Computational Solutions to Large-Scale Data Management and Analysis Challenges in Personalized Health

Semantic Interoperability Challenges for Sharing and Reusing Large Amounts of Heterogeneous Data

Marie-Christine Jaulent
PhD, INSERM research director, France
Interoperability

- What is interoperability in the healthcare informatics domain?
  - Interoperability is the ability of heterogeneous healthcare information systems and software applications to communicate, to exchange data accurately, effectively, and consistently, and to use the information that has been exchanged.

- ... but simply
What is interoperability in the healthcare informatics domain?

Interoperability is the ability of heterogeneous healthcare information systems and software applications to communicate, to exchange data accurately, effectively, and consistently, and to use the information that has been exchanged.

... but simply

The ability of a system to access and use the parts of another system

Syntactic Interoperability

Semantic Interoperability
Interoperability

- What is interoperability in the healthcare informatics domain?
- Interoperability is the ability of heterogeneous healthcare information systems and software applications to communicate, to exchange data accurately, effectively, and consistently, and to use the information that has been exchanged.

- Semantic Interoperability
- Practical Interoperability
Heterogeneous data understandable for humans...

- Cholestatic hepatitis
- Invasive Ductal Carcinoma
  Grade 2 ER+ PR+ HER2/neu+
- Urinary tract infection caused by E.coli resistant to quinolones
- Drug-eluting stent in left anterior descending artery

- Elevated Bilirubine
- Elevated ALAT

Humans interpret multi-sources heterogeneous clinical informations

E.Coli – Quinolones - R
...and for machines

Heterogeneous « models of use »

Representation of information & knowledge in a computable format « Model of meaning »

This patient suffers of cholestatic hepatitis

Cholestatic hepatitis: yes

Diagnosis:
Cholestatic hepatitis

Data structures + Vocabulary
(Reference terminologies/ Ontologies)
(e.g ICD-10, SNOMED, LOINC, etc)
The rôle of ontologies

Semantic interoperability platform for the integration and the sharing of heterogeneous data
Semantic interoperability challenges in the context of pharmacovigilance

- Pharmacovigilance
  - “the science and activities relating to the detection, assessment, understanding, and prevention of adverse effects or any other drug-related problems” (World Health Organization)
  - Begins during clinical trials and continues after the drug is released onto the market
Case description

- Report ID & Date
  - reportid - receivedate
- Patient Data
  - sex - onset age - weight
- Seriousness
  - serious (Yes/NO) – hospitalization – death – deathdate – lifethreatening - disabling – seriousnessother
- Reaction
  - reactionmeddрапт – reactionoutcome – actiondrug

WHO ATC classification system

- Basic Drug Information
  - medicinalproduct – drugcharacterization – drugindication – dosagetext – drugstartdate - drugenddate
- Administration Information
  - route – drugseparatedosagenumb - drugdosageform
- Report Information (Origin / Duplicates, etc.)
- Additional Drug Information
  - substance_name - generic_name
Large-scaled data in pharmacovigilance

Spontaneous reporting

Reporting Activity

National/european Repositories Pharmaceutical companies

Signal Detection

AERS (FDA)

Vigibase (WHO)

Over 16 million reports In Vigibase
Large-scaled data in pharmacovigilance

Reporting Activity
Spontaneous

- Bayesian Confidence Propagation Neural Network
- Multi-item Gamma Poisson Shrinker
- Reporting Odds Ratio
- Proportional Reporting Ratio

Bayesian Method
Frequentist Method
SignalDetectionMethod

Vigibase (WHO)

Over 16 million reports in Vigibase

- AERS (FDA)
- Vigibase (WHO)

National/european Repositories
Pharmaceutical compagnies

Signal Detection

"any reported information on a possible causal relationship between a drug and an adverse drug reaction, the relationship being unknown or incompletely documented previously" [WHO]

Over 16 million reports in Vigibase

OntoADR for retrieval and reasoning

? Retrieve cases of Hepatobiliary diseases with Amineptine?

{Amineptine – cholestatic hepatitis}
{Amineptine – cytolytic icterus}
{Amineptine – High ALAT}
{Amineptine – hyperbilirubinemia}
{Amineptine – hepatotoxicity}
{Amineptine – cytolytic hepatitis}

Pharmacovigilance databases

Peripherical Neuropathy - Efalizumab

Individual MedDRA terms (PT)

Groupings of terms based on OntoADR

Under reporting challenges

Spontaneous reporting

Reporting Activity

National/european Repositories
Pharmaceutical companies

AERS (FDA)

Vigibase (WHO)

Signal Detection

Under reporting
Under reporting challenges

- Estimated between 78% and 99% in France
- Concurrent exploration of diverse data sources
  - Spontaneous reporting systems,
  - Electronic health records,
  - Bibliographic databases,
  - Social media
- Semantic interoperability challenges
  Natural language (layman language)
Under reporting challenges

- Estimated between 78% and 99% in France
- Concurrent exploration of diverse data sources
  - Spontaneous reporting systems,
  - Electronic health records,


- Semantic interoperability challenges
  - Natural language (layman language)
Under reporting challenges

- Estimated between 78% and 99% in France
- Concurrent exploration of diverse data sources
  - Spontaneous reporting systems,
  - Electronic health records,
  - Bibliographic databases,


Natural language (layman language)
Social media for Pharmacovigilance

1. Extract messages’ content from forum (french language)

2. Identify messages with drug names and medical terms for adverse events (AEs)

3. Evaluate the relevance and quality of the information shared by patients on web forums about drug safety

4. Evaluate its potential utility for pharmacovigilance.

5. Set up a data processing workflow for routine monitoring of health product uses
## Websites of interest

<table>
<thead>
<tr>
<th>Site</th>
<th>Size before anonymisation</th>
<th>Size after anonymisation</th>
<th>Number of pages</th>
<th>Number of conversation thread</th>
<th>Number of messages</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.allodocteurs.fr/">http://www.allodocteurs.fr/</a></td>
<td>84 M</td>
<td>56 M</td>
<td>7 083</td>
<td>6 012</td>
<td>22 476</td>
</tr>
<tr>
<td><a href="http://www.atoute.org/">http://www.atoute.org/</a></td>
<td>16 G</td>
<td>12 G</td>
<td>294 611</td>
<td>157 267</td>
<td>4773 039</td>
</tr>
<tr>
<td><a href="http://www.carenity.com/">http://www.carenity.com/</a></td>
<td>4,8 M</td>
<td>4,6 M</td>
<td>149</td>
<td>135</td>
<td>320</td>
</tr>
<tr>
<td><a href="http://www.comprendrechoisir.com/">http://www.comprendrechoisir.com/</a></td>
<td>40 M</td>
<td>33 M</td>
<td>3 121</td>
<td>3 013</td>
<td>6 678</td>
</tr>
<tr>
<td><a href="http://www.docteurlic.com">http://www.docteurlic.com</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><a href="http://www.doctissimo.fr">http://www.doctissimo.fr</a></td>
<td>16 G</td>
<td>12 G</td>
<td>294 611</td>
<td>157 267</td>
<td>4773 039</td>
</tr>
<tr>
<td>http://www.doctissimo.fr_sante</td>
<td>53 G</td>
<td>33 G</td>
<td>780</td>
<td>538</td>
<td>16 345 311</td>
</tr>
<tr>
<td>http://www.doctissimo.fr_grossesse</td>
<td>138 G</td>
<td>96 G</td>
<td>1 113</td>
<td>1 283</td>
<td>34 854 099</td>
</tr>
<tr>
<td><a href="http://www.e-sante.fr/">http://www.e-sante.fr/</a></td>
<td>91 M</td>
<td>71 M</td>
<td>1 232</td>
<td>13</td>
<td>35 575</td>
</tr>
<tr>
<td><a href="http://www.famili.fr/">http://www.famili.fr/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><a href="http://sante-az.aufeminin.com/">http://sante-az.aufeminin.com/</a></td>
<td>3,9 G</td>
<td>2,2 G</td>
<td>152 839</td>
<td>140 735</td>
<td>849 655</td>
</tr>
<tr>
<td><a href="http://forums.futura-sciences.com/">http://forums.futura-sciences.com/</a></td>
<td>237 M</td>
<td>193 M</td>
<td>12 001</td>
<td>9 759</td>
<td>76 459</td>
</tr>
<tr>
<td><a href="http://sante.journaldesfemmes.com/">http://sante.journaldesfemmes.com/</a></td>
<td>58 M</td>
<td>45 M</td>
<td>4 543</td>
<td>4 217</td>
<td>25 844</td>
</tr>
<tr>
<td><a href="http://sante-medecine.commentcamarche.net/">http://sante-medecine.commentcamarche.net/</a></td>
<td>2,3 G</td>
<td>1,7 G</td>
<td>219 440</td>
<td>213 156</td>
<td>1 112 084</td>
</tr>
<tr>
<td><a href="http://afm-telethon.fr/">http://afm-telethon.fr/</a></td>
<td>8,2 M</td>
<td>6,6 M</td>
<td>1 175</td>
<td>1 012</td>
<td>3 502</td>
</tr>
<tr>
<td><a href="http://www.entrepatients.net/fr">http://www.entrepatients.net/fr</a></td>
<td>11 M</td>
<td>7,7 M</td>
<td>968</td>
<td>861</td>
<td>3 485</td>
</tr>
<tr>
<td><a href="http://www.notrefamille.com">http://www.notrefamille.com</a></td>
<td>976 M</td>
<td>586 M</td>
<td>34 909</td>
<td>29 751</td>
<td>273 021</td>
</tr>
<tr>
<td><a href="http://www.baclofene.fr/portal.php">http://www.baclofene.fr/portal.php</a></td>
<td>106 M</td>
<td>79 M</td>
<td>2 819</td>
<td>1 819</td>
<td>36 232</td>
</tr>
<tr>
<td><a href="http://www.hepatites.net/">http://www.hepatites.net/</a></td>
<td>452 M</td>
<td>404 M</td>
<td>16 079</td>
<td>11 783</td>
<td>177 891</td>
</tr>
<tr>
<td><a href="http://www.seronet.info/">http://www.seronet.info/</a></td>
<td>14,5 M</td>
<td>13,7 M</td>
<td>278</td>
<td>147</td>
<td>175 049</td>
</tr>
</tbody>
</table>

**Posts Extraction (Vigi4Med Scrapper tool)**

- Over 60 million posts extracted from 22 French health related web forums
- From 01/01/2000 to 07/03/2015
Social media for Pharmacovigilance

1. Extract messages’ content from forum (french language)

2. Identify messages with drug names and medical terms for adverse events (AEs)

3. Evaluate the relevance and quality of the information shared by patients on web forums about drug safety

4. Evaluate its potential utility for pharmacovigilance.

5. Set up a data processing workflow for routine monitoring of health product uses
Message annotation

- Natural language processing (NLP)
  - Detection of drugs (e.g. Lutenyl), adverse effect (e.g. weight gain) and causal relation
  - Detection of spelling errors (e.g. Luthenyl/Lutenyl)

Automatic natural language processing
  - Detection of at least one drug and one AE in 6,569,555 posts
  - Detection of 55,350,564 drug-AE couples

Au début luthenyl a bien fonctionné mais j'ai du doubler les doses car il n'était plus efficace puis je l'ai pris tout le long du cycle (au début de j10 à j22). De là sont venus la douleur dans les jambes et la prise de poids.
Relevance and quality study

- Agomelatine, baclofen, duloxetine, exenatide, strontium ranelate, and tetrazepam
- Harmonisation of study period between forums et French PharmacoVigilance Database (FPVD)
- 1,284 messages posted by patients in forums versus 2,521 spontaneous declarations stored in FPVD

Comparaison of the two data sources
- % patient profile and treatment
- % information about adverse effect
- % information about impact
Relevance and quality study

Agomelatine, baclofen, duloxetine, exenatide, strontium ranelate, and tetrazepam

Harmonisation of study period between forums et French PharmacoVigilance Database (FPVD)

1,284 messages posted by patients in forums versus 2,521 spontaneous declarations stored in FPVD

Comparaison of the two data sources

- % patient profile and treatment
- % information about adverse effect
- % information about impact

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Cases from forums N = 1,284</th>
<th>Cases from French DB N = 2,521</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>4.8%</td>
<td>93.3%</td>
</tr>
<tr>
<td>Gender</td>
<td>47.7%</td>
<td>99.5%</td>
</tr>
<tr>
<td>Indication</td>
<td>60.7%</td>
<td>37.1%</td>
</tr>
<tr>
<td>traitement duration</td>
<td>22.6%</td>
<td>54.7%</td>
</tr>
</tbody>
</table>
Relevance and quality study

- Agomelatine, baclofen, duloxetine, exenatide, strontium ranelate, and tetrazenapm

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Cases from forums N = 1284</th>
<th>Cases from French DB N = 2521</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apparition delay</td>
<td>19.9%</td>
<td>83.5%</td>
</tr>
<tr>
<td>Evolution</td>
<td>6.4%</td>
<td>82.3%</td>
</tr>
</tbody>
</table>

Comparaison of the two data sources
- % patient profile and treatment
- % information about adverse effect
- % information about impact
Relevance and quality study

Agomelatine, baclofen, duloxetine, exenatide, strontium ranelate, and tetrazepam

Harmonisation of study period between forums et French PharmacoVigilance Database (FPVD)

1,284 messages posted by patients in forums versus 2,521 spontaneous declarations stored in FPVD

Comparaison of the two data sources

- % patient profile and treatment
- % information about adverse effect
- % information about impact

A preliminary study

- Signal detection for Baclofen in Social Media
  - Baclofen is a mainstay in the management of spasticity in several neurological diseases since the mid-70s
  - Off-label use of high-dose baclofen for alcohol-dependence has increased in the last 5 years in France
    - Temporary recommendation for use to frame it
  - Recent signals emerged on other off-label use of baclofen
    - Interesting drug for a study on web forums

- Normalisation
  - AEs : UMLS-MedDRA (V19.1) mapping database
  - Drug
    - Perfect correspondence to the RacinePharma dictionary (5,164 drug terms)
    - Linked to the Anatomical Therapeutic Chemical (ATC) classification system (for 83.5%)
Extracted material

- 2,622,445 couples detected from 747,500 posts (from 07/02/2000 to 07/02/2015)
- 5,997 corresponded to a perfect match with « BACLOFENE »
- 40,158 to one of the lexical variations
  - 46 155 couples with baclofen
    - 999 distincts AEs with baclofen
    - The 10 most frequents AEs represented 54% of the AEs found with baclofen
    - Among the 10 most frequent AEs detected with Baclofen, 2 were indications (15%)
Signal detection

- Based on frequentist and bayesian methods
  - 36 of the 37 AEs found in baclofen’ Summary of Product Characteristics (SPC) were detected with baclofen in forums
  - Known AEs can be detected from web forums data using classical signal detection algorithms
  - 2,621 messages were analysed, 782 were validated as cases
- Cross-checking for 4 AEs of interest (Glaucoma, Boulimia nervosa, Sleep Apnea Syndrome (SAS), Nightmare)
  - Logistic regression based method (Class-Imbalanced Subsampling Lasso)
  - Boulimia was found in 7 validated messages
  - Nightmare was found in 22 validated messages
  - Sleep Apnea Syndrome was found in 5 validated messages
  - Signals were confirmed

Signal Detection for Baclofen in Web Forums: A Preliminary Study.
Karapetiantz P, Audeh B, Lillo-Le Louët A, Bousquet C.
Thank you for your attention

Marie-Christine Jaulent

Cédric Bousquet

Bissan Audeh

Pierre Karapetiantz

Gunnar Declerk

Julien Souvignet