NLM resources for research and practice: an overview and an R&D application

Dina Demner-Fushman
IGERT: From Data to Solutions
October 3, 2014
Disclaimer

The views and opinions expressed do not necessarily state or reflect those of the U.S. Government, and they may not be used for advertising or product endorsement purposes.
America's two greatest gifts to the world are jazz and Medline.


• MEDLINE® is the National Library of Medicine® (NLM®) journal citation database. Started in the 1960s and online since 1971, it now provides over 21 million references to biomedical and life sciences journal articles back to 1946.

• MEDLINE includes citations from over 5,600 worldwide journals in about 40 languages; about 60 languages for older journals.
  – The journals are selected based on the recommendation of the Literature Selection Technical Review Committee (LSTRC), an NIH-chartered advisory committee of external experts analogous to the committees that review NIH grant applications.
  – The LSTRC considers the quality of the scientific content of a journal, including originality and the importance of the content for the MEDLINE global audience.

• Since 2005, between 10,000-20,000 completed references are added each week.
• Available since 1996

• Over 23 million references include
  – the MEDLINE database
  – In-process citations
  – Citations to articles that are out-of-scope (e.g., covering astrophysics) from MEDLINE journals
  – "Ahead of Print" citations that precede the article's final publication in a MEDLINE indexed journal.
  – Citations that precede the date that a journal was selected for MEDLINE indexing (when supplied electronically by the publisher).
  – Citations to author manuscripts of articles published by NIH-funded researchers.
  – Citations for the books available on the NCBI Bookshelf
  – (a citation for the book and in some cases each chapter of the book).
How do Citations get into MEDLINE/PubMed?

Publishers

Medical Article Record System

Data Creation and Maintenance System

MeSH data
Serials Data
PubMed Related Citations

Data Creation and Maintenance System

MEDLINE/DCMS Database (Oracle Tables)

Indexing Life Cycle

MTI MTIFL

Medline Literature
Enhancing Biomedical Information
MeSH Suggestions

MEDLINE Indexing

MEDLINE/DCMS Database (Oracle Tables)

FTP site

U.S. National Library of Medicine

NCBI

Courtesy of George Thoma, Lou Knecht, John Rozier, Sara Tybaert, David Gillikin, Joe Thomas and James Mork
Human readable citation formats
XML Publisher DTD tags and XML for the MEDLINE DTD example

The following is a glossary of the tags defined in the PubMed DTD. Click on each of the tag names below for more information. You can also view an Example of a Standard XML File.

Data Tags (R = Required, O = Optional, O/R = Optional or Required). Tag names are case sensitive. Required tags must be included; optional tags must be included only if the data requested appears in the print or electronic article. Optional or Required tags are dependent on the use of other tags.

<table>
<thead>
<tr>
<th>File Header (R)</th>
<th>Replaces (O)</th>
<th>ArticleTitle (O)</th>
<th>Affiliation (O)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArticleSet (R)</td>
<td>ArticleTitle (O)</td>
<td>VernacularTitle (O)</td>
<td>Identifier (O)</td>
</tr>
<tr>
<td>Article (R)</td>
<td>FirstPage (O/R)</td>
<td>LocationID (O/R)</td>
<td>GroupList (O/R)</td>
</tr>
<tr>
<td>Journal (R)</td>
<td>LastPage (O)</td>
<td>Language (O)</td>
<td>Group (R)</td>
</tr>
<tr>
<td>PublisherName (R)</td>
<td>ELocationID (O/R)</td>
<td>AuthorList (O/R)</td>
<td>GroupName (R)</td>
</tr>
<tr>
<td>JournalTitle (R)</td>
<td>Language (O)</td>
<td>Author (R)</td>
<td>IndividualName (O)</td>
</tr>
<tr>
<td>Issn (R)</td>
<td>FirstName (O/R)</td>
<td>FirstPage (R)</td>
<td>PublicationType (O)</td>
</tr>
<tr>
<td>Volume (O/R)</td>
<td>LastName (O/R)</td>
<td>LastPage (O)</td>
<td>ArticleIdList (O/R)</td>
</tr>
<tr>
<td>Issue (O/R)</td>
<td>Suffix (O)</td>
<td>CollectiveName (O)</td>
<td>ArticleId (R)</td>
</tr>
<tr>
<td>PubDate (R)</td>
<td>CollectiveName (O)</td>
<td>ObjectList (O)</td>
<td>History (O)</td>
</tr>
<tr>
<td>Year (R)</td>
<td>Object (O)</td>
<td>Abstract (O)</td>
<td>OtherAbstract (O)</td>
</tr>
<tr>
<td>Month (O/R)</td>
<td>Param (O)</td>
<td>CopyrightInformation (O)</td>
<td></td>
</tr>
<tr>
<td>Season (O)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day (O)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Publisher DTD

• A free archive for full-text biomedical and life sciences journal articles launched in 2000.

• A repository for journal literature deposited by
  – participating publishers
  – author manuscripts submitted in compliance with the NIH Public Access Policy

• Reciprocal links between the full text in PMC and citations in PubMed.
  – PubMed citations are created for content not already in MEDLINE
  – Some PMC content, e.g., book reviews, is not cited in PubMed

• The PMC Open Access Subset under a Creative Commons or similar license
  – FTP service to download a complete set of files — XML, images, PDF, and supplementary data files.
    • Articles: 852,041; Images: 2,618,629; Articles containing no images: 246,066
  – An API to find articles by ID, or those that have been updated recently.
Entrez

Entrez is NCBI’s primary text search and retrieval system that integrates the PubMed database of biomedical literature with 39 other literature and molecular databases including DNA and protein sequence, structure, gene, genome, genetic variation and gene expression.
Entrez Programming Utilities (E-utilities)

• A set of nine programs that provide an application programming interface (API) to Entrez
  
  – **EInfo (database statistics)** Provides the number of records indexed in each field of a given database, the date of the last update of the database, and the available links from the database to other Entrez databases.
  
  – **ESearch (text searches)** Responds to a text query with the list of matching UIDs in a given database (for later use in ESummary, EFetch or ELink), along with the term translations of the query.
  
  – **EPost (UID uploads)** Accepts a list of UIDs from a given database, stores the set on the History Server, and responds with a query key and web environment for the uploaded dataset.
  
  – **ESummary (document summary downloads)** Responds to a list of UIDs from a given database with the corresponding document summaries.
  
  – **EFetch (data record downloads)** Responds to a list of UIDs in a given database with the corresponding data records in a specified format.
  
  – **ELink (Entrez links)** Responds to a list of UIDs in a given database with list of external and internal links.
  
  – **EGQuery (global query)** Responds to a text query with the number of records matching the query in each Entrez database.
  
  – **ESpell (spelling suggestions)** Retrieves spelling suggestions for a text query in a given database.
  
  – **ECitMatch (batch citation searching in PubMed)** Retrieves PubMed IDs (PMIDs) corresponding to a set of input citation strings.
Platforms for “crowdsourcing”

• **dbGAP**: the database of Genotypes and Phenotypes archives and distributes the results of studies that have investigated the interaction of genotype and phenotype.


DRUG AND CHEMICALS
INFORMATION
Identify or Search for a Pill

Imprint: letters or numbers on either side of the pill. Pill does not have an imprint.

Shape: Select Shape

Color: Select Color

Size: Select Size (search +/- 2mm)

Score: Unknown 1 2 3 4

Drug Name or Ingredient(s):

Inactive Ingredient(s):

Find pills WITHOUT this ingredient.

Label Author:

DEA Schedule:

Select DEA Schedule

Product Code:

Search Clear

Discover
There's more to a pill than how it looks. What's inside the pill other than the drug? Is it a controlled substance?

Connect
Learn more than the pill's name. Pillbox links you to the drug label, clinical trials, breastfeeding safety, and more.

Explore
Pillbox's image explorer is a photo album for pills. Sort by color, shape, size, scoring, and the text printed on the pill.

(requires Adobe Flash and will not run on some mobile devices)

Data updated: November 4, 2013
PubChem provides information on the biological activities of small molecules. PubChem substance contains more than 140 million records. Pccompound contains more than 51 million unique structures. PCBioAssay contains more than 1 million BioAssays. Each BioAssay contains a various number of data points.
RxNav

RxNav is a browser for several drug information sources, including RxNorm, RxTerms and NDF-RT. RxNav finds drugs in RxNorm from the names and codes in its constituent vocabularies.

Launch RxNav

Drug Interaction API

A drug interaction API is now available. The API serves as a replacement for the drug interactions removed from the NDF-RT data set in September.

The RxClass API is here!

The RxClass API is a web service for accessing drug classes and associated drug members. The RxClass Browser uses the API for exploring and navigating through the class hierarchies to find the RxNorm drug members.
• Information **for patients** from the National Institutes of Health and other trusted sources on over 900 diseases and conditions, drugs and wellness issues.

• A medical encyclopedia and a medical dictionary, easy-to-understand tutorials.

• Health information in Spanish and other languages.

• Health information from the media updated daily.

• No advertisement and no endorsement for any company or product.

http://www.nlm.nih.gov/medlineplus/
• Linking Patient Portals and EHRs to Consumer Health Information

• MedlinePlus Connect accepts requests for information on diagnoses (problem codes), medications, and lab tests, and returns related MedlinePlus information. It is available as a Web application or a Web service.
  – For problem code requests, MedlinePlus Connect supports: ICD-9-CM (International Classification of Diseases, 9th edition, Clinical Modification); ICD-10-CM, SNOMED CT® (Systematized Nomenclature of Medicine, Clinical Terms)
  – For medication requests, MedlinePlus Connect supports RXCUI (RxNorm Concept Unique Identifier); NDC (National Drug Code)
  – For lab test requests, MedlinePlus Connect supports: LOINC® (Logical Observation Identifiers Names and Codes)

• May help an EHR achieve one of the criteria for Meaningful Use of Health Information Technology.

• **Consumer** information about genetic conditions and the genes or chromosomes associated with those conditions.
  – Genetic Condition Summaries
  – Gene Summaries
  – Gene Family Summaries
  – Chromosome Summaries
  – Handbook
    • explanations of how genes work and how mutations cause disorders;
      information about genetic testing, and gene therapy.
  – Glossary

Clinical Trials Registry

ClinicalTrials.gov is a registry and results database of publicly and privately supported clinical studies of human participants conducted around the world. Learn more about clinical studies and about this site, including relevant history, policies, and laws.

ClinicalTrials.gov currently lists 175,684 studies with locations in all 50 states and in 187 countries.
Clinical Trials Results

ClinicalTrials.gov
A service of the U.S. National Institutes of Health

Search for studies: [Search]
Advanced Search | Help | Studies by Topic | Glossary

Find Studies ▼ About Clinical Studies ▼ Submit Studies ▼ Resources ▼ About This Site ▼

Home ▶ Find Studies ▶ Search Results

14563 studies found for: ( NOT NOTEXT ) [FIRST-RECEIVED-RESULTS-DATE]
Modify this search | How to Use Search Results

List ▶ By Topic ▶ On a Map ▶ Search Details

+ Show Display Options

Download ▶ Subscribe to RSS

- Include only open studies □ Exclude studies with unknown status

<table>
<thead>
<tr>
<th>Rank</th>
<th>Status</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Terminated</td>
<td>Oxandrolone to Heal Pressure Ulcers</td>
</tr>
<tr>
<td></td>
<td>Has Results</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Condition: Pressure Ulcer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Interventions: Drug; Oxandrolone; Drug; Placebo</td>
</tr>
<tr>
<td>2</td>
<td>Terminated</td>
<td>Study of Pegylated Interferon-Alfa 2b in Combination With PUVA Therapy In CTCL</td>
</tr>
<tr>
<td></td>
<td>Has Results</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Condition: Lymphoma</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Interventions: Biological: Pegylated interferon α-2b; Other: Psoralens with ultraviolet light A; Other: Narrowband-ultraviolet light B</td>
</tr>
</tbody>
</table>
### Outreach Through Quality Information, Capacity Building and Community Engagement

#### Web Sites For and About Specific Populations

<table>
<thead>
<tr>
<th>Web Site Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Indian Health</td>
<td>Addresses the health and well being of American Indians in the United States</td>
</tr>
<tr>
<td>GeneEd</td>
<td>Links to vetted genetic Web sites based on high school science curriculum</td>
</tr>
<tr>
<td>Arctic Health</td>
<td>Addresses the issues affecting the health of our planet’s northern-most inhabitants</td>
</tr>
<tr>
<td>HIV/AIDS Information for Specific Populations</td>
<td>Comprehensive HIV/AIDS information for scientists, physicians, educators, and consumers</td>
</tr>
<tr>
<td>Asian American Health</td>
<td>Addresses the health and well being of Asian Americans in the United States</td>
</tr>
<tr>
<td>K-12 Science and Health Education</td>
<td>Working with teachers and science experts to provide free reliable resources to help introduce, reinforce, and supplement education</td>
</tr>
<tr>
<td>Community Health Maps</td>
<td>Low cost and easy-to-use mapping and mobile communication tools</td>
</tr>
<tr>
<td>Multi-Cultural Resources for Health Information</td>
<td>Information about cultural competency, tools, health literacy, research, and policy</td>
</tr>
<tr>
<td>Environmental Health Student Portal</td>
<td>Web information resources for middle school students and teachers</td>
</tr>
<tr>
<td>NLM 4 Caregivers</td>
<td>Health information outreach to caregiver networks through social media</td>
</tr>
<tr>
<td>ToxTown</td>
<td>Environmental health concerns where you live, work, and play</td>
</tr>
<tr>
<td>ToxMystery</td>
<td>Interactive Game for ages 7-11, household chemical hazards</td>
</tr>
<tr>
<td>NLM_OSP Outreach Listserv</td>
<td>Multilingual health information resources for those providing care to resettled refugees and asylum seekers</td>
</tr>
</tbody>
</table>
The UMLS was initiated in 1986 as a long-term research & development project of the National Library of Medicine to reduce system development problems caused by differences in medical terminologies.

- **Unified**
- **Medical**
- **Language**
- **System**

UMLS®
Unified Medical Language System®

UMLS Knowledge Sources & Tools

- **Metathesaurus**
  - Clusters terms into concepts with unique identifier
    - Over 140 source vocabularies (82 in Active Release) in 21 languages
    - 2,973,458 million concepts (8,646,656 distinct normalized concept names)

- **Semantic Network**
  - Defines relationships between concepts, organizes concepts into categories
    - 133 Semantic types (broad categories)
    - 54 Semantic relations (between categories)

- **Lexical resources**
  - SPECIALIST Lexicon (over 476,000 records)
    - morphology (inflection, derivation); spelling variants; part of speech....
  - Lexical tools (programs and databases)
    - Variant generation; spell check; POS tagger...

- **MetamorphoSys** (installation and customization)

- **UMLS Terminology Services (UTS)**
  - provide web interfaces and web services to search and retrieve UMLS data

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

• It is too broad for a specific purpose
• You are interested in only one of the 21 languages
• UMLS preserves the meanings, hierarchical connections, and other relationships between terms present in its source vocabularies
  — some of these make no sense in your context
  — some are generally suppressible
• You need to add your terms or relations
• You don’t have licenses for copyrighted sources
UMLS-Based Resources & Use Examples

• Named Entity Recognition
  – MetaMap

• Automated indexing assistance
  – MTI

• Information retrieval
  – ClinicalTrials.gov
  – Open-i
MetaMap

• Purpose: Concept Identification (Named-Entity Recognition)
  – Identify UMLS Metathesaurus concepts in text

Examples:

Text gap, lexical variation: assay/assays

\textit{sister-chromatid exchange genotoxicity assays}

C1519335:Sister Chromatid Exchange Assay

Composite phrases, word order, text gaps

\textit{pain on the left side of the chest}

C0541828:chest pain left side(Left sided chest pain)


MetaMap slides courtesy of Dr. Lan Aronson, François Lang, and James Mork
The Algorithm

• Parsing
  – SPECIALIST minimal-commitment parser
  – SPECIALIST lexicon
  – MedPost part-of-speech tagger

• Variant generation
  – SPECIALIST lexicon
  – Lexical Variant Generation (LVG)

• Candidate retrieval from the Metathesaurus

• Candidate evaluation

• Mapping construction
Use example

• With the help of MetaMap concept positional information XML tags we can now extract more than 130 different variations of the “kidney stone” string that are being documented within our sample corpus (Past Medical History section of 640,000 notes).

• These 137 trigger strings are matched by 10 UMLS synonyms.

Example:

– C0022650: Kidney Calculi:
  • calculi in bilateral kidneys; calculi in both kidneys; calculi in the left renal; calculi within the calices of the left kidney....

Courtesy of Sina Madani, MD Anderson Cancer Center
The NLM Medical Text Indexer (MTI)

- Product of Indexing Initiative
- Principal developer James Mork
- Assists NLM Indexers
- In use since mid-2002
- Uses article Title and Abstract
- Semi-Automatic MeSH Indexing Recommendations
- Performs initial indexing for small subset of journals (MTIFL)
• Combines UMLS-based
  – Text processing
  – Information retrieval

• With image processing

• To:
  – Prepare citation
  – Index with search engine

• Provides:
  – User interface for searching biomedical literature
    • By text and/or example images
    • Search filters (image type, limits, etc.)
Preparing citations for indexing

**Image and Text Acquisition**
- Articles & images
- MEDLINE® citations
- Links to Health Topics

**Text Processing**
- Bibliographic and meta-information extraction
- Caption and mention extraction
- UMLS terms extraction
- Image type detection
- Image feature extraction

**Image Processing**
- Image and text acquisition

**Enriched Citation**
- Image and text acquisition
- Text processing
- Image processing

Links to Health Topics: MEDLINE® citations, Articles & images.
Enriched citations

- XML for indexing
- `<EnrichedCitation>`
  - `<Title>`...
  (other MEDLINE citation fields)
- `<ImageDocument>`
- `<Modality>`
- `<Caption>`
- `<Mention>`
- `<Panel>`
- `<VisualWords>`
- `<Modality>`
- `<ROI>`

On Optical Detection of Densely Labeled Synapses in Neuronal and Mapping Connectivity with Combinatorial Multiplexed Fluorescent Synaptic Markers


**Bottom Line:** We simulate fluorescence from a population of densely labeled synapses in a block of hippocampal neurons, completely reconstructed from electron microscopy data, and show that high-end LM is able to detect such patterns with over 95% accuracy. We conclude therefore, that with the described approach neural connectivity in macroscopically large neural circuits can be mapped with great accuracy, in scalable manner, using fast optical tools, and straightforward image processing. Relying on an electron microscopy dataset, we also derive and explicitly enumerate the conditions that should be met to allow synaptic connectivity studies with high-resolution optical tools.

Affiliation: Department of Statistics and Center for Theoretical Neuroscience, Columbia University, New York, New York, USA. 

yanyan@chelnko.com

Abstract: We propose a new method for mapping neural connectivity optically, utilizing CreLox system Brainsbow to tag synapses of different neurons with random mixtures of different fluorophores, such as GFP, YFP, etc., and then detecting patterns of fluorophores at different synapses using light microscopy (LM). Such patterns will immediately report the pre- and post-synaptic cells at each synaptic connection, without tracing neural projections from individual synapses to corresponding cell bodies. We simulate fluorescence from a population of densely labeled synapses in a block of hippocampal neurons, completely reconstructed from electron microscopy data, and show that high-end LM is able to detect such patterns with over 95% accuracy. We conclude therefore, that with the described approach neural connectivity in macroscopically large neural circuits can be mapped with great accuracy, in scalable manner, using fast optical tools, and straightforward image processing. Relying on an electron microscopy dataset, we also derive and explicitly enumerate the conditions that should be met to allow synaptic connectivity studies with high-resolution optical tools.

**Figure 1:** Schematic explanation of synapse detection using co-localization of fluorophores from different pre- and post-synaptic markers. A) Schematic drawing of the synaptic Baldwin, with a red fluorophore on the pre-synaptic side and a green fluorophore on the post-synaptic side of a synaptic cleft. Spatial resolution of the fluorescence from the pre- and post-synaptic fluorophores, occurring due to the proximity of a synaptic cleft, allow detecting synapses optically without explicitly resolving them. B) Due to absence of the fluorophores in the bulk of the animal and dendrite synapses, nearby processes do not interfere with the detection process even when all neurons are labeled, unlike in regular Brainsbow. C) Due to closer spatial co-localization of the pre- and post-synaptic fluorophores across the synaptic cleft, their fluorescence intensity is closely correlated near labeled synapses. In this figure, we show a simulated scatter plot of the fluorescence intensity in the Brainsbow. Total dose of exposure is taken away from one labeled synapse (Red/Blue ~300 µm), and the dose of exposure is taken away from another. (Red/Blue ~300 µm). One can threshold the image with certain thresholds, T1 for the pre-synaptic marker and T2 for the post-synaptic marker (dashed lines). In order to separate the pre-synaptic (Red) from the post-synaptic (Blue) synapses, we use the teacher presence of a synapse. Owing correlation in the fluorescence from the pre- and post-synaptic markers, synapses may be detected even when they cannot be unambiguously resolved into isolated puncta. Labelled here are three synapses, fluorescence from which is identically shown with thin blue, red, and brown lines. These are observed using the fluorescent markers, green and red. First synapse is tagged only with “green” marker second synapse is tagged with “green” and “red” markers, and third synapse is tagged with “red” marker. Combined fluorescence from these synapses is shown with their red and green lines for the two markers respectively. Even though none of the synapses can be seen separately in either green or red channel, by thresholding fluorescence with appropriate thresholds, T1 and T2, three different synapses (black dots) indicate presence of three synapses.

**Mention:** While one can detect synapses with LM by looking for explicitly isolated fluorescent puncta, one can also use a more powerful, yet simpler, procedure for detecting synapses explicitly. Specifically, co-localized synapses labeled with two fluorophores, a fluorophore GFP in the pre-synaptic side and a fluorophore RFP in the post-synaptic side. Because of the spatial proximity of these two fluorophores across the synaptic cleft (i.e., ~50–60 nm apart), the fluorophores attached to these fluorophores will be clearly visible in the region near labeled synapses, Figure 2A and 2B. This situation may be quantified and used to detect synapses even when they cannot be resolved as an isolated punctum, Figure 2C.
Open-i User Interface Functions

- Images as search input
- LinkOut resources
- Multiple views
- Filtering by image modality, journal type, clinical specialties

- Query expansion via UMLS synonymy *
- Re-rank according to specific task *
- Phrase search and complex queries *

* Essie indexing supports these functions. Essie is a search engine developed by Lister Hill Center for ClinicalTrials.gov
CASE STUDY: SUPPORTING NLM CUSTOMER SERVICES
Request Processing System

NLM Main MedlinePlus PubMed GHR

Frequently Asked Questions

Stock Answer Generator

Health Information requests

Question Answering

Customer relationship management system

Siebel
Incoming customers requests

Average 131.5 per day

- **Document Delivery/ILL**
- **Reference Questions**
- **PubMed Corrections**
- **PubMed**
- **MedlinePlus**
- **Drug Products Questions**
- **ClinicalTrials.gov**
- **Other**
Frequent requests

• PubMed corrections (about 20-25 a day)
  – In this pubmed reference:
    http://www.ncbi.nlm.nih.gov/pubmed/23653210 there is a typo in the title from the pubmed database. The title should be "N6-formyladenosine" instead of "N4-formyladenosine". Thanks!

• Consumer health questions (about 15-20 a day)
  – My sister is diabitic since age 8, now age 20. she is not coping well with it, i've been to the nutrition DR.s with her and we have since change our diet at home just for her, but it's still too havy on her. i want to help her, maybe if i can get her a counselor or support group that understands her situation/diabits
Request triage

• Classify incoming requests as PubMed corrections, Consumer health questions (limited to disorders), other
  – MaxEnt, SVM trained on 100K 2011 requests, 80 - 85% F-score
  – Ongoing / Future work
    • More classes: Junk; Other types of CHQ (e.g., medications); UMLS
    • Subclasses: Find a Doctor or Expert for a given disorder
    • Classes and distributions are changing

• PubMed Corrections Assistant:
  – Prepares stock replies for PubMed correction requests

• Question Answering module:
  – Prepares answers for consumer health questions
PubMed Corrections Assistant

• Searches the request for PMID(s) or reference(s)
• Retrieves PubMed citations using E-Utilities
• Extracts the citation status field:
  – PubMed
  – PubMed - in process (now undergoing changes)
  – PubMed - as supplied by Publisher
  – PubMed - indexed for MEDLINE
• Using the status, finds an appropriate stock reply
• Adds the stock reply to the request
• Sends the augmented request to customer service agent for approval/editing
CRC answers in Siebel
# Unexpected consequences: New PubMed corrections request form

<table>
<thead>
<tr>
<th>Menu item</th>
<th>Look up</th>
<th>Citation field value</th>
<th>Form Action(s) Before sending to Siebel</th>
<th>Siebel XML tag AREA</th>
<th>Siebel XML tag SUBAREA</th>
<th>Siebel XML tag Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Misspelling or error in PubMed Record Status</td>
<td>PubMed</td>
<td>MEDLINE Lookup CitationSubset</td>
<td>Add to SRDescriptionNotes: -current text/-correction text -PMID -Summary (text) citation</td>
<td>Quality Control of NLM DB</td>
<td>MEDLINE</td>
<td>Indexing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>For CitationSubset == OM</td>
<td>Same action as above</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Misspelling or error in PubMed Record Status</td>
<td>PubMed</td>
<td>In-Process OR In-data-review</td>
<td>Add to SRDescriptionNotes: -current text/-correction text -PMID -Summary (text) citation</td>
<td>Quality Control of NLM DB</td>
<td>In Process</td>
<td>Indexing</td>
</tr>
<tr>
<td>Misspelling or error in PubMed Record Status</td>
<td>PubMed</td>
<td>Publisher</td>
<td>Add to SRDescriptionNotes: -PMID -Summary (text) citation</td>
<td>Quality Control of NLM DB</td>
<td>Publisher Supplied</td>
<td>Customer Service</td>
</tr>
<tr>
<td>Misspelling or error in PubMed Record Status</td>
<td>PubMed</td>
<td>OLDMEDLINE</td>
<td>Add to SRDescriptionNotes: -current text/-correction text -PMID -Summary (text) citation</td>
<td>Quality Control of NLM DB</td>
<td>MEDLINE</td>
<td>MMS</td>
</tr>
<tr>
<td>Misspelling or error in PubMed Record Status</td>
<td>PubMed</td>
<td>PubMed-not-MEDLINE</td>
<td>Add to SRDescriptionNotes: -current text/-correction text -PMID -Summary (text) citation</td>
<td>Quality Control of NLM DB</td>
<td>PubMed Only</td>
<td>Indexing</td>
</tr>
<tr>
<td>Question about MeSH subject indexing</td>
<td>N/A</td>
<td>N/A</td>
<td>Add to SRDescriptionNotes: -PMID -Summary (text) citation</td>
<td>Quality Control of NLM DB</td>
<td>Assigned MeSH Headings</td>
<td>Indexing</td>
</tr>
<tr>
<td>Copy of Full Text</td>
<td>N/A</td>
<td>N/A</td>
<td>Add to SRDescriptionNotes: -PMID -Summary (text) citation</td>
<td>PubMed</td>
<td>Full-Text/Obtaining Articles</td>
<td>Customer Service</td>
</tr>
<tr>
<td>Forgot MyNCBI Password</td>
<td>N/A</td>
<td>N/A</td>
<td>Send question to Siebel Forward for action to <a href="mailto:info@ncbi.nlm.nih.gov">info@ncbi.nlm.nih.gov</a></td>
<td>PubMed</td>
<td>MyNCBI</td>
<td>Customer Service</td>
</tr>
</tbody>
</table>
Question Answering Assistant

• Represents question in a structured form (frame) – question understanding
• Uses the frame to generate queries to search engines
• Retrieves articles
• Extracts answers
• Ranks answers
Understanding consumer questions is hard

- Different styles
  - Education level
  - Native language

- Mostly informal language
  - Ungrammatical sentences
  - Inconsistent punctuation and capitalization
  - Abbreviations
  - Misspellings
  - Extraneous information
  - Abundance of anaphora and ellipses
<table>
<thead>
<tr>
<th>Question type</th>
<th>Question template(s)</th>
<th>All questions could specify context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information</td>
<td>Information about Disease</td>
<td></td>
</tr>
<tr>
<td>Person/Organization</td>
<td>Geo / internet location of organization / specialist for (testing</td>
<td>treatment</td>
</tr>
<tr>
<td>Anatomy</td>
<td>Body part/function affected by Disease</td>
<td></td>
</tr>
<tr>
<td>Management</td>
<td>[Effects of Drug/Procedure as] treatment/prevention for Disease</td>
<td></td>
</tr>
<tr>
<td>Diagnosis</td>
<td>[Procedure as] diagnosis for Disease</td>
<td></td>
</tr>
<tr>
<td>Cause</td>
<td>Cause of Disease</td>
<td></td>
</tr>
<tr>
<td>Susceptibility</td>
<td>Population affected by Disease; Inheritance patterns</td>
<td></td>
</tr>
<tr>
<td>Complications</td>
<td>Problems caused by Disease</td>
<td></td>
</tr>
<tr>
<td>Prognosis</td>
<td>Outcome/life expectancy/quality of life [in patient] with Disease</td>
<td></td>
</tr>
<tr>
<td>Other Effects</td>
<td>effects of a Disease that are not explicitly Complications or Manifestations</td>
<td></td>
</tr>
<tr>
<td>Manifestations</td>
<td>Signs  and symptoms of Disease</td>
<td></td>
</tr>
</tbody>
</table>
Rule-based question understanding

• Find terms corresponding to entities of interest:
  – UMLS terms corresponding to Diseases, Anatomy, Drugs, Genes
  – Gazetteers for question cues & question type indicators
• Use syntactic patterns that establish relations between terms
• Translate relations to structured form (frames)
• Use methods for anaphora and ellipsis resolution
Example question frames

Are there treatments for trisomy 13?
What is the prognosis?
What is life like for those who have this condition?

• Frame 1
  – Question type: Management
    • Indicated by: treatments
  – Theme (problem): trisomy 13
  – Question cue: are there

• Frame 2
  – Question type: Prognosis
    • Indicated by: prognosis
  – Theme (problem): trisomy 13
    • Ellipsis resolution
  – Question cue: what

• Frame 3
  ❖ Question type: Manifestations
    ❖ Indicated by: life like
  ❖ Theme (problem): trisomy 13
    ❖ Anaphora resolution
      ❖ this condition → trisomy 13
  ❖ Question cue: what
Frame construction

- Type and Theme argument are mandatory
- Stanford dependencies to identify theme and question cue arguments
  - direct object, passive nominal subject, noun compound modifier, relative clause modifier, etc.
- Transitivity on Information triggers
  - Is there information regarding prognosis of . . .
- Transitivity on population group terms
  - What is the life expectancy for a child with trisomy 13? . . .
Trisomy 13

Trisomy 13 (also called Patau syndrome) is a genetic disorder in which an extra copy of chromosome 13 is present in the body. This condition occurs when an extra copy of chromosome 13 is present in place of one of the usual two copies of chromosome 13. This extra copy of chromosome 13 can result in developmental problems and health issues.

Causes

Trisomy 13 occurs when an extra copy of chromosome 13 is present in the body. This can happen for a variety of reasons, including:

- **An extra chromosome 13**
- **A translocation involving chromosome 13**
- **A partial deletion of chromosome 13**

These causes can lead to the extra copy of chromosome 13, which can interfere with normal development.

Symptoms

Symptoms of trisomy 13 can include:

- **Cleft lip or palate**
- **Clenched hands**
- **Close-set eyes**
- **Decreased muscle tone**
- **Extra fingers or toes**
- **Hernias**
- **Hole in the iris (coloboma)**
- **Micophthalmia**
- **Muscle weakness**
- **Partial deletion of chromosome 13**

What is trisomy 13?

Trisomy 13, also known as Patau syndrome, is a chromosomal condition that occurs when an extra copy of chromosome 13 is present in the body. This can lead to a range of health issues and developmental problems. In many cases, individuals with trisomy 13 have heart defects, brain or spinal cord abnormalities, and other physical and intellectual disabilities. Some common symptoms include:

- **Heart defects**
- **Brain and spinal cord abnormalities**
- **Intellectual disability**
- **Muscle weakness**
- **Muscle tone problems**

How common is trisomy 13?

Trisomy 13 occurs in about 1 in 16,000 newborns. Although women of any age can have a child with trisomy 13, the chance of having a child with this condition increases as a woman gets older.

What are the genetic changes related to trisomy 13?

Most cases of trisomy 13 result from having three copies of chromosome 13 in each cell of the body instead of the usual two copies. The extra genetic material disrupts the normal course of development, causing the characteristic features of trisomy 13.

Trisomy 13 can also occur when part of chromosome 13 becomes attached (translocated) to another chromosome during the formation of reproductive cells (eggs and sperm) or very early in fetal development. Affected people have two normal copies of chromosome 13 plus an extra copy of chromosome 13 attached to another chromosome. In rare cases, only part of chromosome 13 is present in three copies. The physical signs and symptoms in these cases may not show all the symptoms of full trisomy 13.

A small percentage of people with trisomy 13 have an extra copy of chromosome 13 in only some of the body’s cells. In these people, the condition is called mosaic trisomy 13. The severity of mosaic trisomy 13 depends on the type and number of cells that have the extra chromosome. The physical features of mosaic trisomy 13 are often milder than those of full trisomy 13.
Translating structured questions to search queries

- Search for the theme (problem) in titles
  - If found, return an appropriate section for question type

![Example Screen Shot]

**Treatment**

Treatment varies from child to child and depends on the specific symptoms.

**Support Groups**

Support Organization for Trisomy 18, 13 and Related Disorders (SOFT) -- www.trisomy.org

**Outlook (Prognosis)**

More than 80% of children with trisomy 13 die in the first year.

**Possible Complications**

Complications begin almost immediately. Most infants with trisomy 13 have congenital heart
Selecting the best answer

• Rank articles and sections

• From the best sections, select paragraphs containing more answer type indicators
  – trim long paragraphs

• Due to the presence of several life-threatening medical problems, many infants with trisomy 13 die within their first days or weeks of life. Only five percent to 10 percent of children with this condition live past their first year.

• More than 80% of children with trisomy 13 die in the first year.
Unexpected consequences: Quality assurance for web resources

• **GHR**
  – Added AddThis links

• **MedlinePlus**
  – Revised topics:
    • Tuberculosis testing
    • Shingles
    • Tubal ligation ...
Anaphora, Coreference, Bridging Inferences, Ellipsis...

- my question is this: I was born w/a esophagus atresia w/dextrocardia. While the heart hasn't caused problems, the other has [caused problems]. I get food caught all the time. My question is... is there anything that can fix it cause I can't eat anything lately without getting it caught. I need help or will starve!

- Pronominal anaphora
  - it → esophagus atresia
  - it → food

- Other-anaphora (Comparative anaphora)
  - the other → esophagus atresia

- Ellipsis
  - Omission of caused problems: VP ellipsis
Anaphora resolution for frame extraction

Dependency-based frame extraction with simple dictionary lookup provides good precision but low recall.

Anaphora/ellipsis resolution helps significantly in recovering relevant frames with little negative effect on precision.

Better term identification → better anaphora/ellipsis resolution → better question understanding.

<table>
<thead>
<tr>
<th></th>
<th># of frames</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>14</td>
<td>0.32</td>
<td>0.93</td>
<td>0.48</td>
</tr>
<tr>
<td>Anaphora</td>
<td>26</td>
<td>0.54</td>
<td>0.85</td>
<td>0.66</td>
</tr>
</tbody>
</table>
Ongoing & future work

• Expand knowledge base
  – Link to relevant ClinicalTrials.gov
  – Mine NIH ICs, CDC, FDA consumer Web sites
  – Identify PubMed and PubMed Central articles for consumers

• Answer generation

• Understanding questions beyond “known disease”

• Addressing remaining request types