Semantic indexing in PubMed

Olivier Bodenreider

Lister Hill National Center for Biomedical Communications
Bethesda, Maryland - USA
Orientation

- NLM is the world's largest biomedical library
  - Located in Bethesda, Maryland, near Washington, DC
- PubMed provides access to MEDLINE, NLM’s bibliographic database of over 20M citations
  - MEDLINE covers 5600 journals and adds almost 1M new citations each year
  - PubMed is part of the Entrez system of the National Center for Biotechnology Information (NCBI)
Outline

- Anatomy of a MEDLINE citation
- Types of PubMed searches
  - Simple text search
  - Search based on MeSH indexing
- Automatic indexing
- Beyond topics
Anatomy of a MEDLINE citation

**Title**

Fluoroquinolone therapy and Achilles tendon rupture.

Vanek D, Saxena A, Boggs JM.
California School of Podiatric Medicine, San Francisco, CA, USA.

**Abstract**

Fluoroquinolones have been associated with tendinopathies. The authors present three cases of Achilles tendinopathy in which the patients' symptoms were preceded by treatment for unrelated bacterial infections with ciprofloxacin. Although the exact mechanism of the relationship is not understood, those who engage in sports or exercise should be advised of the risk of quinolone-induced tendinopathy.

PMD: 12869605 [PubMed - indexed for MEDLINE]

**Indexing**

Publication Types, MeSH Terms, Substances

LinkOut - more resources
<table>
<thead>
<tr>
<th>Publication Types, MeSH Terms, Substances</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Publication Types</strong></td>
</tr>
<tr>
<td>Case Reports</td>
</tr>
<tr>
<td>Review</td>
</tr>
<tr>
<td><strong>MeSH Terms</strong></td>
</tr>
<tr>
<td>Achilles Tendon/drug effects*</td>
</tr>
<tr>
<td>Adult</td>
</tr>
<tr>
<td>Aged</td>
</tr>
<tr>
<td>Anti-Infective Agents/adverse effects*</td>
</tr>
<tr>
<td>Anti-Infective Agents/therapeutic use</td>
</tr>
<tr>
<td>Ciprofloxacin/adverse effects*</td>
</tr>
<tr>
<td>Ciprofloxacin/therapeutic use</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Humans</td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Middle Aged</td>
</tr>
<tr>
<td>Muscular Diseases/chemically induced</td>
</tr>
<tr>
<td>Risk Factors</td>
</tr>
<tr>
<td>Rupture</td>
</tr>
<tr>
<td>Tendon Injuries/chemically induced*</td>
</tr>
<tr>
<td><strong>Substances</strong></td>
</tr>
<tr>
<td>Anti-Infective Agents</td>
</tr>
<tr>
<td>Ciprofloxacin</td>
</tr>
</tbody>
</table>
Types of PubMed searches

Non-semantic search

- PubMed does not require the use of MeSH for querying
  - Supports "Google-like" text searches
    - "no librarian required"
  - But can identify MeSH terms even if they are not labeled as such
Non-semantic search Example

- Find articles about the cheese Gruyère
  - Gruyère

PubMed.gov
US National Library of Medicine
National Institutes of Health

Show additional filters
Article types
Review
More ...

Display Settings: Summary, 20 per page, Sorted by Recently Added
Send to: Filters: Manage Filters

Titles with your search terms
[Symptoms and respiratory function in a group of Gruyère] [Rev Mal Respir. 1990]


Occurrence of the angiotensin-converting enzyme inhibiting tripeptides Val-Pro-Pro and Ile-Pro-Pro in different cheese varieties of Swiss origin.

Agroscope Liebefeld-Posieux Research Station ALP, Schwarzenburgstrasse 161, 3003 Bern, Switzerland.

Abstract
The contents of the 2 antihypertensive peptides Val-Pro-Pro (VPP) and Ile-Pro-Pro (IPP) were determined in 101 samples from 10 different Swiss cheese varieties using HPLC with subsequent triple mass spectrometry. In the category of extra hard and hard cheeses, the Protected Denomination of Origin cheeses Berner Alpkäse and Berner Hobelkäse, L'Etivaz à rebibes, Le Gruyère, Sbrinz, Emmentaler (organic and conventional) and in the category of semihard cheeses, the varieties Tilziter, Appenzeller 1/4 fat and full fat, Tête de Moine, and Vacherin fribourgeois were screened in the study. The average concentration of the sum of VPP and IPP in the screened cheese varieties varied to a large extent, and substantial variations
MeSH (semantic) search

- Medical Subject Headings (MeSH)
  - Controlled vocabulary developed at NLM for indexing and retrieval of MEDLINE citations
  - ~26,000 descriptors (main headings)
  - <100 qualifiers (subheadings)
  - 214,000 supplementary concept records

- Hierarchical structure ("tree numbers")
  - Supports query expansion ("explosion")
    - Search for a descriptor or any of its descendants

Simple MeSH search Example

◆ Find articles about drug-induced psychoses
  ● "Psychoses, Substance-Induced"[Mesh]
Search with “Explosion”

- By default, PubMed retrieves articles indexed with a descriptor or any of its descendants
- Use **mesh:noexp** to prevent “explosion” from happening

Quinolones [D03.438.810.835]

4-Quinolones [D03.438.810.835.055] + Carteolol [D03.438.810.835.188]

Fluoroquinolones [D03.438.810.835.322]

Ciprofloxacin [D03.438.810.835.322.186] + Enoxacin [D03.438.810.835.322.280]

Norfloxacin [D03.438.810.835.322.374]

Ofloxacin [D03.438.810.835.322.500]

Pefloxacin [D03.438.810.835.322.750]

PQQ Cofactor [D03.438.810.835.661]
“Explosion” Example

- Find articles about fluoroquinolones (or desc.)
  - "fluoroquinolones"[Mesh]
Search leveraging synonymy in MeSH

- MeSH descriptors include related concepts (Entry terms)
  - Synonyms
  - Closely related (and clustered or indexing and retrieval purposes)

- All terms from a descriptor and its entry terms are used for retrieval in PubMed
Autoimmune Addison’s disease.

Mitotane therapy in adrenocortical cancer induces CYP3A4 and inhibits 5α-reductase, explaining the need for personalized glucocorticoid and androgen replacement.

Idiopathic intracranial hypertension: a unifying neuroendocrine hypothesis through the adrenal-brain axis.
## Entry terms for “Addison Disease”

<table>
<thead>
<tr>
<th>MeSH Heading</th>
<th>Entry Term</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Addison Disease</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tree Number</th>
<th>Scope Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>C19.053.500.263</td>
<td>An adrenal disease characterized by the progressive destruction of the ADRENAL CORTEX, resulting in insufficient production of ALDOSTERONE and HYDROCORTISONE. Clinical symptoms include ANOREXIA; NAUSEA; WEIGHT LOSS; MUSCLE WEAKNESS; and HYPERPIGMENTATION of the SKIN due to increase in circulating levels of ACTH precursor hormone which stimulates MELANOCYTES.</td>
</tr>
<tr>
<td>C20.111.163</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Entry Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addison’s Disease</td>
</tr>
<tr>
<td>Primary Adrenal Insufficiency</td>
</tr>
<tr>
<td><strong>Primary Adrenocortical Insufficiency</strong></td>
</tr>
<tr>
<td>Primary Hypoadrenalism</td>
</tr>
</tbody>
</table>

Lister Hill National Center for Biomedical Communications
Search leveraging UMLS Synonymy

- **Unified Medical Language System (UMLS)**
  - Terminology integration system
  - ~130 biomedical terminologies
  - Synonymous terms clustered into concepts

- **UMLS synonymy used in PubMed**
  - Query translation happens “behind the scenes”
  - E.g., search on “primary adrenocortical insufficiency”
    - Retrieves articles about “Addison’s disease”
Results: 21 to 40 of 192998

   Li T, Wei X, Watkins AC, Sanchez PG, Wu ZJ, Griffith BP.
   PMID: 23773712 [PubMed - as supplied by publisher]
   Related citations

22. Drug-Eluting Stents Versus Bare-Metal Stents in Patients With Decreased GFR: A Meta-analysis.
   PMID: 23773369 [PubMed - as supplied by publisher]
   Related citations

23. Targeting inflammatory pathways in myocardial infarction.
   Christia P, Frangogiannis NG.
No entry term for “Heart attack”

<table>
<thead>
<tr>
<th>MeSH Heading</th>
<th>Myocardial Infarction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree Number</td>
<td>C14.280.647.500</td>
</tr>
<tr>
<td>Tree Number</td>
<td>C14.907.585.500</td>
</tr>
<tr>
<td>Annotation</td>
<td>do not coordinate with ACUTE DISEASE for &quot;acute infarct&quot;</td>
</tr>
<tr>
<td>Scope Note</td>
<td>NECROSIS of the MYOCARDIUM caused by an obstruction of the</td>
</tr>
<tr>
<td>Entry Term</td>
<td>Myocardial Infarct</td>
</tr>
<tr>
<td>See Also</td>
<td>Heart Rupture, Post-Infarction</td>
</tr>
</tbody>
</table>

Search details:

"myocardial infarction"[MeSH Terms]
OR ("myocardial"[All Fields] AND "infarction"[All Fields])
OR "myocardial"
Query Translation:

"myocardial infarction"[MeSH Terms] OR ("myocardial"[All Fields] AND "infarction"[All Fields]) OR "myocardial infarction"[All Fields] OR ("heart"[All Fields] AND "attack"[All Fields]) OR "heart attack"[All Fields]
Subheading restrictions

- Subheadings represent the context of use of a particular descriptor
  - Ciprofloxacin/Adverse effects
  - Mood Disorders/Chemically induced
- Assigned during indexing
- Can be queried in PubMed
Subheading restrictions  Example

� Find articles about drugs involved in adverse events
  ● "Chemicals and Drugs Category"/adverse effects[MeSH]
Recapitulative example

Find articles about drugs involved in adverse events and drug-induced manifestations

(("Chemicals and Drugs Category"[Mesh]) AND (adverse effects[sh] OR contraindications[sh] OR mortality[sh]))) AND (chemically induced[sh] OR ("Drug-Induced Liver Injury"[Mesh:noexp]) OR ("Drug Eruptions"[Mesh:noexp]) OR ("Epidermal Necrolysis, Toxic"[Mesh]) OR ("Drug-Induced Liver Injury, Chronic"[Mesh]) OR ("Erythema Nodosum"[Mesh]) OR ("Serotonin Syndrome"[Mesh]) OR ("Hand-Foot Syndrome"[Mesh]) OR ("Neuroleptic Malignant Syndrome"[Mesh]) OR ("MPTP Poisoning"[Mesh]) OR ("Dyskinesia, Drug-Induced"[Mesh]) OR ("Neurotoxicity Syndromes"[Mesh:noexp]) OR ("Psychoses, Substance-Induced"[Mesh]) OR ("Akathisia, Drug-Induced"[Mesh]))) AND (medline[sb])
Automatic indexing
Automatic indexing Motivation

- Indexing by humans is costly and has limited reproducibility
- Natural language processing can effectively support named entity recognition
- Automatic indexing can produce
  - Suggestions for human indexers
  - Final indexing for some journals
Automatic indexing Principles

◆ Hybrid approach
  ● Concepts extracted from title and abstract
    ▪ Mapped from UMLS to MeSH
  ● MeSH descriptors extracted from related citations

◆ Post-processing
  ● Clustering and ranking
  ● Integrate indexing rules
    ▪ E.g., “rule of 3”
      – Index with a higher-level descriptor rather than with 3 or more lower-level descriptors
Automatic indexing workflow:

- Medical Text Indexer

Automatic indexing Applications

◆ MEDLINE indexing
  ● Support MEDLINE indexing at NLM
    ■ 3600 new citations processed every weeknight
    ■ Suggestions displayed in the indexing environment
  ● “First-line” indexing
    ■ For 75 journals
    ■ MTI recommendations are used as an indexer
    ■ Simply reviewed by a senior indexer

◆ Cataloging and History of Medicine
  ● Assisted indexing
Beyond topics
Beyond concepts... relations

◆ Also known as
  ● Facts
  ● Predications
  ● Nano-publications
  ● ...

◆ Relation extraction
  ● Usually based on natural language processing (NLP)
    ■ E.g., SemRep
  ● Relations stored in \((subject, predicate, object)\) form
    ■ With provenance information
Experimental application Semantic MEDLINE

- Multi-document summarization
- Based on a database of 60M predications extracted from MEDLINE
- Entities normalized to the UMLS Metathesaurus
- Relations aligned with the UMLS Semantic Network
- Interfaced with PubMed (for retrieving PMIDs) on a given topic
  - Forms the basis for summarization

Title: Multiple binding sites revealed by interaction of relaxin family peptides with native and chimeric relaxin family peptide receptors 1 and 2 (LGR7 and LGR8).

Abstract:
Relaxin family peptide 1 (RXFP1) receptor (LGR7) and RXFP2 receptor (LGR8) were recently identified as the receptor targets for h2 relaxin and insulin-like peptide 3 (INSL3), respectively. In this study, we define the pharmacology of these two receptors by using a number of receptor chimeras and relaxin family peptides. We have identified two binding sites on these receptors: one primary, high-affinity site within the ectodomain and a secondary, lower affinity site within the transmembrane region. The primary site was found to dictate receptor binding characteristics, although the lower affinity site also exerts some influence and modulates ligand affinity for the primary site in a manner dependent upon the peptide in question. Not all relaxin peptides were able to bind to the RXFP2 receptor, indicating that the relaxin-RXFP2 receptor interaction is species-specific.

INSL3 was found to exhibit characteristics of a partial agonist at the RXFP2 and chimeric RXFP1/2 receptors, with low maximal cAMP response but high potency in coupling to this pathway. cAMP accumulation studies also revealed that the binding sites couple to cAMP signaling pathways with differing efficiency: the high-affinity site signals with high efficiency, whereas the lower affinity site signals with little to no efficiency. Comparisons between RXFP1, RXFP2, the chimeric receptors, and the truncated receptor revealed that the interaction between receptor sites is critical for optimal ligand binding and signal transduction and that the ectodomain is essential for signalling. Evidence obtained in this study supports...
Relation extraction Applications

- **Enhanced information retrieval**
  - Indexing on relations in addition to concepts or association main heading/subheading

- **Multi-document summarization**
  - Extract and visualize the facts extracted from 250 recent abstracts on the treatment of Parkinson’s disease

- **Question answering**
  - Clinical and biological questions

- **Knowledge discovery**
  - Connect facts from heterogeneous resources