Can Wikipedia be used to derive an open Clinical Terminology?

Hoa Ngo, PhD

*Health Data Interoperability @ Australian E-Health Research Centre, CSIRO*
Healthcare Interoperability: an example

A 75 year old female
Chest pain & difficult breathing

Send EHR

Request patient’s EHR

Family Clinic

INTERNET

Hospital Emergency
Healthcare Interoperability: an example

Age: 75; Gender: F
Coding: xxx
History: xxx congestive heart failure xxx
xxx occasional MI xxx
Diagnosis: xxx severe pain in back xxx
xxx pain in shoulder xxx
Prescription: Ibuprofen 400 Mg

Knowledge Database

Pain

Back pain

Heart disease

reduce

strengthen

Chest pain

cause

Shortness Of Breathing

Congestive heart failure

subclass

 subclass

NSAID

instance

MI

ibuprofen
Clinical terminology: What is it?

**Parents**
- Disorder of cardiac ventricle (disorder)
- Heart failure (disorder)

**Children**
- Benign hypertensive heart disease with congestive cardiac failure (disorder)
- Biventricular congestive heart failure (disorder)
- Chronic congestive heart failure (disorder)
- Congestive heart failure as early postoperative complication (disorder)
Clinical terminology: Whom & Why?

Interoperability

- Physicians
- Health Plans
- Labs
- Healthcare Associations
- Hospitals
- Registries
- Practices
Development of a clinical terminology is HARD due to various practical issues.
Ms. A, 75 years old. She has history of congestive heart disease and occasional MI. She has severe pain in back and shoulder. She is prescribed Ibuprofen 400mg for pain relief.
Lack of terms & Lack of explanation

Age: 75; Gender: F
History: I50.9 (congestive heart failure) I21.A9 (occasional MI)
Diagnosis: M54.5 (severe pain in back) M25.519 (pain in shoulder)
Prescription: Ibuprofen 400 Mg
Maintenance & update issues

- Dynamic contents
- Slowly periodic update

SNOMED CT: every 6 months; OPCS: every 4 years; ICD: every 10 years
Open clinical terminology: Research questions

distributed collaborative authoring

standardization

Concept

Definition
A one-piece, handheld phone that includes battery power and may be used without any peripheral power or antenna. (Nokia)

Object

term is verbal designation

Designations (terms)
Handy (DE)
cellular phone, cell phone (US)
mobile (UK) (two variants)

Term comprehension
Open clinical terminology: Research questions

Hypothesis:

Wikipedia can be used to derive Open Clinical Terminology

Why?

✓ Wikipedia platform supports:
  ▪ Distributed collaborative authoring
  ▪ Term standardization
  ▪ Fully term definition

✓ Semantic Media Wiki extension\cite{1} supports Knowledge Graph

Question:

How well does Wikipedia cover existing clinical terminology like SNOMED CT?

\[1\] https://en.wikipedia.org/wiki/Semantic_MediaWiki
Statistical Experiments

Design and Evaluation
Experimental design

Resources:

- 12,000 clinical SNOMED CT concepts extracted from real patients’ EHR
- Fully indexed Wikipedia from dump archive – version 2016
- 02 clinical terminology experts

Specification:

04 types of matching relationships

- Exact Match
- More Specific
- More General
- Not Found
Experimental design

Inter-rater reliability:

Measure by *Krippendorff's alpha*: agreement $= \frac{50+37+1+2}{100} = 0.90$

Determining sample size[^2]:

\[
n = \frac{N \times X}{N + X - 1}
\]

where

\[
X = \frac{Z_{\alpha/2}^2 \times p \times (1 - p)}{MOE^2}
\]

- confidence value 95% $\rightarrow$ MOE = 0.05 $\rightarrow$ $Z_{\alpha/2} = 1.96$
- $p$ : proportional match = 0.52

\[
n = 373
\]

Experimental evaluation

The proportion of Exact Match is: \( \frac{185}{400} = 0.463 \)

\( \rightarrow \) the coverage of Wikipedia over SNOMED CT is 46.3 ± 5% with 95% confidence

Further analysis:
### Experimental evaluation

**Common patterns in creating disease names**

<table>
<thead>
<tr>
<th>Modifier types</th>
<th>Composite disease names</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anatomical location</td>
<td>Breast cancer</td>
</tr>
<tr>
<td>Symptoms</td>
<td>Trigeminal nerve inflammation</td>
</tr>
<tr>
<td>Treatment</td>
<td>Dopa-responsive dystonia</td>
</tr>
<tr>
<td>Causative agent</td>
<td>Staph infection</td>
</tr>
<tr>
<td>Biomolecular etiology</td>
<td>G6PD deficiency</td>
</tr>
<tr>
<td>Heredity</td>
<td>X-linked agammaglobulinemia</td>
</tr>
<tr>
<td>Eponyms</td>
<td>Schwartz-Jampel syndrome</td>
</tr>
<tr>
<td>Adjectives</td>
<td>Severe malaria</td>
</tr>
</tbody>
</table>

---

Experimental evaluation

Composite matching

➢ a SNOMED concept = {several Wikipedia articles}

➢ Study on 220 More Specific SNOMED concepts over all 100 + 400 selected concepts

<table>
<thead>
<tr>
<th>Combination types</th>
<th>Found</th>
<th>Combination types</th>
<th>Found</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anatomical location</td>
<td>86</td>
<td>Symptoms + Adjective</td>
<td>2</td>
</tr>
<tr>
<td>Anatomical location + Adjectives</td>
<td>4</td>
<td>Symptoms</td>
<td>13</td>
</tr>
<tr>
<td>Anatomical location + Symptoms</td>
<td>3</td>
<td>Treatment</td>
<td>2</td>
</tr>
<tr>
<td>Anatomical location + Causative</td>
<td>1</td>
<td>Adjectives</td>
<td>36</td>
</tr>
<tr>
<td>Causative agent</td>
<td>12</td>
<td>Causative agent + Adjective</td>
<td>1</td>
</tr>
<tr>
<td>Not found</td>
<td>60</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Combination of both Exact Match and Composite Match

\[ \text{coverage} = \frac{52 + 185 + 160}{500} = 0.794 \]
Experimental discussion

“The more common concepts are more likely to have an article in Wikipedia”

Analysis on average frequency of SNOMED CT concepts

<table>
<thead>
<tr>
<th>Coverage</th>
<th>Number of concepts</th>
<th>Avg.Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single match</td>
<td>225</td>
<td>8.33</td>
</tr>
<tr>
<td>More specific &amp; composite match</td>
<td>160</td>
<td>5.10</td>
</tr>
<tr>
<td>More specific &amp; no composite match</td>
<td>60</td>
<td>1.70</td>
</tr>
<tr>
<td>More general</td>
<td>11</td>
<td>2.20</td>
</tr>
<tr>
<td>No match</td>
<td>24</td>
<td>1.73</td>
</tr>
</tbody>
</table>
Experimental discussion

“Different granularity is a main cause of coverage gap between Wikipedia and SCT”
Conclusions

Our statistical experiments showed:

With 95% confidence: the coverage of a clinically-relevant subset of SNOMED CT is high

❖ Wikipedia covers 46.3 ± 5% of SNOMED CT disease concepts

❖ Wikipedia can describe 79% of SNOMED CT disease concepts through both exact and composite matching

Therefore:

→ Wikipedia is a good candidate to seed an Open Clinical Terminology
THANK YOU

Come and visit us at the CSIRO booth # 35

Our researchers and scientists would love to share more with you about how their work is enabling digital health in Australia and around the world.