Grouping pharmacovigilance terms with semantic distance

M Dupuch\textsuperscript{1,2}, M Lerch\textsuperscript{3}, A Jamet\textsuperscript{2,4}, MC Jaulent\textsuperscript{1,2}, RFescharek\textsuperscript{5} and N Grabar\textsuperscript{6}

(1) Université Pierre et Marie Curie - Paris6, Paris, F-75006 France;  
(2) INSERM, U872 eq. 20, Paris, F-75006 France;  
(3) Consulting & Coaching, Berlin, Germany;  
(4) HEGP, AP-HP, Paris, France;  
(5) CSL Behring GmbH, Marburg, Germany;  
(6) CNRS UMR 8163 STL, Université Lille 3, France
Plan of the presentation

- Context
- Objective
- Material
- Methods
- Results and Discussion
- Conclusion and Perspectives
Pharmacovigilance

- Pharmacovigilance: activity related to the collection, analysis and prevention of adverse drug reactions (ADRs)
- ADRs are coded with 2 terminologies:
  - MedDRA (Medical Dictionary for Drug Regulatory Activities)
  - WHO-ART (World Health Organization - Adverse Reaction Terminology)
- Signal: unexpected or not well documented relation between drugs and ADRs
  - Exploitation of statistical methods for detection of signals (Bate & al, 1998; Meyboom & al, 2002)
  - Exploitation of SMQs (Standardized MedDRA Queries)
    - Groupings of terms associated to a given safety topic
  - Grouping of related cases of pharmacovigilance
Exploitation and evaluation of SMQs

- SMQs are created manually by experts
  - Structure of MedDRA
  - Scientific literature
  ⇒ Long and tedious process
- Exploitation and evaluation of SMQs:
  - Consequent filtering and evaluation of cases is required:
    - High sensibility and over-inclusiveness (Mozzicato, 2007; Pearson & al, 2009)
  - Silences:
    - Important terms may not be included (Pearson & al, 2009)
⇒ Propose automatic methods for the creation of SMQs or of new groupings
Objective

- Hypothesis: exploitation of the semantic distance
  - Depends on the number of edges (the shortest path)

Previous exploitation of semantic distance with pharmacovigilance terms:
- subsets of MedDRA terms (Bousquet & al, 2005)
- subsets of WHO-ART terms (Iavindrasana & al, 2006)
  ⇒ No evaluation of the groupings with the SMQs

Objective: continue to adapt the semantic distance to PV terms
- Creation of groupings of terms and of new SMQs
- Exploitation of the whole set of MedDRA terms
- Comparison with the SMQs
## Material

**MedDRA**

<table>
<thead>
<tr>
<th>Level</th>
<th>Expanded form</th>
<th>Nb Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOC</td>
<td>System Organ Class</td>
<td>26</td>
</tr>
<tr>
<td>HLGT</td>
<td>High Level Group Terms</td>
<td>332</td>
</tr>
<tr>
<td>HLT</td>
<td>High Level Terms</td>
<td>1,688</td>
</tr>
<tr>
<td>PT</td>
<td>Prefered Terms</td>
<td>18,209</td>
</tr>
<tr>
<td>LLT</td>
<td>Lowest Level Terms</td>
<td>66,587</td>
</tr>
</tbody>
</table>

- Two kinds of material derived from MedDRA:
  - ontoEIM resource
  - SMQs
OntoEIM resource (Alecu & al, 2008):
- Projection of MedDRA terms on SNOMED CT
  - Currently 46% of MedDRA terms are aligned
- Finer-grained hierarchy of MedDRA terms
- Formal definitions (FD) as extracted from SNOMED CT:
  - Morphology M, topography T, causality C and expression E
- Arsenical keratosis (diagnosis ADR)
  - M: Squamous cell neoplasm; Morphologically abnormal structure
  - T: Skin structure; Structure of skin and or surface epithelium
  - C: Arsenic AND OR arsenic compound
  - E: Abnormal keratinization
Material

SMQs (Standartized MedDRA Queries)

- Groupings of terms related to a safety topic
  - *Acute renal failure, Hepatic disorders, Thrombocytopenia*...
- Helpful for the searching of close pharmacovigilance cases
- Created manually by experts
- 84 SMQs exist currently
- SMQs include PT and LLT terms
- Gold standard: 9 SMQs (ADRs leading to hospitalization, vital prognosis and death)
  - *Acute renal failure, Agranulocytosis, Anaphylactic reaction, Cytopenia, Gastrointestinal haemorrhages, Peripheral neuropathy, Rhabdomyolysis, Severe cutaneous adverse reaction, Thrombocytopenia*
Methods

Tree-step method:

1. Optimisation of the alignment and of formal definitions
2. Computing of the semantic distance and grouping of terms
3. Evaluation of the generated groupings of terms
1. Optimisation of the alignment and of formal definitions

- 46% of MedDRA terms are aligned with SNOMED CT
- 51.6% of PTs and 35.4% of LLTs

⇒ Optimize the alignment and improve the coverage

PTs and LLTs aligned

\[ \text{dist}_{\text{OntoEIM}} \]
- FD PT
- FD LLT

PTs aligned; LLTs not aligned

Transfer FD PTs \Rightarrow LLTs

\[ sp(\text{LLT}) = sp(\text{PT}) + 1 \]
1. Optimisation of the alignment and of formal definitions

- **LLTs aligned; PTs not aligned**
- Transfer FD LLTs $\Rightarrow$ PTs
  - $sp(PT) = sp(LLT) - 1$
- **LLTs and PTs not aligned**
  - $\Rightarrow$ Semantic distance not computed

- **Improvement of the alignment of MedDRA terms:**
  - $+10\%$ for PTs (61.6\%)
  - $+30\%$ for LLTs (65.4\%)
2. Computing of semantic distance and grouping of terms

- Computing of the semantic distance (Rada et al, 1989)
- The shortest path between two concepts \( \text{dist}_{Rada} = sp(c_1, c_2) \)
- The shortest path is the sum of all its edges
- Each edge is equal to 1
- Parameters tested:
  - One axis (ADRs)
  - Three axes (ADR + FD)
  - All SMQ terms
  - Aligned SMQ terms
  - Best grouping
  - Merging of \( n \) best groupings

<table>
<thead>
<tr>
<th>One axis (ADRs)</th>
<th>Three axes (ADR + FD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All SMQ terms</td>
<td>Aligned SMQ terms</td>
</tr>
<tr>
<td>Best grouping</td>
<td>Merging of ( n ) best groupings</td>
</tr>
</tbody>
</table>

\[
sp_{ADR} = 4, \quad sp_T = 10 \quad \text{et} \quad sp_M = 0
\]

\[\text{pcc}_D = 4, \quad \text{pcc}_T = 10 \quad \text{et} \quad \text{pcc}_M = 0\]
2. Computing of semantic distance and grouping of terms

- Each axis is weighted (Petiot & al, 1996)
  \[ dist_{ontoEIM}(A, B) = \frac{\sum_{i \in \{ADR, M, T\}} W_i \times sp(A_i, B_i)}{\sum_{j \in \{ADR, M, T\}} W_j} \]

  \[ W_{ADR} = 1, \quad W_M = 2 \quad \text{et} \quad W_T = 1 \]

  \[ \Rightarrow M \text{ is the most important} \]

- Generation of semi-matrix
  - \( dist_{ontoEIM} \) between all PT and LLT terms

  \[ dist_{ontoEIM}(\text{Abdominal abscess, Pharyngeal abscess}) = 3.5 \]

- Grouping: distance \( \leq 2 \)
3. Evaluation of the generated groupings of terms

- Three evaluation measures:
  - precision \( P \): number of relevant grouped terms as a percentage of the total number of the grouped terms
  - recall \( R \): number of relevant grouped terms as a percentage of the number of terms in the corresponding SMQ
  - f-measure \( F \): the harmonic mean of \( P \) and \( R \)

- Association between the generated groupings and the SMQs
  \( \Rightarrow \) precision
Merging of groupings: increasing of the overall performance
1 axis vs 3 axes: best performances with 1 axis
   incompleteness of the formal definitions
Set of aligned terms: $R$ and $F$ increase but $P$ decreases
Min-max intervals very large: variability between the SMQs
   various strategies needed according to the safety topics
Best results: 1 axis, merged groupings
   high precision (expectation of the experts)
Conclusion and Perspectives

- Method for the creation of groupings of ADR terms
- Optimization of alignment: +10% for PTs, +30% for LLTs
- Different parameters of the method:
  - Best results: 1 axis, merged groupings
- Evaluation: high precision (expectation of the experts)
- Perspectives:
  - Improvement of the alignment of the MedDRA terms
  - Broad and narrow versions of SMQs
  - Adjustment of variables (edge weights, coefficient of axes...)
  - Other measures for semantic distance (Leacock, Zhong...)
  - Other clustering methods (hierarchical, partitionning)
  - NLP methods for enriching and refining the groupings
  - Definition of different strategies for different safety topics