Bioclinical Data Warehouses & Translational Research: the HEGP Case

Patrice Degoulet & Paul Avillach

Faculté de Médecine René Descartes, Université Paris 5
Hôpital Européen Georges Pompidou (HEGP)

INSERM UMR_S 872 eq22

August 2012
Why should we develop Bioclinical Data Warehouses

1. EHR driven research
   – Patient selection for CR studies (e.g., EHR4CR)
   – Biomarker discovery (phenotype-genotype)
   – Personalized medicine

2. Phenotypic augmentation for clinical research studies
   – Additional source of bioclinical data for cohort studies
   – Reuse of EHR data to feed a CR study

• HEGP is the most recent acute care hospital within the 37 AP-HP hospitals

• HEGP meets the needs of the 600,000 inhabitants of the Paris south-west
HEGP background

Shared Biobank (2008-)
HEGP BDW

Production environment

Evaluation/Research environment

ETL suite (Talend Open Studio)

EHR/BDW integration

EHR: Operational Database (ODS)

EHR: Mirrored Database

Biomedical Data Warehouse (BDW)

External Databases

Real time requests

Data Analysis
Data Mining

i2b2/tranSMART tools

Business Object
IBM Ilog Rules

R
## BDW content

### i2b2 CDW content (July 2012)

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Categories</th>
<th>Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td></td>
<td>432 033</td>
</tr>
<tr>
<td>Concept dimension</td>
<td>ICD10 classification</td>
<td>21 356</td>
</tr>
<tr>
<td></td>
<td>Laboratory results classification</td>
<td>8 272</td>
</tr>
<tr>
<td></td>
<td>Drug classification (ATC)</td>
<td>33 612</td>
</tr>
<tr>
<td></td>
<td>EHR forms concepts</td>
<td>5 950</td>
</tr>
<tr>
<td>Observation facts</td>
<td>ICD10 Diagnosis</td>
<td>2 537 633</td>
</tr>
<tr>
<td></td>
<td>Laboratory results</td>
<td>85 598 854</td>
</tr>
<tr>
<td></td>
<td>Drug prescriptions</td>
<td>2 474 985</td>
</tr>
<tr>
<td></td>
<td>EHR forms items</td>
<td>28 641 547</td>
</tr>
<tr>
<td></td>
<td>Text reports</td>
<td>902 747</td>
</tr>
</tbody>
</table>
• 158 MD + Pharm trained
• 1386 requests - 15 cohorts created (i2b2 datamarts)
## Rate of inappropriate prescriptions

6 alternating 2-month phases: control vs. intervention (Aug. 2006- Aug. 2007)

<table>
<thead>
<tr>
<th>Physicians</th>
<th>Alerting off</th>
<th>Alerting on</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Junior</td>
<td>21.5%</td>
<td>16.3%</td>
<td>p=0.88</td>
</tr>
<tr>
<td>Senior</td>
<td>20.9%</td>
<td>29.3%</td>
<td>p=0.01</td>
</tr>
<tr>
<td>Total</td>
<td>21.3%</td>
<td>19.9%</td>
<td>0.63 (NS)</td>
</tr>
</tbody>
</table>
In silico evaluation of decision rules (BDW)

Initial set
(CPOE $v_n$)

280 rules
(10 drugs)

Final set
(CPOE $v_{n+1}$)

371 rules
(10 drugs)

Suppressed: 45 (16.1%)

Modified: 105 (37.5%)

Added: 136 (48.2%)

The Transmart integration process

1. Define an Ontology
   - The ontology dictates how data is displayed within tranSMART

2. Convert data into ontology-formatted specifications
   - For example, tab-delimited text (.txt) files

3. Transform text files into the standard format
   - Processes differ between clinical data, gene expression data, and SNP data

4. Copy files to i2b2 and DEAPP tables
   - Loading scripts populate relevant tables

http://www.transmartproject.org
Analysis of *PTEN*, *BRAF*, and *EGFR* Status in Determining Benefit From Cetuximab Therapy in Wild-Type *KRAS* Metastatic Colon Cancer

Pierre Laurent-Puig, Anne Cayre, Gilles Manceau, Emmanuel Buc, Jean-Baptiste Bachet, Thierry Lecomte, Philippe Rougier, Astrid Lievre, Bruno Landi, Valérie Boige, Michel Ducreux, Marc Ychou, Frédéric Bibeau, Olivier Bouché, Julia Reid, Steven Stone, and Frédérique Penault-Llorca

TranSMART-based Ontology definition

Biomarker data

Non-omics

Omics

Immunological data

SNP analysis

Mutations detection

KRAS

Yes

No

N/A

Data Label

Data Value

Category Code

BRAF

Yes

No

N/A

NRAS

Yes

No

N/A

Hirsch Score

Positive

Negative

N/A

EGFR copy number

PTEN Cyto

PTEN MBR

PTEN Nx

Concept inexistant dans une terminologie contrôlée
Proof of concept

- R module in tranSMART
- Published figure in JCO
Achievements

• A methodology to export concepts and data from an integrated CIS (2009-)
• An operational CDW directly used by MD and Pharm (2011-)
• Installation & evaluation of a tranSMART platform to augment clinical data with omic information (2012-)

Benefits of the approach

• Availability of clinical data has generated a virtuous cycle at the HEGP end-user level (e.g., improved standardized questionnaires)
• tranSMART is based on the i2b2 model for clinical data
Limits of the approach

- Lack of semantic integration tools to merge i2b2-tranSMART concepts
- i2b2-tranSMART data model heterogeneity

Perspectives

- Semantic integration within the i2b2-tranSMART platform to enable:
  1) meta-analysis from multiple cohorts
  2) phenotypic augmentation in genomic driven research
Acknowledgments

Informatics & Public Health Dept.
Pierre Durieux, MD
Eric Zapletal, PhD
Kevin Zarca, MD
Vincent Canuel, Resident

Contacts:
patrice.degoulet@egp.aphp.fr
paul.avillach@egp.aphp.fr

www.i2b2.org
www.transmartproject.org
www.recomdata.com