T29: UMLS Concept Identification Using the MetaMap System

AMIA Fall Symposium
Tutorial 29, Methods Series
Sunday, November 14, 2010
8:30am – 12:00pm

Alan R. Aronson, Dina Demner-Fushman, François-Michel Lang, James G. Mork
Tutorial Outline

• Background: why concept identification? (Lan)
• Introduction to MetaMap (Lan)
• The MetaMap algorithm (Lan)
• Input/output formats (François, Jim)
• MetaMap options in depth (Lan)
• MetaMap processing modes (Lan)
• Usage warnings (Lan)
• Access methods (Jim)
• Research projects using MetaMap (Jim, Dina)
• Future directions (Lan)
Tutorial Outline

- **Background: why concept identification? (Lan)**
  - Introduction to MetaMap (Lan)
  - The MetaMap algorithm (Lan)
  - Input/output formats (François, Jim)
  - MetaMap options in depth (Lan)
  - MetaMap processing modes (Lan)
  - Usage warnings (Lan)
  - Access methods (Jim)
  - Research projects using MetaMap (Jim, Dina)
  - Future directions (Lan)
Why Concept Identification?

- Structured data vs. text
- Concept identification is useful/essential for many tasks including
  - Information extraction/Data mining
  - Classification/Categorization
  - Text summarization
  - Question answering
  - Literature-based knowledge discovery
Motivation for Creation of MetaMap

• Our original concept identification task, Information Retrieval (IR):
  • retrieval of MEDLINE records based on textual queries by
  • identifying biomedical concepts occurring in both queries and MEDLINE, and
  • leveraging knowledge about concepts contained in UMLS Metathesaurus
Tutorial Outline

• Background: why concept identification? (Lan)
  ➢ **Introduction to MetaMap (Lan)**
• The MetaMap algorithm (Lan)
• Input/output formats (François, Jim)
• MetaMap options in depth (Lan)
• MetaMap processing modes (Lan)
• Usage warnings (Lan)
• Access methods (Jim)
• Research projects using MetaMap (Jim, Dina)
• Future directions (Lan)
Concept Identification Programs

- Selected programs that map biomedical text to a thesaurus
  - SAPHIRE (*Hersh et al., 1990*)
  - CLARIT (*Evans et al., 1991*)
  - **MetaMap** (*Aronson et al., 1994*)
  - Metaphrase (*Tuttle et al., 1998*)
  - **MMTx (2001)**
  - KnowledgeMap (*Denny et al., 2003*)
  - Mgrep (*Meng, 2009--unpublished*)

- Characteristics of MetaMap
  - Linguistic rigor
  - Flexible partial matching
  - Emphasis on thoroughness rather than speed
  - Restricted to English syntax and vocabulary
Example (best mappings)

- PMID – 9339686
- AB – Cerebral blood flow (CBF) in newborn infants is often below levels necessary to sustain brain viability
  - Cerebrovascular Circulation
  - Infant, Newborn
  - CEREBRAL BLOOD FLOW IMAGING
  - Frequent
  - Levels (qualifier value)
  - Sustained
  - Brain
  - Viable
  - Entire brain
  - in adults.
  - Adult

U. S. NATIONAL LIBRARY OF MEDICINE
Example (best mappings with WSD)

- PMID – 9339686
- AB – *Cerebral blood flow (CBF) in newborn infants* is often below levels necessary to sustain brain viability in adults.

CEREBRAL BLOOD FLOW IMAGING

Frequent Levels (qualifier value) Sustained Brain Viable

Entire brain

U. S. NATIONAL LIBRARY OF MEDICINE
MetaMap Examples (1/7)

• “inferior vena caval stent filter” maps to
  • ‘Inferior Vena Cava Filter’ (‘Vena Cava Filters’) and
  • ‘Stent’

• “medicine” with --allow_overmatches maps to
  • ‘Alternative Medicine’ or
  • ‘Medical Records’ or
  • ‘Nuclear medicine procedure, NOS’ or ...

• “pain on the left side of the chest” with
  --quick_composite_phrases maps to
  • ‘Left sided chest pain’ (under development)
Example: Normal processing (2/7)

Phrase: “lung cancer.”

Meta Candidates (8):
1000 Lung Cancer (Malignant neoplasm of lung) [Neoplastic Process]
1000 Lung Cancer (Carcinoma of lung) [Neoplastic Process]
  861 Cancer (Malignant Neoplasms) [Neoplastic Process]
  861 Lung [Body Part, Organ, or Organ Component]
  861 Cancer (Cancer Genus) [Invertebrate]
  861 Lung (Entire lung) [Body Part, Organ, or Organ Component]
  861 Cancer (Specialty Type - cancer) [Biomedical Occupation or Discipline]
  768 Pneumonia [Disease or Syndrome]

Meta Mapping (1000):
1000 Lung Cancer (Carcinoma of lung) [Neoplastic Process]

Meta Mapping (1000):
1000 Lung Cancer (Malignant neoplasm of lung) [Neoplastic Process]
Example: Variants (−v) (3/7)

Phrase: “lung cancer.”

lung cancer [noun] variants (n=1):
  lung cancer{{noun}, 0=[]}

  lung [noun] variants (n=9):
  lung{{noun}, 0=[]}  lungs{{noun}, 1="i"}  pneumonia{{noun}, 5="ds"}
    pneumonias{{noun}, 5="ds"}
    pneumonic{{adj}, 2="s"}  pulmonic{{adj}, 4="ss"}  pulmonary{{adj}, 2="s"}
  pulmonic{{adj}, 2="s"}

  cancer [noun] variants (n=4):
  cancer{{noun}, 0=[]}  cancerous{{adj}, 3="d"}  cancers{{noun}, 1="i"}
    carcinomatous{{adj}, 2="s"}
...

U.S. NATIONAL LIBRARY OF MEDICINE
Phrase: “obstructive sleep apnea.”
Meta Candidates (8):

... 

Meta Mapping (1000):
1000 Obstructive sleep apnoea (Sleep Apnea, Obstructive) [Disease or Syndrome]

Meta Mapping (901):
827 Obstructive (Obstructed) [Functional Concept]
901 Apnea, Sleep (Sleep Apnea Syndromes) [Disease or Syndrome]

Meta Mapping (851):
827 Obstructive (Obstructed) [Functional Concept]
827 Sleep [Organism Function]
827 APNOEA (Apnea) [Pathologic Function]

...
Example: Show sources (-G) (5/7)

Phrase: “scorpion sting.”

Meta Candidates (4):
1. 1000 Scorpion sting \{MDR, DXP\} [Injury or Poisoning]
2. 861 Sting (Sting Injury \{MTH, MSH, MDR, RCD, SNM, SNOMEDCT, SNMI, WHO\}) [Injury or Poisoning]
3. 694 Scorpion (Scorpions \{LCH, MSH, MTH, SNM, SNOMEDCT, SNMI, CSP, RCD, NCBI\}) [Invertebrate]
4. 694 SCORPION (Scorpion antigen \{MTH, LNC\}) [Immunologic Factor]

Meta Mapping (1000):
1. 1000 Scorpion sting \{MDR, DXP\} [Injury or Poisoning]
Example: Restrict to sources (–GR  LCH) (6/7)

Phrase: “scorpion sting.”

Meta Candidates (1):
   694 Scorpion (Scorpions {LCH}) [Invertebrate]

Meta Mapping (694):
   694 Scorpion (Scorpions {LCH}) [Invertebrate]
Example: Restrict to STs (¬J neop) (7/7)

Phrase: "lung cancer."

Meta Candidates (3):
1000 Lung Cancer (Malignant neoplasm of lung) [Neoplastic Process]
1000 Lung Cancer (Carcinoma of lung) [Neoplastic Process]
861 Cancer (Malignant Neoplasms) [Neoplastic Process]

Meta Mapping (1000):
1000 Lung Cancer (Carcinoma of lung) [Neoplastic Process]
Meta Mapping (1000):
1000 Lung Cancer (Malignant neoplasm of lung) [Neoplastic Process]
MetaMap Demo

Tutorial Outline

• Background: why concept identification? (Lan)
• Introduction to MetaMap (Lan)
  ➢ The MetaMap algorithm (Lan)
• Input/output formats (François, Jim)
• MetaMap options in depth (Lan)
• MetaMap processing modes (Lan)
• Usage warnings (Lan)
• Access methods (Jim)
• Research projects using MetaMap (Jim, Dina)
• Future directions (Lan)
The Algorithm

- Parsing
  - Using SPECIALIST minimal commitment parser, SPECIALIST lexicon, MedPost part of speech tagger

- Variant generation
  - Using SPECIALIST lexicon, Lexical Variant Generation (LVG)

- Candidate retrieval
  - From the Metathesaurus

- Candidate evaluation

- Mapping construction
Ocular complications of myasthenia gravis.
Variant Generation

- Variants of adjective *ocular* (13, 9 occur in UMLS):
  - *ocular*[{adj}, 0=[]]
  - *eye*[{noun}, 2="s"]
  - *eyes*[{noun}, 3="si"]
  - *optic*[{adj}, 4="ss"]
  - *ophthalmic*[{adj}, 4="ss"]
  - *ophthalmia*[{noun}, 7="ssd"]
  - *ophthalmias*[{noun}, 8="ssdi"]
  - *ophthalmiac*[{noun}, 7="ssd"]
  - *ophthalmiacs*[{noun}, 8="ssdi"]
  - *oculus*[{noun}, 3="d"]
  - *oculi*[{noun}, 4="di"]
  - *ocularity*[{noun}, 3="d"]
  - *ocularities*[{noun}, 4="di"]
Evaluation Function

- Weighted average of
  - **centrality** (is the head involved?)
  - **variation** (average of all individual word variations)
  - **coverage** (how much of the text is matched?)
  - **cohesiveness** (in how many pieces?)
Evaluation Results

Phrase: **Ocular complications**

Meta Candidates (8):

- 861 Complications (Complication) [patf]
- 861 complications (Complication Aspects) [patf]
- 777 Complicated [ftcn]
- 694 Ocular (Eye) [bpoc]
- 638 Eye (Entire Eye) [bpoc]
- 611 Optic (Optics) [ocdi]
- 611 Ophthalmic [spco]
- 588 Ophthalmia (Endophthalmitis) [dsyn]
Mapping Construction

Phrase: *Ocular complications*

Meta Mapping (888):

- Ocular (Eye) [bpoc]
- Complications (Complication) [patf]

Meta Mapping (888):

- Ocular (Eye) [bpoc]
- complications (Complication Aspects) [patf]
Mapping Construction (with WSD)

Phrase: *Ocular complications*

Meta Mapping (888):

- Ocular (Eye) [bpoc]
- Complications (Complication) [patf]

Meta Mapping (888):

- Ocular (Eye) [bpoc]
- complications (Complication Aspects) [patf]
Tutorial Outline

• Background: why concept identification? (Lan)
• Introduction to MetaMap (Lan)
• The MetaMap algorithm (Lan)

➢ Input/output formats (François, Jim)
• MetaMap options in depth (Lan)
• MetaMap processing modes (Lan)
• Usage warnings (Lan)
• Access methods (Jim)
• Research projects using MetaMap (Jim, Dina)
• Future directions (Lan)
Input Formats

• ASCII only input
• Unformatted English free text
• MEDLINE Citations
• Input records delimited by blank line
• Single-line delimited input (via job Scheduler)

heart attack
lung cancer

• Single-line delimited input with ID (Scheduler)

000001|heart attack
000002|lung cancer
Input Should Have Syntactic Structure

- Lack of structure → long phrases
  → combinatorial explosion in mappings

- protein-4 FN3 fibronectin type III domain GSH lutathione GST glutathione S-transferase hIL-6 human interleukin-6 HSA human serum albumin IC(50) half-maximal inhibitory concentration Ig immunoglobulin IMAC immobilized metal affinity chromatography K(D) equilibrium constant

- from filamentous bacteriophage f1 PCR polymerase-chain reaction PDB Protein Data Bank PSTI human pancreatic secretory trypsin inhibitor RBP retinol-binding protein SPR surface plasmon resonance TrxA
## Input Format Syntax Limitation

- Phrase length is negligible constraint in MEDLINE:
  - \( \mu = 1.31 \)
  - \( \sigma = 1.18 \)
  - 99.9% of MEDLINE phrases \( \leq 8 \) tokens
- --phrases_only MetaMap option
Be careful of bulleted lists!

<table>
<thead>
<tr>
<th>Agency</th>
<th>NSF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Biotechnology, Biochemical, and Biomass Engineering (BBBE)</td>
</tr>
<tr>
<td>Code</td>
<td>PD 10-1491</td>
</tr>
<tr>
<td>Link</td>
<td><a href="http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=501024&amp;ndldId=USNSF_39">http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=501024&amp;ndldId=USNSF_39</a></td>
</tr>
<tr>
<td>Abstract</td>
<td>The BBBE program emphasizes basic engineering and biological research that advances the fundamental knowledge base that contributes to a better understanding of cellular and biomolecular processes (in vivo, in vitro, and/or ex vivo) and eventually to the development of generic enabling technology and practical application. Research projects supported through the BBBE program include, but are not limited to:</td>
</tr>
</tbody>
</table>

- Fermentation technology
- Enzyme technology
- Recombinant DNA technology
- Cell culture technology
- Ex vivo and therapeutic stem cell culture technology
- Metabolic engineering
- Tissue engineering
- Nanobiotechnology
- Quantitative systems biotechnology

The duration of unsolicited awards is generally one to three years. The average annual award size for the program is $100,000 for individual investigators and $200,000 for multiple investigators. Any proposal received outside the announced dates will be returned without review.
| Description | The Biotechnology, Biochemical, and Biomass Engineering (BBBE) program deals with problems involved in economic processing and manufacturing of products of economic importance by effectively utilizing renewable resources of biological origin and bioinformatics originating from genomic and proteomic information. The BBBE program emphasizes basic engineering and biological research that advances the fundamental knowledge base that contributes to a better understanding of biomanufacturing processes (in vivo, in vitro, and/or ex vivo) and eventually to the development of generic enabling technology and practical application. Quantitative assessments of bioprocesses and their rates at the levels of gene regulation and expression, signal transduction pathways, posttranslational protein processing, enzymes in reaction systems, metabolic pathways, cells and tissues in cultivation, and biological systems including animal, plant, microbial and insect cells, etc. are considered vital to the successful research outcomes in the BBBE program. Research proposals supported through the BBBE program include, but are not limited to, Fermentation technology, Enzyme technology, Recombinant DNA technology, Cell culture technology, Ex vivo and therapeutic stem cell culture technology, Metabolic engineering, Biosensor development, Food processing with special focus on the safety of the nation's food supply, Tissue engineering, Nanobiotechnology, Quantitative systems biotechnology. The duration of unsolicited awards is generally one to three years. Additional information about this program can be accessed. |
| URL | NSF Program Description 08-1491 |
Output Formats: Summary

- Human-readable output
- MetaMap Machine Output (MMO)
- XML output
- Colorized MetaMap output (MetaMap 3D)
- Fielded (MMI) Output
Phrase: "heart attack"

Meta Candidates (8):

1. 1000 Heart attack (Myocardial Infarction) [Disease or Syndrome]
2. 861 Heart [Body Part, Organ, or Organ Component]
3. 861 Attack, NOS (Onset of illness) [Finding]
4. 861 Attack (Attack device) [Medical Device]
5. 861 attack (Attack behavior) [Social Behavior]
6. 861 Heart ( Entire heart) [Body Part, Organ, or Organ Component]
7. 861 Attack (Observation of attack) [Finding]
8. 827 Attacked (Assault) [Injury or Poisoning]

Meta Mapping (1000):

1. 1000 Heart attack (Myocardial Infarction) [Disease or Syndrome]
Human Readable: Metathesaurus String

Phrase: "heart attack"
Meta Candidates (8):

1000 Heart attack (Myocardial Infarction) [Disease or Syndrome]
861 Heart [Body Part, Organ, or Organ Component]
861 Attack, NOS (Onset of illness) [Finding]
861 Attack (Attack device) [Medical Device]
861 attack (Attack behavior) [Social Behavior]
861 Heart (Entire heart) [Body Part, Organ, or Organ Component]
861 Attack (Observation of attack) [Finding]
827 Attacked (Assault) [Injury or Poisoning]

Meta Mapping (1000):

1000 Heart attack (Myocardial Infarction) [Disease or Syndrome]
Human Readable: Preferred Name

Phrase: "heart attack"

Meta Candidates (8):

1. 1000 Heart attack (Myocardial Infarction) [Disease or Syndrome]
2. 861 Heart [Body Part, Organ, or Organ Component]
3. 861 Attack, NOS (Onset of illness) [Finding]
4. 861 Attack (Attack device) [Medical Device]
5. 861 attack (Attack behavior) [Social Behavior]
6. 861 Heart (Entire heart) [Body Part, Organ, or Organ Component]
7. 861 Attack (Observation of attack) [Finding]
8. 827 Attacked (Assault) [Injury or Poisoning]

Meta Mapping (1000):

1. 1000 Heart attack (Myocardial Infarction) [Disease or Syndrome]
Human Readable: MetaMap Scores

Phrase: "heart attack"

Meta Candidates (8):

1000 Heart attack (Myocardial Infarction) [Disease or Syndrome]
861 Heart [Body Part, Organ, or Organ Component]
861 Attack, NOS (Onset of illness) [Finding]
861 Attack (Attack device) [Medical Device]
861 attack (Attack behavior) [Social Behavior]
861 Heart (Entire heart) [Body Part, Organ, or Organ Component]
861 Attack (Observation of attack) [Finding]
827 Attacked (Assault) [Injury or Poisoning]

Meta Mapping (1000):

1000 Heart attack (Myocardial Infarction) [Disease or Syndrome]
Human Readable: Semantic Types

Phrase: "heart attack"

Meta Candidates (8):

1000 Heart attack (Myocardial Infarction) [Disease or Syndrome]
861 Heart [Body Part, Organ, or Organ Component]
861 Attack, NOS (Onset of illness) [Finding]
861 Attack (Attack device) [Medical Device]
861 attack (Attack behavior) [Social Behavior]
861 Heart (Entire heart) [Body Part, Organ, or Organ Component]
861 Attack (Observation of attack) [Finding]
827 Attacked (Assault) [Injury or Poisoning]

Meta Mapping (1000):

1000 Heart attack (Myocardial Infarction) [Disease or Syndrome]
Phrase: "heart attack"
Meta Candidates (8):
   1000 Heart attack (Myocardial Infarction)
   861 Heart
   861 Attack, NOS (Onset of illness)
   861 Attack (Attack device)
   861 attack (Attack behavior)
   861 Heart (Entire heart)
   861 Attack (Observation of attack)
   827 Attacked (Assault)
Meta Mapping (1000):
   1000 Heart attack (Myocardial Infarction)
Human Readable with CUIs (—I)

Meta Candidates (8):
- 1000 C0027051: Heart attack (Myocardial Infarction) [Disease or Syndrome]
- 861 C0018787: Heart [Body Part, Organ, or Organ Component]
- 861 C0277793: Attack, NOS (Onset of illness) [Finding]
- 861 C0699795: Attack (Attack device) [Medical Device]
- 861 C1261512: attack (Attack behavior) [Social Behavior]
- 861 C1281570: Heart (Entire heart) [Body Part, Organ, or Organ Component]
- 861 C1304680: Attack (Observation of attack) [Finding]
- 827 C0004063: Attacked (Assault) [Injury or Poisoning]

Meta Mapping (1000):
- 1000 C0027051: Heart attack (Myocardial Infarction) [Disease or Syndrome]
Human Readable with Sources (−G)

Meta Candidates (8):

1000 Heart attack (Myocardial Infarction {MEDLINEPLUS}) [Disease or Syndrome]
861 Heart {AIR, BI, PNDS} [Body Part, Organ, or Organ Component]
861 Attack, NOS (Onset of illness {MTH, SNOMEDCT, AOD}) [Finding]
861 Attack (Attack device {MTH, MMSL}) [Medical Device]
861 attack (Attack behavior {MTH, PSY, AOD}) [Social Behavior]
861 Heart (Entire heart {MTH, SNOMEDCT}) [Body Part, Organ, or Organ Component]
861 Attack (Observation of attack {MTH, SNOMEDCT}) [Finding]
827 Attacked (Assault {ICD10AM,ICPC2P}) [Injury or Poisoning]

Meta Mapping (1000):

1000 Heart attack (Myocardial Infarction {MEDLINEPLUS}) [Disease or Syndrome]
Output Formats: Human Readable

Phrase: "heart attack"

Meta Candidates (8):

1000 Heart attack (Myocardial Infarction) [Disease or Syndrome]
861 Heart [Body Part, Organ, or Organ Component]
861 Attack, NOS (Onset of illness) [Finding]
861 Attack (Attack device) [Medical Device]
861 attack (Attack behavior) [Social Behavior]
861 Heart (Entire heart) [Body Part, Organ, or Organ Component]
861 Attack (Observation of attack) [Finding]
827 Attacked (Assault) [Injury or Poisoning]

Meta Mapping (1000):

1000 Heart attack (Myocardial Infarction) [Disease or Syndrome]
Output Formats: Machine Output

Prolog terms (pretty-printed & condensed!)

candidates([
    ev(-1000, 'C0027051', 'Heart attack', 'Myocardial Infarction', [heart,attack],
        [dsyn], [[[1,2],[1,2],0]], yes, no, ['MEDLINEPLUS'], [0/12]),
    ev(-861, 'C0018787', 'Heart', 'Heart', [heart],
        [bpoc], [[[1,1],[1,1],0]], yes, no, ['AIR'], [0/5]),
    ev(-861, 'C0277793', 'Attack', 'Onset of illness', [attack],
        [fndg], [[[2,2],[1,1],0]], yes, no, ['MTH'], [6/6]),
    ev(-861, 'C0699795', 'Attack device', [attack],
        [medd], [[[2,2],[1,1],0]], yes, no, ['MTH','MMSL'], [6/6]),
    ev(-861, 'C1261512', attack, 'Attack behavior', [attack],
        [socb], [[[2,2],[1,1],0]], yes, no, ['MTH','PSY','AOD'], [6/6]),
    ev(-861, 'C1281570', 'Entire heart', [heart],
        [bpoc], [[[1,1],[1,1],0]], yes, no, ['MTH','SNOMEDCT'], [0/5]),
    ev(-861, 'C1304680', 'Observation of attack', [attack],
        [fndg], [[[2,2],[1,1],0]], yes, no, ['MTH','SNOMEDCT'], [6/6]),
    ev(-827, 'C0004063', 'Attacked', 'Assault', [attacked],
        [inpo], [[[2,2],[1,1],1]], yes, no, ['ICD10AM'], [6/6]])).
Output Formats: Machine Output

Prolog terms (pretty-printed & condensed!)

candidates(
    ev(-1000, 'C0027051', 'Heart attack', 'Myocardial Infarction', [heart,attack], [dsyn], [[[1,2],[1,2],0]], yes, no, ['MEDLINEPLUS'], [0/12]),
    ev(-861, 'C0018787', 'Heart', 'Heart', [heart], [bpoc], [[[1,1],[1,1],0]], yes, no, ['AIR'], [0/5]),
    ev(-861, 'C0277793', 'Attack', 'ONS', 'Onset of illness', [attack], [fndg], [[[2,2],[1,1],0]], yes, no, ['MTH'], [6/6]),
    ev(-861, 'C0699795', 'Attack', 'Attack device', [attack], [medd], [[[2,2],[1,1],0]], yes, no, ['MTH','MMSL'], [6/6]),
    ev(-861, 'C1261512', 'Attack behavior', [attack], [socb], [[[2,2],[1,1],0]], yes, no, ['MTH','PSY','AOD'], [6/6]),
    ev(-861, 'C1281570', 'Heart', 'Entire heart', [heart], [bpoc], [[[1,1],[1,1],0]], yes, no, ['MTH','SNOMEDCT'], [0/5]),
    ev(-861, 'C1304680', 'Attack', 'Observation of attack', [attack], [fndg], [[[2,2],[1,1],0]], yes, no, ['MTH','SNOMEDCT'], [6/6]),
    ev(-827, 'C0004063', 'Attacked', 'Assault', [attacked], [inpo], [[[2,2],[1,1],1]], yes, no, ['ICD10AM'], [6/6]))
)
<Candidate><CandidateScore>1000</CandidateScore><CandidateCUI>C0027051</CandidateCUI><CandidateMatched>Heart attack</CandidateMatched><CandidatePreferred>Myocardial Infarction</CandidatePreferred><MatchedWords Count=2><MatchedWord>heart</MatchedWord><MatchedWord>attack</MatchedWord></MatchedWords><SemTypes Count=1><SemType>dsyn</SemType></SemTypes><MatchMaps Count=1><MatchMap><TextMatchStart>1</TextMatchStart><TextMatchEnd>12</TextMatchEnd><ConcMatchStart>1</ConcMatchStart><ConcMatchEnd>12</ConcMatchEnd><LexVariation>0</LexVariation></MatchMap></MatchMaps><IsOverMatch>no</IsOverMatch><IsHead>yes</IsHead><IsSource>yes</IsSource><Sources Count=24><Source>MEDLINEPLUS</Source></Sources><ConceptPIs Count=1><ConceptPI><StartPos>0</StartPos><Length>12</Length></ConceptPI></ConceptPIs></Candidate>
Output Formats: Formatted XML

<Candidate>
  <CandidateScore>-1000</CandidateScore>
  <CandidateCUI>C0027051</CandidateCUI>
  <CandidateMatched>Heart attack</CandidateMatched>
  <CandidatePreferred>Myocardial Infarction</CandidatePreferred>
  <MatchedWords
    Count=2><MatchedWord>heart</MatchedWord><MatchedWord>attack</MatchedWord></MatchedWords>
  <SemTypes Count=1><SemType>dsyn</SemType></SemTypes>
  <MatchMaps Count=1>
    <MatchMap>
      <TextMatchStart>1</TextMatchStart>
      <TextMatchEnd>2</TextMatchEnd>
      <ConcMatchStart>1</ConcMatchStart>
      <ConcMatchEnd>2</ConcMatchEnd>
      <LexVariation>0</LexVariation>
    </MatchMap>
  </MatchMaps>
  <IsHead>yes</IsHead>
  <IsOverMatch>no</IsOverMatch>
  <Sources Count=24><Source>MEDLINEPLUS</Source></Sources>
  <ConceptPIs Count=1><ConceptPI><StartPos>0</StartPos><Length>12</Length></ConceptPI></ConceptPIs>
</Candidate>
North Carolina women were surveyed to examine whether women’s disability status was associated with their risk of being assaulted within a year. Women’s violence experiences were classified into three groups: violence, physical assault only (without sexual assault), and assault (with or without physical assault). Multivariable analysis revealed that women with disabilities were not significantly more than women without disabilities to have experienced physical assault within the past year (odds ratio [OR] = 1.18, 95% confidence interval [CI] = 0.62 to 2.27); however, women with disabilities had more than the odds of experiencing sexual assault in the past year compared with women without disabilities (OR = 4.89, 95% CI = 2.21 to 10.83).

Department of Maternal and Child Health, University of North Carolina, Chapel Hill, NC, USA.
Tutorial Outline

• Background: why concept identification? (Lan)
• Introduction to MetaMap (Lan)
• The MetaMap algorithm (Lan)
• Input/output formats (François, Jim)
  ➢ **MetaMap options in depth (Lan)**
    • MetaMap processing modes (Lan)
    • Usage warnings (Lan)
• Access methods (Jim)
• Research projects using MetaMap (Jim, Dina)
• Future directions (Lan)
MetaMap Options

- Word Sense Disambiguation (WSD, \(-y\))
  - Based on Susanne Humphrey’s Journal Descriptor Indexing (*Humphrey et al., 1998, 2006*)
  - Provides modest improvement in results
- Negation (\(--\text{negex}\))
  - Important for clinical text
  - Based on Wendy Chapman’s NegEx algorithm (*Chapman et al., 2001*)
- Behavior options
- Output/Display options
WSD Examples (1/4)

• “Fifteen (6.4%) of 234 colds treated with placebo…”
  
  Cold (cold temperature) [npop]
  Cold (Common Cold) [dsyn]
  Cold (Cold Sensation) [phsf]
WSD Examples (2/4)

• “… the drugs were compared in two four-point, double-blind bioassays.”

  double (Diplopia) [dsyn] vs. Double (Duplicate) [ftcn]
  Blind (Blind Vision) [dsyn] vs. BLIND (Blinded) [resa] vs. Blind (Visually Impaired Persons) [podg]
  Bioassays (Biological Assay) [lbpr]
WSD Examples (3/4)

• “More neuroactive substances were prescribed for patients with superior mentation and minimal physical disability; the difference between low and high groups was 1.7 (mentation) and 2.8 (physical status).”

  High (Euphoric mood) [menp] vs. High [alco] WRONG
  Groups [inpr]
WSD Examples (4/4)

“The authors conclude that these cases of progressive hepatic disease with histologic changes simulating those found in livers of alcoholic patients offer evidence that heavy alcohol consumption may affect the liver in an indirect fashion.”

Liver (Entire liver) [bpoc] vs. Liver [bpoc] vs. LIVER (Liver Extract) [orch,phsu] WRONG
Negation Example

• “There is no focal infiltrate or pleural effusion.”

• --negex output (in addition to normal output):

NEGATIONS:
Negation Type: nega
Negation Trigger: no
Negation PosInfo: 9/2
Negated Concept: C0332448:Infiltrate
Concept PosInfo: 18/10

Negation Type: nega
Negation Trigger: no
Negation PosInfo: 9/2
Negated Concept: C2073625:pleural effusion, C0032227:Pleural Effusion
Concept PosInfo: 32/16
Behavior Options (1/4)

• Data model options
  -A --strict_model (the default; focused on concepts likely to be found in text)
  -C --relaxed_model (includes most Metathesaurus content)

• Major options highlighted earlier
  -y --word_sense_disambiguation
    --negex

• Other major options
  -Q --quick_composite_phrases (experimental, for well-behaved larger phrases: pain on the left side of the chest)
  -i --ignore_word_order
Example: Default (with word order)

Phrase: “Jurkat T cells”

Meta Candidates (8):
913 Jurkat Cells [Cell]
901 T-Cells (T-Lymphocyte) [Cell]
827 Cells [Cell]
793 Cell (Entire cell) [Cell]
793 Cell (Cell Device Component) [Medical Device]
793 Cell (Cell (compartment)) [Spatial Concept]
743 Cellular [Functional Concept]
721 Cellularity [Qualitative Concept]

Meta Mapping (913):
913 Jurkat Cells [Cell]
Example: Ignore Word Order (−i)

Phrase: “Jurkat T cells”

Meta Candidates (8):
882 T-Cells (T-Lymphocyte) [Cell]
858 Jurkat Cells [Cell]
790 Cells [Cell]
756 Cell (Entire cell) [Cell]
756 Cell (Cell Device Component) [Medical Device]
756 Cell (Cell (compartment)) [Spatial Concept]
706 Cellular [Functional Concept]
684 Cellularity [Qualitative Concept]

Meta Mapping (882):
882 T-Cells (T-Lymphocyte) [Cell]
Behavior Options (2/4)

• Browse mode options (example below)
  -z --term_processing
  -o --allow_overmatches
  -g --allow_concept_gaps
  -m --hide_mappings

• Inference mode options
  -Y --prefer_multiple_concepts

With --term_processing,
With --allow_overmatches
With --allow_concept_gaps, \textit{i.e.}
“mouse protein” $\rightarrow$
‘Ly6d protein, mouse’
among over 8,000 results
Behavior Options (3/4)

• Parsing/lexical options (not often used)
  -t --no_tagging
  -d --no_derivational_variants
  -D --all_derivational_variants
  -a --all_acros_abbrs
  -u --unique_acros_abbrs_only

• List truncation options (reduces tenuous matches and saves processing time)
  -r --threshold <integer>
Behavior Options (4/4)

• Source/ST limitation options
  -R --restrict_to_sources <list>
  -e --exclude_sources <list>
  -J --restrict_to_sts <list>
  -k --exclude_sts <list>
Output/Display Options (1/2)

• Human-readable output/display options
  -p --hide_plain_syntax
  -x --syntax
  -T --tagger_output
  -v --variants
  -c --hide_candidates
  -m --hide_mappings
  -I --show_cuis
  -s --hide_semantic_types
  -G --sources

Useful for explaining MetaMap’s behavior
Output/Display Options (2/2)

- Other output/display options (override human-readable options)
  - `--machine_output`
  - `--fielded_mmi_output` (mmi = MetaMap Indexing)
  - `--XML <none>` (format or noformat)
Tutorial Outline

• Background: why concept identification? (Lan)
• Introduction to MetaMap (Lan)
• The MetaMap algorithm (Lan)
• Input/output formats (François, Jim)
• MetaMap options in depth (Lan)

➢ MetaMap processing modes (Lan)
• Usage warnings (Lan)
• Access methods (Jim)
• Research projects using MetaMap (Jim, Dina)
• Future directions (Lan)
MetaMap Processing Modes

• **Semantic mode** (the *normal, default* mode)
  • Seeks *correct*, or *best*, answer
  • Uses MetaMap’s strict data model

• **Inference mode** (\(-Y, \ --prefer\_multiple\_concepts\))
  • Similar to semantic mode except
  • Prefers multiple concepts to facilitate inferencing

• **Browse mode** (\(-zogm, \ --term\_processing\)
  \(--allow\_overmatches \ --allow\_concept\_gaps\)
  \(--hide\_mappings\))
  • Seeks all answers, even tenuous ones
  • Often uses MetaMap’s relaxed data model (\(-C\))
  • Often includes \(-i, \ --ignore\_word\_order\)
Example: Semantic Mode (with WSD)

> metamap -y

-y, --word_sense_disambiguation

Phrase: "bladder cancer"
Meta Candidates (10):

1000 bladder cancer (Malignant neoplasm of urinary bladder) [Neoplastic Process]
1000 Bladder Cancer (Carcinoma of bladder) [Neoplastic Process]
861 Bladder (Urinary Bladder) [Body Part, Organ, or Organ Component]
861 Cancer (Malignant Neoplasms) [Neoplastic Process]
861 Cancer (Neoplasm) [Neoplastic Process]
861 Cancer (Cancer Genus) [Eukaryote]
861 Bladder (Entire bladder) [Body Part, Organ, or Organ Component]
861 Cancer (Primary malignant neoplasm) [Neoplastic Process]
861 Cancer (Cancer::Point in time:^Patient:-) [Clinical Attribute]
805 Vesical (Vesico-) [Spatial Concept]

Meta Mapping (1000):

1000 Bladder Cancer (Carcinoma of bladder) [Neoplastic Process]
Example: Inference Mode (with WSD)

metamap -yY
...
Phrase: "bladder cancer"
Meta Candidates (10):
  694 Bladder (Urinary Bladder) [Body Part, Organ, or Organ Component]
  694 Cancer (Malignant Neoplasms) [Neoplastic Process]
  694 Cancer (Neoplasm) [Neoplastic Process]
  694 Cancer (Cancer Genus) [Eukaryote]
  694 Bladder (Entire bladder) [Body Part, Organ, or Organ Component]
  694 Cancer (Primary malignant neoplasm) [Neoplastic Process]
  694 Cancer (Cancer:::Point in time:^Patient:-) [Clinical Attribute]
  666 bladder cancer (Malignant neoplasm of urinary bladder) [Neoplastic Process]
  666 Bladder Cancer (Carcinoma of bladder) [Neoplastic Process]
  638 Vesical (Vesico-) [Spatial Concept]
Meta Mapping (777):
  694 Bladder (Entire bladder) [Body Part, Organ, or Organ Component]
  694 Cancer (Malignant Neoplasms) [Neoplastic Process]
Example 1/2: Browse Mode

> metamap -zogm

... 
Phrase: "bladder cancer"
Meta Candidates (5597):

1000 bladder cancer (Malignant neoplasm of urinary bladder) [Neoplastic Process]
   Cancer of Bladder
   Malignant Bladder Neoplasm

1000 Bladder Cancer (Carcinoma of bladder) [Neoplastic Process]
   BLADDER CARCINOMA
   Cancer of Bladder

861 Bladder (Urinary Bladder) [Body Part, Organ, or Organ Component]
861 Cancer (Malignant Neoplasms) [Neoplastic Process]

...

583 gall bladder (Gallbladder) [Body Part, Organ, or Organ Component]
583 Cancer Hospital (Hospitals, Cancer) [Health Care Related Organization, Manufactured Object]

...
Example 2/2: Browse Mode

> metamap -zogm

... 
Phrase: "achilles reflex"
Meta Candidates (815):

861 Achilles (Structure of achilles tendon) [Body Part, Organ, or Organ Component]
861 Reflex (Reflex action) [Organ or Tissue Function]
861 reflex (Reflex motion descriptor) [Organ or Tissue Function]
861 Reflex (Observation of reflex) [Finding]
827 Achilles tendon reflex (Ankle reflex) [Clinical Attribute]
  Ankle reflex
827 reflexes (Examination of reflexes) [Diagnostic Procedure]
722 examination of Achilles reflex [Diagnostic Procedure]
  ankle reflex exam
679 intensity of left Achilles tendon reflex [Finding]
  left ankle jerk reflex
679 intensity of right Achilles tendon reflex [Finding]
  right ankle jerk reflex
583 Gag reflex (Gagging) [Finding]

...
Tutorial Outline

• Background: why concept identification? (Lan)
• Introduction to MetaMap (Lan)
• The MetaMap algorithm (Lan)
• Input/output formats (François, Jim)
• MetaMap options in depth (Lan)
• MetaMap processing modes (Lan)

➤ Usage warnings (Lan)
• Access methods (Jim)
• Research projects using MetaMap (Jim, Dina)
• Future directions (Lan)
Candidates vs. Mappings

- The mappings are MetaMap’s final answer to text input
- The candidate list is an intermediate result
  - Often contains many bad matches among the good ones (similar to Browse mode results vs. Semantic mode results)
  - Should be used judiciously/selectively and only when mappings are found to be inadequate
Candidates vs. Mappings Example

Phrase: "heart attack"

Meta Candidates (8):

1000 Heart attack (Myocardial Infarction) [Disease or Syndrome]
861 Heart [Body Part, Organ, or Organ Component]
861 Attack, NOS (Onset of illness) [Finding]
861 Attack (Attack device) [Medical Device]
861 attack (Attack behavior) [Social Behavior]
861 Heart (Entire heart) [Body Part, Organ, or Organ Component]
861 Attack (Observation of attack) [Finding]
827 Attacked (Assault) [Injury or Poisoning]

Meta Mapping (1000):

1000 Heart attack (Myocardial Infarction) [Disease or Syndrome]
Tutorial Outline

- Background: why concept identification? (Lan)
- Introduction to MetaMap (Lan)
- The MetaMap algorithm (Lan)
- Input/output formats (François, Jim)
- MetaMap options in depth (Lan)
- MetaMap processing modes (Lan)
- Usage warnings (Lan)

- Access methods (Jim)
  - Research projects using MetaMap (Jim, Dina)
  - Future directions (Lan)
MetaMap Availability

- **Web access** *(start here)*
  - Interactive and batch (file) processing via Scheduler

- **MMTx (MetaMap Transfer) ***(becoming obsolete)*
  - Java-based implementation of MetaMap
  - MetaMap vs. MMTx

- **MetaMap itself**
  - Initial release (for Linux): September, 2008

*All usage requires UMLS license agreement*
MetaMap APIs

- Java MetaMap API

- Java API to SKR Scheduler

- MetaMap UIMA Annotator
About MetaMap

MetaMap is a highly configurable program developed by Dr. Alan (Liam) Aronson at the National Library of Medicine (NLM) to map biomedical text to the UMLS Metathesaurus or, equivalently, to discover Metathesaurus concepts referred to in text. MetaMap uses a knowledge-intensive approach based on symbolic, natural language processing (NLP) and computational linguistic techniques. Besides being applied for both IR and data mining applications, MetaMap is one of the foundations of NLM's Medical Text Indexer (MTI) which is being applied to both semiautomatic and fully automatic indexing of biomedical literature at NLM. For more information on MetaMap and related research, see the the SKR Research Information Site (http://skr.nlm.nih.gov/papers/index.shtml)

What's New in MetaMap?

October 2010 - MetaMap 2010 Release

With the 2010 Release of MetaMap, we are retiring three previous versions of MetaMap, namely MetaMap07, MetaMap08, and MetaMap08V2. Only the MetaMap binary executables are being retired; the MetaMap UMLS datasets corresponding to these releases (2007AA, 2008AA, and 2008AB) will remain available.

MetaMap 2010 includes less new functionality than previous releases because the bulk of our development efforts since MetaMap08V2 have focused on converting MetaMap from Quintus Prolog to SICStus Prolog, which will henceforth be the principal implementation vehicle of MetaMap. We also converted MetaMap 2010 to version 4.8.24 of Berkeley DB, as recommended by SICStus.

New functionality and enhancements delivered in MetaMap 2010 include the following:

- De-Normalized Data Tables,
- Minimum Concept Length,
- Changes in Numerical Output Format,
- Silent Mode, and
- Variants Bug Fix

Please see the Release Notes for more information.

April 2010 - UIMA Annotator

The MetaMap UIMA Annotator encodes MetaMap named entries in a format usable by UIMA components. The annotator is based on the MetaMap UIMA Wrapper (http://sourceforge.net/projects/metamap-uima/) authored by Kai Schlap. The annotator utilizes classes from the 2009v2 version of the MetaMap Java API which is required for use of the annotator.

Differences from Kai Schlap's MetaMap UIMA Wrapper include a modified UIMA type system which includes Acronyms and Abbreviations and the addition of a MatchMap structure to the Candidates and Mappings. Use of MetaMap XML and the SKR API are currently not supported.
## MetaMap/MMTx Distribution Modes

### Avenues to MetaMap:

<table>
<thead>
<tr>
<th>Access</th>
<th>Description</th>
<th>Go To</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Web Access</strong></td>
<td>Our Semantic Knowledge Representation (SKR) website provides both Interactive and Batch facilities that allow users to send text to our internal machines and run various programs including the MetaMap program. The Interactive facility is designed for testing options and running small amounts of text. The Batch facility runs large amounts of text through our Scheduler program which distributes the workload over a large pool of clients.</td>
<td>GO TO SKR</td>
</tr>
<tr>
<td><strong>MetaMap</strong></td>
<td>Distributable version of the original Prolog MetaMap program. Currently only includes binary distribution for Solaris and Linux platforms.</td>
<td>GO TO MetaMap</td>
</tr>
<tr>
<td><strong>SKR API</strong></td>
<td>Java-based API to the SKR Scheduler facility was created to provide users with the ability to programmatically submit jobs to the Scheduler Interactive and Batch facilities instead of using the web-based interfaces. We have tried to reproduce full functionality for all of the programs under the SKR Scheduler umbrella. The SKR API has been tested on the Solaris, Linux, and Windows XP platforms.</td>
<td>GO TO SKR API</td>
</tr>
</tbody>
</table>

**NOTE:** MMTx is no longer supported except for major bug fixes. We recommend all users switch to the downloadable MetaMap (described above) if possible.

| MMTx | MetaMap Transfer (MMTx) is a Java-based distributable version of the MetaMap program. Includes binary and source distributions and is supported on Solaris, Linux, Windows, and Mac platforms. MMTx was an early attempt at providing a distributable version of MetaMap and is currently being phased out in favor of the original Prolog version of MetaMap. There are two reasons for the phase out of MMTx: 1) The original Prolog version of MetaMap is much faster, especially now with the new speed enhancements (V2). 2) We were never able to make the results the same between MMTx and MetaMap - there was always about a 20% difference in the overall results MMTx would produce. | GO TO MMTx |


U.S. National Library of Medicine
Tutorial Outline

• Background: why concept identification? (Lan)
• Introduction to MetaMap (Lan)
• The MetaMap algorithm (Lan)
• Input/output formats (François, Jim)
• MetaMap options in depth (Lan)
• MetaMap processing modes (Lan)
• Usage warnings (Lan)
• Access methods (Jim)

➢ Research projects using MetaMap (Jim, Dina)
• Future directions (Lan)
NLM Applications using MetaMap: MTI (Medical Text Indexer)

- Product of Indexing Initiative
- Assists NLM Indexers
- Production since mid-2000
- Uses article Title and Abstract
- Semi-Automatic MeSH Indexing Recommendations
- Automatic Keyword Indexing
MTI Uses

• Assisted indexing of MEDLINE/PubMed articles (DCMS)
  – Citations processed nightly

• Assisted indexing of Cataloging (TSD) and History of Medicine Division (HMD) records
  – Production mid-2007
  – Easy modification to MTI to accommodate differences
  – Tightly integrated into their workflow

• Automatic indexing of NLM Gateway meeting abstracts
How MTI Uses MetaMap

- Dominate input pathway (weighted 7 – 2 over PRC)
- Calls MetaMap twice
- Pass I – cast a wide net to identify as many UMLS concepts as possible, while balancing time constraints
  - `metamap –iDN`
  - Ignore Word Order (-i)
  - All Derivational Variants (-D)
  - Fielded MMI Output (-N)
How MTI Uses MetaMap (contd.)

• Pass II – more focused review for actual MeSH Headings in the text, reinforces Pass I items
  • `metamap -dN -R 'MSH'`
  • No Derivational Variants (-d)
  • Fielded MMI Output (-N)
  • Restrict to Sources (-R) restricted to MeSH
**MetaMap Fielded MMI Output (-N)**

<table>
<thead>
<tr>
<th>PMID</th>
<th>Path Name</th>
<th>Score</th>
<th>UMLS Concept Found Preferred Name</th>
<th>List of Semantic Types</th>
<th>List of Entry Term Quartets</th>
<th>Location(s), boost scores for TI</th>
</tr>
</thead>
<tbody>
<tr>
<td>17285228</td>
<td>Homocystine</td>
<td>430.78</td>
<td>Homocystine</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Homocysteine</td>
<td>&quot;Homocysteine&quot;-ab-3-&quot;Homocysteine&quot;, &quot;Homocystine&quot;-ab-2-&quot;homocysteine&quot;, &quot;Homocystine&quot;-ab-1-&quot;Homocysteine&quot;, &quot;Homocystine&quot;-ti-1-&quot;homocysteine&quot;</td>
<td>TI;AB:406:12</td>
<td>227:12</td>
</tr>
</tbody>
</table>

- **Entry Term Quartet**: "Homocystine"-ab-2-"homocysteine"
- "Homocystine" – UMLS Concept Preferred Name
  - **ab** – Found in Abstract
  - 2 – Found in section’s second utterance
- "homocysteine" – Actual text used for mapping

- **PMID**: 17285228
- **Path Name**: Homocystine
- **Score**: 430.78
- **UMLS Concept Unique Identifier**: C0019879
- **List of Semantic Types**: [aapp,bacs]
- **List of Entry Term Quartets**: 
  - "Homocystine"-ab-3-"Homocysteine", "Homocystine"-ab-2-"homocysteine", "Homocystine"-ab-1-"Homocysteine", "Homocystine"-ti-1-"homocysteine"
- **Location(s), boost scores for TI**: 406:12|227:12|74:12|35:12

U. S. NATIONAL LIBRARY OF MEDICINE
Filtering Using Entry Term Information

• Homocystine vs Homocysteine

0000000|MM|424.55|Homocystine|C0019879|[aapp,bacs]
 |["Homocystine"-ti-1-"Homocystine"]|TI|20:11
0000000|MM|424.55|homocysteine|C0019878|[aapp,bacs]
 |["homocysteine"-ti-1-"Homocystine"]|TI|20:11

• Borne → Bear

0000000|MM|734.96|Ursidae Family|C0004897|[mamm]
 |["Bear"-ti-1-"borne"]|TI|32:5
0000000|MM|112.74|Bearing Device Component|C1704689|[medd]
 |["Bearing"-ti-1-"borne"]|TI|32:5
0000000|MM|112.74|Caliber|C1301886|[qnco]
 |["Bore"-ti-1-"borne"]|TI|32:5
Example

Text: preventing hemorrhage pre-operatively

- No Derivational Variants (-d)
  812 HAEMORRHAGE NOS (Hemorrhage) [Finding]

- All Derivational Variants + Ignore Word Order (-iD)
  783 Operative haemorrhage (Blood Loss, Surgical) [Pathologic Function]
Ambiguity Example

Text: respirable particulate matter

- No Derivational Variants (-d)
  Nothing for “respirable”

- All Derivational Variants + Ignore Word Order (-iD)
  523 respiration (Cell Respiration) [Cell Function]
  523 Respirator (Mechanical Ventilator) [Medical Device]
  523 Respiration [Physiologic Function]
  523 respirator (Treatment with respirator) [Therapeutic or Preventive Procedure]
  523 respiration (respiratory gaseous exchange in organisms) [Biologic Function]

Examples from metamap10 with 2010AA UMLS
NLM Applications using MetaMap: RIDeM
(Repository for Informed Decision Making)

- Summarization
- Meta-analysis, Reviews
- Drugs (indications, interactions, etc.)
- Clinical question answering (CQA 1.0)
- Information for patients
- Linking evidence to patient records (InfoBot)
- Image retrieval (iMEDLINE)
- Annotation of Interactive Publications
- Clinical research (HDiscovery)
- Translational research: linking of basic research to clinical information
Example data flow: InfoBot

1. Local modules automatically extract selected patient data
2. Transmit to InfoBot
3. Format and send requested data
4. Requested data
5. Local modules automatically display requested data
<table>
<thead>
<tr>
<th>Patient Name</th>
<th>Assigned Location</th>
<th>Visit Type</th>
<th>Visit Status</th>
<th>Temporary Location</th>
<th>Provider</th>
<th>Check Orde...</th>
<th>Flag New</th>
<th>Unack Alerts</th>
<th>New Result</th>
<th>New Ord...</th>
<th>New Do...</th>
<th>To Si...</th>
<th>Admitting Protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEST, PATIENT MAR</td>
<td>CADRE-CC</td>
<td>Outpatient</td>
<td>DSC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>01-AR-0227</td>
</tr>
<tr>
<td>TEST, PATIENT SEVEN</td>
<td>CADRE-CC</td>
<td>Exp Inf...</td>
<td>DSC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>00-CC-0096</td>
</tr>
<tr>
<td>TEST, PATIENT SIX</td>
<td>CADRE-CC</td>
<td>Inpatient</td>
<td>DSC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>00-CC-0096</td>
</tr>
<tr>
<td>CONIHTTEST, PATIENT BL MIS</td>
<td>CC-CADRE</td>
<td>Outpatient</td>
<td>DSC</td>
<td></td>
<td>Luxenberg, Ste...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>00-I-0099</td>
</tr>
<tr>
<td>NIHTEST, PATIENT A</td>
<td>CC-CADRE</td>
<td>Inpatient</td>
<td>DSC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>01-CC-0135</td>
</tr>
<tr>
<td>NIHTESTCC, PATIENT BPM MIS</td>
<td>CC-CADRE</td>
<td>Inpatient</td>
<td>DSC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>00-CH-0093</td>
</tr>
<tr>
<td>SILVEY-WILKES, DELORES MERV</td>
<td>CRCOH-FSWS</td>
<td>Outpatient</td>
<td>DSC</td>
<td></td>
<td>Nussenblatt, R...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>00-EI-0204</td>
</tr>
<tr>
<td>TESTER, BILLY JOE JR</td>
<td>OP-9-CC</td>
<td>Outpatient</td>
<td>ADM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>01-N-0139</td>
</tr>
</tbody>
</table>
**Chief Complaint:** SEIZURES

**CRIS Data**

- **CHIEF COMPLAINT:** NONE
- Diabetes. Pt awake, alert, following commands well. MAEs. Strong grasps bilaterally. Non verbal communication. Nodding head to yes/no questions. No sedation used through the night. Medicated for pain X1 with morphine. Remains restrained until extubation this AM. Pt remains orally intubated.

**Protocol(s)**

- 01-M-0152

**Medications**

- aspirin 325 mg
- diltiazem
- esomeprazole magnesium
- fluoxetine
- olanzapine

**Take Home Medications**

**Evidence Based Practice Search Results**

- Diabetes OR morphine OR extubation OR extubate OR CPAP OR Neosynephrin

- □ check spelling □ has abstract □ English □ Human □ Subsets □ Preference

- Modify the above query and [Search]

- Severe hypoglycaemia and glycaemic control in Type 1 diabetes: meta-analysis of multiple daily insulin injections compared with continuous subcutaneous insulin infusion.

- Non-invasive positive pressure ventilation (CPAP or bilevel NPPV) for cardiogenic pulmonary edema.
### Text types processed for RIDeM

<table>
<thead>
<tr>
<th>Text types</th>
<th>Format</th>
<th>Encoding</th>
<th>Length</th>
<th>Idiosyncrasies</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEDLINE abstracts</td>
<td>XML</td>
<td>UTF-8</td>
<td>~300 words</td>
<td>structured /not, well-formed</td>
</tr>
<tr>
<td>Clinical questions</td>
<td>unknown</td>
<td>unknown</td>
<td>~10</td>
<td>Structured /not</td>
</tr>
<tr>
<td>Clinical notes</td>
<td>“ASCII art”</td>
<td>unknown</td>
<td>20--600</td>
<td>ungrammatical</td>
</tr>
<tr>
<td>Case descriptions</td>
<td>unknown</td>
<td>unknown</td>
<td>~700</td>
<td>well-formed/not</td>
</tr>
<tr>
<td>Full-text journal articles</td>
<td>HTML, PDF, XML…</td>
<td>UTF-8</td>
<td>~3,000</td>
<td>structured/not, well-formed</td>
</tr>
<tr>
<td>Interactive Publications</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DailyMed drug package inserts</td>
<td>XML+SPL</td>
<td>UTF-8</td>
<td>~800</td>
<td>structured, well-formed</td>
</tr>
</tbody>
</table>
Text pre-processing

- Required (batch file submission and API)
  - ASCII encoding
  - Size under 2,000 - 3,000 characters
  - Format (MEDLINE, free text, single line, etc.)
- Sensible
  - Isolate sections/passages for entity extraction
    - Indication section of a drug label, image description in the article
  - Remove mark-up tags
  - Expand colloquial abbreviations
    - In the NIH CC notes MAN=Multiple Endocrine Neoplasia
- Optional
  - Split HTML lists into items
  - Use a third party parser to extract phrases
MetaMap settings

• Access mode
  • Batch file submission
    • DailyMed package inserts processing
    • Full-text processing for image text indexing
  • API
    • Note and question processing (back-off to local table look-up)
• Options
  • Default for experiments
  • Subset semantic types to disorders, interventions, anatomy
• Output format
  • Machine (trade-off between ease of processing and volume)
MetaMap output processing

- Extract into RIDeM concept container:
  - Lexical match
  - Concept unique identifier
  - Concept preferred name
  - Semantic types list
  - Negation status
  - Phrase type
  - POS list
  - First character offset
  - Length
Target entity extraction: InfoBot clinical notes

- UMLS-based recognition of the elements of a well-formed clinical question (Patient/Problem-Intervention)
- Example text
  1st Research Participation Problem: Hodgkin's Disease; Lymphoma, Post MRD HSCT; HLA-matched related donor
  Allogeneic Hematopoietic Stem Cell Transplantation
  Dual Energy X-ray Absorptiometry
  CT scan/PET/DEXA scans
  Licensed independent practitioner

Goal/s: Pt will verbalize understanding of role in research participation/protocol and about the disease process.

Planned Interventions: Obtain/Monitor Labs per protocol. Review results of tests such as CT scan/PET/DEXA scans. Have LIP or MD discuss protocol with pt and keep patient up-to-date on progress. Pt to keep appointments with follow-up clinics and tests. Provide pt with reference materials, internet access, information about medications and education materials. Pt to follow plan of care with LIP and ask questions to address concerns.
1st Research Participation Problem: Hodgkin's Disease; Lymphoma, Post MRD HSCT;
Goal/s: Pt will verbalize understanding of role in research participation/protocol and about the disease process.
Planned Interventions: Obtain/Monitor Labs per protocol. Review results of tests such as CT scan/PET/DEXA scans. Have LIP or MD discuss protocol with pt and keep patient up-to-date on progress. Pt to keep appointments with follow-up clinics and tests. Provide pt with reference materials, internet access, information about medications and education materials. Pt to follow plan of care by LIP and ask questions to address concerns.
correct sense; correct sense, but high level; wrong sense (FP and FN); ignored
Entity extraction

- Discard lexical matches under four characters (unless MetaMap identified an author-defined abbreviation – scientific publications, or the abbreviation is on the clinical institution local list)
- Mark high-level terms (using a local look-up list)
- Discard stop-words (using a local look-up list)
- Apply document-specific extraction methods
  - MEDLINE abstracts – position in the document, frequency, co-occurrence, classifiers
72 yo male admitted from Location Hosp s-p seizure activity. Pt has pmh of hodgkins dx, s-p chemo, and HTN. Pt had gone to hospital on Date for confusion and unsteady gait, head CT was negative and pt sent home on ASA. Returned on Date with focal seizure and then grand mal and was admitted to Location Hospital ICU. Started on Ativan gtt. MRI done showing diffuse lesions consistent with encephalitis. Head CT with ? of embolic stroke. Pt continued with seizures and transferred to Location for further workup.
Allogeneic hematopoietic stem-cell transplantation in patients with hematologic malignancies after dose-escalated treosulfan/fludarabine conditioning.

PURPOSE: Treosulfan was introduced recently as a conditioning agent for allogeneic blood stem-cell transplantation. The favorable nonhematologic toxicity profile at 3 x 10 g/m(2) was the basis for dose escalation in this prospective, multicenter trial.

PATIENTS AND METHODS: Fifty-six patients with various hematologic malignancies who were not eligible for standard conditioning were treated with one of three doses: 10 g/m(2), 12 g/m(2), or 14 g/m(2) of intravenous treosulfan, which was administered on days -6 to -4 combined with fludarabine 30 mg/m(2) on days -6 to -2. Patients in complete remission (CR; 42%) or non-CR (58%) received grafts from matched related (47%) or matched unrelated (51%) donors; one patient had a mismatched related donor (2%).

RESULTS: No engraftment failure occurred. Overall, extramedullary toxicity and the nonrelapse mortality rate at 2 years (20%) were low and did not increase with dose. Cumulative incidence of relapse/progression reached 31%. The overall survival and progression-free survival rates were 64% and 49%, respectively, in the total study population. An inverse dose dependency of relapse incidence was indicated in the subgroup receiving transplantations from matched related donors (P = .0568).

CONCLUSION: Treosulfan-based conditioning was feasible at all three doses. The 3 x 14 g/m(2) dose was selected for additional studies, because it combines desired characteristics of low toxicity and a low relapse rate.
MetaMap UMLS clinical content views

- **Experiments**
  - 2008 Extract PICO frames from MEDLINE abstracts
  - 2009 Extract PICO elements from clinical notes
- **Data**
  - 2008 LNCV document collection
  - 2009 LNCV clinical text collection
- **Conclusions**
  - Different tasks need different MetaMap views
    - Filters
    - Semantic type subsets

U. S. NATIONAL LIBRARY OF MEDICINE
Tutorial Outline

- Background: why concept identification? (Lan)
- Introduction to MetaMap (Lan)
- The MetaMap algorithm (Lan)
- Input/output formats (François, Jim)
- MetaMap options in depth (Lan)
- MetaMap processing modes (Lan)
- Usage warnings (Lan)
- Access methods (Jim)
- Research projects using MetaMap (Jim, Dina)
  ➢ Future directions (Lan)
Ongoing MetaMap Development (1/2)

- Technical algorithm enhancements resulting in at least 3x speedup in MetaMap execution
  - MetaMap is now up to 10 times faster than MMTx
- Environment development
  - Migration from Sun/Solaris to Linux environment
  - Update to current Berkeley DB to coordinate with migration from Quintus to SICStus Prolog
  - Release of MetaMap for OS X and soon for Windows
- MetaMap 3D (colorized MetaMap output)
- Detection of user-defined acronyms
Ongoing MetaMap Development (2/2)

• Higher-order tokenization
  • Detection of author-defined acronyms/abbreviations
  • To be augmented with recognition of chemical names, bibliographic references, numeric quantities, etc.

• Negation detection (negex)

• Word sense disambiguation (WSD, -Y)
  • Currently based on Journal Descriptor Indexing (JDI)
  • To be augmented and combined with other, Machine Learning approaches (Antonio Jimeno-Yepes)
Future MetaMap Development (1/2)

- Release MetaMap for Windows
- Further develop API services
- Enhance MetaMap’s accuracy with additional WSD algorithms
- Augment tokenization with recognition of chemical names, bibliographic references, numeric quantities, etc.
- Complete composite phrase processing (e.g., pain on the left side of the chest)
Future MetaMap Development (2/2)

- Enhance of processing short words, including acronyms/abbreviations
- Handle space/hyphen/null alternation (e.g., breast feed/breast-feed/breastfeed)
- Tune MetaMap’s evaluation metric to improve accuracy
- Technical enhancements in graceful back-off for long phrases
- Incorporate user-suggested improvements
Web Pointers*

- NLM Indexing Initiative: http://ii.nlm.nih.gov/ (general II or MTI information)

* All MetaMap access requires a UMLS license: http://www.nlm.nih.gov/research/umls/license.html
Tutorial Faculty Pointers

- Alan (Lan) R. Aronson: alan@nlm.nih.gov
- Dina Demner-Fushman: ddemner@mail.nih.gov
- François-Michel Lang: flang@mail.nih.gov
- James G. Mork: mork@nlm.nih.gov
Comments or Questions?