REPOSITORIES WITH MEDICAL RESOURCES, THEIR INDEXING AND RETRIEVAL

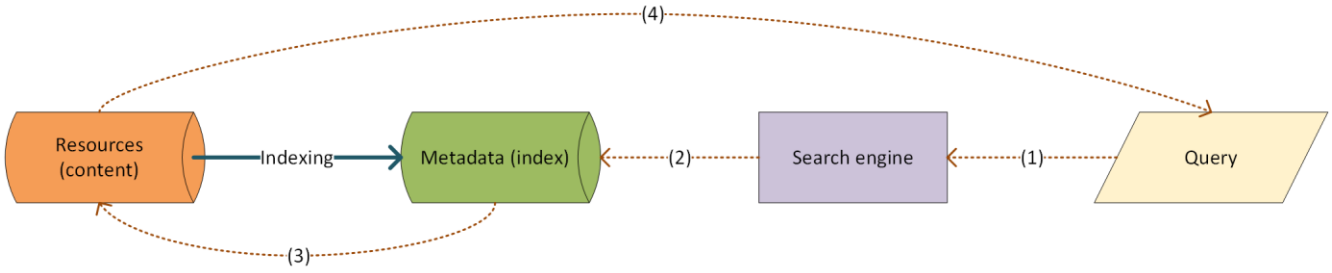
Szymon Wilk
Institute of Computing Science
Poznan University of Technology
Poznan, Poland

INTRODUCTION

Types of information

- **Patient-specific** information – applies to specific patients, associated with healthcare provision and stored in EPR/HIS
- **Knowledge-based** information – information (knowledge) derived from research, published in books, journals …

Information retrieval (IR) is the field concerned with acquisition, organization and searching of knowledge-based information [Hersh, 2009]



Hersh, W. (2014). Information retrieval and digital libraries. In: Shortliffe, E.H., Cimino, J. (eds.): *Biomedical Informatics. Computer Applications in Health Care and Biomedicine*, Springer, 626-654.

Some history…

- 1878 – Index Medicus
  - Index of medical articles
  - Metadata: title, authors, keywords (*subject headings*)
- 1966 – MEDLARS (Medical Literature Analysis and Retrieval System), NLM
  - Computer-based version of Index Medicus
  - Metadata: similar to paper version (disk space limitations)
  - Medical Subject Headings (MeSH) used as keywords for indexing
  - Running queries sent by regular mail
- 1990s – bibliographic databases managed by NLM made available on the Web (→ MEDLINE/PubMed)

KNOWLEDGE-BASED INFORMATION AND EBM

Information needs

Possible states of information needs in clinical context

1. **Unrecognized** – clinician unaware of information need
2. **Recognized** – clinician aware of need but may or may not pursue it
3. **Pursued** – information seeking occurs but may or may not be successful
4. **Satisfied** – „found necessary informaiton”

Unrecognized information needs result in non-adherence to up-to-date clinical practice and sub-optimal quality of care!

Information needs in practice

- On average 3 questions per 4 patients (primary care)
  - What drug to prescribe? What causes observed symptom? ...
- Answer seeking in 12-36% cases
  - No explicit formulation of a question
  - Discarding the search ($\rightarrow$ too time-consuming, acceptable searching time = 2 min or less)
- Integration of IR mechanisms with EPR/HIS systems for automatic provision of links to knowledge-based resources

Timpka T, Arborelius E. The GP's dilemmas: a study of knowledge need and use during health care consultations. *Methods Inf Med.* 1990; 29(1):23-9.
Ely JW, Osheroff JA, Ebell MH, et al. Analysis of questions asked by family doctors regarding patient care. *BMJ* 1999;319:358-361.
Hersh WR, Crabtree MK, Hickam DH, et al. Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions. *J Am Med Inform Assoc* 2002;9(3):283-93.

Evidence-based Medicine (EBM)

Integration of **experience** (art) with **best clinical results** (knowledge-based information) for **informed clinical decision making**

- 3-step process
  1. Phrasing a clinical question that is relevant and answerable
  2. Identifying evidence (studies in articles) that address the question
  3. Critically appraising the evidence – does it apply to the patient?
- Categories of considered questions
  1. Therapy or intervention – benefits?
  2. Diagnosis – diagnosing test?
  3. Harm – detrimental health effect?
  4. Prognosis – outcome of disease course?

Most valuable evidence comes from RCTs, but they are not always possible (e.g., harm and ethical issues)

Sackett DL, Rosenberg WM, Gray JA, et al. Evidence based medicine: what it is and what it isn't. *BMJ* 1996;312:71-2.
Ely JW., Osheroff JA, Ebell MH, et al. Obstacles to answering doctors' questions about patient care with evidence: qualitative study. *BMJ* 2002;324:1-7.

REPOSITORIES OF KNOWLEDGE-BASED INFORMATION

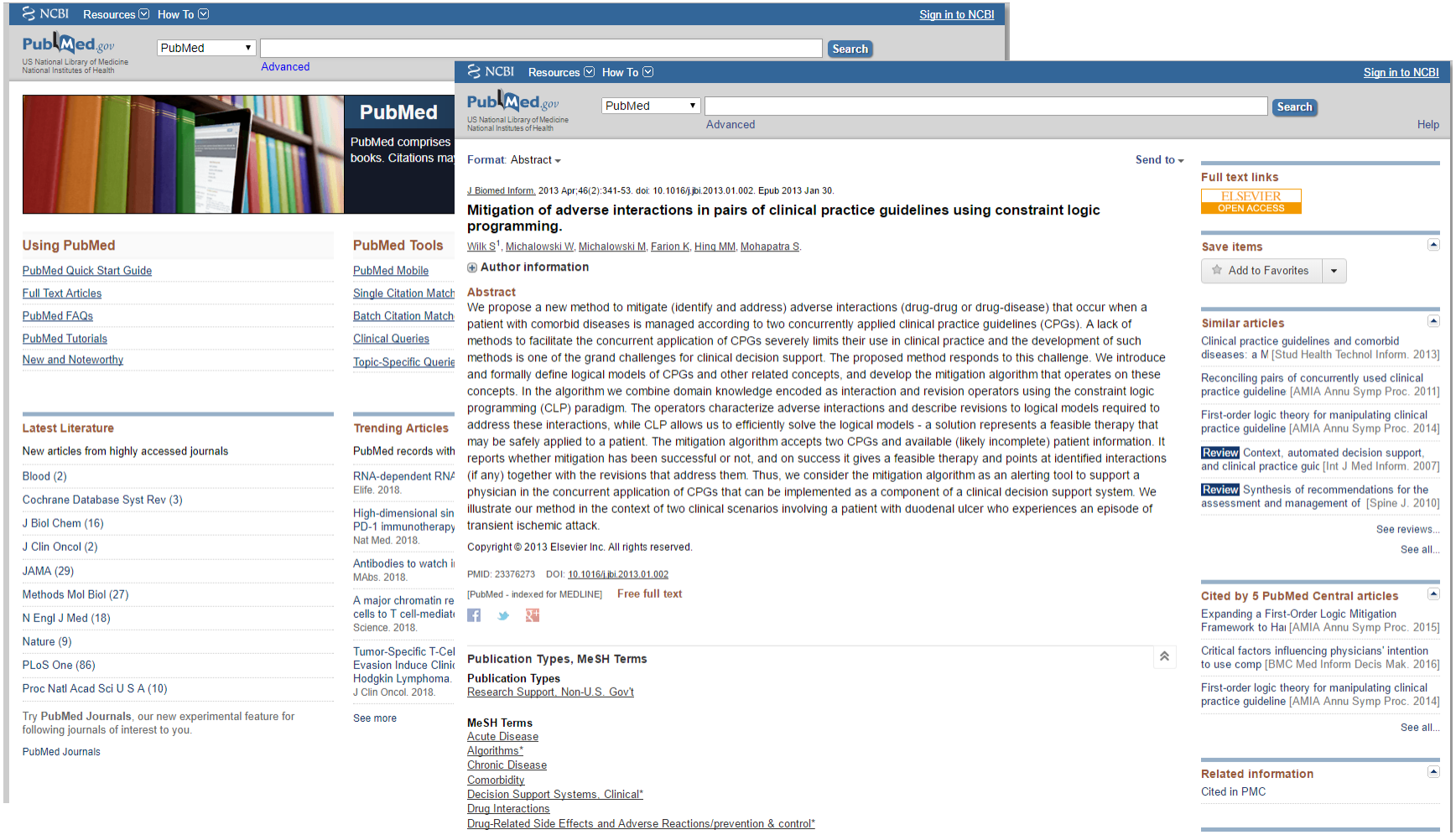
Types of repositories

1. Bibliographic content
   - Citations or pointers to medical literature, no content
   - Examples: MEDLINE/PubMed (NLM), EMBASE ("European MEDLINE"), National Guidelines Clearinghouse (NGC) with clinical guidelines
2. Full-text content
   - Full publication content, often linked from a bibliographic repository (e.g., PubMed $\rightarrow$ publisher website)
   - Usually maintained by commercial publishers, PubMed Central for publications prepared within NIH grants

Types of repositories

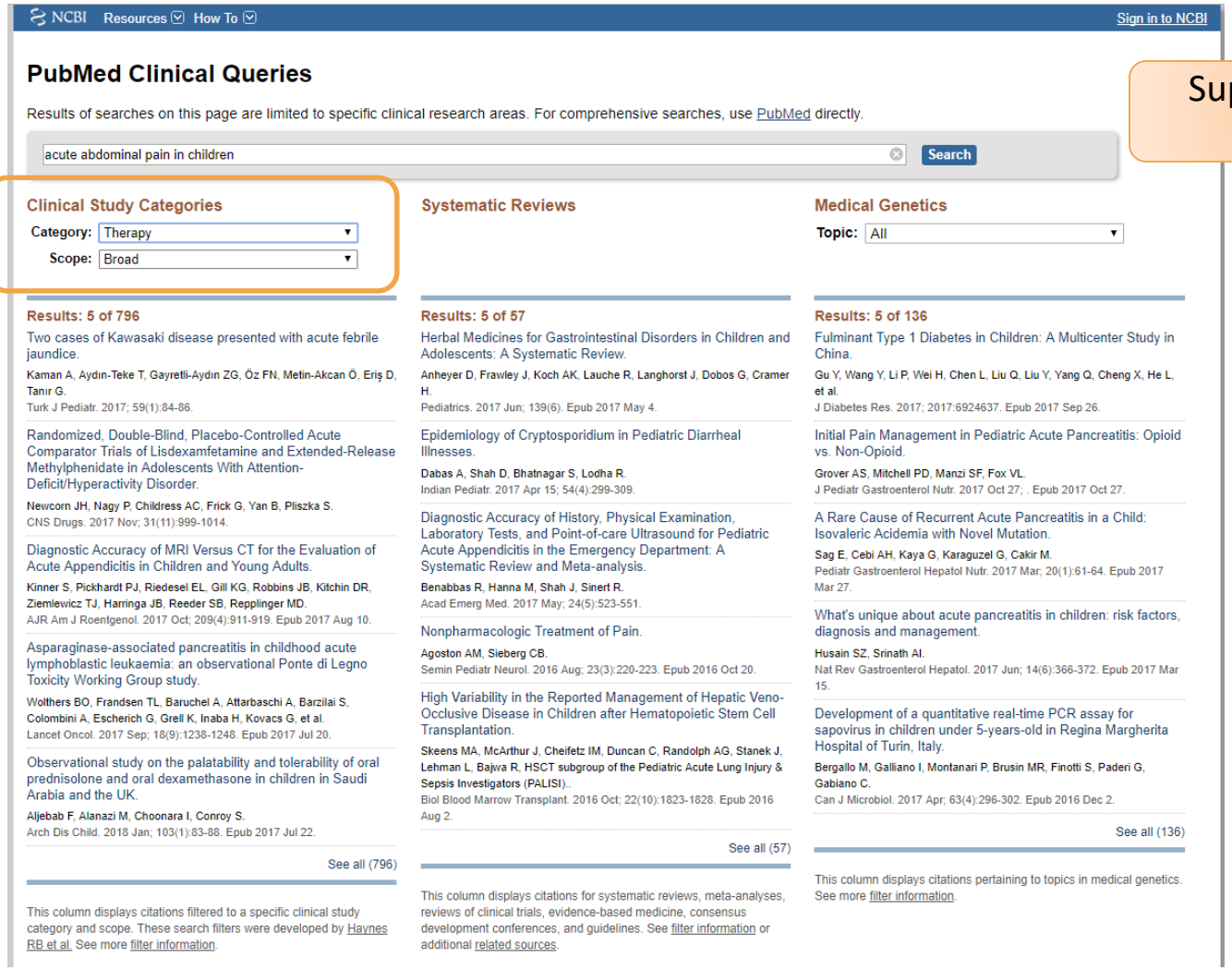
3. Annotated content
   - Resources stored in specialized databases depending on the format (image, genomics, citation, EBM…)
   - Examples: Visible Human Project, PEIR, The Cochrane Library (Database of Systematic Reviews)
4. Aggregated content
   - Aggregation of content from the first three categories (copies or links to original repositories)
   - Examples: MedlinePlus, Merck Manual

PubMed



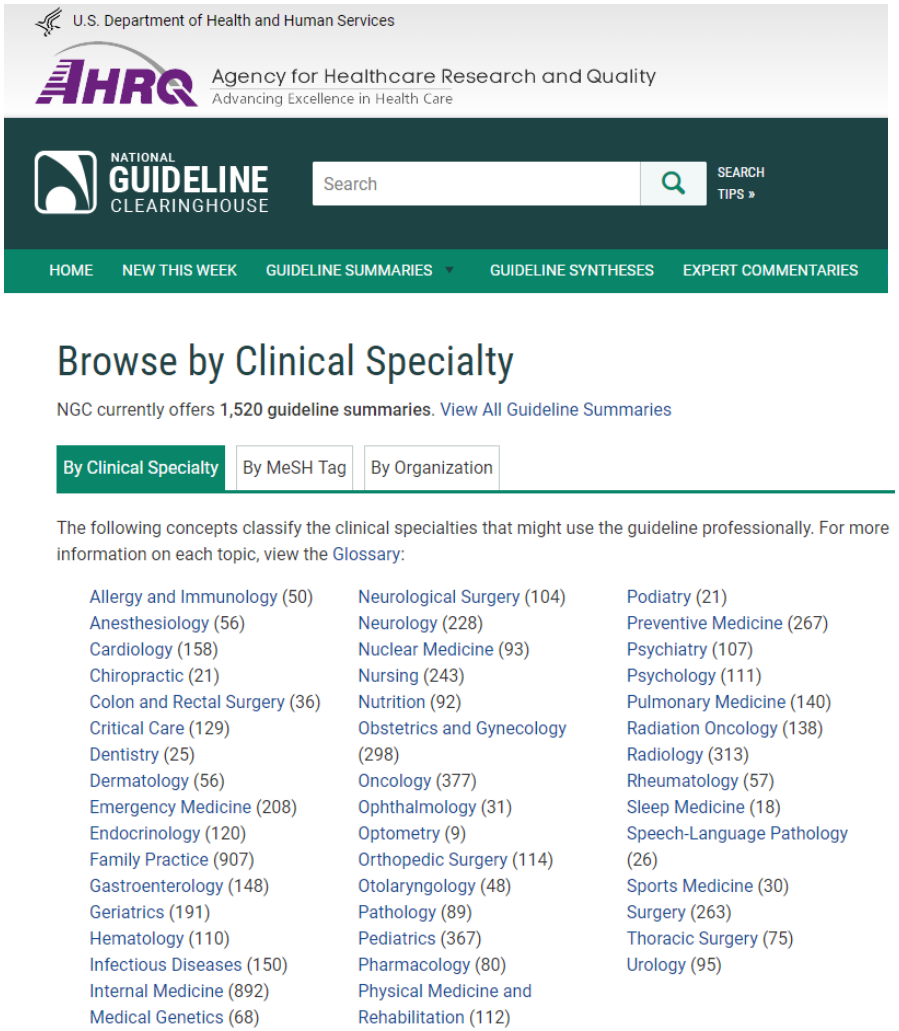
https://www.ncbi.nlm.nih.gov/pubmed

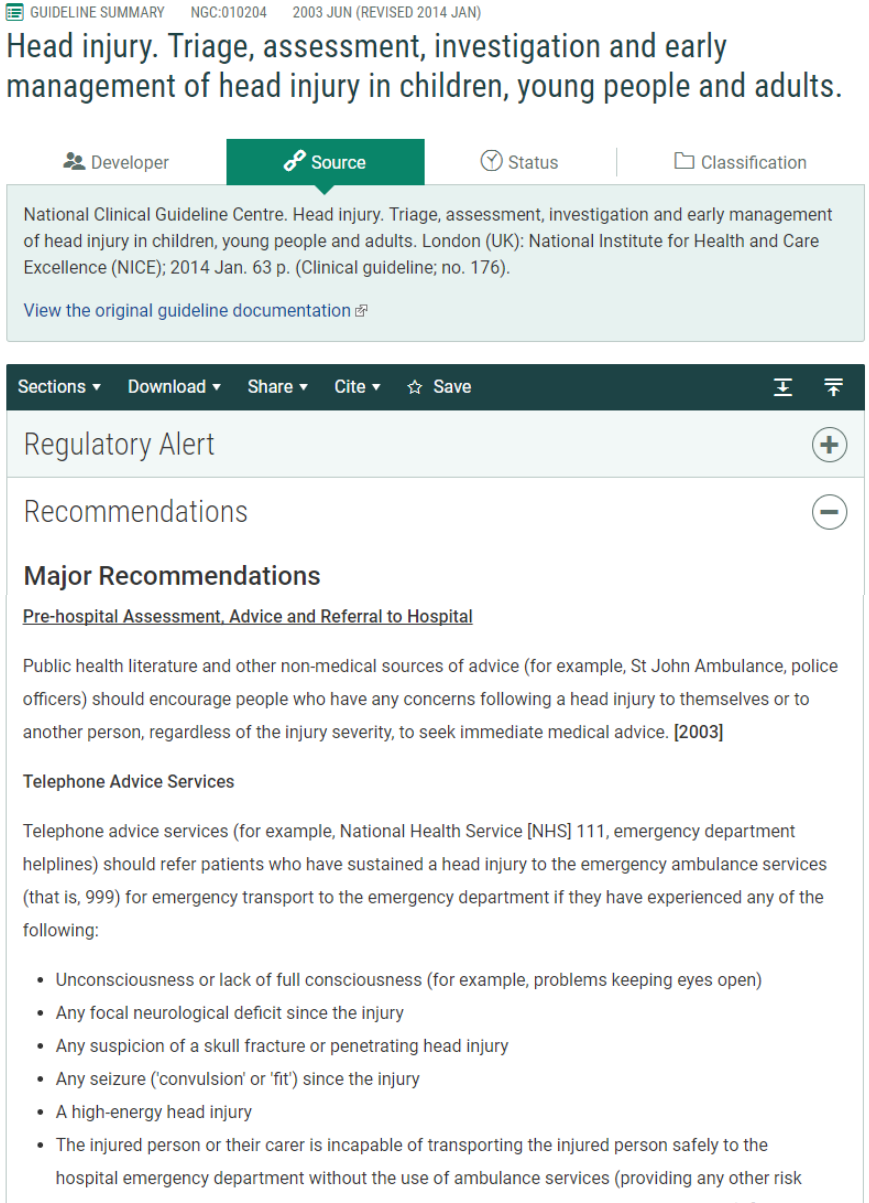
PubMed – Clinical Queries



Support for EBM questions

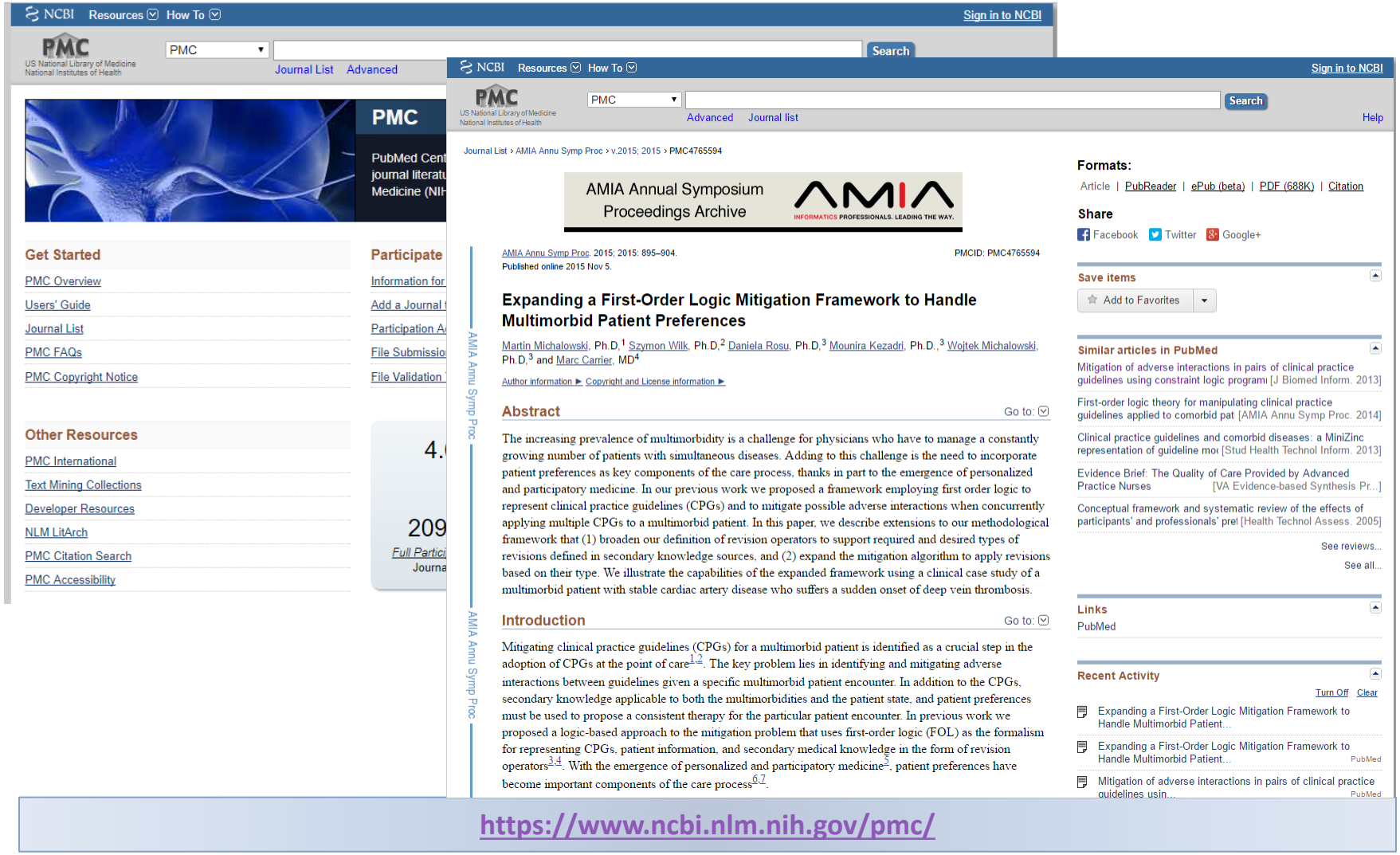
NGC





https://www.guideline.gov/

PubMed Central



https://www.ncbi.nlm.nih.gov/pmc/

Visible Human Project

Home

The Visible Human Project®

Overview

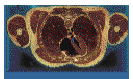
The Visible Human Project® is an outgrowth of the NLM's 1986 Long-Range Plan. It is the creation of complete, anatomically detailed, three-dimensional representations of the normal male and female human bodies. Acquisition of transverse CT, MR and cryosection images of representative male and female cadavers has been completed. The male was sectioned at one millimeter intervals, the female at one-third of a millimeter intervals.

The long-term goal of the Visible Human Project® is to produce a system of knowledge structures that will transparently link visual knowledge forms to symbolic knowledge formats such as the names of body parts.

The National Library of Medicine thanks the men and the women who will their body to science, thereby enabling medical research and development.

Further Information

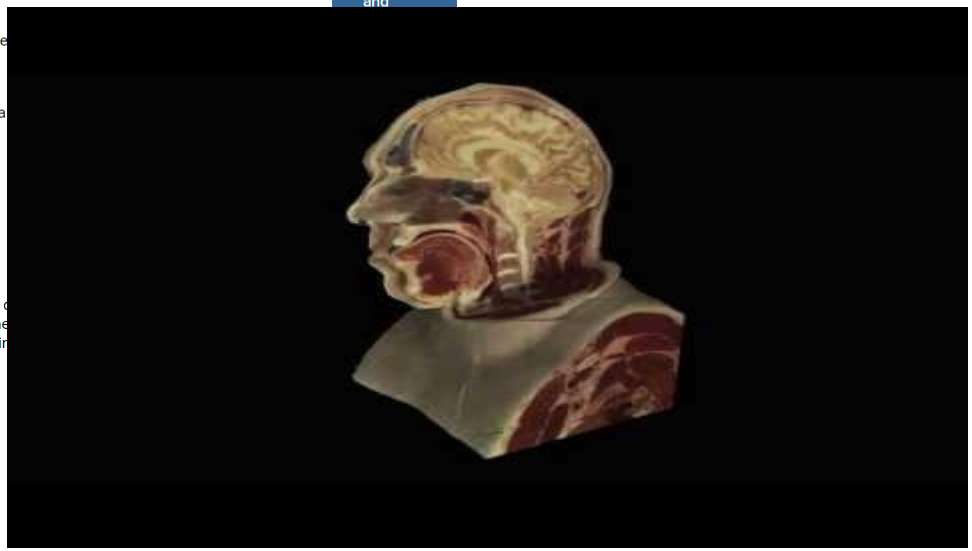
- **General Information**
  - A description of The Visible Human Project® image data and how to obtain it (include
  - The Visible Human Project® FactSheet.
  - The Visible Human Project® From Wikipedia, the free encyclopedia
  - The Visible Human Project®: From Data to Knowledge: An update of ongoing Nationa
  - Digitally encoded videos - requires RealPlayer.
  - A sampler of images and animations from the Project.
  - Belarusian translation of The Visible Human Project® Overview
  - German translation of The Visible Human Project® Overview
  - Russian translation of The Visible Human Project® Overview
  - Finnish translation of The Visible Human Project Overview
- **NLM Initiatives**
  - Cryosection, MRI and CT image data of the head of a 72 year old male. Cryosections
    photographed at a resolution of 1056 x 1528 pixels. Work done at Brigham and Wome
    contract to NLM. Available only to VHP license holders. These images can be found i
    to the NLM image server.



Projects Based on the Visible Human Data Set

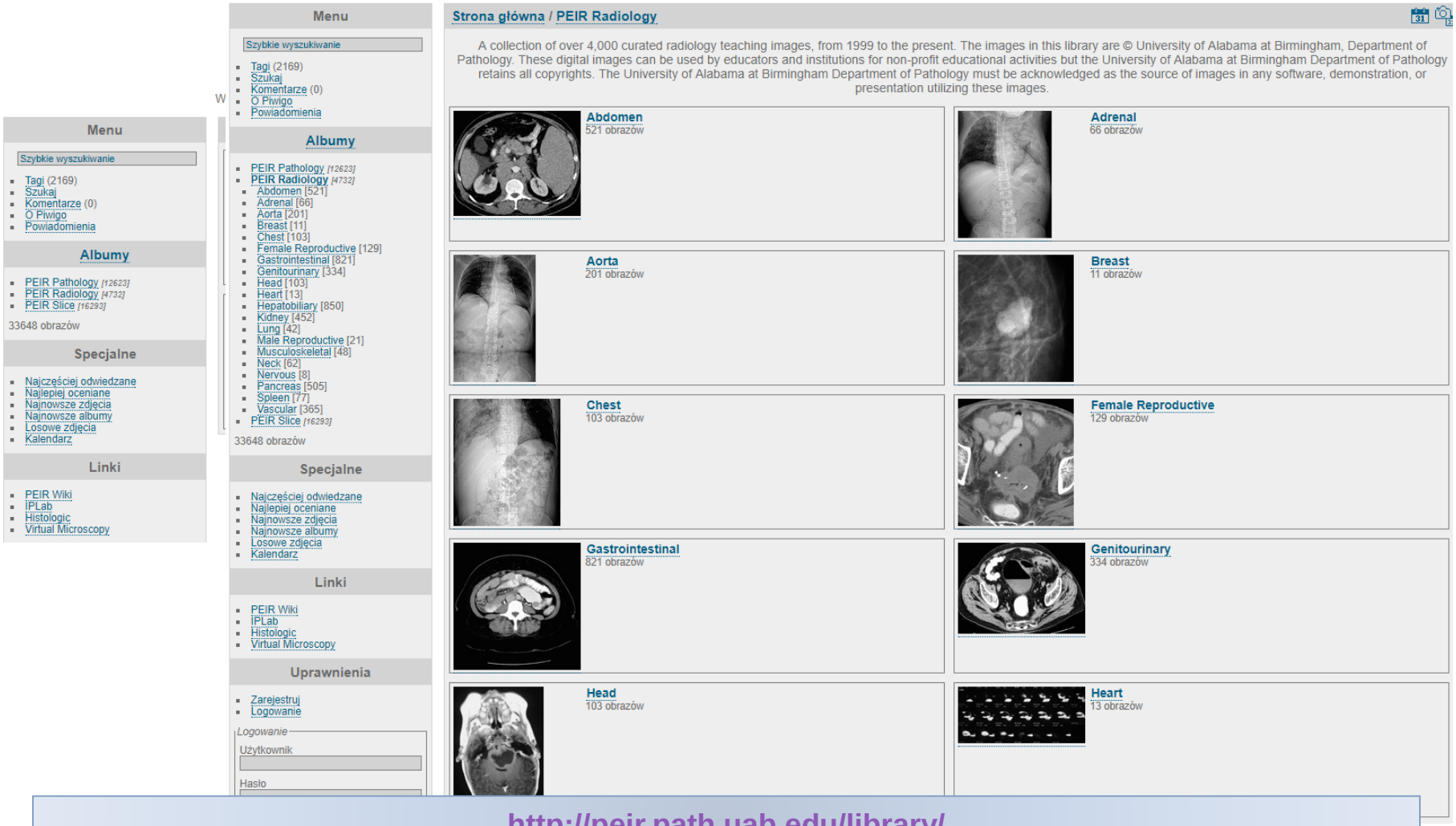
Applications for viewing images

Sources of images and



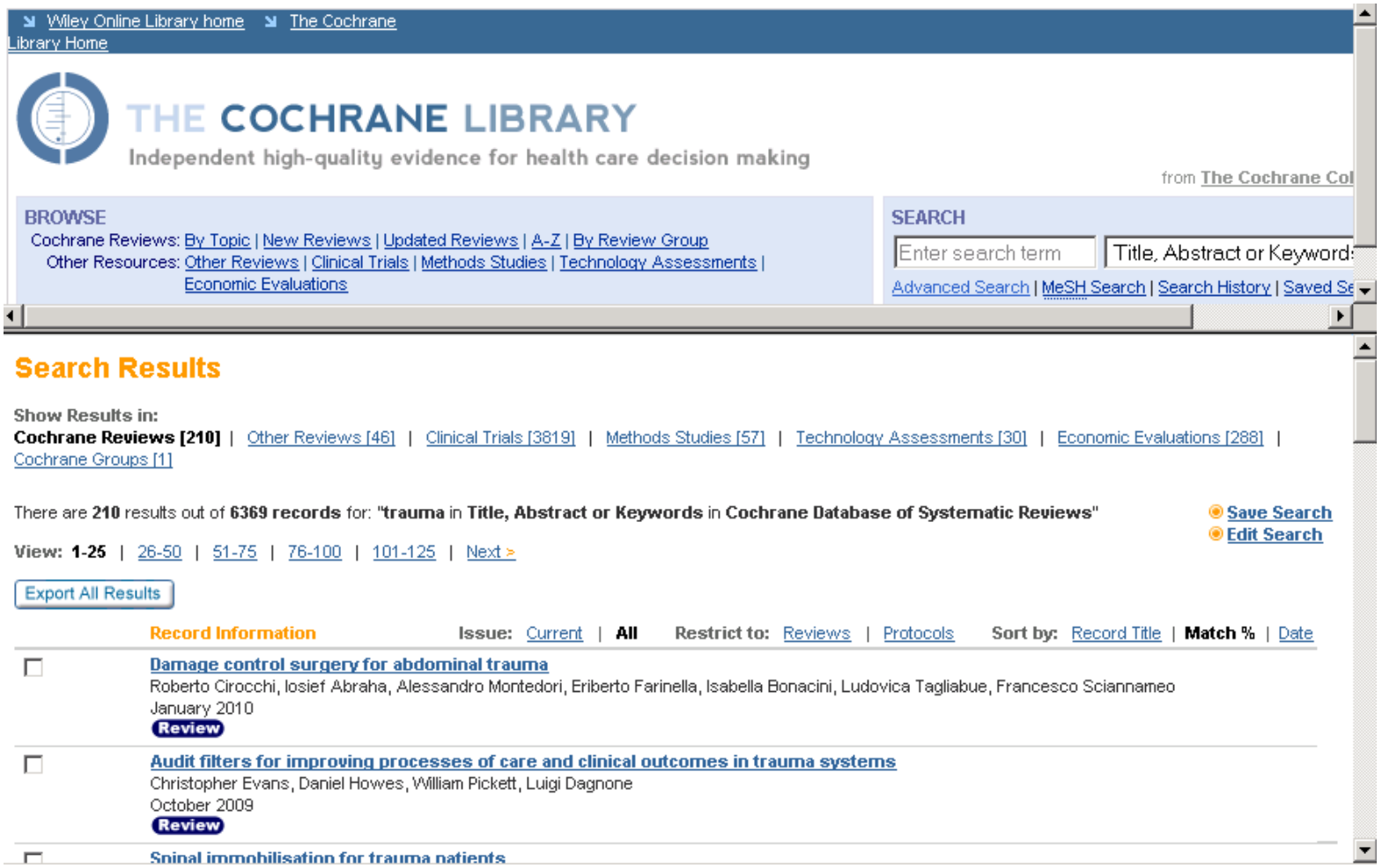
https://www.nlm.nih.gov/research/visible/visible_human.html

PEIR (Pathology Information Educational Resource) Digital Library



http://peir.path.uab.edu/library/

The Cochrane Library



http://www.cochranelibrary.com

Systematic review

PLAIN LANGUAGE SUMMARY

Family therapy for asthma in children

Psychological factors may have an effect on asthma in children, or its severity. As some children with families who are having problems have severe asthma, family therapy has been tried. The aim is to resolve any problems there might be in a family, in case they are causing the child stress and then making asthma worse. The review found some evidence from two trials that family therapy (in addition to standard asthma treatments) might help reduce a child's asthma symptoms, but more research is needed to be certain.

BACKGROUND

The incidence of childhood asthma has increased annually at a rapid rate over the last 20 years. A variety of factors have been implicated in the pathogenesis of bronchial asthma including allergies, infections, endocrinological disorders, genetic predisposition and, more recently, psychological elements. Although psychological factors are widely acknowledged to play a part both in precipitating episodes of asthma, and in the control of symptoms, pharmacological treatment alone continues to be the main treatment, and therefore the focus of most research. Recently, however, traditional medical models upholding a split between the "psyche" and the "soma" are being replaced by those which recognise the powerful influences of the mind on the body. Similarly, the need for integrated treatment models which consider behavioural or psychological interventions in addition to pharmacotherapy are well documented (Cluss 1986; Lehrer 1992; Onnis 1984; Molinari 1994; Towns 1994)

Studies which do acknowledge psychosocial aspects to the disease range from cognitive behavioural approaches, to education programmes, relaxation techniques, autogenic therapy, and rational emotive behaviour modification as an adjunct to medication. However, such studies, although recognising that asthma has strong associations with emotional disturbances, have been conducted mainly with adults.

Of the few studies which examine psychological influences in childhood asthma, psychosocial and emotional factors are regarded as important in the most severely ill of asthmatic children (Gustafsson 1986; Lask 1979). A child's chronic illness can place psychological burdens on both child and family (Newacheck 1991; Pless 1991; Steinhauer 1974). In addition, Weil (Weil 1999) concluded that two psychological variables, child and carer mental health, were important predictors of subsequent asthma morbidity. The earliest research conducted in the 1970s indicated the presence of disturbed family relations in the families of children with severe asthma (Liebman 1974; Liebman 1976). This prompted a limited number of trials to be conducted in the area of family therapy.

The theoretical basis which underpins the systemic view of family therapy presupposes that "symptoms" can be the product of a dysfunctional family system. Therapeutic intervention is used to alert the family system to its dysfunctional behaviours and to empower it to adaptively overcome the difficulties which give rise to the symptoms. The body of research undertaken so far has been based on a structural family therapy perspective which assumes that relationships between family members adhere to certain patterns, which are often maladaptive to the current life situation. Systemic theories regard phenomena in terms of circularity, rather than in the linear terms of cause and effect inherent to the medical model of illness. Therefore asthma is assumed to be both a symptom of the family dysfunction, and a contributor to it (Lask 1979). Family therapy has been used to decrease the symptoms and impact of asthma.

OBJECTIVES

The objective of this review is to test whether family therapy as an adjunct to traditional medication can be shown to have a significant effect in reducing the symptoms and impact of asthma in children.

METHODS

Criteria for considering studies for this review

Types of studies

Clinical trials included were randomised and controlled. Authors have been contacted to ascertain randomisation techniques. Trials included in this review compared family therapy + pharmacological intervention vs pharmacological intervention alone.

Types of participants

Chronically asthmatic children receiving medication who were treated at hospital outpatient departments or clinics, and their families.

Types of interventions

Any family therapy based on systemic theories which focus on the whole family and which aim to arrive at an understanding of

Copyright © 2008 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd.

DATA AND ANALYSES

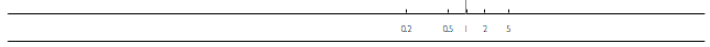
Comparison 1. FAMILY THERAPY & DRUG THERAPY vs DRUG THERAPY ALONE

Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1 General paediatric assessment - no improvement			Peto Odds Ratio (Peto, Fixed, 95% CI)	Totals not selected
2 General paediatric assessment - deterioration			Peto Odds Ratio (Peto, Fixed, 95% CI)	Totals not selected
3 Functionally impaired days - no improvement			Peto Odds Ratio (Peto, Fixed, 95% CI)	Totals not selected
4 Functionally impaired days - deterioration			Peto Odds Ratio (Peto, Fixed, 95% CI)	Totals not selected
5 Nights B2 inhalers used - no improvement			Peto Odds Ratio (Peto, Fixed, 95% CI)	Totals not selected
6 Nights B2 inhalers used - deterioration			Peto Odds Ratio (Peto, Fixed, 95% CI)	Totals not selected
7 Peak expiratory flow rate - no improvement			Peto Odds Ratio (Peto, Fixed, 95% CI)	Totals not selected
8 Peak expiratory flow rate - deterioration			Peto Odds Ratio (Peto, Fixed, 95% CI)	Totals not selected
9 Peak expiratory flow rate - pre-medication, morning			Mean Difference (IV, Fixed, 95% CI)	Totals not selected
10 Peak expiratory flow rate - pre-medication, evening			Mean Difference (IV, Fixed, 95% CI)	Totals not selected
11 Forced expiratory volume - (in 0.75 seconds)			Mean Difference (IV, Fixed, 95% CI)	Totals not selected
12 Thoracic gas volume			Mean Difference (IV, Fixed, 95% CI)	Totals not selected
13 Day wheeze			Mean Difference (IV, Fixed, 95% CI)	Totals not selected
14 Activity			Mean Difference (IV, Fixed, 95% CI)	Totals not selected

Analysis 1.1. Comparison 1 FAMILY THERAPY & DRUG THERAPY vs DRUG THERAPY ALONE, Outcome 1 General paediatric assessment - no improvement.

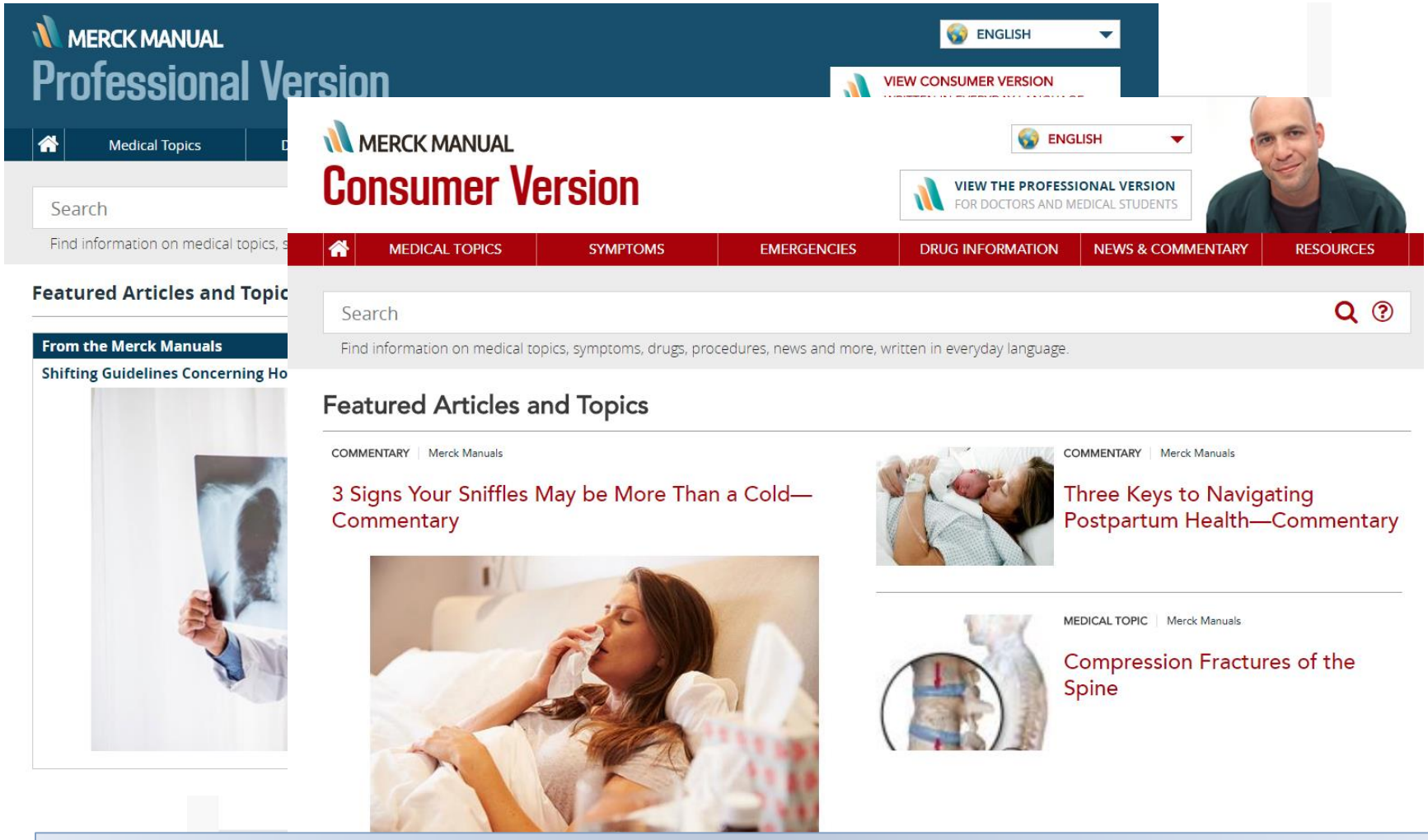
Review: Family therapy for asthma in children
Comparison: 1 FAMILY THERAPY & DRUG THERAPY vs DRUG THERAPY ALONE
Outcome: 1 General paediatric assessment - no improvement

Study or subgroup	Treatment n/N	Control n/N	Peto Odds Ratio Peto,Fixed,95% CI	Peto Odds Ratio Peto,Fixed,95% CI
Gustafsson 1986	3/9	7/8		0.12 [0.02, 0.80]



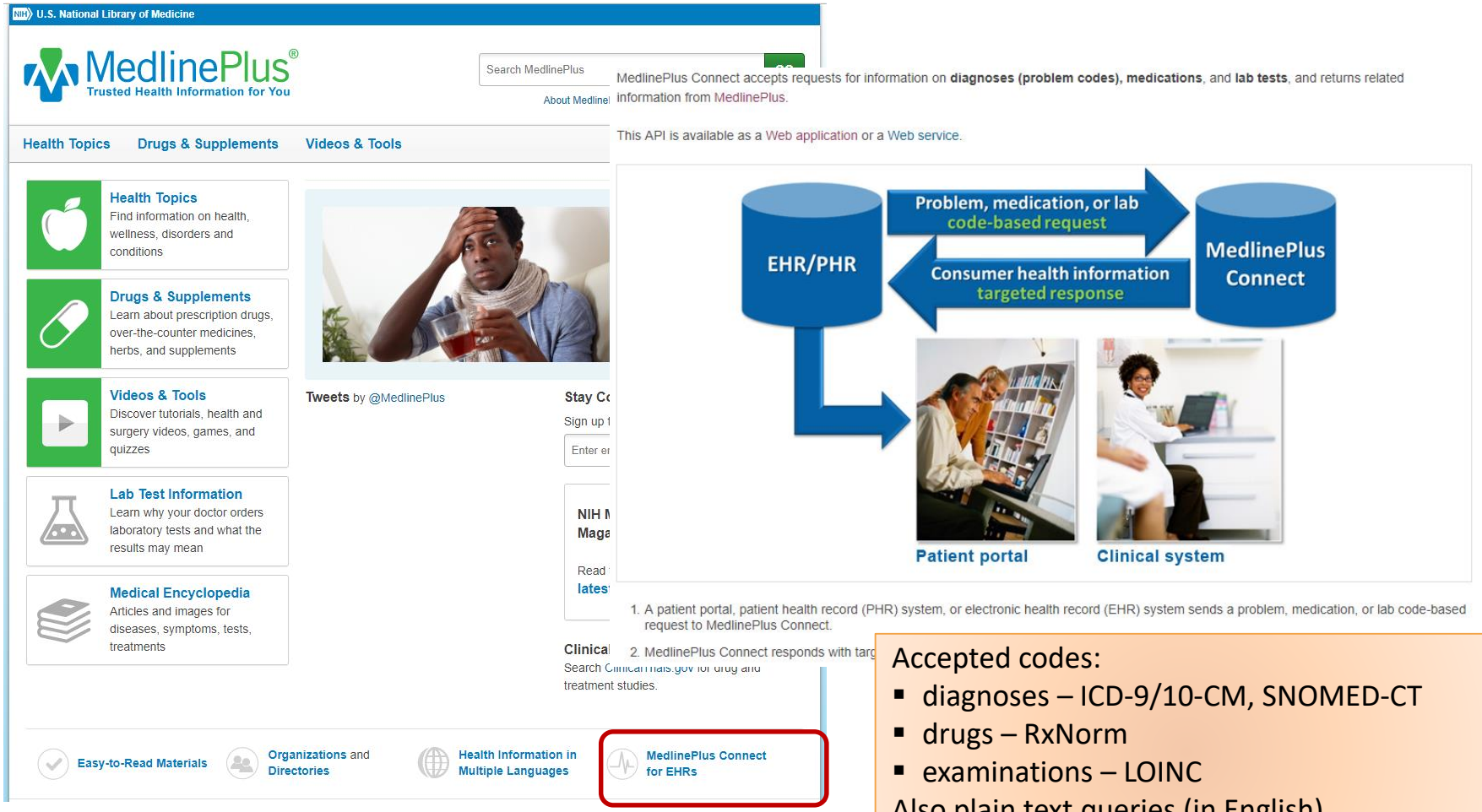
Copyright © 2008 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd.

Merck Manual



http://www.merckmanuals.com/

MedlinePlus



Accepted codes:

- diagnoses – ICD-9/10-CM, SNOMED-CT
- drugs – RxNorm
- examinations – LOINC

Also plain text queries (in English)

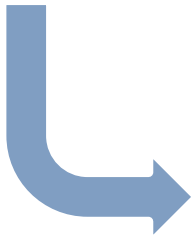
Results returned in JSON

https://medlineplus.gov/

MedLinePlus

ICD-9-CM code = 250.33

https://apps2.nlm.nih.gov/medlineplus/services/mpconnect_service.cfm?mainSearchCriteria.v.cs=2.16.840.1.113883.6.103&mainSearchCriteria.v.c=250.33&knowledg




```
"entry":[
    {
        "lang": "en",
        "title": {
            "_value": "Coma",
            "type": "text"
        },
        "link": [
            {
                "title": "Coma",
                "rel": "alternate",
                "type": "html",
                "href": "https://medlineplus.gov/coma.html?utm_source=mplusconnect&
                utm_medium=service"
            }
        ],
        "id": {
            "_value": "tag: https:, 2018-12-01:https://medlineplus.gov/coma.html?
            utm_source=mplusconnect&utm_medium=service"
        },
        "updated":{
            "_value": "2018-01-12T00:01:27Z"
        },
        "summary": {
            "_value": "<p>A coma is a deep state of unconsciousness.   An individual in a coma
            is alive but unable to move or respond to his or her environment. Coma may occur
            as a complication of an underlying illness, or as a result of injuries, such as <a
            href=\"https://medlineplus.gov/traumaticbraininjury.html?utm_source=mplusconnect&
            amp;utm_medium=service\">brain injury</a>.</p><p>A coma rarely lasts more than 2
            to 4 weeks. The outcome for coma depends on the cause, severity, and site of the
            damage. People may come out of a coma with physical, intellectual, and
            psychological problems.  Some people may remain in a coma for years or even
            decades. For those people, the most common cause of death is infection, such as
            pneumonia.</p><p class=\"NLMattribution\">   NIH: National Institute of
            Neurological Disorders and Stroke</p> <p class=\"NLMrelatedLinks\"><ul><li><a
            href=\"https://medlineplus.gov/ency/article/003931.htm?utm_source=mplusconnect&amp;
            utm_medium=service\">EEG</a> (Medical Encyclopedia)</li></ul></p>",
            "type": "html"
        }
    },
```

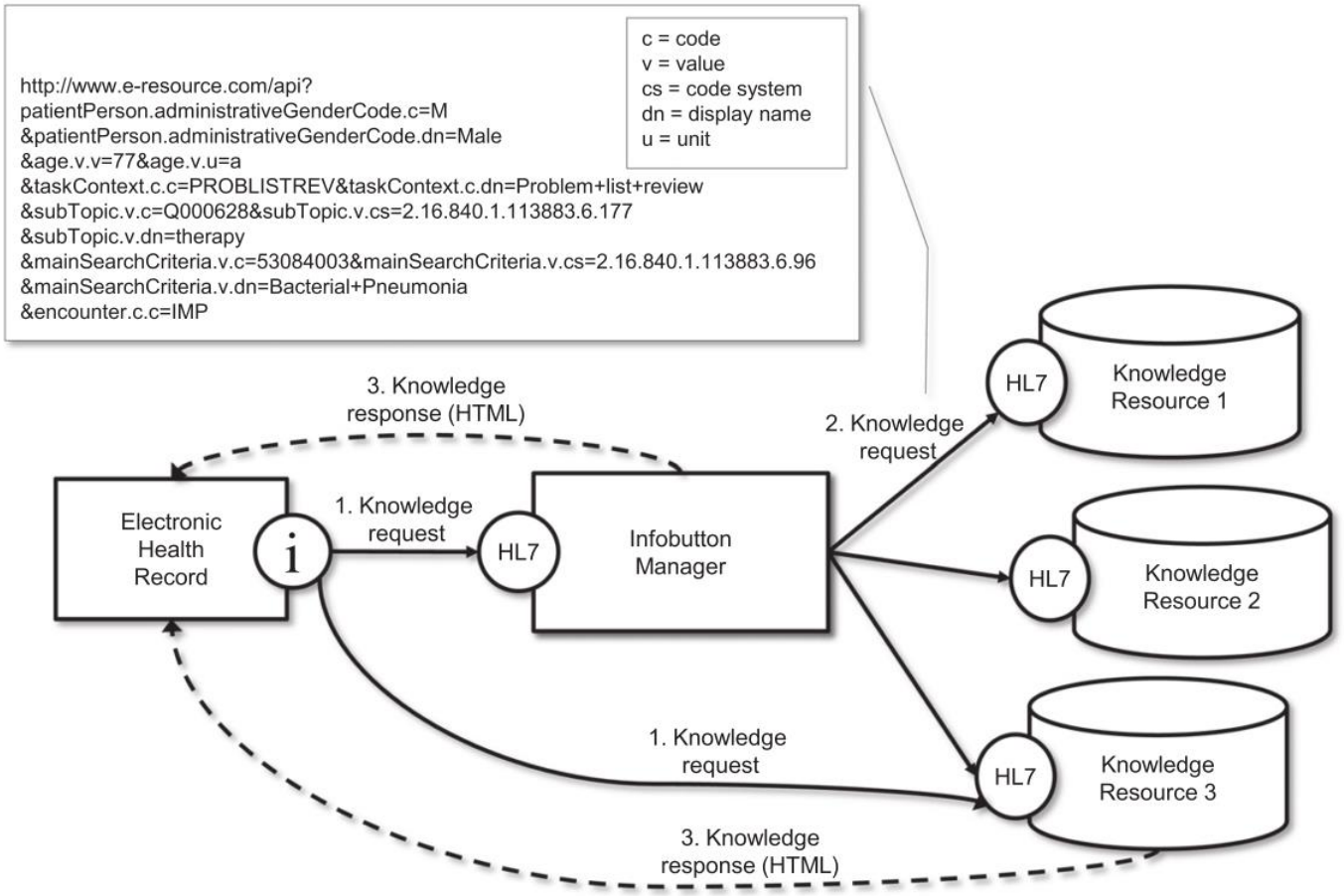
INFOBUTTONS

Infobuttons

- Links from EPR/HIS to (potentially) relevant resources in external repositories with knowledge-based information
- **Infobutton Manager** – selection of a specific information source depending on the context (e.g., patient data, query topic, "consumer" of a response)
- Detailed specification in the HL7 standard (*context-aware information retrieval*)
  - Based on HL7 V3 RIM
  - Query and its content defined as RMIM
  - Invocation via URL, SOAP or RESTful
- Open-source implementation of the infrastructure

http://www.infobuttons.org/ and http://www.openinfobutton.org/

Infobutton Manager – schemat działania

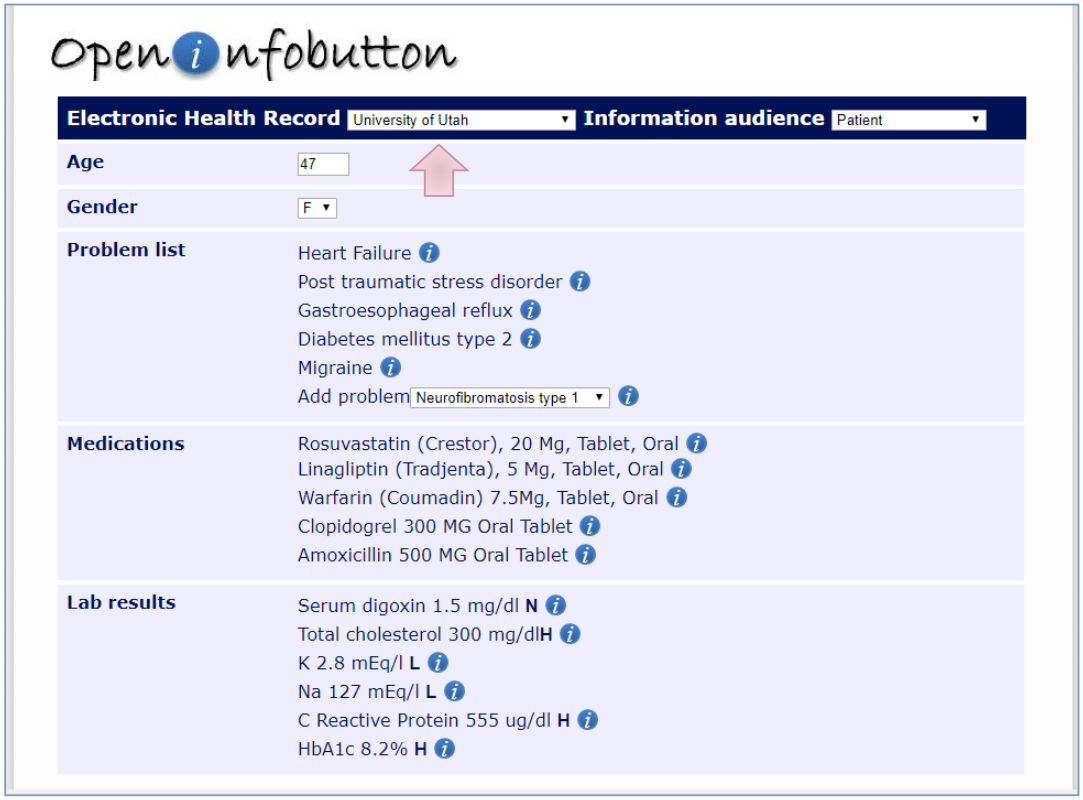


Del Fiol G, Huser V, Strasberg HR, Maviglia SM, Curtis C, Cimino JJ. Implementations of the HL7 Context-Aware Knowledge Retrieval (“Infobutton”) Standard: challenges, strengths, limitations, and uptake. *J Biomed Inform* 2012;45(4):726–35.

Infobutton - context

http://lite.bmi.utah.edu/InfobuttonQA.html

Infobuttons – demo



http://lite.bmi.utah.edu/OpenInfobuttonDemo.html

INDEXING AND SEARCHING

Indexing of resources

- **Manual** indexing
  - Conducted by human experts
  - Using some controlled terminology and an approved protocol (e.g., for solving disagreement between experts)
  - Applied (usually) to bibliographic and annotated resources
- **Automatic** indexing
  - Using terms appearing in indexed resources and/or some controlled terminology
  - Applied typically to full text resources
- **Hybrid** indexing → automatic indexes "cleaned" later by human experts (e.g., MEDLINE)

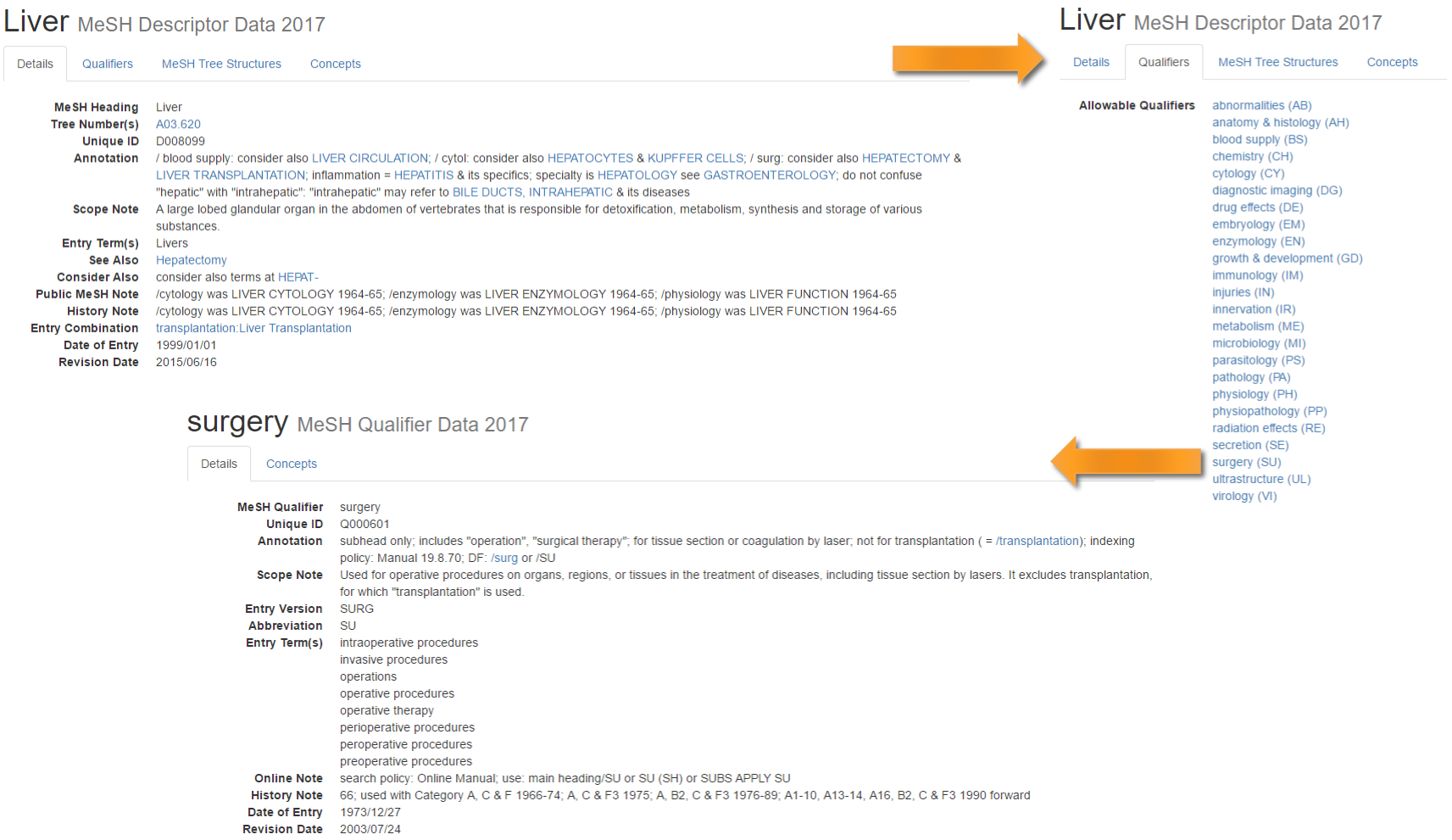
Controlled terminologies

Controlled terminology is a set of concepts (a canonical term and alternatives) and relations between them (hierarchy, synonym, related) used for indexing and for encoding resources.

- MeSH (Medical Subject Heading)
  - Terminology developed by NLM (MEDLINE), 1st edition in 1960
  - Annual updates and extensions: 5,700 concepts in 1st edition and 26k+ in the most recent (170k+ terms)
  - Types of entries
    - *Descriptor* (main heading) → a biomedical concept
    - *Qualifier* (subheading) → additional characteristic of a descriptor
    - *Supplementary concept record*) → drug or other chemical substance
- EMTREE (for EMBASE)

http://www.nlm.nih.gov/mesh/

MeSH descriptors and qualifiers



Liver MeSH Descriptor Data 2017

Details | Qualifiers | MeSH Tree Structures | Concepts

MeSH Heading	Liver
Tree Number(s)	A03.620
Unique ID	D008099
Annotation	/ blood supply: consider also LIVER CIRCULATION; / cytol: consider also HEPATOCYTES & KUPFFER CELLS; / surg: consider also HEPATECTOMY & LIVER TRANSPLANTATION; inflammation = HEPATITIS & its specifics; specialty is HEPATOLOGY see GASTROENTEROLOGY; do not confuse "hepatic" with "intrahepatic": "intrahepatic" may refer to BILE DUCTS, INTRAHEPATIC & its diseases
Scope Note	A large lobed glandular organ in the abdomen of vertebrates that is responsible for detoxification, metabolism, synthesis and storage of various substances.
Entry Term(s)	Livers
See Also	Hepatectomy
Consider Also	consider also terms at HEPAT-
Public MeSH Note	/cytology was LIVER CYTOLOGY 1964-65; /enzymology was LIVER ENZYMOLOGY 1964-65; /physiology was LIVER FUNCTION 1964-65
History Note	/cytology was LIVER CYTOLOGY 1964-65; /enzymology was LIVER ENZYMOLOGY 1964-65; /physiology was LIVER FUNCTION 1964-65
Entry Combination	transplantation:Liver Transplantation
Date of Entry	1999/01/01
Revision Date	2015/06/16

Liver MeSH Descriptor Data 2017

Details | Qualifiers | MeSH Tree Structures | Concepts

Allowable Qualifiers	abnormalities (AB) anatomy & histology (AH) blood supply (BS) chemistry (CH) cytology (CY) diagnostic imaging (DG) drug effects (DE) embryology (EM) enzymology (EN) growth & development (GD) immunology (IM) injuries (IN) innervation (IR) metabolism (ME) microbiology (MI) parasitology (PS) pathology (PA) physiology (PH) physiopathology (PP) radiation effects (RE) secretion (SE) surgery (SU) ultrastructure (UL) virology (VI)

surgery MeSH Qualifier Data 2017

Details | Concepts

MeSH Qualifier	surgery
Unique ID	Q000601
Annotation	subhead only; includes "operation", "surgical therapy"; for tissue section or coagulation by laser; not for transplantation (= /transplantation); indexing policy: Manual 19.8.70; DF: /surg or /SU
Scope Note	Used for operative procedures on organs, regions, or tissues in the treatment of diseases, including tissue section by lasers. It excludes transplantation, for which "transplantation" is used.
Entry Version	SURG
Abbreviation	SU
Entry Term(s)	intraoperative procedures invasive procedures operations operative procedures operative therapy perioperative procedures peroperative procedures preoperative procedures
Online Note	search policy: Online Manual; use: main heading/SU or SU (SH) or SUBS APPLY SU
History Note	66; used with Category A, C & F 1966-74; A, C & F3 1975; A, B2, C & F3 1976-89; A1-10, A13-14, A16, B2, C & F3 1990 forward
Date of Entry	1973/12/27
Revision Date	2003/07/24

Organization of descriptors

- Descriptors are divided into 16 categories
- Categories are represented as multi-level trees
  - From the most general to specific descriptors
  - Up to 13 levels of hierarchy
- A descriptor may appear in several categories (trees)

1. − 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] +
2. − Organisms [B]
   - Animals [B01] +
   - Algae [B02] +
   - Bacteria [B03] +
   - Viruses [B04] +
   - Fungi [B05] +
   - Plants [B06] +
   - Archaea [B07] +
   - Mesomycetozoea [B08] +
3. + Diseases [C]
4. + Chemicals and Drugs [D]
5. + Analytical, Diagnostic and Therapeutic
6. + Psychiatry and Psychology [F]
7. + Phenomena and Processes [G]
8. + Disciplines and Occupations [H]
9. + Anthropology, Education, Sociology
10. + Technology, Industry, Agriculture [J]
11. + Humanities [K]

Virus Diseases [C02]
- RNA Virus Infections [C02.782]
  - Orthomyxoviridae Infections [C02.782.620]
    - ➤ Influenza, Human [C02.782.620.365]
    - Influenza in Birds [C02.782.620.375]

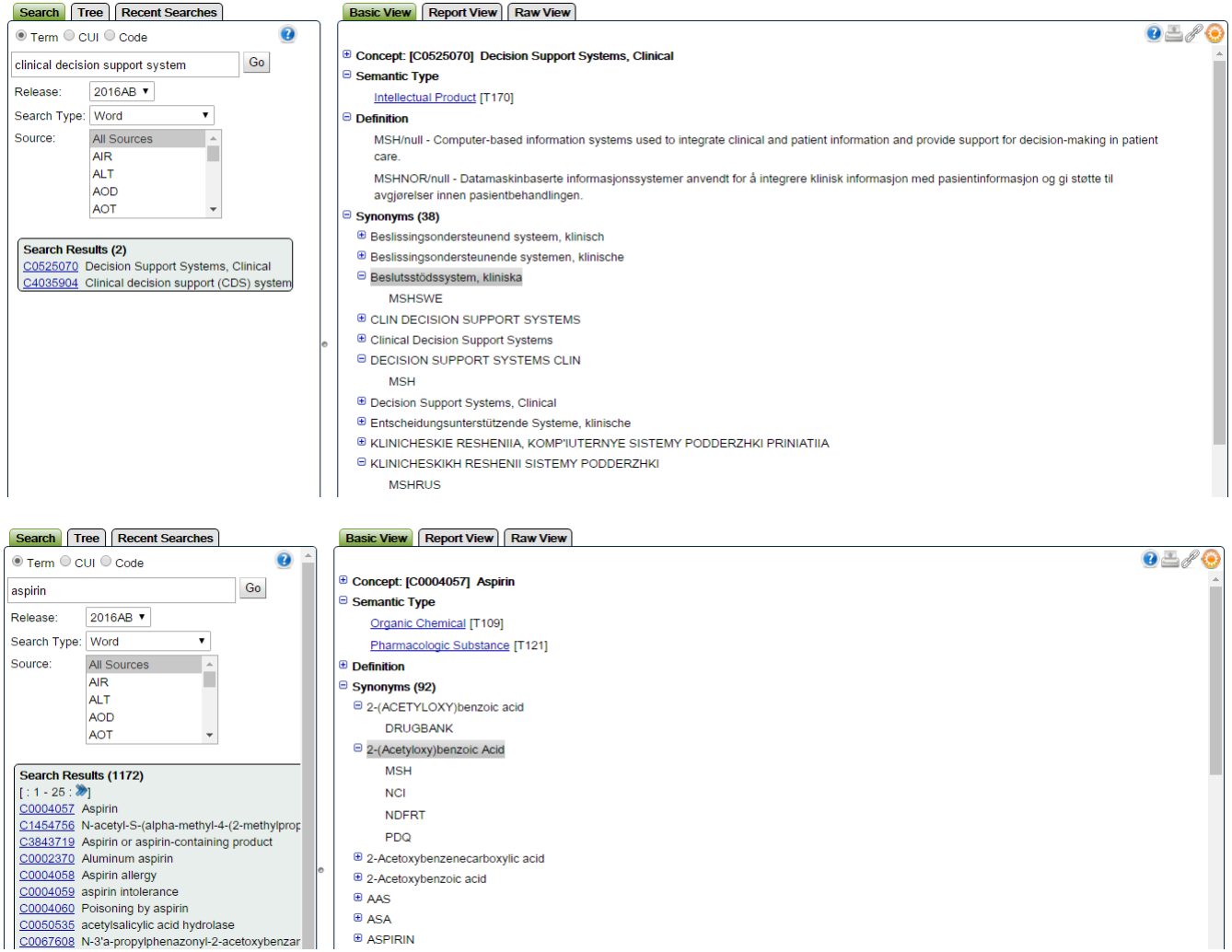
Respiratory Tract Diseases [C08]
- Respiratory Tract Infections [C08.730]
  - Bovine Respiratory Disease Complex [C08.730.085] +
  - Bronchitis [C08.730.099] +
  - Common Cold [C08.730.162]
  - Empyema, Pleural [C08.730.265] +
  - ➤ Influenza, Human [C08.730.310]
  - Laryngitis [C08.730.368] +
  - Legionellosis [C08.730.382] +
  - Lung Abscess [C08.730.407]
  - Lung Diseases, Fungal [C08.730.435] +

Unified Medical Language System (UMLS)

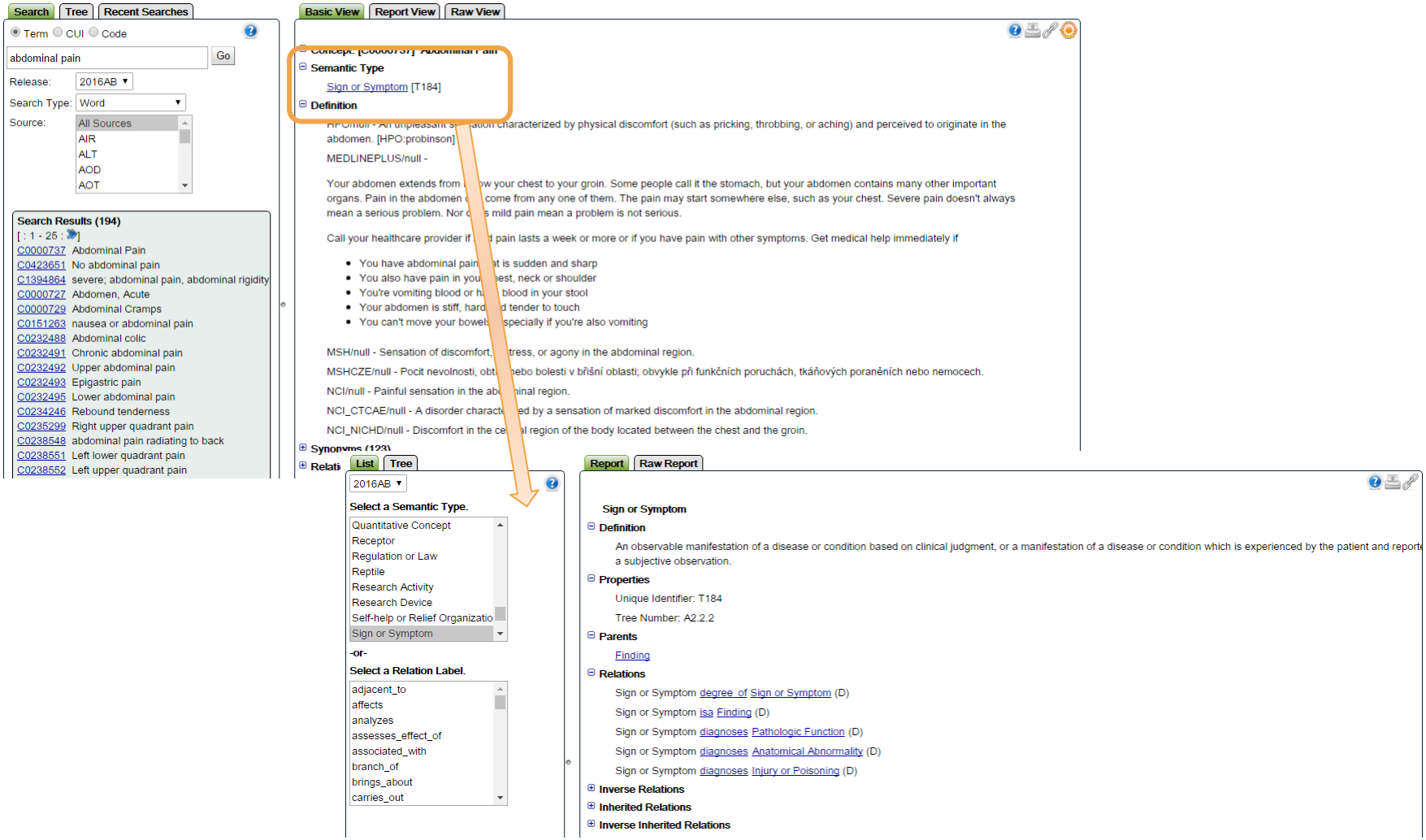
- A set of tools and resources to support automatic processing of biomedical texts developed by NLM
- Metathesaurus
  - Combines multiple existing dictionaries and terminologies
  - Hierarchy and relations between concepts coming from different sources and specified in different languages
- Semantic Network
  - Hierarchy and relations between semantic types associated with specific concepts defined in Metathesaurs
- API for invoking UMLS functions from external applications

http://www.nlm.nih.gov/research/umls/

Metathesaurus



Semantic Network



MetaMap

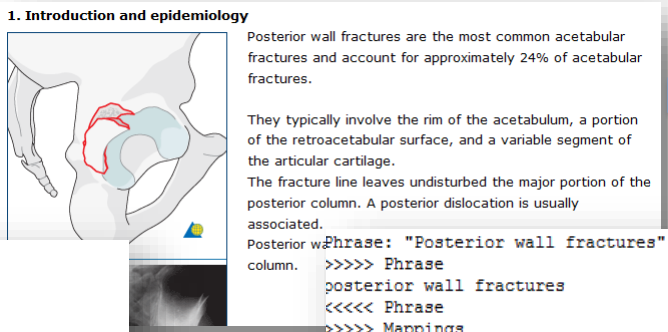
- System for analyzing biomedical texts (in English) and recognizing concepts from Metathesaurus
- Core element of Medical Text Indexer (MTI) used by NLM for automatic indexing of medical publications
- Available for free (open source), can be used to built custom solutions; also available online

http://metamap.nlm.nih.gov/

MetaMap

MeSH only

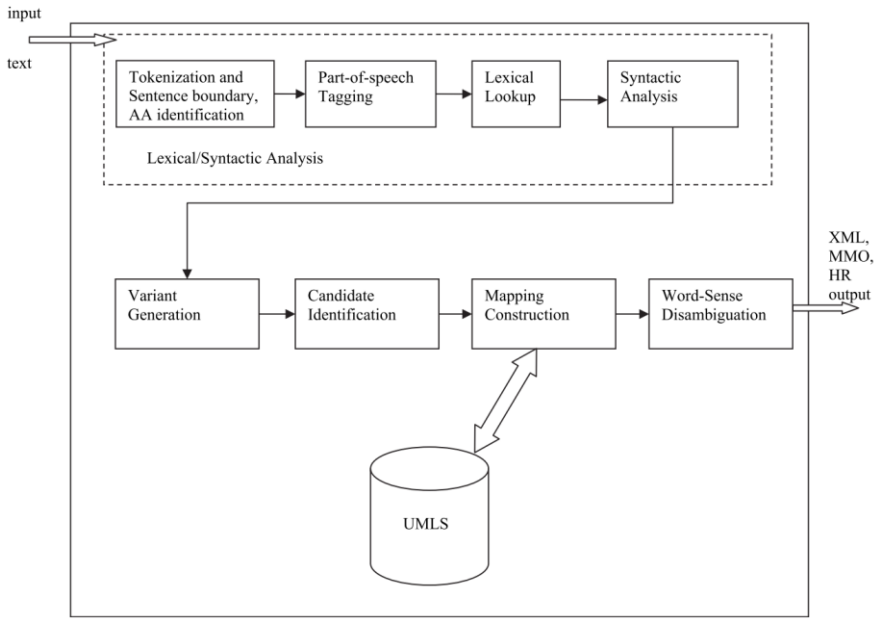
All available terminologies




```
Phrase: "Posterior wall fractures"
>>>>> Phrase
posterior wall fractures
<<<<< Phrase
>>>>> Mappings
Meta Mapping (827):
   827 Fractures (Fractures, Bone) [Injury or Poisoning]
<<<<< Mappings

Phrase: "are"
>>>>> Phrase
<<<<< Phrase

Phrase: "the most common acetabular fractures"
>>>>> Phrase
most common acetabular fractures
<<<<< Phrase
>>>>> Mappings
Meta Mapping (704):
   574 Acetabulum [Body Part, Organ, or Organ Component]
   812 Fractures (Fractures, Bone) [Injury or Poisoning]
<<<<< Mappings

Phrase: "and"
>>>>> Phrase
<<<<< Phrase

Phrase: "account"
>>>>> Phrase
account
<<<<< Phrase
>>>>> Mappings
Meta Mapping (900):
   900 Accountability [Idea or Concept]
<<<<< Mappings
```

```
Phrase: "Posterior wall fractures"
>>>>> Phrase
posterior wall fractures
<<<<< Phrase
>>>>> Mappings
Meta Mapping (981):
   981 Posterior wall fracture [Intellectual Product]
<<<<< Mappings

Phrase: "are"
>>>>> Phrase
<<<<< Phrase

Phrase: "the most common acetabular fractures"
>>>>> Phrase
most common acetabular fractures
<<<<< Phrase
>>>>> Mappings
Meta Mapping (815):
   645 Common (Common (qualifier value)) [Quantitative Concept]
   842 Acetabular Fracture [Injury or Poisoning]
Meta Mapping (815):
   645 Common (shared attribute) [Functional Concept]
   842 Acetabular Fracture [Injury or Poisoning]
<<<<< Mappings

Phrase: "and"
>>>>> Phrase
<<<<< Phrase

Phrase: "account"
>>>>> Phrase
account
<<<<< Phrase
>>>>> Mappings
Meta Mapping (1000):
  1000 Account # (Account number:Identifier:Point in time:^Patient:Nominal) [Clinical Attribute]
Meta Mapping (1000):
  1000 account (account - ActClass) [Idea or Concept]
<<<<< Mappings
```

Text processing in MetaMap



1. Tokenization, sentence boundary determination and acronym/abbreviation identification
2. Part-of-speech tagging
3. Lexical lookup of input words in the SPECIALIST lexicon (UMLS)
4. Syntactic analysis – identification of phrases
5. Variant generation (with variants of all phrase words)
6. Candidate identification and evaluation
7. Mapping construction by combining best candidates
8. Word sense disambiguation by checking semantic consistency of mappings and surrounding text

Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;17(3):229-36.

Automatic indexing

- Preprocessing of indexed documents
  - Splitting the text into words/tokens (*tokenization*)
  - Removal of “insignificant” words (stopwords)
  - Normalization of the format (e.g., all lowercase)
  - Transformation of words into their canonical form (*terms*) through *stemming* (a common part) lub *lemmatization* (base form)

Stemming: am, are, is $\rightarrow$ am, ar, is
Lemmatization: am, are, is $\rightarrow$ be

- **Bag of words** representation – a collection of all terms that appear in the text
  - Direct construction of the index
  - Computation of additional metrics (TF-IDF)

TD-IDF representation

TF = *term frequency*
IDF = *inverse document frequency*

- Each document represented as a vector in *m*-dimensional space (*m* = number of terms, also called *vector-space model*)
- Specific vector elements defined as

$$TFIDF(t,d) = TF(t,d) \times IDF(t)$$

$$TF(t,d) = \text{number of occurrences of term } t \text{ in document } d$$

$$IDF(t) = log\left(\frac{\text{number of all documents}}{\text{number of documents containing term } t}\right)$$

- Combination of local (TF) and global (IDF) perspectives when evaluating specific terms and documents

Problems with automatic indexing

- **Synonymy** – different words, same meaning
- **Polysemy** – the same word, different meanings
- **Content** – words in a document may not reflect its major focus (digressions, references to other concepts)
- **Context** – words take on meaning based on their surrounding (e.g., high blood pressure)
- **Granularity** – queries and documents describe concepts at different levels (e.g., classes of drugs and specific drugs)

Possible solution (to some of the problems): index or query expansion, using n-grams or word embedding…

Searching and retrieval

- Two approaches for matching documents and queries
  - **Exact match** – queries given as logical expressions (AND, OR , NOT) → full match of a document an query is required
  - **Partial match** – application of various matching measures, selection of best matching documents → *relevance ranking*
- Exact match traditionally aimed for bibliographic and annotated repositories, while partial match at full-text ones

Exact match is preferred by more advanced users (perception of better control), however, no significant differences in obtained results

Computing partial matching

- **Jaccard** similarity – for the *bag of words* representation

$$sim(d_i, d_j) = \frac{|d_i \cap d_j|}{|d_i \cup d_j|}$$

- **Cosine** similarity – dla the TFIDF (vector) representation

$$sim(d_i, d_j) = \frac{d_i \cdot d_j}{\|d_i\| \times \|d_j\|} = \frac{\sum_t TFIDF(t, d_i) \times TFIDF(t, d_j)}{\|d_i\| \times \|d_j\|}$$

When computing similarity between a document and a query,
then TF is established from the query, and IDF from the set of documents.

Evaluation measures for IR systems

Direct or **system-oriented** evaluation

- *Precision* and *recall*

$$precision(q) = \frac{\left|retrieved_q \cap relevant_q\right|}{\left|retrieved_q\right|}$$

$$recall(q) = \frac{\left|retrieved_q \cap relevant_q\right|}{\left|relevant_q\right|}$$

- *Precision@k* – established for the first *k* retrieved documents
- *Average precision* (taking into account the order of returned documents) and *mean average precision* (for a set of queries)

$$average_precision(q) = \frac{\sum_{k=1}^{n} precision@k_q}{\left|retrieved_q \cap relevant_q\right|}$$

Problem with direct evaluation

- Relies on the "ground truth" established by experts
- For each query there should be a (separate) set of relevant documents (→ significant workload imposed on experts)
- Significant differences between experts in evaluations of documents for a given query (κ – Cohen's kappa)
  - OHSUMED collection κ = 0.41
  - Our experiment κ = 0.30

Poor agreement between observers!

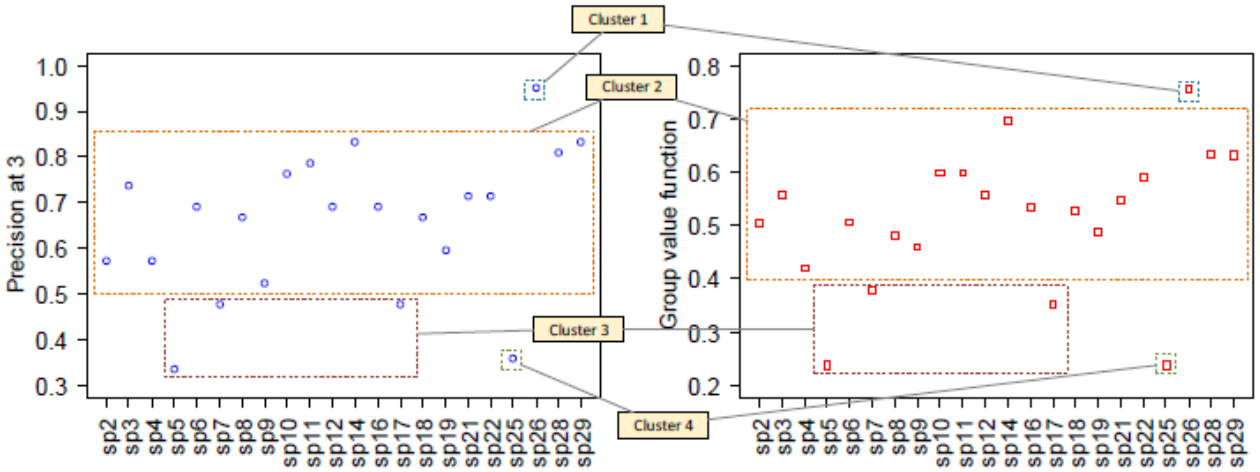
Our experiment…

Appendix 3. Coded triples representing relevance evaluations by physicians

	Vignette		
Participant	1	2	3
sp2	XXX	XXX	YX
sp3	XNN	XXX	XN
sp4	NNY	XXX	NX
sp5	XXX	XXX	XX
sp6	XNX	YXN	XX
sp7	XXN	XXX	XX
sp8	XXX	NXN	XY
sp9	NNY	YXX	XX
sp10	NNY	YXN	NX
sp11	NNY	YXN	NX
sp12	NNN	XXX	NX
sp14	NYN	YXX	NN
sp16	NNN	YXX	XX
sp17	XXN	NXX	XX
sp18	NYN	XXX	NN
sp19	XYN	XXX	YX
sp21	NXN	YXX	XX
sp22	NNN	XNN	NN
sp25	NNX	XXX	NX
sp26	YYY	YNN	NN
sp28	YNN	YNN	NN
sp29	NNY	NXN	NX

There are major differences b
vignette. For example, physi
(considered all of them to be
(considered all of them not o

Appendix 6. Selected clustering of physicians in the context of precision at 3 and group value function



Physician sp5 from cluster 3 and sp25 from cluster 4 are very similar in terms of these two measures. However, a closer look at their coded triples (Appendix 3) reveals differences in evaluations across vignettes. For example, reviews retrieved for vignette 1 were evaluated as *XXX* by sp5 and as *NNX* by sp25, while for vignette 10 the evaluations were *NNX* for sp5 and *XXX* for sp25. While these differences were compensated after averaging values of both measures over all vignettes, they were captured by the kappa coefficient (that indicated the lack of agreement between sp5 and sp25) and resulted in placing these two physicians in two different clusters.

Other approaches to evaluation

- **Indirect** or **user-oriented** evaluation – impact of using the IR system of the outcome of tasks conducted by its users
- Typical scenario – answering questions before and after using an IR system to search for information
  - No differences between exact and partial query matching techniques
  - Diversified increase in accuracy depending on the class of users (larger for nurses, smaller for physicians), however, comparable final results

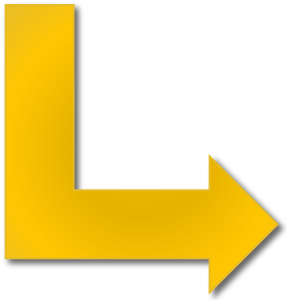
Hersh, W. (2014). Information retrieval and digital libraries. In: Shortliffe, E.H., Cimino, J. (eds.): *Biomedical Informatics. Computer Applications in Health Care and Biomedicine*, Springer, 626-654.

Impact on EBM and IR systems on decisions

Table 3 Nine observational studies reporting cognitive impact of clinical information-retrieval technology on physicians (sorted by frequency of searches for information with positive impact)

Reference number	Searches with positive impact (%)	Number of searches	Number of participants	Recall	Design	Platform
Hayward et al. [28]	20	20	9	Up to 1 month	Cross-sectional	Multiple databases on CD-ROM
Jousimaa et al. [21]	36	2036	102	None	Cohort	Finnish guidelines on CD-ROM
Lindberg et al. [29]	36	1158	552	CIT[a]: Up to 12 months	Cross-sectional	Bibliographic database on CD-ROM
Swinglehurst et al. [30]	39	60	22	Up to 1 month	Case series	Multiple databases (device not reported)
Haynes et al. [20]	41	280	158	Up to 8 months	Cohort	Bibliographic database on the internet
Gorman et al. [14]	51	60	48	Up to 14 months	Cross-sectional	Bibliographic database on the internet
Veenstra [31]	59	261	30	Up to 12 months	Cross-sectional	Bibliographic database on the internet
Schwartz et al. [17]	70	92	3	Not specified	Cohort	Multiple databases on the internet
Crowley et al. [27]	82	625	82	None	Cohort	Multiple databases on the internet

[a] CIT: critical incident technique. This technique is known to be reliable and valid, and may reduce recall bias.



Changed patient management
Information influenced decision
An impact on clinical problem solving
Increased understanding/knowledge or provided reassurance
Would have had an impact on doctors or their practice
Had an impact
Would affect the treatment of future patients
Confirmed patient care decisions or changed patient management

Pluye P, Grad RM, Dunikowski LG, Stephenson R. Impact of clinical information-retrieval technology on physicians: a literature review of quantitative, qualitative and mixed methods studies. *Int J Med Inform* 2005;74(9):745-68.