Leveraging Terminological Resources for Mapping between Rare Disease Information Sources

Bastien Rance, Michelle Snyder, Janine Lewis, Olivier Bodenreider

MedInfo 2013
Copenhagen, Denmark
Objectives

• **Rare disease information sources are incompletely and inconsistently cross-referenced to one another**
  – Difficult for information seekers to navigate across them
  – Development of such cross-references established manually by experts is generally labor intensive and costly

• **Objectives: To develop an automatic mapping between two rare diseases information sources by leveraging terminological resources**
  – Genetic and Rare Diseases Information Center (GARD)
  – Orphanet
# Sources of rare disease terms

<table>
<thead>
<tr>
<th>Source</th>
<th>Genetic and Rare Diseases Information Center (GARD)</th>
<th>Orphanet</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Funding</strong></td>
<td>U.S. National Institutes of Health</td>
<td>Based in Europe (UE, France)</td>
</tr>
<tr>
<td><strong>Number of concepts</strong></td>
<td>6,316 preferred terms and 12,627 synonyms at the time of the study</td>
<td>6,578 preferred terms and 7,552 synonyms at the time of the study</td>
</tr>
<tr>
<td><strong>Cross-references</strong></td>
<td>Cross-references to OMIM</td>
<td>Cross-references to OMIM</td>
</tr>
<tr>
<td><strong>Misc.</strong></td>
<td>extensive information about 1,100 diseases</td>
<td>Cross-references to various reference terminologies including ICD-10-CM, MeSH, SNOMED-CT, MedDRA, and the UMLS</td>
</tr>
</tbody>
</table>
Rare disease terms in the UMLS®

• The Unified Medical Language System®
  – Terminology integration system
    • Over 160 medical terminologies
  – Concept oriented representation
    • One concept groups synonyms from various source terminologies

• Contains rare diseases terms from OMIM, MeSH and other terminologies (SNOMED CT, ICD, etc.)
Methods – overview

Mapping through UMLS

UMLS concept

Consistency between mapping through UMLS and cross-references to OMIM?

Rare Disease Term from GARD

OMIM concept

Cross-reference to OMIM

Rare Disease Term from Orphanet
Mapping between GARD and Orphanet through UMLS

Direct mapping

Indirect mapping (through subsumption)
Mapping to UMLS

Rare Disease Term

Pre-processing (remove acronyms)

- Exact match
- Normalized match
- Extended normalization

Equivalent match to the UMLS

Broader / narrower match to the UMLS

Semantic Group filtering

UMLS concept
Evaluation – True positives

- Automatic mapping from GARD and Orphanet terms to the **same UMLS concept** (or to two hierarchically related UMLS concepts)
- Cross-reference from GARD and Orphanet terms to the **same OMIM concept**

**True positive (direct)**

**True positive (indirect)**
Evaluation – False positives

• Automatic mapping from GARD and Orphanet terms to the **same UMLS concept**

• **Inconsistent cross-reference to OMIM** from GARD and Orphanet terms
  • Only one term is cross-referenced to OMIM
  • No cross-reference to OMIM for either term
  • Cross-reference to different OMIM concepts
Evaluation – False negatives

- Cross-reference from GARD and Orphanet terms to the same OMIM concept
- Inconsistent automatic mapping to UMLS from GARD and Orphanet terms
  - Only one term is mapped to UMLS
  - No mapping to UMLS for either term
  - Mapping to different UMLS concepts
Evaluation – True negatives

- **No automatic mapping to UMLS** from GARD and Orphanet terms
- **No cross-reference to OMIM** from GARD and Orphanet terms
Results (1) – Performance of the automatic mapping to the UMLS

<table>
<thead>
<tr>
<th></th>
<th>GARD</th>
<th>Orphanet</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>6,316</td>
<td>6,578</td>
</tr>
<tr>
<td><strong>Direct mappings</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exact match</td>
<td>4744</td>
<td>3163</td>
</tr>
<tr>
<td>UMLS Norm.</td>
<td>397</td>
<td>826</td>
</tr>
<tr>
<td>Extended Norm.</td>
<td>22</td>
<td>99</td>
</tr>
<tr>
<td><strong>sub-total</strong></td>
<td>5,163 (81%)</td>
<td>4,088 (62%)</td>
</tr>
<tr>
<td><strong>Indirect mappings</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All norm.</td>
<td>198</td>
<td>363</td>
</tr>
<tr>
<td><strong>sub-total</strong></td>
<td>5,361 (85%)</td>
<td>4,451 (68%)</td>
</tr>
</tbody>
</table>
Results (2) – Direct mapping only

<table>
<thead>
<tr>
<th>Direct only</th>
<th>Mapping through OMIM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Mapping through the UMLS</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>No</td>
</tr>
</tbody>
</table>

recall: 61.94%, precision: 89.94%, $F_1$: 73.36%
(Sharp decrease in precision with indirect mappings)
Failure analysis

• We manually reviewed the false positives
  – 207 direct mappings obtained through the UMLS but not corroborated by a mapping through OMIM
  – 50 were classified as correct
  – 54 were classified as acceptable

– Corrected performance of our method:
  • recall: 63.05%, precision: 94.24%, $F_1$: 75.55%
  Compared to “direct only”
  recall: 61.94%, precision: 89.94%, $F_1$: 73.36%
Summary

- Automatic method for mapping rare diseases terminologies
  - Leverages the UMLS
  - Performance
    - Insufficient for completely automatic mapping
    - Can effectively support manual mapping and enrich cross-references
  - Gold standard through OMIM is imperfect
Medical Ontology Research

Contact: olivier@nlm.nih.gov
Web: mor.nlm.nih.gov

Olivier Bodenreider
Lister Hill National Center for Biomedical Communications
Bethesda, Maryland - USA
Methods (2) – Pre-processing

Removing acronyms

• Due to their high ambiguity acronyms are removed from the data sets
  – Example: BBS is excluded from our processing. In addition to *Bardet-Biedl syndrome*, it would also be mapped (incorrectly) to *Berlin Breakage Syndrome*
Methods (3) – Semantic filtering

• Restrict concepts mapped to through the UMLS to those of **Semantic Group Disorders only** (including such **semantic types as Disease or Syndrome and Congenital Abnormality**).

• Provides some level of word sense disambiguation

• **Example**: NF2 can be mapped to both a **disease** (*neurofibromatosis type 2*) or to a **gene** (*NF2*, on chromosome 22, whose mutation causes *neurofibromatosis type 2*).
Methods (6) – Extended Normalization

- eXtended Normalization (XN) – domain specific

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Extended normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transforming Roman numerals</td>
<td>ixc, iii</td>
<td>9c, 3</td>
</tr>
<tr>
<td>intro Arabic numerals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extended stop word list</td>
<td>Familial restrictive cardiomyopathy type 2</td>
<td>Familial restrictive cardiomyopathy 2</td>
</tr>
<tr>
<td>Karyotype normalization</td>
<td>48, XXXY</td>
<td>XXXY</td>
</tr>
</tbody>
</table>

Example:
Familial restrictive cardiomyopathy type 2 [ORPHANET] → CARDIOMYOPATHY, FAMILIAL RESTRICTIVE, 2 [OMIM]
Methods (8) – Subsumption relations (Broader / Narrower)

• Example:

  Ehlers-Danlos syndrome, classic type (Orphanet)
  maps:
  – Ehlers-Danlos Syndrome, **Severe** Classic Type (a synonym for Ehlers-Danlos syndrome type 1 [C0268335]) and
  – Ehlers-Danlos syndrome, **mild** classic type (a synonym for Ehlers-Danlos syndrome type 2 [C0268336])