltrials.gov/.

[59] National Library of Medicine (US). Efetch entrez utility, 2005. http://www.ncbi.nlm.nih.gov/corehtml/query/static/efetch_help.htmlTool. Last access July 2010.

[60] National Library of Medicine (US). Board of regents minutes - september 2007, September 2007. http://www.nlm.nih.gov/od/bor/9-07bor.pdf. Last access July 2010.

[61] National Library of Medicine (US). Board of regents minutes - february 2008, February 2008. http://www.nlm.nih.gov/od/bor/2-08bor.pdf. Last access July 2010.

[62] National Library of Medicine (US). Introduction to the UMLS – UMLS Reference Manual. *NCBI Bookshelf*, September 2009. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=nlmumls&part=ch01. Last access July 2010.

[63] National Library of Medicine (US). Key medline indicators, December 2009. http://www.nlm.nih.gov/bsd/bsd_key.html. Last access July 2010.

[64] National Library of Medicine (US). Pubmed clinical queries, February 2009. http://www.ncbi.nlm.nih.gov/entrez/query/static/clinical.shtml. Last Visited 2009.

[65] National Library of Medicine (US). Semantic Network – UMLS Reference Manual. *NCBI Bookshelf*, September 2009. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=nlmumls&part=ch05. Last access July 2010.

[66] National Library of Medicine (US). Statistics - 2009ab release, November 2009. http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/ release/statistics.html. Last access July 2010.

[67] National Library of Medicine (US). UMLS - Metathesaurus License Agreement Appendix, November 2009. http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/ release/license_agreement_appendix.html. Last access July 2010.

[68] National Library of Medicine (US). Data, news and update information, January 2010. http://www.nlm.nih.gov/bsd/revup/revup_pub.htmlmed_update. Last Visited 2010.

[69] National Library of Medicine (US). Number of titles currently indexed for index medicus and medline on pubmed, July 2010. http://www.nlm.nih.gov/bsd/num_titles.html. Last Visited 2010.

[70] OneLook. Onelook dictionary search, January 2010. http://www.onelook.com/about.shtml. Last access July 2010.

[71] Bob Phillips and Chris Ball. Levels of evidence, May 2007. http://www.cebm.net/index.aspx?o=1025. Last access July 2010.

[72] W S Richardson, M C Wilson, J Nishikawa, and R S Hayward. The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, 123:A12–3, November 1995. PMID: 7582737.

[73] Connie Schardt, Martha B Adams, Thomas Owens, Sheri Keitz, and Paul Fontelo. Utilization of the pico framework to improve searching pubmed for clinical questions. *BMC Medical Informatics and Decision Making*, 7, 2007. PMC1904193.

[74] Ralf W. Schlosser, Rajinder Koul, and John Costello. Asking well-built questions for evidence-based practice in augmentative and alternative communication. *Journal of Communication Disorders*, 40:225–238, 2007.

[75] Stacey D. Scott, Neal Lesh, and Gunnar W. Klau. Investigating human-computer optimization. pages 155–162, Minneapolis, Minnesota, USA, 2002. ACM.

[76] Ryan Singel. Asylum-seeker rejected based on wikipedia, appeals court reverts. *Wired*, September 2008. http://blog.wired.com/27bstroke6/2008/09/asylum-seeker-r.html. Last Visited 2010.

[77] Adam Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Public Domain, March 1776.

[78] Richard Smith. What clinical information do doctors need? *BMJ (Clinical Research Ed.)*, 313(7064):1062–1068, October 1996. PMID: 8898602.

[79] Charles A. Sneiderman, Dina Demner-Fushman, Marcelo Fiszman, Nicholas C. Ide, and Thomas C. Rindflesch. Knowledge-based methods to help clinicians find answers in medline. *J Am Med Inform Assoc*, 14:772–780, November 2007.

[80] C. E. Snow. Research on industrial illumination: A discussion of the relation of illumination intensity to productive efficiency. *The Tech Engineering News*, pages 257–82, 1927.

[81] Irena Spasic, Sophia Ananiadou, John McNaught, and Anand Kumar. Text mining and ontologies in biomedicine: Making sense of raw text. *Brief Bioinform*, 6:239–251, 2005.

[82] Debbie L. Stone, Caroline Jarrett, Mark Woodroffe, and Shailey Minocha. *User Interface Design and Evaluation*. Morgan Kaufmann, 1st edition, 2005.

[83] The Cochrane Collaboration Web Team. The Cochrane Collaboration - Cochrane entities, 2010. http://www.cochrane.org/contact/entities.htm#CRGLIST.

[84] Princeton University. WordNet - About WordNet, 2009. http:// wordnet.princeton.edu/wordnet/. Last access July 2010.

[85] Princeton University. WordNet 3.0 database statistics, 2010. http:// wordnet.princeton.edu/ wordnet/man/wnstats.7WN.html. Last access July 2010.

[86] Johanna I Westbrook, Enrico W Coiera, and A Sophie Gosling. Do online information retrieval systems help experienced clinicians answer clinical questions? *Journal of the American Medical Informatics Association: JAMIA*, 12(3):315–21, 2005. PMID: 15684126.

[87] Wikipedia. Paired difference test - Wikipedia, the free encyclopedia, May 2009. http://en.wikipedia.org/wiki/Paired_difference_test. Last visited July 2010.

[88] Wikipedia. PubMed - Wikipedia, the free encyclopedia, December 2009. http://en.wikipedia.org/wiki/PubMed. Last access July 2010.

[89] N L Wilczynski, K A McKibbon, and R B Haynes. Enhancing retrieval of best evidence for health care from bibliographic databases: calibration of the hand search of the literature. *Medinfo. MEDINFO*, 10:390–3, 2001. PMID: 11604770.

[90] The GRADE working group. Grade working group, 2009. http://www.gradeworkinggroup.org/intro.htm. Last access July 2010.

[91] Hong Yu, Minsuk Lee, David Kaufman, John Ely, Jerome A. Osheroff, George Hripcsak, and James Cimino. Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. *Journal of Biomedical Informatics*, 40:236–251, June 2007.

[92] Pierre Zweigenbaum. Question answering in biomedicine. In *Proc. EACL2003, workshop on NLP for Question Answering*, Budapest, 2003.

Appendix A

Author's Note

Notes on Typographic Style

In this section I would like to take the opportunity to clarify my typographical choices for the reader.

Use of Small Caps Font

Small Caps font is used for acronyms/abbreviations over 2 letters, titles of public statement, relation names, function names and government bills.

Use of Bold Font

Bold font is used for Table, Appendix, Figure and Equation names mentioned in the main body text. Bold is also used for emphasis and when defining a term.

Use of Sans Serif Font

Sans Serif font is used for concept and category names.

Use of Teletype Font

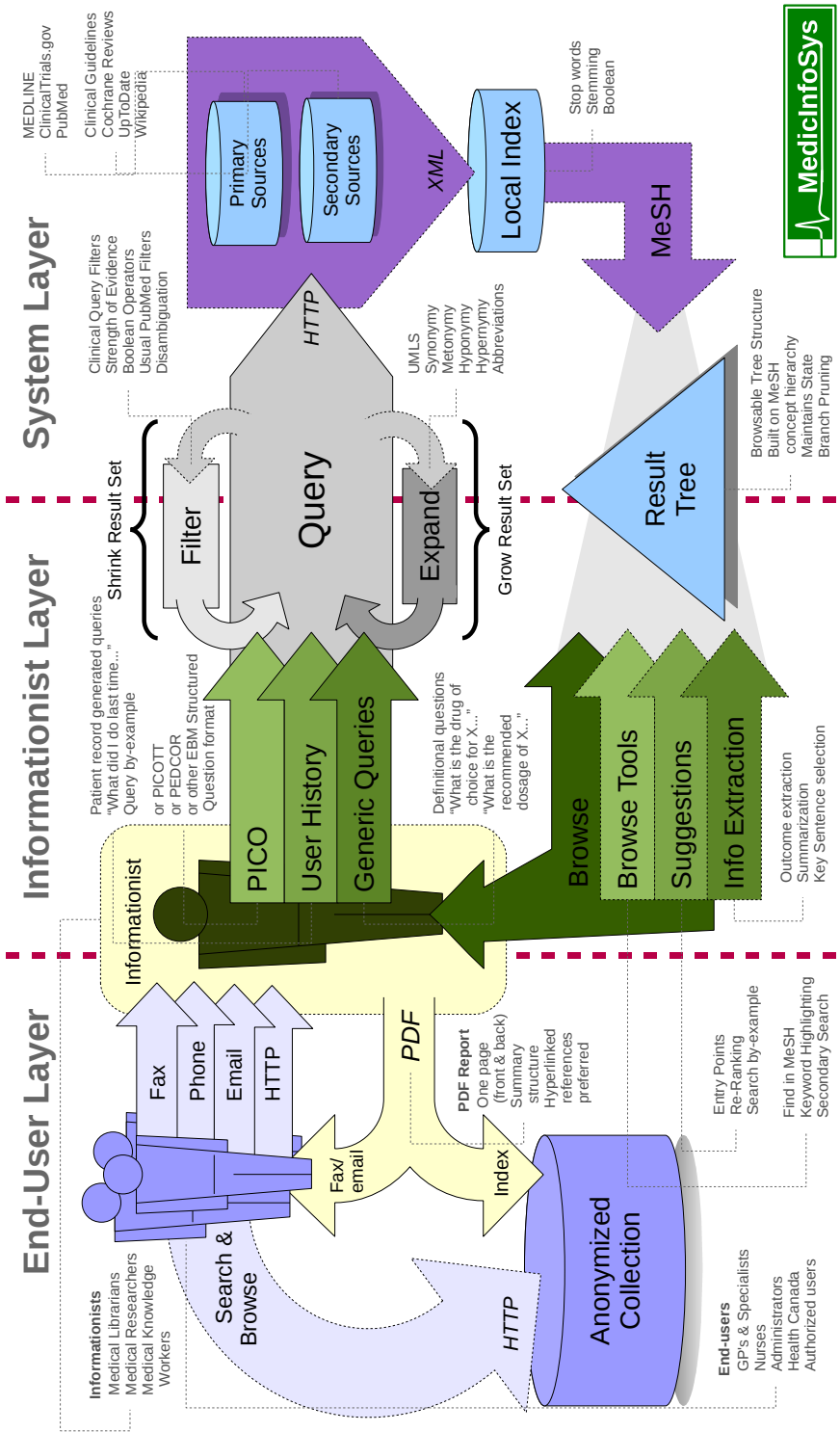
`Teletype` font is used for code, field names, computer input and computer output.

Use of Italic Font

Italic font is used for for questions, long quotes, Latin words and user comments. Italic is occasionally used for emphasis.

Appendix B

Annotated MedicInfoSys Diagram


End-User Layer
Informationist Layer
System Layer
MedicInfoSys
MEDLINE
ClinicalTrials.gov
PubMed
Clinical Guidelines
Cochrane Reviews
UpToDate
Wikipedia
Primary Sources
Secondary Sources
XML
Local Index
Stop words
Stemming
Boolean
MeSH
Clinical Query Filters
Strength of Evidence
Boolean Operators
Usual PubMed Filters
Disambiguation
HTTP
UMLS
Synonymy
Metonymy
Hyponymy
Hypernymy
Abbreviations
Shrink Result Set
Filter
Query
Expand
Grow Result Set
Result Tree
Browsable Tree Structure
Built on MeSH
concept hierarchy
Maintains State
Branch Pruning
Patient record generated queries
"What did I do last time..."
Query by-example
or PICOTT
or PEDCOR
or other EBM Structured
Question format
PICO
User History
Generic Queries
Definitional questions
"What is the drug of
choice for X..."
"What is the
recommended
dosage of X..."
Browse
Browse Tools
Suggestions
Info Extraction
Outcome extraction
Summarization
Key Sentence selection
Informationist
Fax
Phone
Email
HTTP
Fax/
email
PDF
PDF Report
One page
(front & back)
Summary
structure
Hyperlinked
references
preferred
Index
Entry Points
Re-Ranking
Search by-example
Find in MeSH
Keyword Highlighting
Secondary Search
Search &
Browse
HTTP
Anonymized
Collection
Informationists
Medical Librarians
Medical Researchers
Medical Knowledge
Workers
End-users
GP's & Specialists
Nurses
Administrators
Health Canada
Authorized users

Appendix C

MeSH Qualifier List

abnormalities (AB)
administration & dosage (AD)
adverse effects (AE)
agonists (AG)
analogs & derivatives (AA)
analysis (AN)
anatomy & histology (AH)
antagonists & inhibitors (AI)
biosynthesis (BI)
blood (BL)
blood supply (BS)
cerebrospinal fluid (CF)
chemical synthesis (CS)
chemically induced (CI)
chemistry (CH)
classification (CL)
complications (CO)
congenital (CN)
contraindications (CT)
cytology (CY)
deficiency (DF)
diagnosis (DI)
diagnostic use (DU)
diet therapy (DH)
drug effects (DE)
drug therapy (DT)
economics (EC)
education (ED)
embryology (EM)
enzymology (EN)
epidemiology (EP)
ethics (ES)
ethnology (EH)
etiology (ET)
genetics (GE)
growth & development (GD)
history (HI)
immunology (IM)
injuries (IN)
innervation (IR)
instrumentation (IS)
legislation & jurisprudence (LJ)
manpower (MA)
metabolism (ME)
methods (MT)
microbiology (MI)
mortality (MO)
nursing (NU)
organization & administration (OG)
parasitology (PS)
pathogenicity (PY)
pathology (PA)
pharmacokinetics (PK)
pharmacology (PD)
physiology (PH)
physiopathology (PP)
poisoning (PO)
prevention & control (PC)
psychology (PX)
radiation effects (RE)
radiography (RA)
radionuclide imaging (RI)
radiotherapy (RT)
rehabilitation (RH)
secondary (SC)
secretion (SE)
standards (ST)
statistics & numerical data (SN)
supply & distribution (SD)
surgery (SU)
therapeutic use (TU)
therapy (TH)
toxicity (TO)
transmission (TM)
transplantation (TR)
trends (TD)
ultrasonography (US)
ultrastructure (UL)
urine (UR)
utilization (U)
veterinary (VE)
virology (VI) isolation & purification (IP)

Appendix D

User Study

D.1 Questionnaire

D.1.1 Part I: Population Identification

```
 1. Please rate your level of familiarity with:

================================================================
a. Computers
 ---------------------------------------------------------
(Never used one)->  1  2  3  4  5  6  7  <-(use everyday)

================================================================
b. Computer Search
 ---------------------------------------------------------
(Never done one)->  1  2  3  4  5  6  7  <-(use everyday)

================================================================
c. MeSH
 ---------------------------------------------------------
(Never heard of it)->  1  2  3  4  5  6  7  <-(use everyday)

================================================================
d. PubMed
 ---------------------------------------------------------
(Never heard of it)->  1  2  3  4  5  6  7  <-(use everyday)

================================================================
```

e. Medicine

```
 ------------------------------------------------------------
(I know very little)->  1  2  3  4  5  6  7  <-(medical degree)
```

==========

f. Biology

```
 ------------------------------------------------------------
(I know very little)->  1  2  3  4  5  6  7  <-(biology degree)
```

==========

g. Psychology

```
 ------------------------------------------------------------
(I know very little)->  1  2  3  4  5  6  7  <-(psyc. degree)
```

==========

h. Health Informatics

```
 ------------------------------------------------------------
(I know very little)->  1  2  3  4  5  6  7  <-(HINF Student)
```

==========

i. Bioinformatics

```
 ------------------------------------------------------------
(I know very little)->  1  2  3  4  5  6  7  <-(BioINF Student)
```

D.1.2 Part II: System Ratings

2. For each of these questions you are first asked to rate MeSHLINE and then to compare MeSHLINE to PubMed. Reference the following chart as a guide for the comparison of the two systems:

Guide for comparison of PubMed vs. MeSHLINE

---

1 = Always use PubMed for this reason

2 = Usually use PubMed for this reason

3 = Liked PubMed one a little more for this reason

4 = No Preference. Each performed equally in this regard

5 = Liked MeSHLINE one a little more for this reason

6 = Usually use MeSHLINE for this reason

7 = Always use MeSHLINE for this reason

---

a. Effectiveness: Did MeSHLINE give you relevant results?

(no relevant)-> 1 2 3 4 5 6 7 <-(all relevant)

(preferred PubMed)-> 1 2 3 4 5 6 7 <-(preferred MeSHLINE)

b. Efficiency: Did MeSHLINE respond quickly?

(waay to slow)-> 1 2 3 4 5 6 7 <-(very responsive)

(preferred PubMed)-> 1 2 3 4 5 6 7 <-(preferred MeSHLINE)

c. Engaging: Did MeSHLINE encourage you to explore the results?

(Frustrating)-> 1 2 3 4 5 6 7 <-(very interesting)

(preferred PubMed)-> 1 2 3 4 5 6 7 <-(preferred MeSHLINE)

d. Error Tolerance: Did you notice any errors in MeSHLINE?

(Full of bugs)-> 1 2 3 4 5 6 7 <-(found no errors)

(preferred PubMed)-> 1 2 3 4 5 6 7 <-(preferred MeSHLINE)

e. Easy to learn: Was MeSHLINE easy to learn and understand?

(I'm still confused)-> 1 2 3 4 5 6 7 <-(I got it right away)

(preferred PubMed)-> 1 2 3 4 5 6 7 <-(preferred MeSHLINE)

f. Effectiveness: Did the MeSHLINE help you make up your mind

on what you were looking for (help you focus query)?
(not at all)-> 1 2 3 4 5 6 7 <-(very much so)
(preferred PubMed)-> 1 2 3 4 5 6 7 <-(preferred MeSHLINE)

g. Error Tolerance: How would you rank your confidence in the results?
(missing a lot)-> 1 2 3 4 5 6 7 <-(complete)
(preferred PubMed)-> 1 2 3 4 5 6 7 <-(preferred MeSHLINE)

h. Easy to learn: Do you feel like you have a good understanding
of the capabilities of MeSHLINE?
(no idea)-> 1 2 3 4 5 6 7 <-(complete)
(preferred PubMed)-> 1 2 3 4 5 6 7 <-(preferred MeSHLINE)

i. Effectiveness: Rate the ease (or difficulty) in retracing your steps.
(Frustrating)-> 1 2 3 4 5 6 7 <-(Simple)
(preferred PubMed)-> 1 2 3 4 5 6 7 <-(preferred MeSHLINE)

j. Engaging: Did MeSHLINE help you browse to interesting papers
you did not expect?
(Never)-> 1 2 3 4 5 6 7 <-(Every time)
(preferred PubMed)-> 1 2 3 4 5 6 7 <-(preferred MeSHLINE)

k. Overall: Please rate the ease of using MeSHLINE overall.
(Frustrating)-> 1 2 3 4 5 6 7 <-(Intuitive)
(preferred PubMed)-> 1 2 3 4 5 6 7 <-(preferred MeSHLINE)

D.1.3 Part III: Comments

3. What was your least favorite feature of MeSHLINE?

4. What was your favorite feature of MeSHLINE?

5. Suggestions, Improvements, Comments:

D.2 Results

D.2.1 Full Results from Part I & II of the Questionnaire.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	Min	Med	Max	Ave	Questions
7	7	7	7	7	7	7	7	7	7	7	5	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	5	7	7	6.9	1.a) Computers
7	7	7	7	6	7	7	7	7	7	6	5	7	7	6	7	7	7	7	7	7	7	7	7	7	7	6	5	7	7	6.8	1.b) Computer Search
1	2	2	4	6	3	2	4	1	5	5	1	5	4	1	1	3	1	4	3	1	1	1	1	1	7	5	1	2	7	2.8	1.c) MeSH
2	2	1	1	6	4	3	5	1	5	6	2	7	7	5	1	3	5	7	3	4	3	1	3	1	7	6	1	3	7	3.7	1.d) PubMed
2	3	4	1	6	2	4	3	3	3	4	2	5	7	3	2	1	5	6	2	3	2	3	4	4	7	5	1	3	7	3.6	1.e) Medicine
2	4	4	3	2	3	4	2	3	4	4	2	7	2	4	1	2	7	7	2	3	4	4	1	4	5	3	1	3	7	3.4	1.f) Biology
2	3	3	1	2	1	4	4	4	2	4	2	2	1	4	1	1	7	3	2	2	3	1	3	6	6	1	1	2	7	2.8	1.g) Psychology
1	5	3	1	4	4	5	3	2	4	3	2	2	5	2	1	1	5	4	5	2	5	5	1	2	7	7	1	3	7	3.4	1.h) Health Informatics
2	1	3	1	4	3	3	3	2	5	3	1	1	2	2	1	2	4	7	5	2	5	6	1	2	6	5	1	3	7	3.0	1.i) Bioinformatics
6	6	6	4	6	4	5	6	4	7	7	6	7	6	3	7	6	4	6	6	5	6	4	6	3	6	4	3	6	7	5.4	2.a) **Effectiveness:**
6	6	5	5	4	5	6	7	4	4	5	6	6	3	5	7	7	6	6	6	6	6	4	5	2	4	5	2	5	7	5.2	Did PifMed give ...?
3	5	6	3	3	3	4	5	6	6	4	4	6	6	2	4	6	3	4	5	3	6	2	7	1	3	4	1	4	7	4.2	2.b) **Efficiency:**
3	6	7	1	5	3	4	4	6	3	2	2	5	4	3	4	7	3	3	3	6	4	1	4	1	3	4	1	4	7	3.7	Did PifMed respond ...?
7	7	7	5	7	5	6	6	4	7	7	6	7	7	6	7	7	6	5	7	5	6	1	7	3	6	6	1	6	7	5.9	2.c) **Engaging:**
5	6	4	6	5	6	6	6	3	6	7	7	6	7	5	7	7	7	7	7	4	7	1	7	2	5	6	1	6	7	5.6	Did PifMed encourage ...?
7	7	7	6	6	5	6	7	7	7	7	7	7	7	7	7	7	6	7	7	6	6	7	4	6	7	7	4	7	7	6.6	2.d) **Error Tolerance:**
4	6	7	4	5	5	4	4	4	5	4	5	7	4	4	7	7	6	4	4	4	4	4	4	4	4	4	4	4	7	4.7	Did you notice ...?
6	7	7	3	6	7	3	5	5	7	7	5	7	7	6	7	7	7	5	7	4	6	6	7	7	7	7	3	7	7	6.1	2.e) **Easy to learn:**
3	7	5	2	5	6	3	3	3	5	5	7	7	4	3	7	7	6	4	4	6	6	4	7	4	4	6	2	5	7	4.9	Was PifMed easy ...?
7	5	7	5	5	6	7	6	4	7	5	7	7	4	1	7	7	4	6	7	3	4	6	7	3	6	5	1	6	7	5.5	2.f) **Effectiveness:**
7	6	5	4	5	5	7	6	4	6	4	7	7	2	4	7	7	4	6	7	5	5	6	7	4	5	6	2	6	7	5.5	Did PifMed help ...?
7	6	6	4	5	4	2	6	5	6	6	6	6	5	6	6	7	6	6	6	5	6	4	4	2	6	5	2	6	7	5.3	2.g) **Error Tolerance:**
6	6	6	6	3	5	4	6	4	4	6	6	7	4	4	6	7	6	7	4	5	7	4	4	2	4	3	2	5	7	5.0	How would you ...?
6	7	7	4	5	4	4	6	3	7	7	5	6	6	7	6	6	6	6	7	4	7	6	4	6	6	5	3	6	7	5.7	2.h) **Easy to learn:**
4	7	7	4	4	5	2	6	4	5	4	6	7	4	3	6	7	6	3	6	3	6	3	4	4	4	4	2	4	7	4.7	Do you feel ...?
3	7	7	2	5	6	5	6	5	7	6	5	7	7	5	6	7	6	7	5	6	7	2	5	7	6	5	2	6	7	5.6	2.i) **Efficiency**:
3	7	4	1	5	4	4	4	3	6	6	6	7	4	4	6	7	5	7	5	6	7	4	5	7	4	6	1	5	7	5.1	Rate the ease ... .
6	7	6	7	6	4	5	6	5	6	7	7	5	4	5	7	6	6	5	7	3	6	2	6	4	5	6	2	6	7	5.5	2.j) **Engaging:**
6	7	5	7	5	3	6	7	4	5	7	7	6	4	5	7	6	5	7	7	6	5	4	7	4	5	4	3	6	7	5.6	Did PifMed help ...?
6	7	6	3	6	7	4	5	4	6	7	6	7	7	6	7	7	7	6	7	5	6	4	5	4	6	6	3	6	7	5.8	2.k) **Overall:**
6	7	5	2	6	5	6	5	5	5	4	7	7	4	5	7	7	6	6	4	4	5	2	6	2	4	7	2	5	7	5.1	Please rate ease... .

Table D.1: The full results from Part I and Part II of the questionnaire.

D.2.2 Full Results from Part III of the Questionnaire.

Question: What was your least favourite feature of PifMed?

- " You have to click the || + || Boxes to expand. I want to click on the text in the row (i.e. the paper name.)"
- "-No on-screen search or 'find'"
- "-Some titles are a bit long."
- "The Categorization helps me alot." (misplaced comment, not negative.)
- "Search box for the keyword search because it does not work the way it was expected and the category labels in the documents is not interactive to explore related categories."

- "- sometimes it is not clear what the MESH terms are representing"
- "- overlap and different MESH terms"
- "- can not define limits, go to full article, limit by clinical trials - all those great defining features in PubMed"
- "- not as able to define - seems like a loss of specificity"
- "- visualization was a bit complex - not easy to see overlap"
- "- less info - deciding on title only - small diff(sp?) but one I use, knowing author and publication"
- "PubMed allows me to see the paper and keywords and abstract but this is not yet available in PifMed!"
- " The fact that I didn't get the results right away. That would be useful for a specific query for which only few results are expected. "
- "That I wasn't familiar with the MESH categories and what might be under them already"
- " Search option buttons - just give me one MESH / Medline? confusing – I have no medical background"
- "It would be nice to give some hints when typing in keywords like PubMed does."
- "Nothing! Great search engine!"
- " That I would have hit 'Search' intuitively instead of "Query Medline"."
- " Need an hourglass to show searching"
- "The search key/button did not initiate the search, nor did the 'enter' key"
- "that I could not open the files immediately, instead I have to press 'keep' "
- " "

- “ Nothing ”
- “ Sometimes the categories didn’t reflect what I was looking for and I had to search around before I could find some results of interest. ”
- “ However I assume that’s just how they were categorized in Medline, to begin with ”
- “The wait. Although, I understood that there is a goof reason for this. ”
- “ Not being able to go back (trace back) ”
- “ It really limits your scope - as in you don’t really see / browse other articles that might inspire other ideas. Though this could be helpful too, I prefer having the option (PubMed reg. search vs. advanced search) ”
- “- retrieval time ”
- “ - several unexpected behaviors (minor thing) ”
- “ tree structure ”
- “ re-name the buttons - Query Medline doesn’t clearly indicate the search function ”
- “ Author should be listed along with the initial title, sometimes I look at an article simply because the author is known to me, not because of the title. ”
- “ Categories are almost too focused. Could see this being a plus for those in the medical profession but not for the general user. ”
- “ a little slow ”
- “Lack of keywords in the abstract section. This would help in narrowing down your search. ”

Question: What was your favourite feature of PifMed?

- "At-a-glance look at all topic categories surrounding my query"
- "- The Categorization of the results"
- "- I did not have to go too deep in the hierarchy to get relevant results."
- "- I loved the choice of colours."
- "- I wish we had something like PifMed for Web search"
- "The flexibality that I can open & close each category to tracking articles."
- "Categorized query results."
- "Browsing and Discovery"
- "- liked the MESH categories up front for searches low retrieval but found that if I was getting high #'s of articles returned it was a bit difficult to differentiate. "
- "Categorization"
- "'Keep' was intuitive better than 'Save' in PubMed"
- "The hierarchy of concepts. It was really helpful to make search more specific."
- "The tree - That I could easily narrow in on what I was looking for and collapse and expand as desired so that I didn't find myself overwhelmed by information on the screen but, as the same time, could quickly return to the information that I had been looking at before. "
- "Categorization of results"
- "Quickly locating results. help me make up my mind on the contents using categories, and very easy to browse the results back and forth"
- "I liked that PifMed broke down results into categories. PifMed provided results that I wasn't necessarily expecting - more interesting then PubMed"

- " That I could browse through the groupings that I wanted to... get to interesting things faster."
- " - easy to use"
- " - intuitive"
- " - for those of us easily distracted is was easier to get to what you are looking for without all the 'noise' of other unrelated articles. (although sometimes you discover stuff – but I think PifMed has this capacity too)"
- "The categorization of articles. Gave me a sense that we could find articles in the range I wanted to search."
- "The fact that I browse over categories rather than over papers. This acts as a refined query. Sort of like, 'I am not sure what I am looking for, but I'll know when I see it."'
- "clustering / grouping of articles "
- "Easy to navigate. Categorization / hierarchy made it easy to find several related papers once I figured out which category to look in. "
- "Much more interesting to browse than PubMed, where I would need to run several queries to accomplish the same thing."
- "Categories helped a lot on 2/3 of the queries. Figuring out where mitgochondrial phylogeny studies would be in the third query was very difficult. Had it not been a study, I would have adjusted my querying strategy accordingly. "
- "tree-like representation of a large number of results "
- "easy to retrace steps and highlight articles "
- " - browse by category"
- " categorizing relevant topics & show keepers "
- " dividing articles into all the different disciplines, very useful! "

- " easy to retrace the search. "
- " - liked that you could see the navigation. "
- " I like categories in detail, I can get all information once. "
- " - The tree structure made it easier to search"
- " - Keep button HELPFUL "

Question: Suggestions, Improvements, Comments:

- "- Progress indicator for search, so I know it's not frozen (a la gmail login), or a spinning gear."
- "- Didn't understand why 'Query MEDLINE' button was not directing next to search box. I wanted to click 'search'."
- "- 'Open Level 1' Button whould be more handy when I get search results rather than Open All – Close level 2 "
- "-Look in 3, 4"
- "+ I would like to see multi-level Filtering of the search results."
- "- Maybe a 'find' on-screen matches feature would be nice."
- "I hope that is will be more efficient & I really love the idea. :)"
- "If the search option for the keywords and category labels in documents of another category could be improved the is would be interactive. Overall the ease of interactivity should be considered more carefully."
- " it would be interesting to use PifMed in addition to PubMed - or in a situation where you are doing a scoping search, some of what I am reacting to is probably what I am used to "
- "1. Implement viewing the paper right from PifMed "
- "2. Categories are better to be sorted by relevance to the query not just alphabetical order. "

- “Given that I use a web browser for all my search tasks switching to a program with a different look and feel was a bit awkward. I would suggest to re-implement the system for a web browser. ”
- “ When there are a lot of results, don’t have everything ’open’ after a certain depth. Ex. I open one level, then open one item at the next and then BOOM - there are tones of results to go through. More overwhelming the PubMed almost. ”
- “ Think PifMed would have been more beneficial / useful if I had an exact (more specific) query in mind before I started. ”
- “ I feel PifMed is more appropriate for researchers that have a clear mind on what they are looking for because PifMed sorts results by categories and users have to clearly know the categories about their queries beforehand. PubMed is more flexible in terms of this However, it needs more time to find relevant results. ”
- “ Is there a way to search by recent publication date? Author? (i.e. advanced searches) ”
- “It’s the kind of thing I would find myself using for fun, not just work. ”
- “ When can we use this :) ”
- “- for some people the titles ‘health economics’ vs ‘health manpower’ vs ‘health admin’ may be confusing if that is not their field - a cheat sheet of what these are would be good. ”
- “PifMed seemed more appropriate than PubMed for these general queries, where I had a category or range of papers that would be appropriate. I would wonder whether the categorization structure might ‘get in the way’ when searching for a specific paper. ”
- “Great tool. I am always skeptical to leave my google comfort zone, but this is definitely something I would like to use for my own daily research. ”
- “ search, find in MESH and query Medline buttons can be replaced. ”

- " Search and find in Medline buttons would be confusing if there were no instructions given to the user beforehand "
- " other than that, good job! "
- "Interface is slow, but that is probably not possible to change "
- " Sometimes, I would have preferred the branches to have stayed collapsed until I actively chose to open one."
- " Overall, neat interface! "
- "The MESH taxonomy has an unintuitive top-level for people interested in molecular biology. Entering through chemical and drugs, the to amino-acids to get to "proteins" was unexpected. Finding 'organelle' as a category eluded me. I guess that a user with a specific background, unfamiliar with MESH, may have a hard time to see what the lower levels contain just from a node name. "
- "- a possible solution would be to keep track of a users traversal of nodes and order the results (the tree) with the most likely categories on top of the page."
- " - I would rather not have sub-category expanding after a category expansion "
- " progress indication during first response "
- " collapse and push down prev. results after ... (can't make out)"
- " Good Luck!"
- " This system needs quick find capability to highlight some words "
- "I really like that the articles are listed according to discipline, very useful for interdisciplinary research, I wish the arts faculty would use a similar format. the layout is less overwhelming, PubMed is visually cluttered and you have to scroll through all the articles. I prefer PifMed. "
- " It would be nice if there was a category under PifMed for general articles on the subject. "

- " It increases more sub-categories, but well done in all. "

- " Maybe an advance search section that would help the user narrow down the search by year, author & country "

D.2.3 User Times

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	Ave	
123	189	289	122	30	101	171	147	294	186	25	35	37	52	54	69	116	88	115	56	57	114	58	38	50	119	67	103.8	Q1: PubMed
104	123	18	144	23	69	74	73	122	238	92	67	92	112	83	58	60	128	92	132	76	88	70	70	101	131	39	91.8	Q1: PifMed
127	38	155	117	61	44	137	18	93	189	38	193	114	47	27	126	177	60	50	148	36	229	72	57	58	19	57	92.1	Q2: PubMed
103	145	21	39	175	126	136	97	167	115	35	127	87	58	103	96	155	80	45	76	73	86	39	22	42	31	31	85.6	Q2: PifMed
96	35	163	166	119	85	36	57	199	131	48	103	289	105	117	122	86	74	27	164	148	111	8	73	46	64	139	104.1	Q3: PubMed
81	31	15	34	105	79	78	63	49	115	42	139	41	93	79	31	58	107	126	142	66	90	38	59	43	37	26	69.1	Q3: PifMed
104	123	18	144	30	69	171	147	294	186	25	67	37	52	54	69	116	88	115	132	57	114	70	38	50	119	67	94.7	Q1: First
123	189	289	122	23	101	74	73	122	238	92	35	92	112	83	58	60	128	92	56	76	88	58	70	101	131	39	100.9	Q1: Second
127	38	155	117	175	126	137	97	167	115	35	127	87	58	103	96	155	80	50	148	73	229	72	22	42	27	31	99.6	Q2: First
103	145	21	39	61	44	136	18	93	189	38	193	114	47	27	126	177	60	45	76	37	86	39	57	58	31	57	78.4	Q2: Second
96	35	163	34	119	85	78	57	199	115	48	103	41	93	117	122	86	74	126	142	66	90	38	59	43	64	26	85.9	Q3: First
81	31	15	166	105	79	36	63	49	131	42	139	289	105	79	31	58	107	27	164	148	111	8	73	46	37	139	87.4	Q3: Second

Table D.2: The full results from use-time comparison.

D.2.4 User Queries

1. 226223 Starting MEDLINE Query – malaria AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 226223 MEDLINE Query — Count = 23053

2. 226381 Starting MEDLINE Query – Bayes Theorm AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 226381 MEDLINE Query — Count = 9946

3. 226664 Starting MEDLINE Query – use of bayes theorem in clinical decision making AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 226664 MEDLINE Query — Count = 190

1. 222338 Starting MEDLINE Query – alzheimer's disease AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 222338 MEDLINE Query — Count = 46931

2. 222577 Starting MEDLINE Query – visual rating AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 222577 MEDLINE Query — Count = 3345

3. 222803 Starting MEDLINE Query – radiology in china and canada AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 222803 MEDLINE Query — Count = 19

1. 133542 Starting MEDLINE Query – bone conduction AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 133542 MEDLINE Query — Count = 1513
2. 133722 Starting MEDLINE Query – brachytherapy AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 133722 MEDLINE Query — Count = 7777
3. 133923 Starting MEDLINE Query – breast cancer AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 133923 MEDLINE Query — Count = 108209

1. 880044 Starting MEDLINE Query – birds AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 880044 MEDLINE Query — Count = 78144
2. 880190 Starting MEDLINE Query – Alzheimer's AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 880190 MEDLINE Query — Count = 47967
3. 880342 Starting MEDLINE Query – John Hay AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 880342 MEDLINE Query — Count = 57

1. 874215 Starting MEDLINE Query – clustering algorithm AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 874215 MEDLINE Query — Count = 5643
2. 875381 Starting MEDLINE Query – subspace clustering AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 875381 MEDLINE Query — Count = 42
3. 875698 Starting MEDLINE Query – evolutionary subspace clustering AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 875698 MEDLINE Query — Count = 1

1. 806526 Starting MEDLINE Query – identity and social theory AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 806526 MEDLINE Query — Count = 865
2. 806714 Starting MEDLINE Query – terrorism AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 806714 MEDLINE Query — Count = 3510
3. 806901 Starting MEDLINE Query – thc and the brain AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 806901 MEDLINE Query — Count = 956

1. 781997 Starting MEDLINE Query – diabetes AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 781997 MEDLINE Query — Count = 166096
2. 782565 Starting MEDLINE Query – muscular dystrophy AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 782565 MEDLINE Query — Count = 10066

3. 782729 Starting MEDLINE Query – research methods in computer science AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 782729 MEDLINE Query — Count = 22136

1. 778067 Starting MEDLINE Query – protein structure evolution AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 778067 MEDLINE Query — Count = 18137

2. 778420 Starting MEDLINE Query – statistical mechanics protein AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 778420 MEDLINE Query — Count = 313

3. 778534 Starting MEDLINE Query – human evolution AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 778534 MEDLINE Query — Count = 24047

1. 869828 Starting MEDLINE Query – music brain perception AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 869828 MEDLINE Query — Count = 653

2. 870439 Starting MEDLINE Query – bacteria basic information AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 870439 MEDLINE Query — Count = 630

3. 870672 Starting MEDLINE Query – tanning AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 870672 MEDLINE Query — Count = 1058

1. 714694 Starting MEDLINE Query – nmda receptor AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 714694 MEDLINE Query — Count = 24696

2. 714927 Starting MEDLINE Query – bayesian AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 714927 MEDLINE Query — Count = 9800

3. 715248 Starting MEDLINE Query – fixatives AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 715248 MEDLINE Query — Count = 11014

1. 712412 Starting MEDLINE Query – text summarization AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 712412 MEDLINE Query — Count = 33

2. 712513 Starting MEDLINE Query – machine learning AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 712513 MEDLINE Query — Count = 29208

3. 713122 Starting MEDLINE Query – swine flu AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 713122 MEDLINE Query — Count = 665

1. 627043 Starting MEDLINE Query – ADHD striatum AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 627043 MEDLINE Query — Count = 302

2. 627305 Starting MEDLINE Query – striatal medium spiny neuron collaterals AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 627305 MEDLINE Query — Count = 14

3. 627688 Starting MEDLINE Query – Ventral frontal cortex hippocampus AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 627688 MEDLINE Query — Count = 445

1. 701840 Starting MEDLINE Query – cocaine addiction AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 701840 MEDLINE Query — Count = 5181

2. 702062 Starting MEDLINE Query – bulimia AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 702062 MEDLINE Query — Count = 3996

3. 702546 Starting MEDLINE Query – uv rays AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 702546 MEDLINE Query — Count = 29187

1. 623469 Starting MEDLINE Query – perfusion in alzheimer disease AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 623469 MEDLINE Query — Count = 363

2. 623871 Starting MEDLINE Query – medial temporal lobe atrophy in alzheimer disease AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 623871 MEDLINE Query — Count = 218

3. 624084 Starting MEDLINE Query – brain atrophy in alzheimer disease AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 624084 MEDLINE Query — Count = 1350

1. 620102 Starting MEDLINE Query – dehydration children AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 620102 MEDLINE Query — Count = 1255

2. 620278 Starting MEDLINE Query – pediatrics AND emergency AND mental health care AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 620278 MEDLINE Query — Count = 66

3. 620593 Starting MEDLINE Query – parents AND uncertainty AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 620593 MEDLINE Query — Count = 379

1. 542187 Starting MEDLINE Query – sports and spirituality AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 542187 MEDLINE Query — Count = 9

2. 542451 Starting MEDLINE Query – manual labour AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 542451 MEDLINE Query — Count = 2759

3. 542930 Starting MEDLINE Query – bicycles AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 542930 MEDLINE Query — Count = 211

1. 538063 Starting MEDLINE Query – barriers to hepatitis C treatment AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 538063 MEDLINE Query — Count = 108

2. 538210 Starting MEDLINE Query – HIV services in Canada AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 538210 MEDLINE Query — Count = 316

3. 538427 Starting MEDLINE Query – harm reduction AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 538427 MEDLINE Query — Count = 1772

1. 528490 Starting MEDLINE Query – oncogene tumor AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 528490 MEDLINE Query — Count = 74972

2. 528873 Starting MEDLINE Query – human blood cell AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 528873 MEDLINE Query — Count = 421757

3. 529303 Starting MEDLINE Query – baby formula AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 529303 MEDLINE Query — Count = 4883

1. 524610 Starting MEDLINE Query – MYCIN AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 524610 MEDLINE Query — Count = 34

2. 524996 Starting MEDLINE Query – web based recommendation system AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 524996 MEDLINE Query — Count = 20

3. 525768 Starting MEDLINE Query – Fuzzy Logic medical diagnosis AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 525768 MEDLINE Query — Count = 237

1. 189092 Starting MEDLINE Query – syllabification AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 189092 MEDLINE Query — Count = 24

2. 189341 Starting MEDLINE Query – mouse droppings AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 189341 MEDLINE Query — Count = 15

3. 189652 Starting MEDLINE Query – melatonin jet lag AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 189652 MEDLINE Query — Count = 133

1. 104074 Starting MEDLINE Query – health problems in nova scotia AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 104074 MEDLINE Query — Count = 94

2. 104386 Starting MEDLINE Query – computer aid for adhd AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 104386 MEDLINE Query — Count = 4

3. 104522 Starting MEDLINE Query – music adhd AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 104522 MEDLINE Query — Count = 14

1. 074610 Starting MEDLINE Query – child molestation psychology AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 074610 MEDLINE Query — Count = 61

2. 074888 Starting MEDLINE Query – eczema male female AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 074888 MEDLINE Query — Count = 2325

3. 075580 Starting MEDLINE Query – depressed crime rate AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 075580 MEDLINE Query — Count = 18

1. 979753 Starting MEDLINE Query – salbutamol AND “time series analysis” AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 979753 MEDLINE Query — Count = 2

2. 979828 Starting MEDLINE Query – “time series analysis” AND “drug use” AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 979828 MEDLINE Query — Count = 20

3. 980292 Starting MEDLINE Query – pediatric AND salbutamol AND emergency AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 980292 MEDLINE Query — Count = 86

1. 966941 Starting MEDLINE Query – pregnancy symptoms AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 966941 MEDLINE Query — Count = 125902

2. 967931 Starting MEDLINE Query – children obesity between ages four to ten AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 967931 MEDLINE Query — Count = 2

3. 968247 Starting MEDLINE Query – h1n1 symptoms of children AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 968247 MEDLINE Query — Count = 226

1. 896064 Starting MEDLINE Query – rheumatoid arthritis AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 896064 MEDLINE Query — Count = 37312

2. 896535 Starting MEDLINE Query – real-time data mining AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 896535 MEDLINE Query — Count = 77

3. 896874 Starting MEDLINE Query – heart murmur AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 896874 MEDLINE Query — Count = 1478

1. 966601 Starting MEDLINE Query – how to prevent a heart attack AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 966601 MEDLINE Query — Count = 2250

2. 967055 Starting MEDLINE Query – signs of a stroke AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 967055 MEDLINE Query — Count = 67274

3. 967379 Starting MEDLINE Query – what is the risk of having twins AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 967379 MEDLINE Query — Count = 3702

1. 976504 Starting MEDLINE Query – visualization AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 976504 MEDLINE Query — Count = 25470

2. 976780 Starting MEDLINE Query – cell visualization AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 976780 MEDLINE Query — Count = 7457

3. 977727 Starting MEDLINE Query – scientific visualization medical surgery AND (hasabstract[text] AND (medline[sb]) AND (English[lang]) AND (full text[sb])) 977727 MEDLINE Query — Count = 101

D.3 Analysis

D.3.1 Detailed Statistics from Part II of the Questionnaire.

	Effective	Efficient	Engaging	Error Tolerant	Easy to Learn
Mean	5.398	4.667	5.667	5.407	5.568
Standard Error	0.209	0.225	0.227	0.166	0.196
Median	5.25	4.714	6	5.25	5.667
Mode	5	5.25	6	6	6.333
Standard Deviation	1.088	1.171	1.179	0.861	1.016
Sample Variance	1.184	1.370	1.389	0.741	1.033
Kurtosis	-0.382	0.632	2.664	-0.246	0.866
Skewness	-0.563	-0.598	-1.503	-0.216	-1.086
Range	4	5	5	3.5	4
Minimum	3	1.75	2	3.5	3
Maximum	7	6.75	7	7	7
Sum	145.75	126	153	146	150.333
Count	27	27	27	27	27
	Effective	Efficient	Engaging	Error Tolerant	Easy to Learn
95%Confidence Interval	4.204	4.809	5.200	5.067	5.166
to	5.829	5.130	6.133	5.748	5.970
	Effective	Efficient	Engaging	Error Tolerant	Easy to Learn
Largest (1)	7	7	7	7	7
Smallest (1)	3	1.5	2	3.5	3

Table D.3: The full descriptive statistics of the General User population in terms of usability. The mean and confidence interval for each aspect show PifMed to be slightly preferred to PubMed. Strongest in Engagement and weakest in Efficiency. The overall preference for PifMed is slightly higher here than for the Target Users

	Effective	Efficient	Engaging	Error Tolerant	Easy to Learn
Mean	5.205	4.784	5.477	5.250	5.352
Standard Error	0.220	0.205	0.260	0.175	0.207
Median	5.25	4.75	5.75	5.25	5.375
Mode	5	5.25	6	4.75	5.5
Standard Deviation	1.031	0.961	1.220	0.820	0.969
Sample Variance	1.063	0.924	1.488	0.673	0.938
Kurtosis	-0.219	0.921	2.082	-0.139	0.414
Skewness	-0.614	-0.562	-1.323	-0.214	-0.509
Range	3.75	4	5	3.25	4
Minimum	3	2.25	2	3.5	3
Maximum	6.75	6.25	7	6.75	7
Sum	114.5	105.25	120.5	115.5	117.75
Count	22	22	22	22	22
95%Confidence Interval	Effective	Efficient	Engaging	Error Tolerant	Easy to Learn
from	4.747	4.358	4.937	4.886	4.923
to	5.662	5.210	6.018	5.614	5.782
Outliers	Effective	Efficient	Engaging	Error Tolerant	Easy to Learn
Largest (1)	6.75	6.25	7	6.75	7
Smallest (1)	3	2.25	2	3.5	3

Table D.4: The full descriptive statistics of the Target User population in terms of usability. The mean and confidence interval for each aspect show PifMed to be slightly preferred to PubMed. Strongest in Engagement and weakest in Efficiency.

Appendix E

Conclusion

E.1 Correspondence with NLM

From: "Rappoport, Marina (NIH/NLM) [E]" [removed]
To: [my email -removed]
CC: "Spina, Fran (NIH/NLM) [E]" [removed],
"Burts, Leonore (NIH/NLM) [C]" [removed]
Date: Wed, 14 Jul 2010 14:37:09 -0400
Subject: RE: MEDLINE MeSH Categorization Suggestions.

Your e:mail was forwarded to the Index Section of the National Library of Medicine. Our Section is responsible for indexing for MEDLINE/PubMed over 700,000 articles per year. We are using controlled vocabulary - Medical Subject Headings. Each article is analyzed and appropriate MeSH terms are assigned.

Our trained indexers analyze the full text of the article and assign MeSH headings using a very complicated system of coordinated MeSH indexing. When assigning MeSH headings, indexers are using our internal interface. . General public does not have access to this system.

If you believe that your article was indexed incorrectly, i.e. certain MeSH headings are missing and/or certain MeSH headings are used inappropriately, you should contact custserv@nlm.nih.gov, indicating PMID and specific MeSH headings, which, in your opinion, are incorrect. We will then re-examine this specific article, and either correct indexing or explain to you why we think the article

was indexed properly.

Sincerely,

Marina Rappoport
Head, Unit A, Index Section, BSD
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894
Voice: 301-[removed]
Fax: 301-[removed]
[email address removed]

------Original Message------

From: [email address removed]
Sent:07/11/2010 15:26:09
To: Custhelp [email address removed]
Subject: MEDLINE MeSH Categorization Suggestions.

SUBJECT: MEDLINE MeSH Categorization Suggestions.
EMAIL: [email address removed]
NAME: Pif
GROUP: Student
STATE:
COUNTRY: Canada
FROM: http://www.nlm.nih.gov/pubs/factsheets/errata.html
DATE: 07/11/2010
MESSAGE: Hi,

Do you take suggestions of the Type:
PMID: XXXXX, shouldn't have MeSH descriptor: YYYY as a major topic.
or
PMID: XXXX, should be indexed with MeSH Term: YYY.
If you do, do you have a page to make these types of suggestions? If not, why that policy?

Thanks a bunch.