REPOZYTORIA ZASOBÓW MEDYCZNYCH, ICH INDEKSOWANIE I PRZESZUKIWANIE

Informatyka w Medycynie, 2017-2018

WPROWADZENIE

Wydobywanie informacji

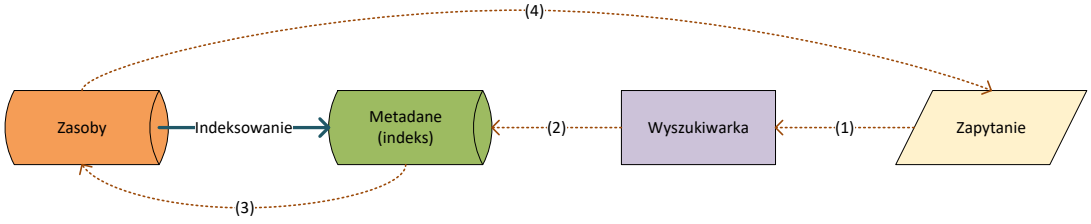
- Rodzaje informacji
  - informacje związane z pacjentem (patient specific) – związane z procesem opieki nad danym pacjentem, przechowywane w systemach EPR/EHR
  - informacje bazująca na wiedzy (knowledge-based) – wyniki i wnioski z eksperymentów i obserwacji, dostępne w książkach, czasopismach, …

Wydobywanie informacji (ang. *information retrieval*, *IR*) to dziedzina zajmująca się pozyskiwaniem (gromadzeniem), organizacją i wyszukiwaniem informacji bazującej na wiedzy

- „Tradycyjne” IR ograniczone do repozytoriów z publikacjami, obecnie rozszerzone na inne formaty (np. multimedia, …)

Proces wydobywania informacji

- Indeksowanie – tworzenie *metadanych* związanych z (poszczególnymi) zasobami
- Wydobywanie – wyszukiwanie na podstawie *metadanych* zasobów odpowiadających zapytaniu użytkownika



Trochę historii…

- 1878 – Index Medicus
  - indeks artykułów medycznych
  - metadane: tytuł, autorzy, słowa kluczowe (*subject headings*)
- 1966 – MEDLARS (Medical Literature Analysis and Retrieval System), NLM
  - komputerowa wersja Index Medicus-a
  - metadane: podobne jak w wersji papierowej (ograniczenia techniczne)
  - wykorzystanie MeSH jako formalnego zestawu słów kluczowych
  - realizacja zapytań przesyłanych tradycyjną pocztą
- lata 90-te –udostępnienie przez WWW baz bibliograficznych prowadzonych przez NLM (→ MEDLINE/PubMed)

INFORMACJA BAZUJĄCA NA WIEDZY I EBM

Informacje bazująca na wiedzy

- Informacja podstawowa (ang. *primary*)
  - nowo odkryta wiedza kliniczna (wraz z powiązanymi danymi)
  - oryginalne wyniki specficznych badań publikowane w literaturze (czasopisma, raporty, …)
- Informacja wtórna (ang. *secondary*)
  - przegląd, podsumowanie lub synteza literatury podstawowej
  - publikowana w różnych formach – artykuły przeglądowe, przeglądy systematyczne (ang. *systematic reviews*), wytyczne i algorytmy kliniczne
  - najczęściej stosowana w praktyce przez lekarzy (→ „bryk” do szybkiego przejrzenia i/lub wykorzystania „przy pacjencie”)

Potrzeby informacyjne

- Różni użytkownicy i różne potrzeby dotyczące informacji → najwięcej badań dotyczących lekarzy
- Stany potrzeby informacyjnej
  1. nierozpoznana – „nie wiem, że czegoś nie wiem”
  2. rozpoznana – „wiem, że czegoś nie wiem”
  3. realizowana – „szukam informacji (nie wiem, czy znajdę)”
  4. wypełniona – „znalazłem potrzebną informację”

Nierozpoznane potrzeby informacyjne są podstawową przyczyną tego, że w praktyce oferowana opieka często nie odpowiada najnowszym zaleceniom czy wytycznym!

Potrzeby informacyjne w praktyce

- Średnio 3 pytania na 4 pacjentów (opieka podstawowa)
  - jaki lek podać? co powoduje dany symptom? jaki test wykonać?
- Poszukiwanie odpowiedzi w 12-36% przypadków
  - brak jawnego sformułowania pytania
  - odstąpienie od poszukiwania ($\rightarrow$ zbyt duża czasochłonność, akceptowalny czas to ok. 2 min.)
  - podstawowe źródło informacji – współpracownicy lub podręczniki
- Integracja mechanizmów IR z systemami typu EMR/EHR w celu automatyzacji procesu wyszukiwania informacji

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)

Rozumne, celowe, proste zastosowanie najnowszych i najbardziej rzetelnych danych naukowych w podejmowaniu decyzji dotyczących opieki zdrowotnej nad indywidualnym pacjentem.

- Połączenie doświadczenia lekarza z najbardziej aktualną informacją bazującą na wiedzy
  - informacja pozwala na weryfikację i aktualizację doświadczenia
  - doświadczenie pozwala na uwzględnienie specyfiki pacjenta
- Modelowy schemat postępowania
  1. Sformułowanie pytania klinicznego (na które można odpowiedzieć)
  2. Identyfikacja dowodów (informacji) zawierających odpowiedź na postawione pytanie
  3. Krytyczna ocena znalezionych dowodów – czy można je zastosować do konkretnego pacjenta?

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.

Zapytania w EBM

- Klasy zapytań klinicznych
  1. ogólne (ang. *background*) – dotyczą natury problemu (niezależne od pacjenta), odpowiedzi w podręcznikach
  2. szczegółowe (ang. *foreground*) – dotyczą postępowania z pacjentem ze specyficznym problemem, typowe dla EBM
- Rodzaje zapytań szczegółowych
  1. terapeutyczne – jaką terapię zastosować
  2. diagnostyczne – jaki test diagnostyczny wybrać
  3. prognostyczne – jak może się rozwijać stwierdzony problem
  4. dotyczące szkodliwego wpływu (ang. *harm*) – jakie są efekty zastosowania danej terapii czy też ekspozycji na jakiś czynnik

Najbardziej wartościowe dowody uzyskiwane na podstawie wyników randomizowanych badań klinicznych (*randomized controlled trial*, *RCT*) → nie zawsze dostępne (np. ze względów etycznych)

REPOZYTORIA INFORMACJI BAZUJĄCEJ NA WIEDZY

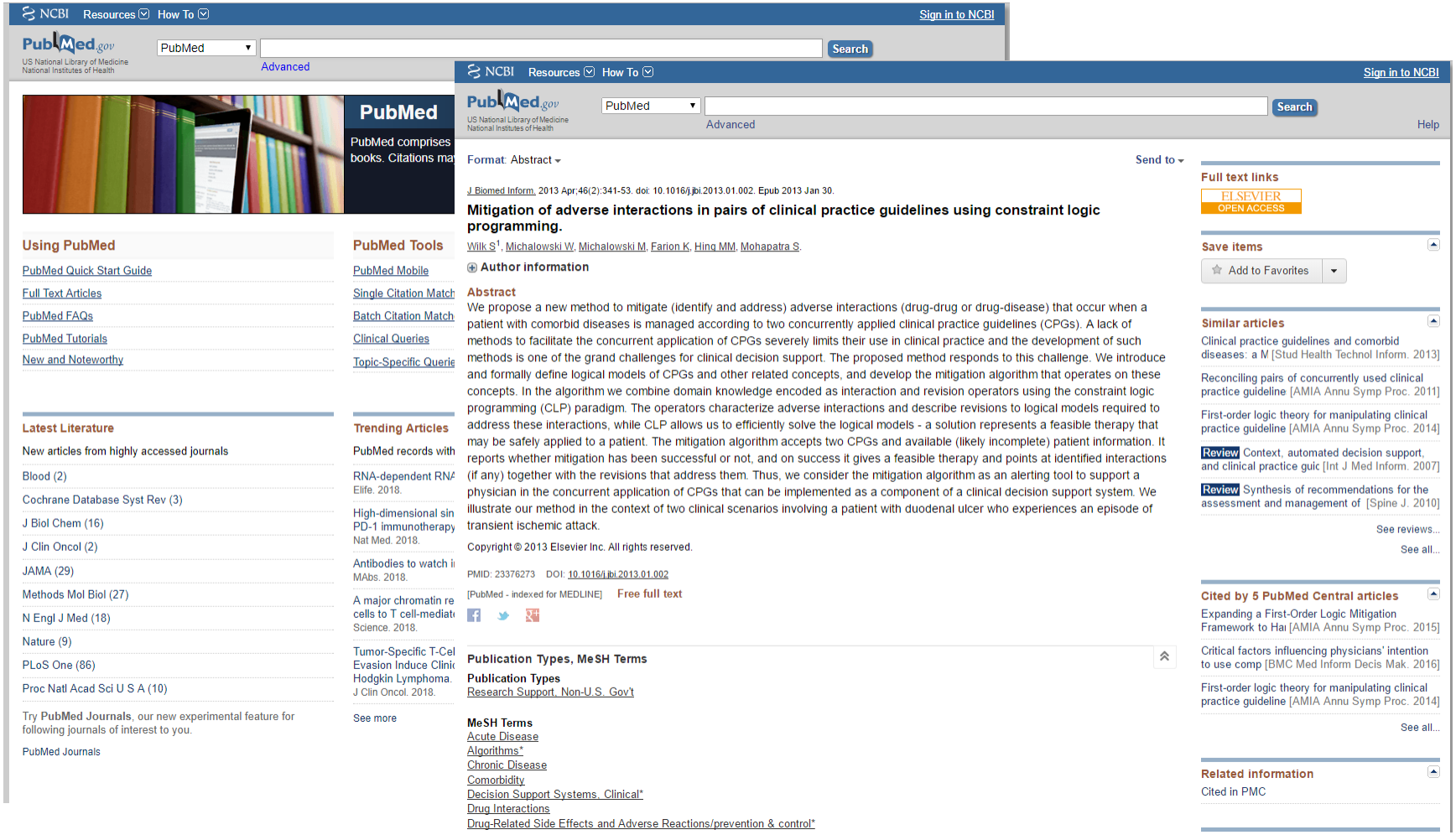
Rodzaje repozytoriów

1. Bazy bibliograficzne
   - podstawowa informacja o publikacjach, bez ich treści
   - przykłady: MEDLINE/PubMed (NLM), EMBASE („europejski MEDLINE”), National Guidelines Clearinghouse (NGC) z wytycznymi klinicznymi
2. Bazy pełnotekstowe
   - pełna treść publikacji, często odnośniki z baz bibliograficznych (np. PubMed $\rightarrow$ strona wydawcy)
   - zazwyczaj utrzymywane przez wydawców komercyjnych, PubMed Central dla publikacji powstałych w ramach grantów NIH

Rodzaje repozytoriów

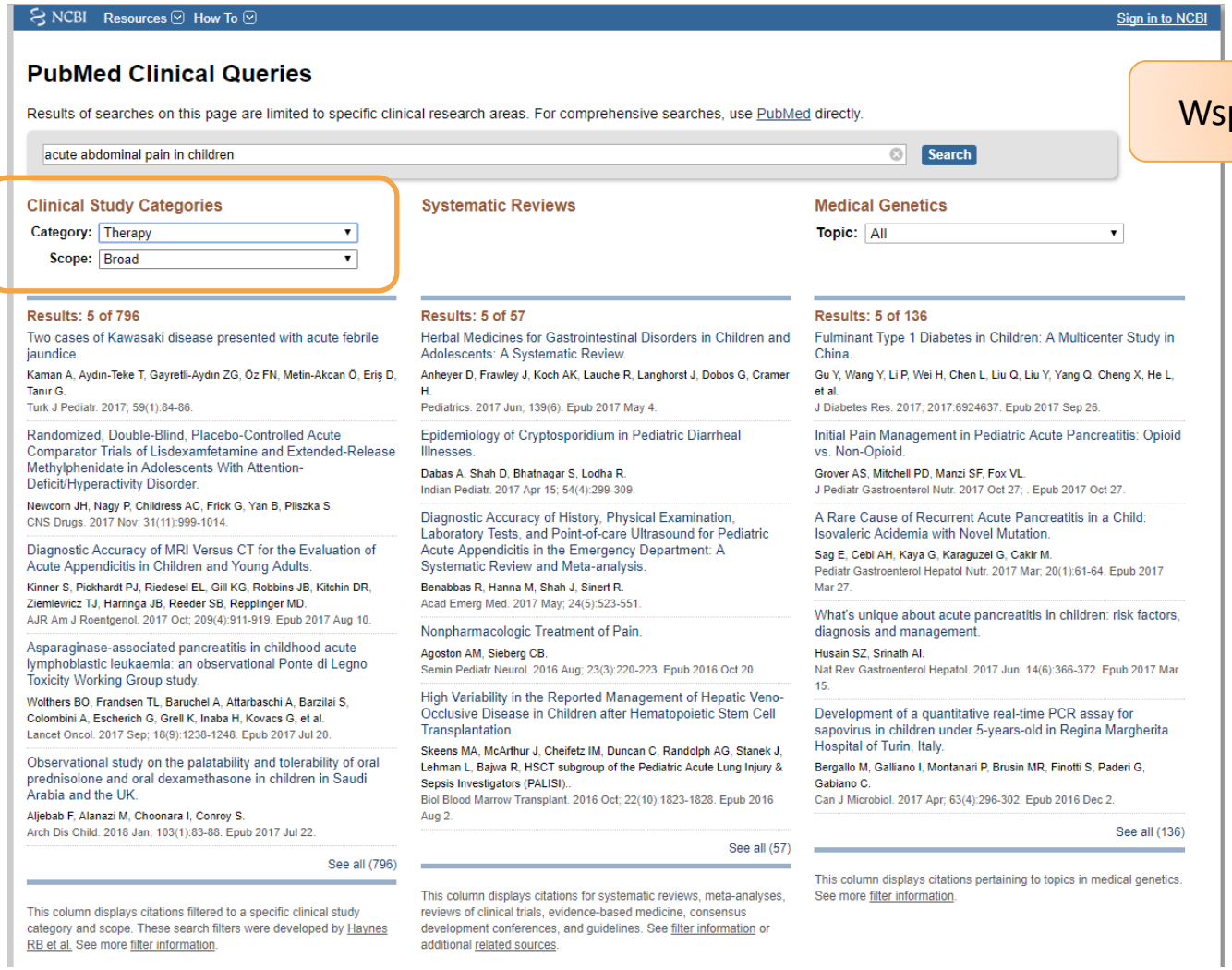
3. Bazy adnotowane (ang. *annotated*)
   - zasoby przechowywana w specjalizowanych bazach danych (obrazowych, biologicznych, cytowań, dowodów EBM, …)
   - przykłady: Visible Human Project, PEIR, The Cochrane Library (Database of Systematic Reviews)
4. Bazy agregujące treści
   - agregacja treści pochodzących z innych typów repozytoriów (kopie treści lub odnośniki do repozytoriów źródłowych)
   - przykłady: MedlinePlus, Merck Manual

PubMed



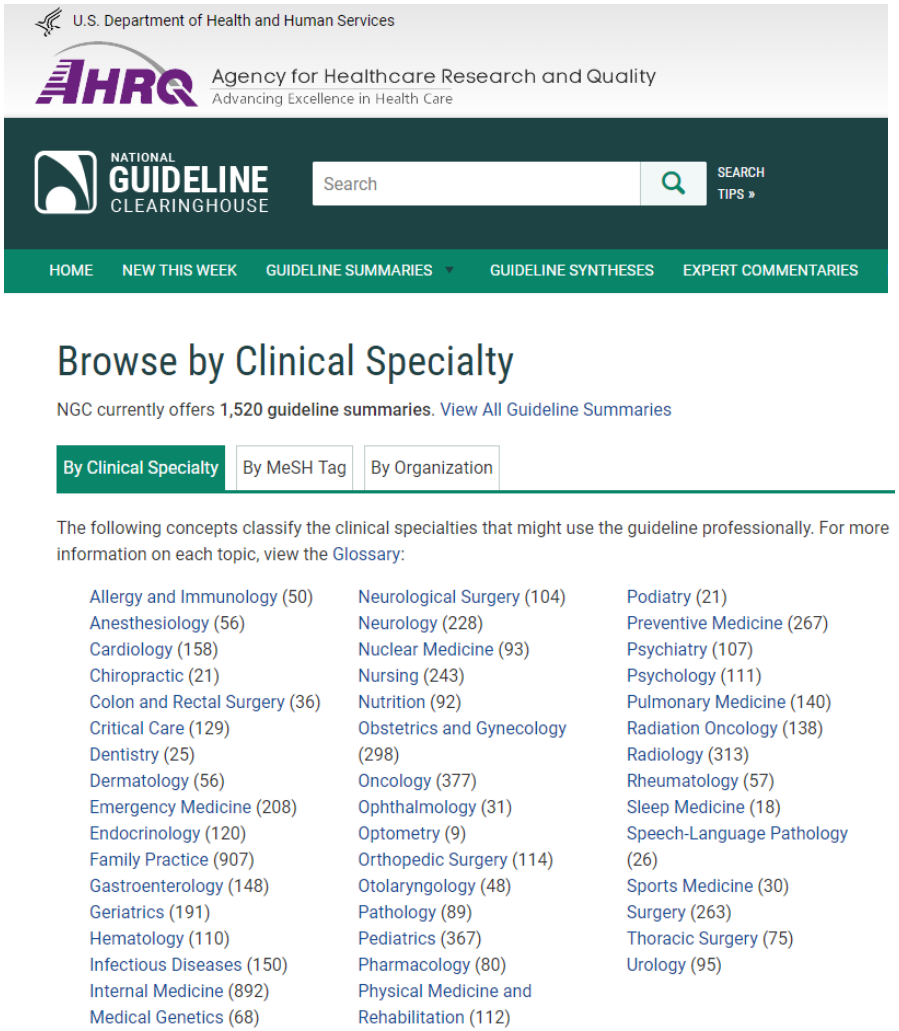
https://www.ncbi.nlm.nih.gov/pubmed

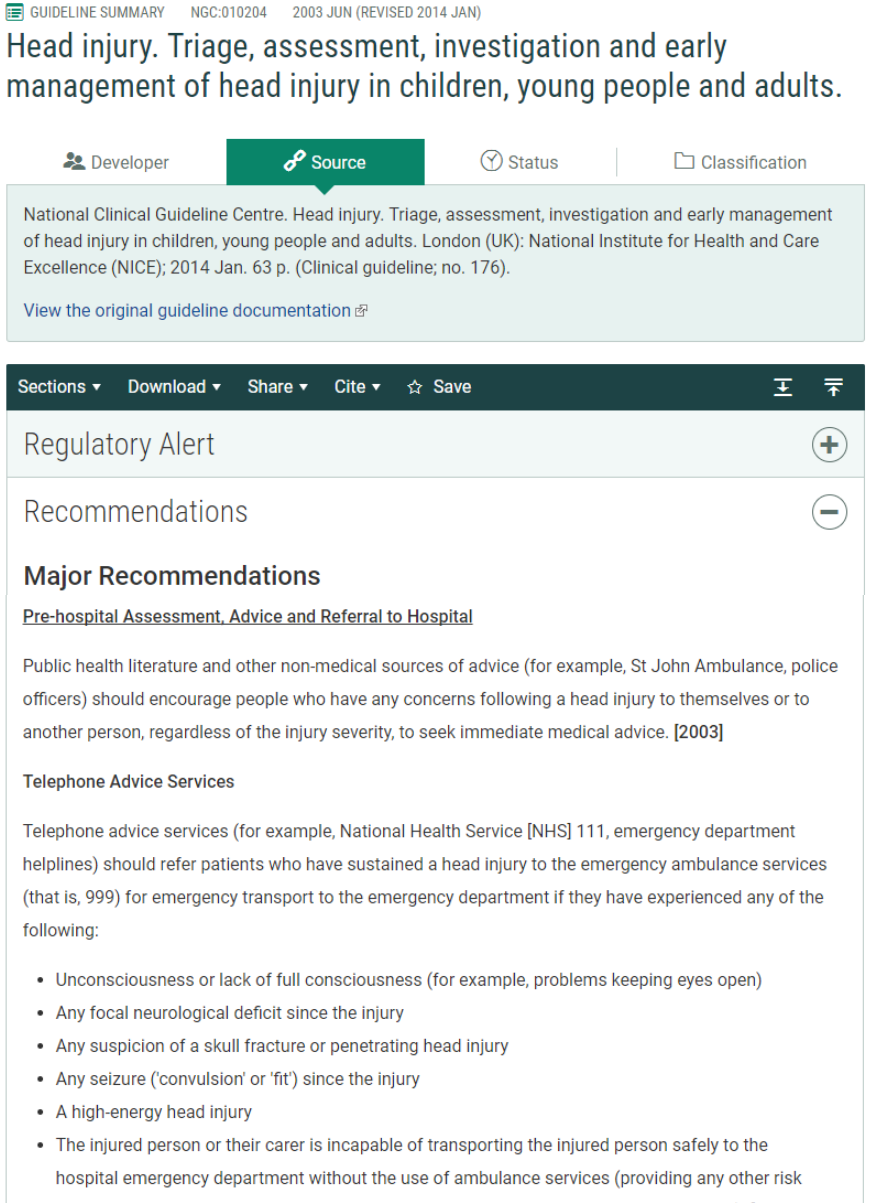
PubMed – Clinical Queries



Wsparcie dla EBM

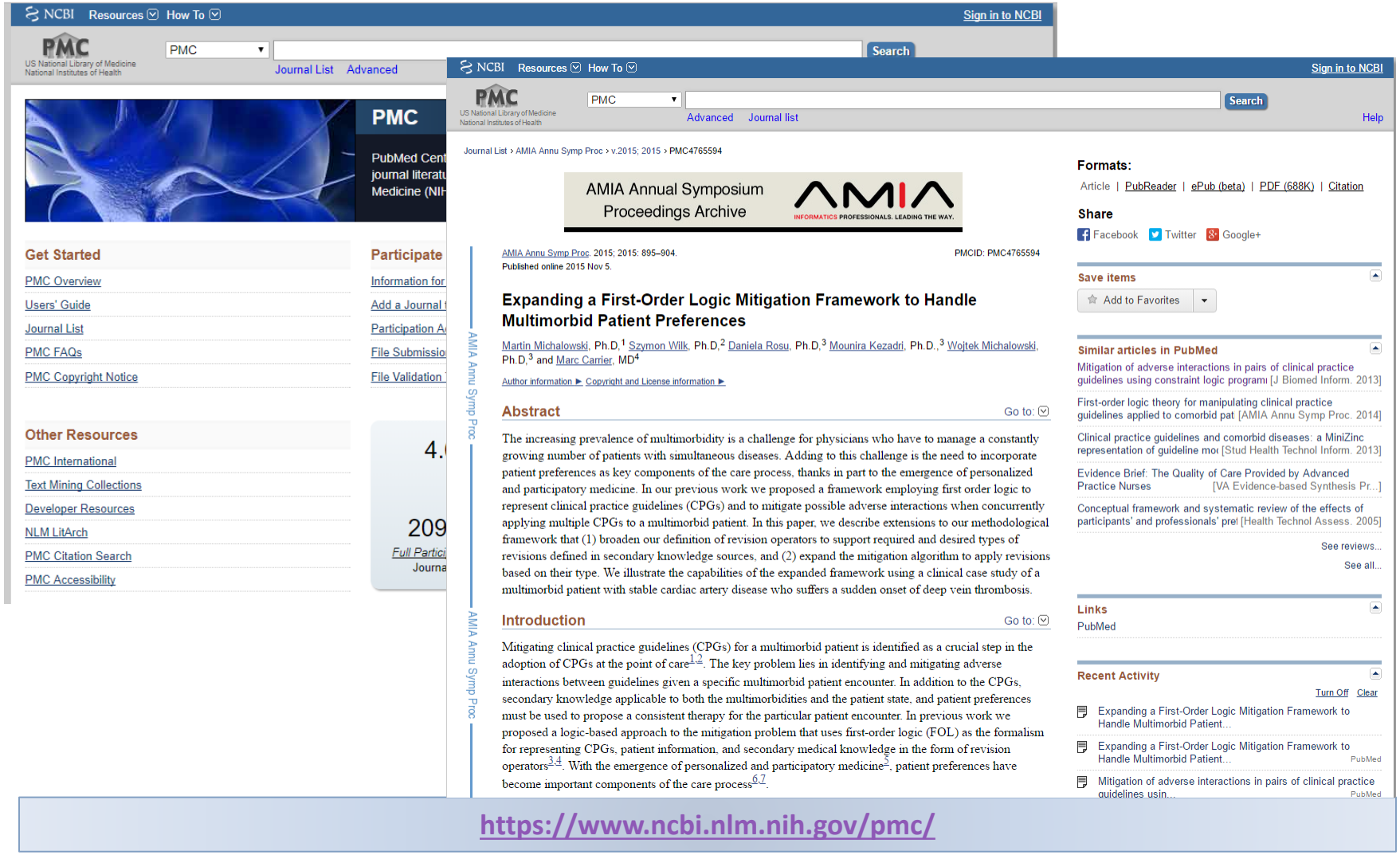
NGC





https://www.guideline.gov/

PubMed Central



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

Visible Human Project

Home

The Visible Human Project®

Overview

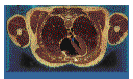
The Visible Human Project® is an outgrowth of the NLM's 1986 Long-Range Plan. It is the creation of complete, anatomically detailed, three-dimensional representations of the normal male and female human bodies. Acquisition of transverse CT, MR and cryosection images of representative male and female 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

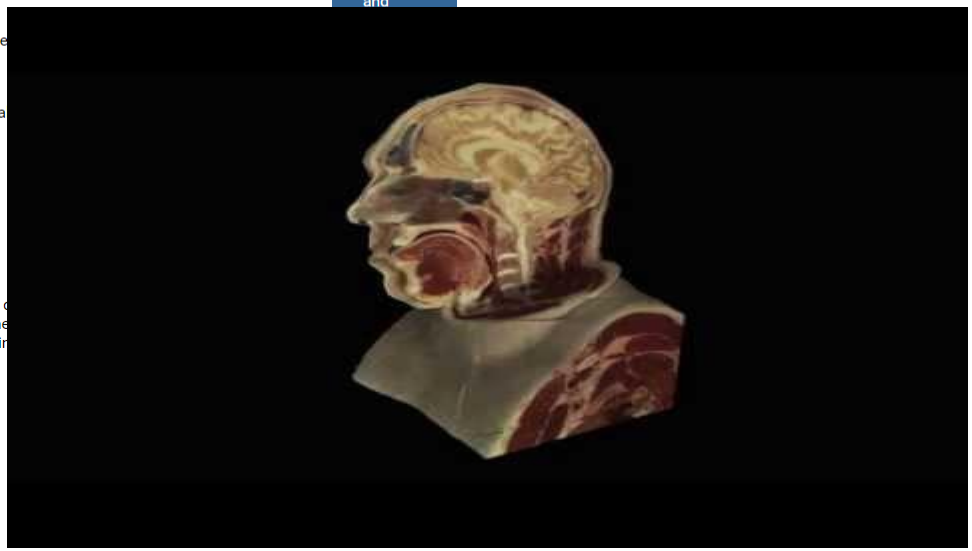
- **General Information**
  - A description of The Visible Human Project® image data and how to obtain it (include
  - The Visible Human Project® FactSheet.
  - The Visible Human Project® From Wikipedia, the free encyclopedia
  - The Visible Human Project®: From Data to Knowledge: An update of ongoing 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

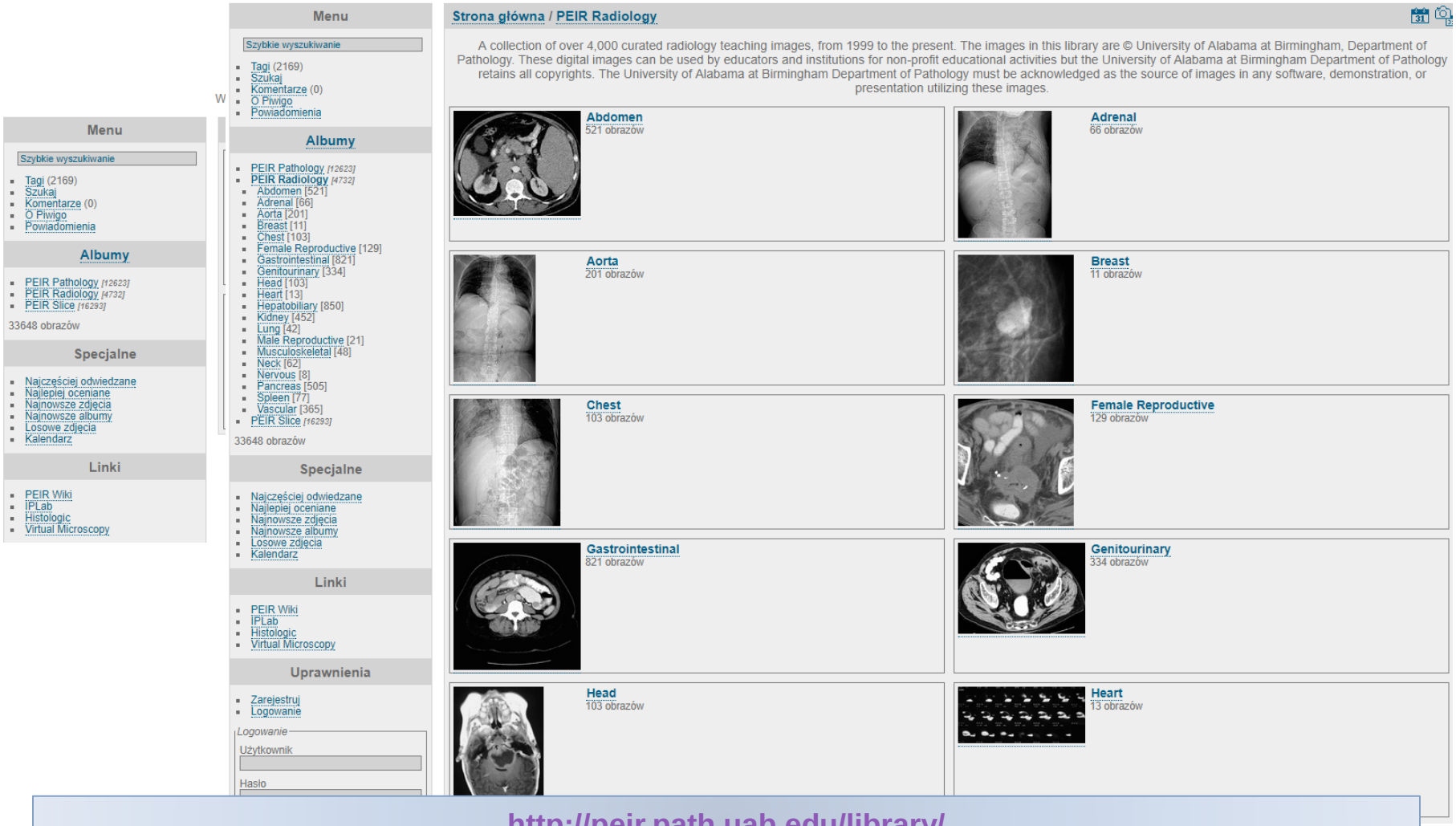
Applications for viewing images

Sources of images and



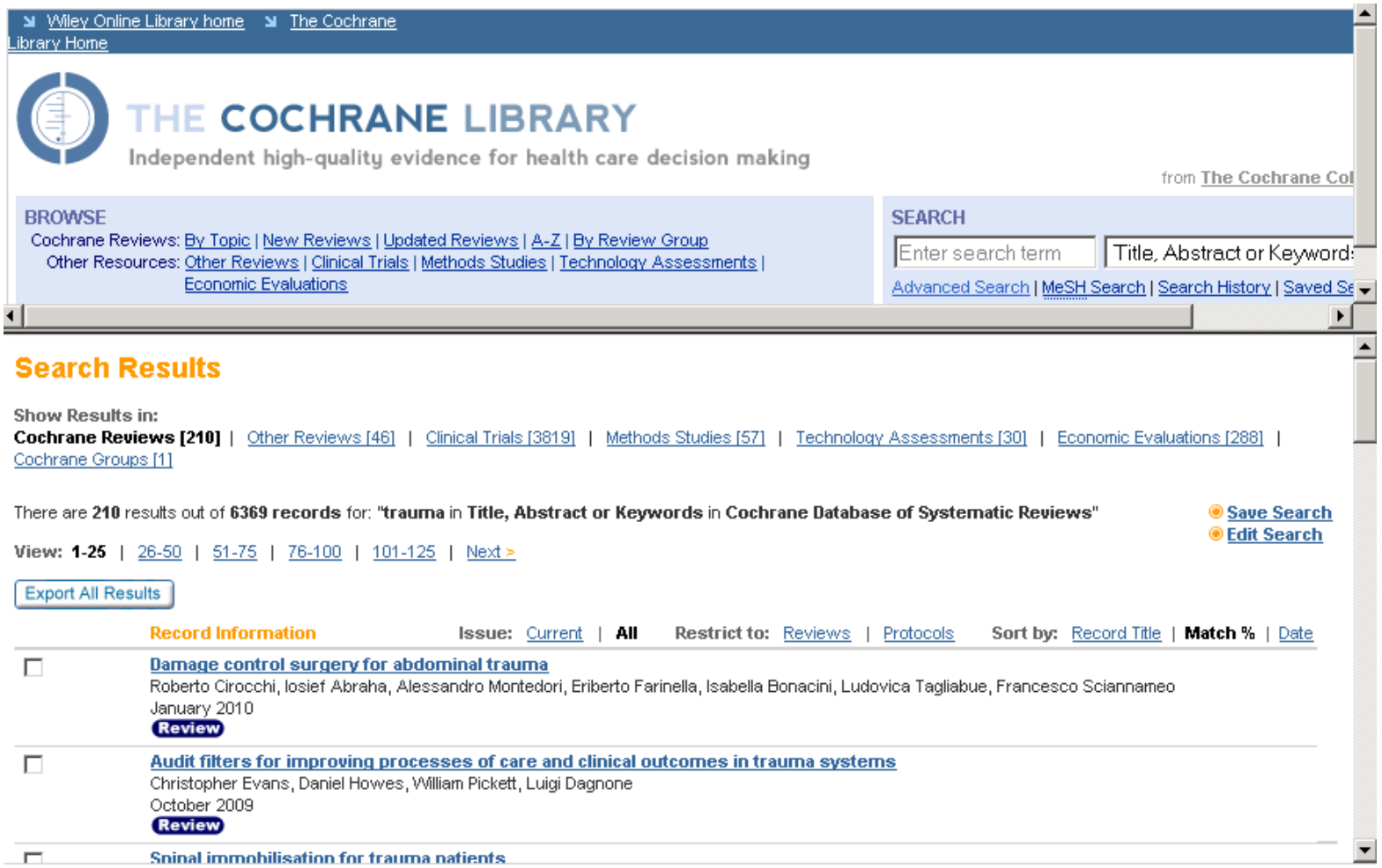
https://www.nlm.nih.gov/research/visible/visible_human.html

PEIR (Pathology Information Educational Resource) Digital Library



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

The Cochrane Library



http://www.cochranelibrary.com

Przegląd systematyczny

- Podsumowanie wyniki z wielu publikacji podstawowych na wybrany i dobrze zdefiniowany temat
- Opracowywane i weryfikowane przez zespół recenzentów
- Stała struktura opracowania
  - cześć tekstowa z prezentacją problemu i podsumowaniem wyników
  - analiza dostępnych danych

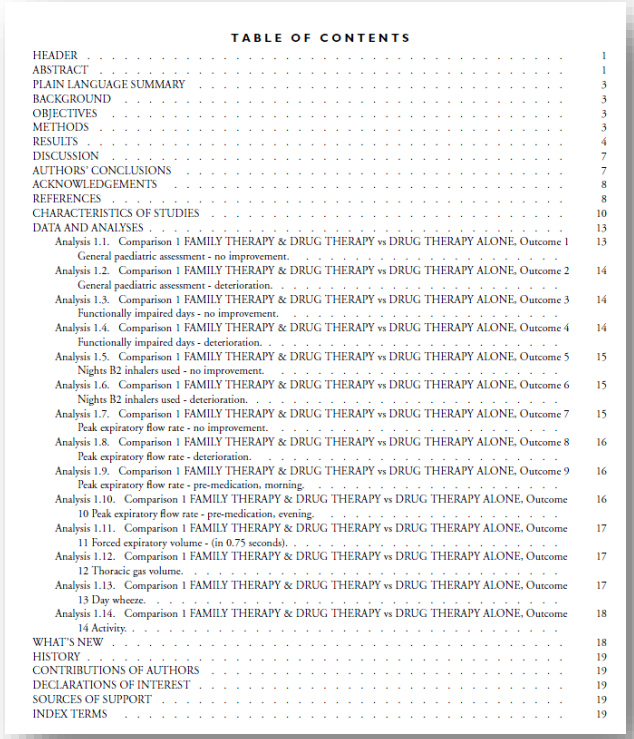


TABLE OF CONTENTS

Przegląd systematyczny

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

DATA AND ANALYSES

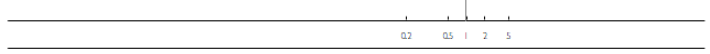
Comparison 1. FAMILY THERAPY & DRUG THERAPY vs DRUG THERAPY ALONE

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

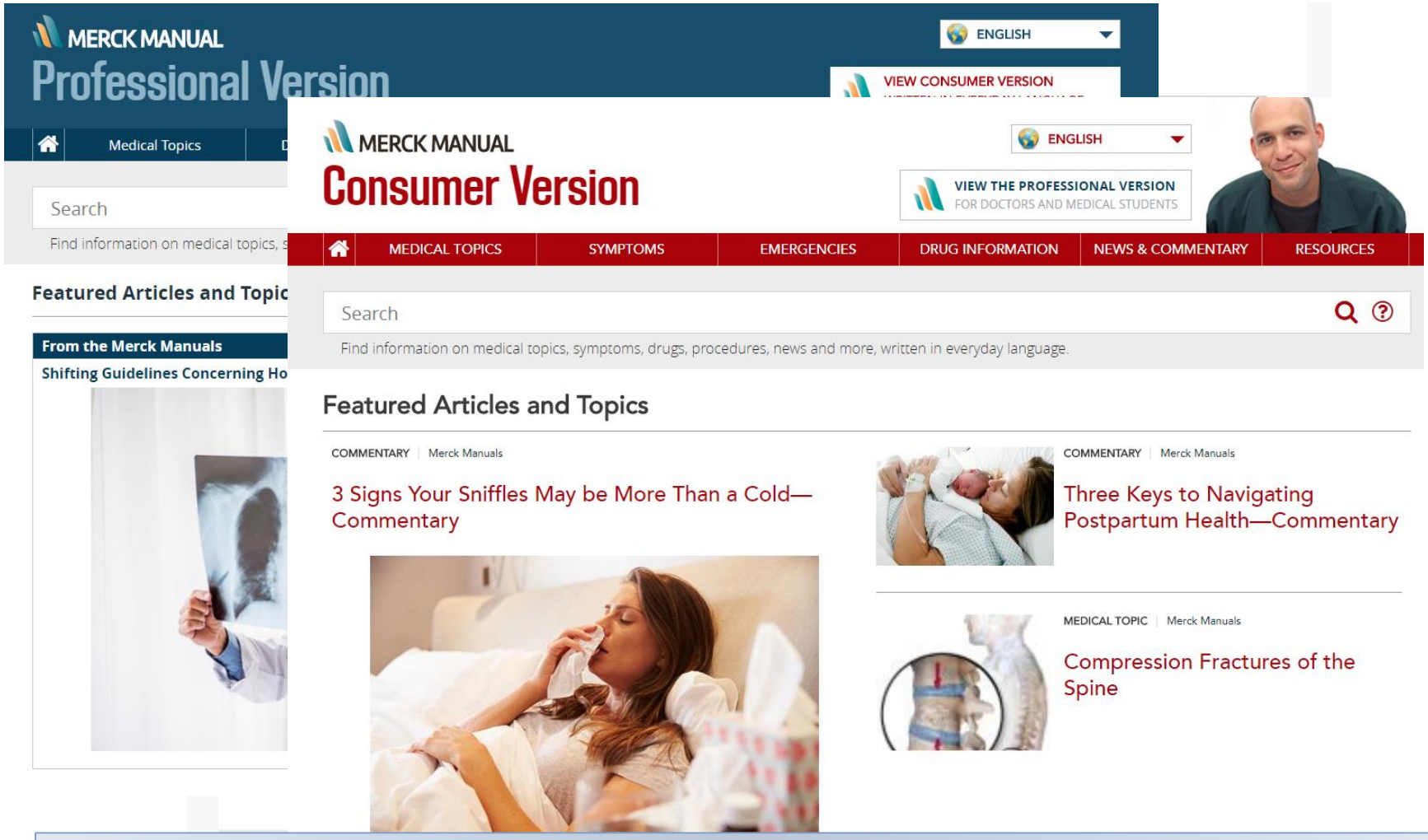
Analysis 1.1. Comparison 1 FAMILY THERAPY & DRUG THERAPY vs DRUG THERAPY ALONE, Outcome 1 General paediatric assessment - no improvement.

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]

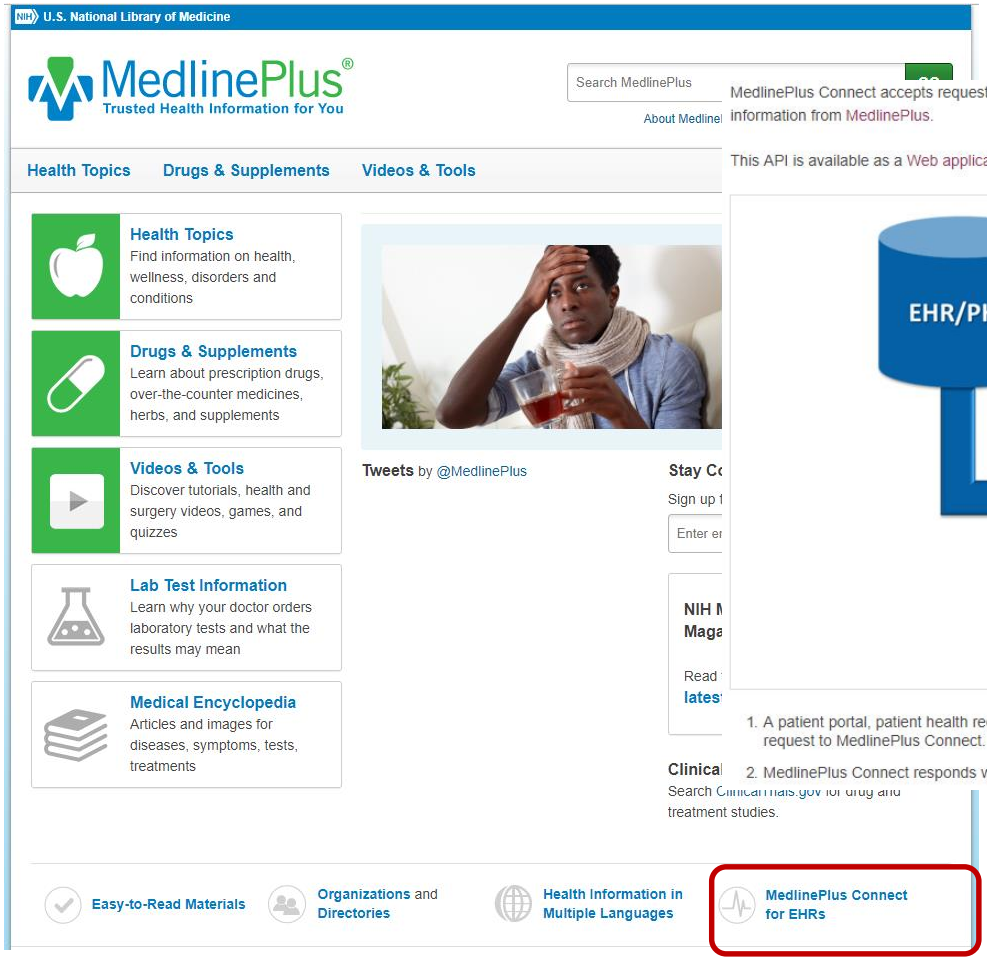


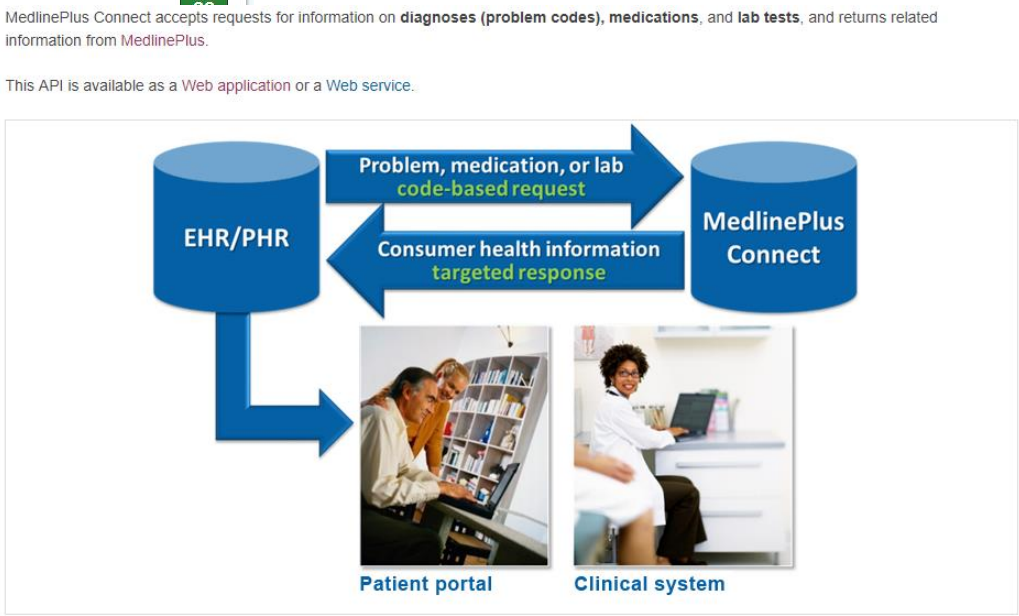
Merck Manual



http://www.merckmanuals.com/

MedlinePlus





Akceptowane kody
- diagnozy – ICD-9/10-CM, SNOMED-CT
- leki – RxNorm
- badania – LOINC

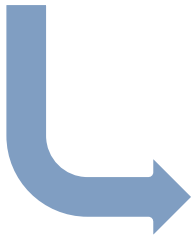
Także zapytanie tekstowe (po angielsku)
Wyniki w formacie JSON

https://medlineplus.gov/

MedLinePlus

Kod ICD-9-CM = 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"
        }
    },
```

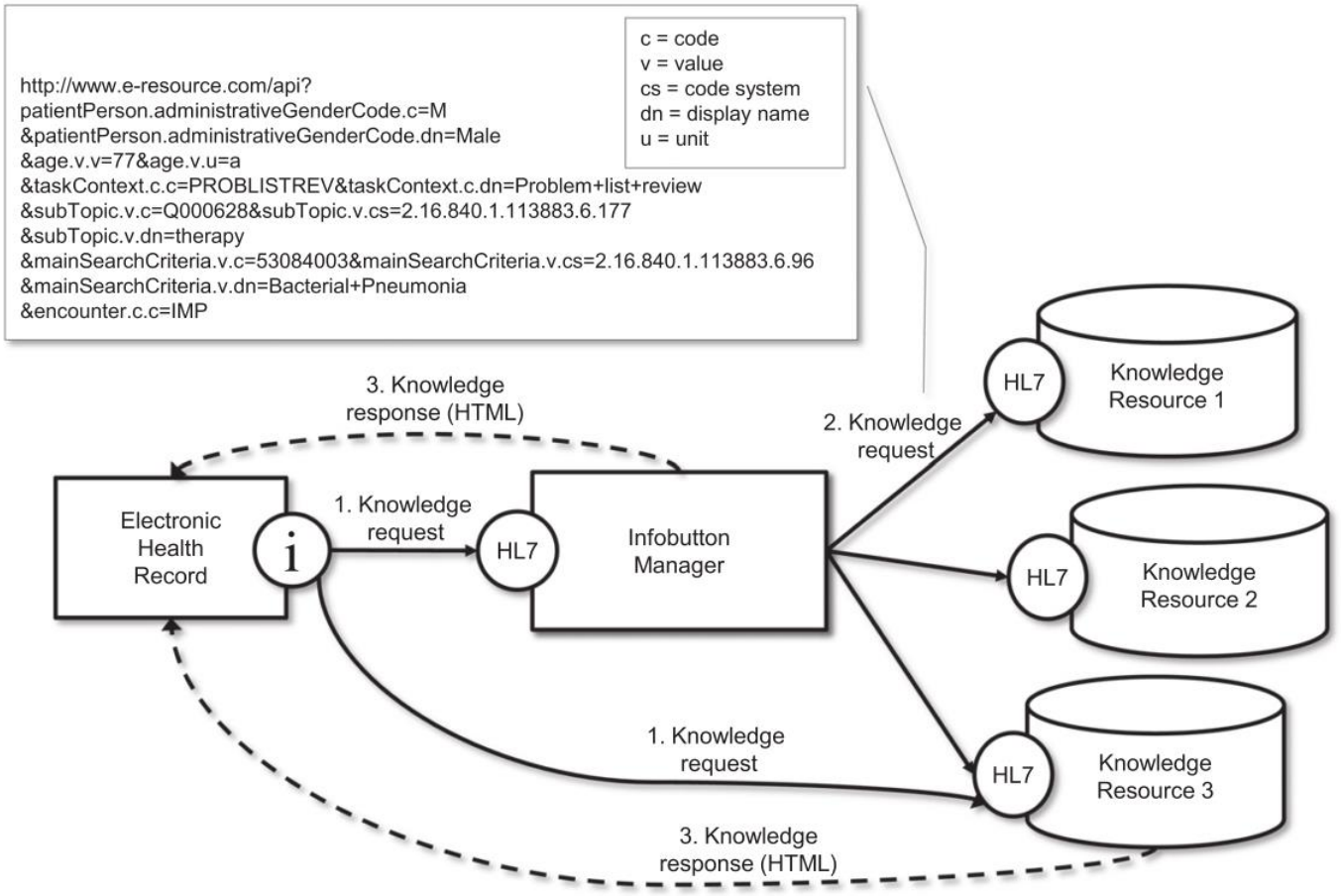
INFOBUTTONS

Infobuttons

- Odnośniki z systemów typu EPR do potencjalnie interesujących informacji w zewnętrznych repozytoriach
- *Infobutton Manager* – dobór odpowiedniego źródła wiedzy w zależności od kontekstu (np. dane pacjenta, temat zapytania, odbiorca wyniku)
- Szczegółowa specyfikacja w ramach standardu HL7 (*context-aware information retrieval*)
  - wykorzystanie modelu HL7 RIM (→ HL7 V3)
  - zapytanie i jego kontekst jako RMIM
  - wywołanie poprzez URL, SOAP albo RESTful
- Dostępne otwarte implementacje infrastruktury

http://www.infobuttons.org/ **oraz** http://www.openinfobutton.org/

Infobutton Manager – schemat działania

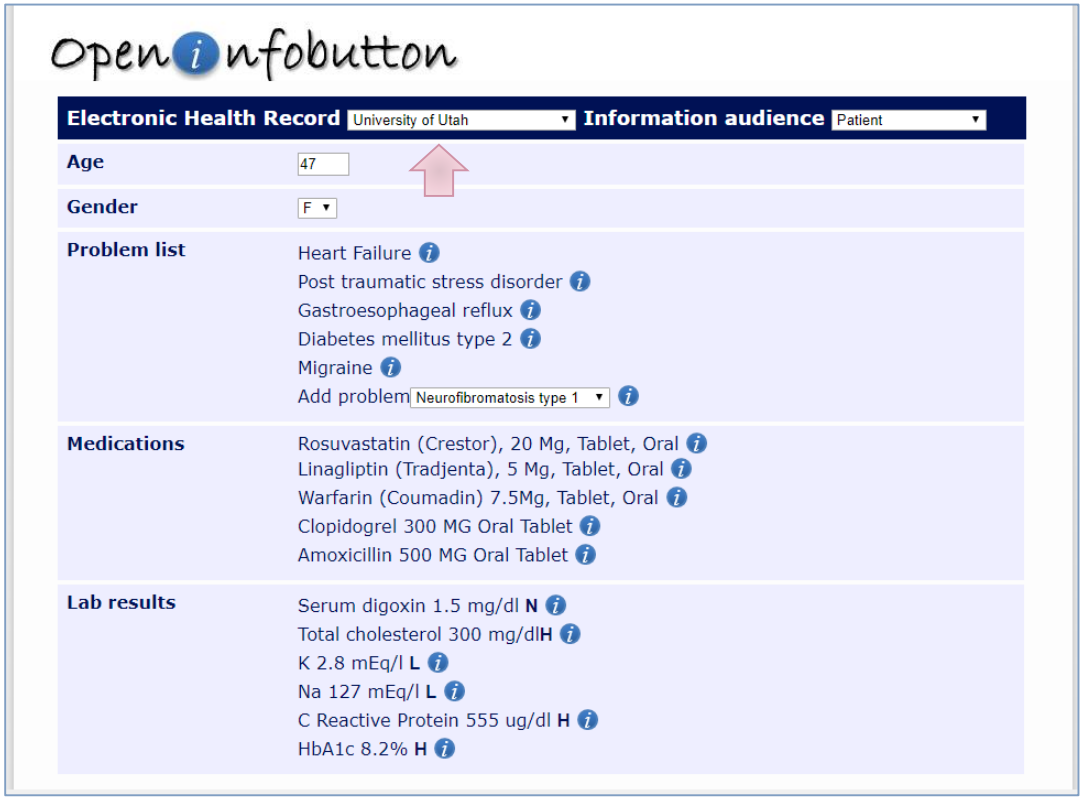


Del Fiol G, Huser V, Strasberg HR, Maviglia SM, Curtis C, Cimino JJ. Implementations of the HL7 Context-Aware Knowledge Retrieval ("Infobutton") Standard: challenges, strengths, limitations, and uptake. *J Biomed Inform* 2012;45(4):726–35.

Infobutton - kontekst

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

Infobuttons – demo



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

INDEKSOWANIE I PRZESZUKIWANIE REPOZYTORIÓW

Indeksowanie zasobów

- Indeksowanie ręczne
  - wykonywane przez ludzi – ekspertów z danej dziedziny
  - zazwyczaj przy użyciu kontrolowanej terminologii i zgodnie z pewnym przyjętym protokołem (np. do rozwiązywania konfliktów)
  - stosowane w repozytoriach bibliograficznych i adnotowanych
- Indeksowanie automatyczne
  - z wykorzystaniem terminów z rozważanych zasobów i/lub kontrolowanej terminologii
  - stosowane w repozytoriach pełnotekstowych
- Podejścia hybrydowe → indeksy automatyczne „czyszczone” przez ekspertów (np. w MEDLINE)

Kontrolowane terminologie

Terminologia kontrolowana (ang. *controlled terminology*) to zbiór terminów (dla każdego wersja kanoniczna i alternatywy) oraz relacji między nimi (hierarchia, synonim, związek), który może zostać wykorzystany do indeksowania zasobów (a także do kodowania informacji).

- MeSH (Medical Subject Heading)
  - terminologia opracowana przez NLM (MEDLINE), I wydanie w 1960 r.
  - corocznie uaktualniany, poprawiany i rozszerzany: 5 700 terminów w pierwszym wydaniu, > 26 tys. w najnowszym)
  - typy terminów
    - deskryptor (*Main Heading*) → pojęcie biomedyczne
    - kwalifikator (*Subheading*) → doprecyzowanie znaczenia deskryptora
    - rekord uzupełniający (*Supplementary Concept Record*) → lek lub inna substancja chemiczna
- EMTREE (dla EMBASE)

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

MeSH – deskryptory i kwalifikatory

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

Organizacja deskryptorów

- Deskryptory podzielone na 16 kategorii
- Kategorie zorganizowane jako wielopoziomowe drzewa
  - od najbardziej ogólnych do najbardziej szczegółowych
  - do 13 poziomów w kategorii
- Deskryptor może wystąpić w więcej niż jednym drzewie

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]

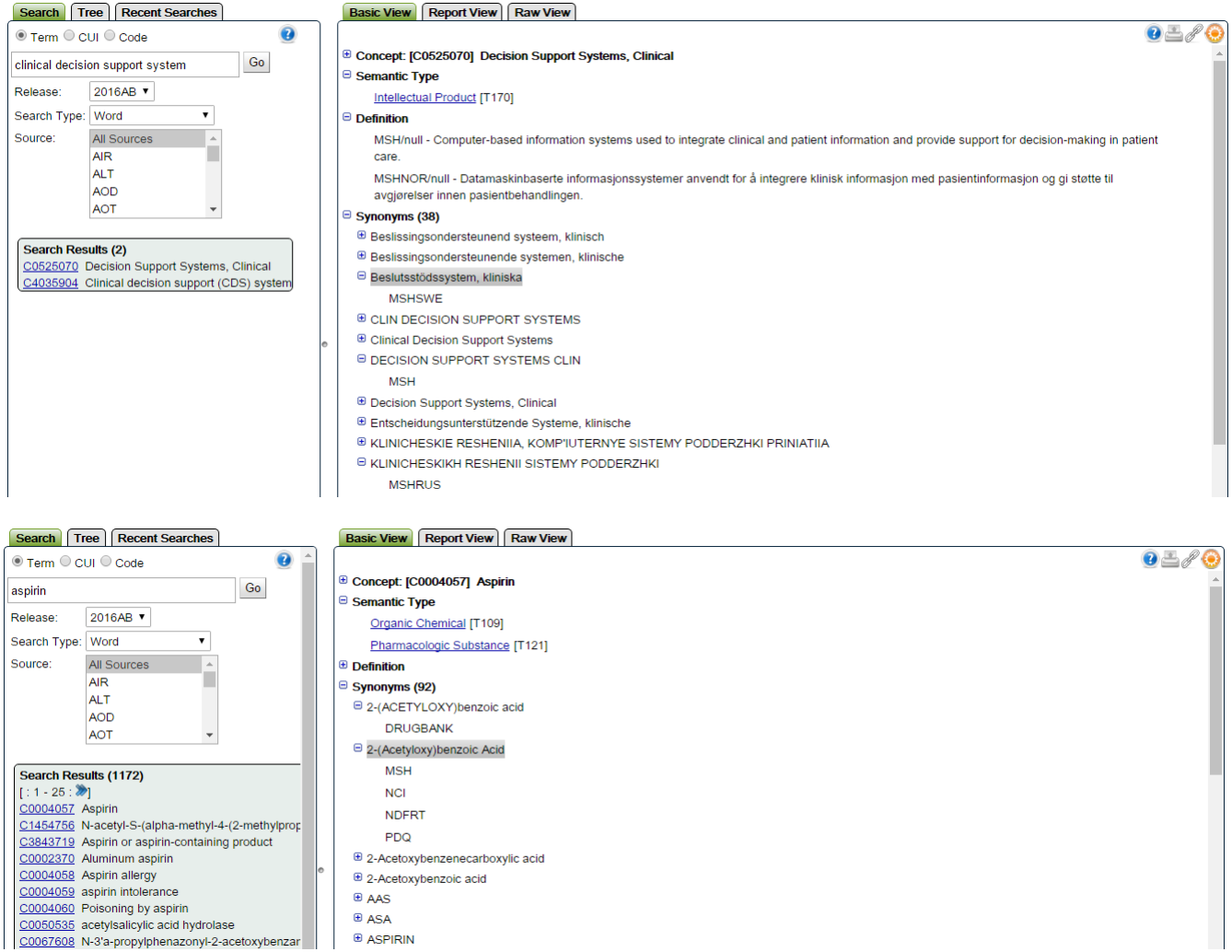
Respiratory Tract Diseases [C08]
Respiratory Tract Infections [C08.730]
Bovine Respiratory Disease Complex [C08.730.085] +
Bronchitis [C08.730.099] +
Common Cold [C08.730.162]
Empyema, Pleural [C08.730.265] +
▶ Influenza, Human [C08.730.310]
Laryngitis [C08.730.368] +
Legionellosis [C08.730.382] +
Lung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

UMLS (Unified Medical Language System)

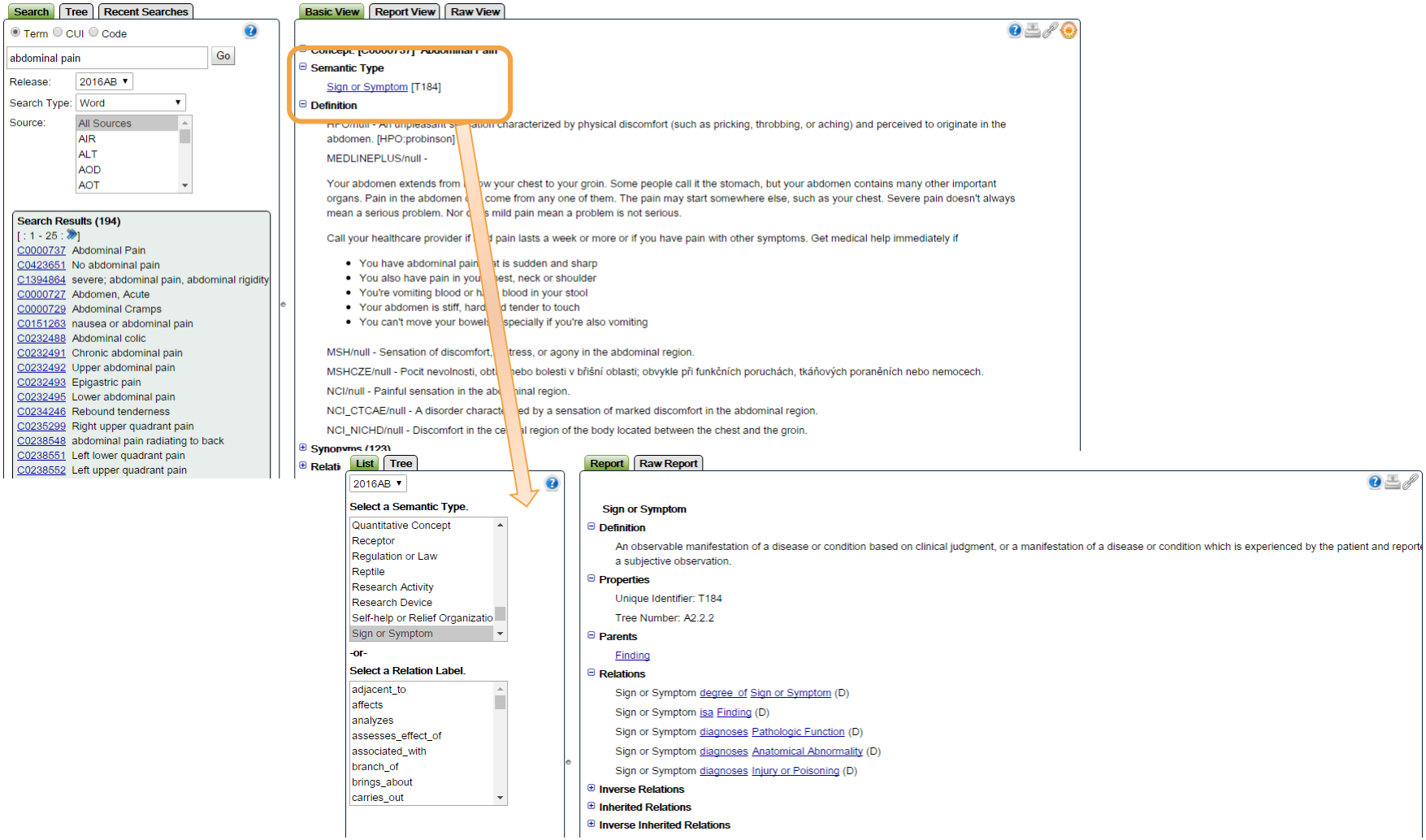
- Zestaw zasobów i narzędzi wspomagających komputerowe przetwarzanie tekstów biomedycznych rozwijany przez NLM
- Metathesaurus
  - meta-słownik łącząca wiele istniejących słowników oraz terminologii
  - hierarchia i powiązania między pojęciami pochodzącymi z różnych źródeł i w różnych językach
- Semantic Network
  - hierarchia i powiązania między typami semantycznymi związanymi z poszczególnymi pojęciami z Metathesaurus-a
- API pozwalające na wykorzystanie funkcjonalności UMLS-a we własnych aplikacjach

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

Metathesaurus



Semantic Network



MetaMap

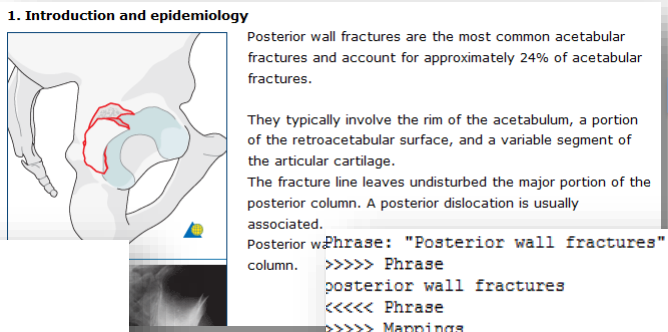
- System do analizy tekstów biomedycznych w języku angielskim rozwijany przez NLM
- Wykorzystuje techniki przetwarzania języka naturalnego do wykrywania w tekstach terminów z UMLS Metathesaurus
- Podstawowy element systemu Medical Text Indexer (MTI) używanego przez NLM do automatycznego indeksowania publikacji medycznych
- Możliwość wykorzystania we własnych zastosowaniach (udostępniony darmowo jako *open source*)

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

MetaMap

Tylko MeSH

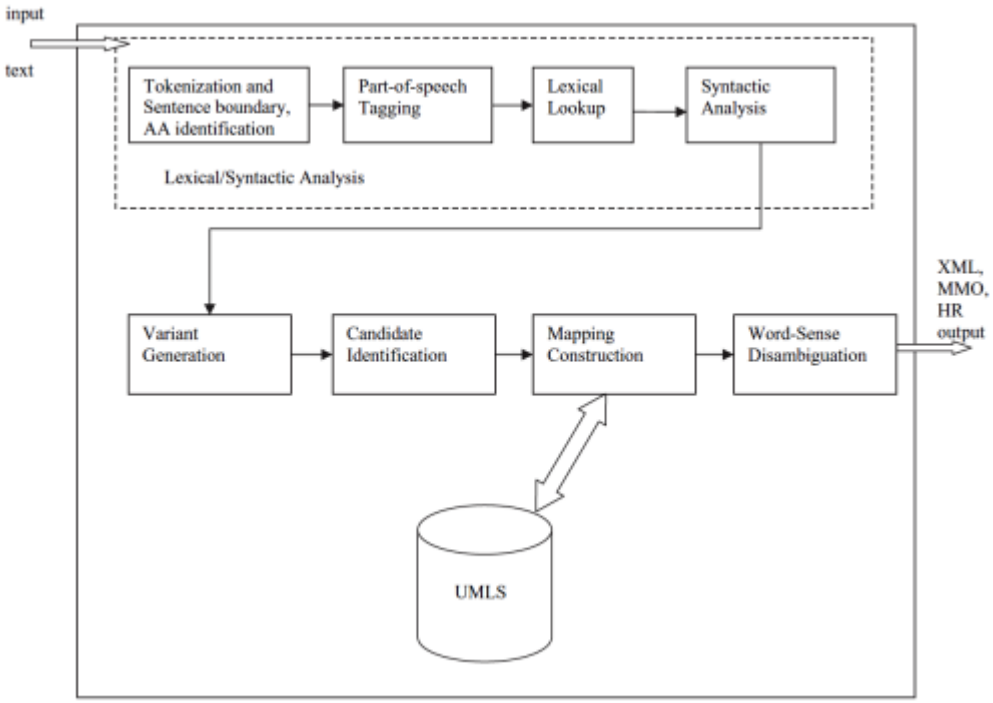
Wszystkie dostępne słowniki




```
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
```

Przetwarzanie tekstu w MetaMap



1. Podział na tokeny, wykrywanie skrótów
2. Oznaczanie części mowy
3. Uzyskanie informacji leksykalnej z leksykonu SPECIALIST (UMLS)
4. Analiza składniowa – wykrywanie fraz
5. Generowanie wariantów fraz
6. Identyfikacja kandydatów do mapowań
7. Konstrukcja mapowań (przez łączenie kandydatów)
8. Usuwanie zanegowanych mapowań
9. Wybór mapowań najlepiej dopasowanych do otaczającego tekstu

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

Indeksowanie automatyczne

- Wstępne przetwarzanie dokumentu
  - podział tekstu na słowa/tokeny (tokenizacja)
  - usuwanie „nieznaczących” słów (tzn. stopwords)
  - normalizacja – ujednolicenie zapisu (np. tylko małe litery)
  - transformacja słów do formy kanonicznej (termów) – *stemming* (wspólna część) lub *lematyzacja* (forma podstawowa)

stemming: am, are, is → am, ar, is
lematyzacja: am, are, is → be

- Reprezentacja w formie **worka słów** (ang. *bag of words*), czyli zbioru termów występujących w dokumencie
  - wykorzystanie bezpośrednie do budowy indeksu
  - budowa dodatkowych metryk na podstawie tej reprezentacji (TF-IDF)

Reprezentacja TF-IDF

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

- Każdy dokument reprezentowany jako wektor w przestrzeni *m*-wymiarowej (*m* – liczba termów)
- Poszczególne elementy wektora zdefiniowane jako

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

$TF(t,d)$ = liczba wystąpień termu *t* w dokumencie *d*

$$IDF(t) = log\left(\frac{liczba\ wszystkich\ dokumentow}{liczba\ dokumentow\ z\ termem\ t}\right)$$

- Połączenie perspektywy „lokalnej” (TF) oraz „globalnej” (IDF) przy ocenie poszczególnych termów i dokumentów

Problemy z indeksowaniem automatycznym

- Synonimy – różne termy, takie samo znaczenie
- Polisemia – takie same termy, różne znaczenie
- Znalezione termy mogą nie być adekwatne do głównego tematu dokumentu („dygresje”)
- Znacznie termów zależne od sąsiedztwa (np. wysokie ciśnienie) – konieczność rozważania większego kontekstu
- Różna „granularność” indeksowanych dokumentów i zapytań – zapytanie dotyczy antybiotyków jako klasy leków, a dokumenty dotyczą konkretnych substancji

Mechanizmy rozszerzania indeksu albo zapytań w celu rozwiązania (części) powyższych problemów

Przeszukiwanie repozytoriów

- Dwa podejścia do dopasowywania dokumentów i zapytań
  - dopasowanie dokładne (*exact match*) – zapytania w formie wyrażeń logicznych (AND, OR , NOT) → wymagane pełne dopasowanie dokumentu i zapytania
  - dopasowanie częściowe (*partial match*) – wykorzystanie różnych miar dopasowania, wybór dokumentów najlepiej (także częściowo) dopasowanych do zapytania (→ *relevance ranking*)
- Dopasowanie dokładne dla indeksów stworzonych ręcznie (bazy bibliograficzne i adnotowane), dopasowanie częściowe dla indeksów automatycznych (repozytoria pełnotekstowe)

Dopasowanie dokładne preferowane przez zaawansowanych użytkowników (wrażenie lepszej kontroli), ale w praktyce brak znaczących różnic w uzyskiwanych wynikach

Wyznaczanie dopasowania częściowego

- Miara Jaccarda – dla reprezentacji *bag of words* ($\rightarrow$ dokument jako zbiór słów)

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

- Miara **cosinusowa** – dla reprezentacji TFIDF ($\rightarrow$ dokument jako wektor liczbowy)

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

W przypadku obliczania podobieństwa między dokumentem a zapytaniem, czynnik TF jest obliczany na podstawie zapytania, a IDF – na podstawie zbioru dokumentów.

Miary oceny systemów IR

- Ocena bezpośrednia (*system-oriented evaluation*)
- Precyzja (*precision*) i przywołanie (*recall*)

$$precision(q) = \frac{\left|wyszukane_q \cap istotne_q\right|}{\left|wyszukane_q\right|}$$

$$recall(q) = \frac{\left|wyszukane_q \cap istotne_q\right|}{\left|istotne_q\right|}$$

- *Precision@k* – precyzja wyznaczona na podstawie *k*-pierwszych odszukanych dokumentów
- *Average precision* (uwzględniająca kolejność zwracanych dokumentów) i *mean average precision* (dla zbioru zapytań)

Problem ze oceną bezpośrednią

- Ocena za pomocą „złotego standardu” (albo *ground truth*)
- Konieczność określenia dokumentów istotnych z punktu widzenia danego zapytania ($\rightarrow$ ogromna pracochłonność)
- Znaczące różnice w ocenach ekspertów wskazujących dokumenty istotne dla zapytania
  - Kolekcja OHSUMED – kappa = 0.41
  - Nasz eksperyment – kappa = 0.30

słaba zgodność ocen

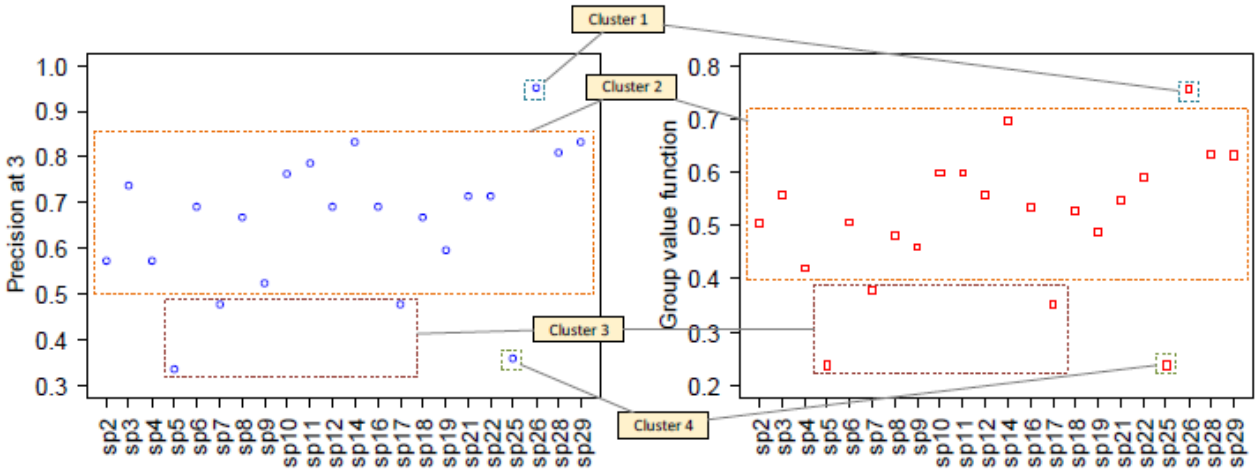
Nasz eksperyment…

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.

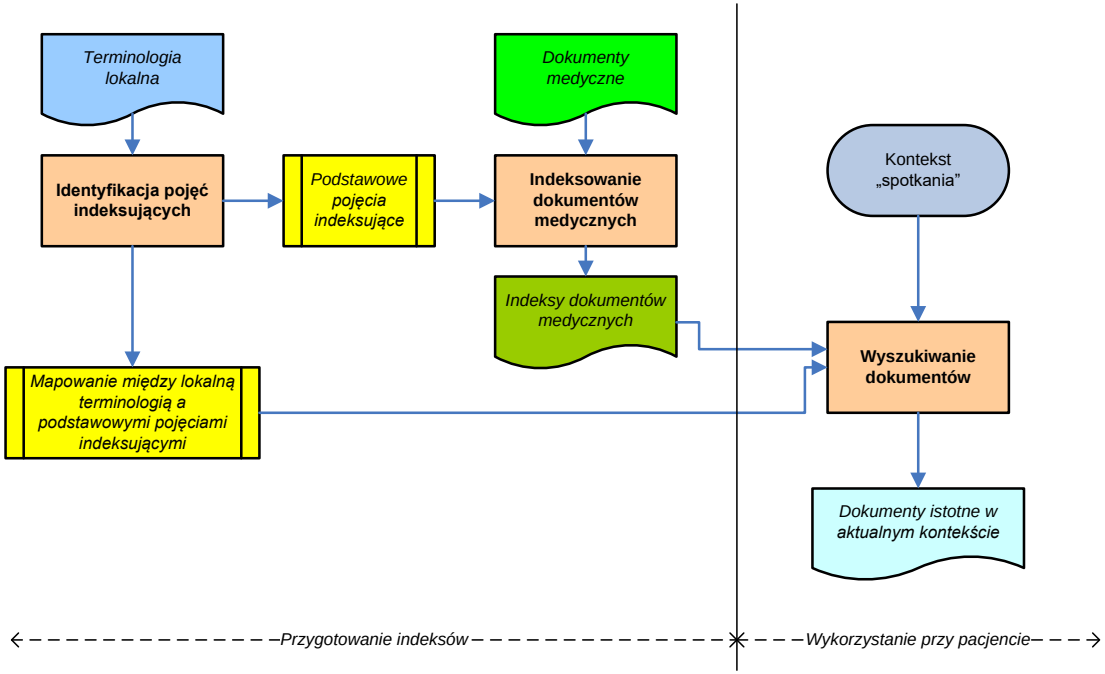
Inne podejścia do oceny

- Ocena pośrednia – wpływ użycia systemu na jakość zadań wykonywanych przez użytkowników (*user-oriented evaluation*)
- Częsty scenariusz – odpowiedzi na pytania przed i po użyciu systemu IR do wyszukania informacji
  - w eksperymentach brak różnic między różnymi podejściami do formułowania zapytań (→ dopasowanie dokładne i częściowe)
  - różny przyrost trafności w zależności od kategorii użytkowników (większy dla pielęgniarek, mniejszy dla lekarzy), ale porównywalny efekt końcowy

ROZSZERZANIE INDEKSÓW I ZAPYTAŃ

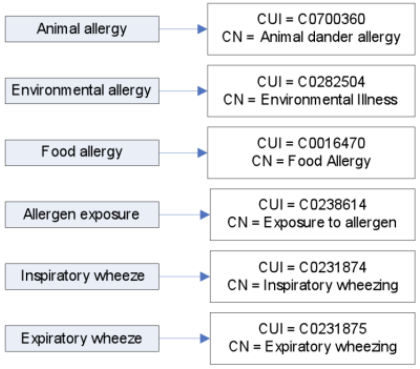
Rozszerzanie indeksów

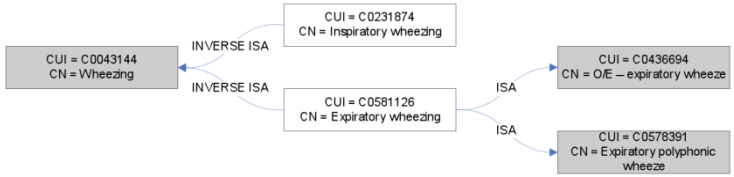
- Automatyczne wyszukiwanie przeglądów systematycznych „dopasowanych” do aktualnego pacjenta
- Uwzględnienie lokalnej terminologii w indeksowaniu i wyszukiwaniu dokumentów



Identyfikacja terminów indeksujących

- Podstawowe terminy indeksujące (Metathesaurus)
- Rozszerzone terminy indeksujące – synonimy, hypernimy i hyponimy dla terminów podstawowych (Metathesaurus)

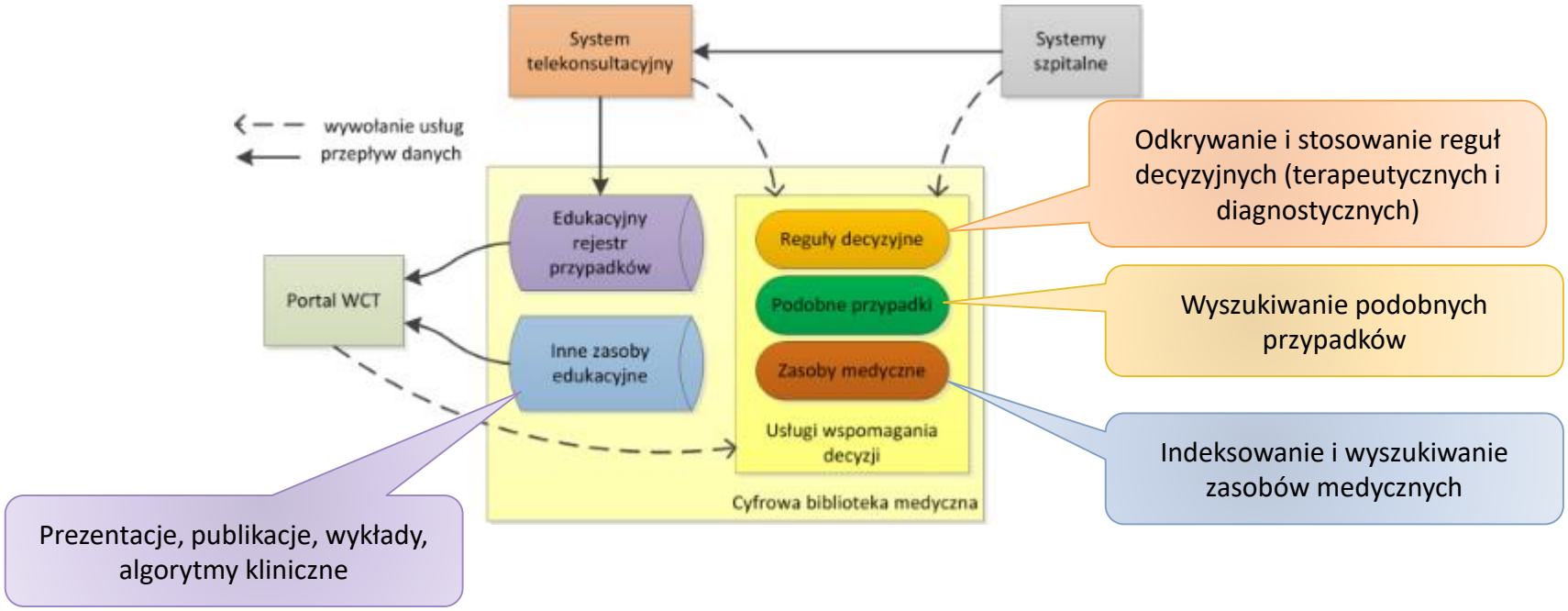




- Wyszukiwanie terminów rozszerzonych w dokumentach (MetaMap) i mapowanie do terminów podstawowych

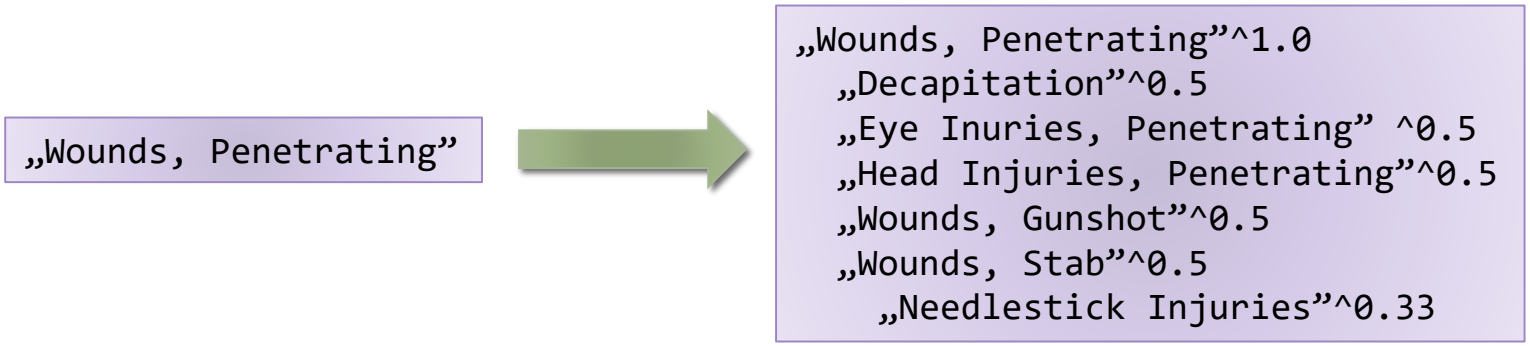
Rozszerzanie zapytań

- Wyszukiwanie zasobów w wybranych serwisach (usługa w ramach Wielkopolskiego Centrum Telemedycyny)
- Indeksowanie zasobów za pomocą terminów MeSH



Rozszerzanie zapytań

- Tłumaczenie zapytań tekstowych użytkownika (PL) na ciąg terminów MeSH – *stemming* i wybór najlepszego dopasowania
- Rozszerzenie wewnętrznego zapytania poprzez dodanie bardziej szczegółowych terminów z mniejszą wagą



- Poprawa wyników (MAP@10, przywołanie) dla rozszerzonych zapytań, ale pogorszenie precyzji (akceptowalne w praktyce)

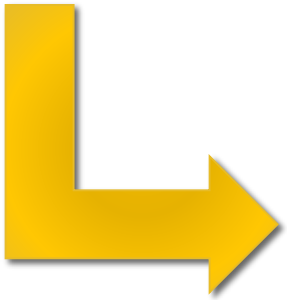
PODSUMOWANIE

Wpływ EBM i systemów IR na podejmowane decyzje

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.

Podsumowanie

- Znaczenie klinicznych systemów IR
  - potrzeby informacyjne personelu medycznego
  - medycyna oparta na dowodach (EBM)
- Różne podejścia do indeksowania i wyszukiwania (dopasowania) zasobów medycznych
- Ocena systemów IR i problemy z „tradycyjnymi” podejściami → zorientowanie na użytkowników i ich zadania
- Integracja systemów IR z EPR → *infobuttons* jako standard HL7 i jego praktyczne wykorzystanie