Sources of data for PASS studies in the EU

Andrew Bate
Senior Director, Analytics Team Lead, Epidemiology
ISoP 2013 Pre conference course- Pisa, 01 October 2013
Disclosures

• I am a full time employee of Pfizer and hold stocks and stock options
Agenda

- Overview
- Data types for observational studies
- Why we need to consider sample size when work with large EU databases
- Overview of EU database
- Using multiple databases to increase sample size
- Recent and ongoing global initiatives on safety surveillance research
  - The example of IMI PROTECT
- Conclusions
Longitudinal records of EMRs and transactional insurance claims data

- Rich data
  - Time stamped diagnoses (without any requirement of clinical suspicion)
  - Recorded exposure; and reliable non-exposure
  - Detailed information on disease history prior to drug exposure
  - Other data: test results, hospital referrals and admissions, surgical procedures, notes, symptoms, signs and administrative data
  - Often linked/can be linked to other healthcare data
  - But challenging for screening that no clinical suspicion link between prescription and outcome
Introduction

• Data types for observational studies
  Secondary use of existing data (EMR data), Claims data
• Primary data collection for the purposes of the study (or studies)
• De novo (Some primary data collection on existing data to make)
• Linking of different sources of observational data
Consisting of anonymous demographic, medical and prescription information at patient level, THIN data provide rich and longitudinal Electronic Health Records (EHR) of each patient which were directly collected from general practitioners in UK health care system.
## Some selected observational databases

<table>
<thead>
<tr>
<th>Database</th>
<th>Country</th>
<th>Characteristic</th>
<th>Population Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>THIN</td>
<td>UK</td>
<td>GP primary care database</td>
<td>10.5 M&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>Danish National Health Service Register Database</td>
<td>Denmark</td>
<td>Healthcare registry of care</td>
<td>5.5 M&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>Premier</td>
<td>US</td>
<td>Clinical data from the hospitals</td>
<td>130 M+ patient discharges&lt;sup&gt;3&lt;/sup&gt;</td>
</tr>
<tr>
<td>Normative Health Information (NHI) Database</td>
<td>US</td>
<td>Transactional claims records of a commercial health insurer</td>
<td>60 M&lt;sup&gt;4&lt;/sup&gt;</td>
</tr>
<tr>
<td>Health Insurance Review and Assessment Service (HIRA)</td>
<td>Korea</td>
<td>Insurance Claims from near universal national system</td>
<td>48 M&lt;sup&gt;5&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

1 Blak et al Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. Informatics in Primary Care 2011;19:251–5
2 Furu K. et. al. The Nordic Countries as a Cohort for Pharmacoepidemiological Research. Basic & Clinical Pharmacology & Toxicology 2009; 106: 86-94
3 Fisher BT et al. In-hospital databases In Pharmacoepidemiology 5th Edn 2011 pp 244-258
5 Kimura T et al. Pharmacovigilance systems and databases in Korea, Japan and Taiwan. Pharmacoepidemiology and Drug Safety. 2011; 20: 1237–1245
Heterogeneity of databases in Europe

- Varying country, population coverage, with differing healthcare systems
- Varying Healthcare system coverage
  - Some are primarily geared to a specific type of care (e.g. primary), others more general coverage
- Capability to follow patients varies
- Reliability of different types of variables
- Some are mature and well established for pharmacoepidemiology research, others are emerging
- Varying sample size
- Some attributes difficult to capture in any secondary use of databases
Sources of variability with EU databases

- Data on what types of patient recorded in the database
  - Generalizability?
- Database size (Number of patients and amount of follow up)
- Data Elements captured in data set
  - What is potentially captured
  - Reliability of their capture (missing data?), accuracy of follow up of patients
- Type of healthcare encounters captured
  - Duplicate records, part of records across multiple databases

Capability for conducting chart review for case validation and for adding additional observations through
Sample size consideration is important

- Feasibility of studies
- Interpretation of study results

Each EU database is smaller than US claims databases

It could on occasion be important to use multiple databases to increase sample size

- Rare event
- Rare exposure
## Examples of EU databases and their size

<table>
<thead>
<tr>
<th>Database</th>
<th>Country</th>
<th>Population Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPRD1 (now known as CPRD)</td>
<td>UK</td>
<td>11 M</td>
</tr>
<tr>
<td>THIN1</td>
<td>UK</td>
<td>7.8 M</td>
</tr>
<tr>
<td>Disease Analyzer UK2</td>
<td>UK</td>
<td>5 M</td>
</tr>
<tr>
<td>Disease Analyzer DE2</td>
<td>Germany</td>
<td>20 M</td>
</tr>
<tr>
<td>Disease Analyzer FR2</td>
<td>France</td>
<td>3.7 M</td>
</tr>
<tr>
<td>GePaRD3</td>
<td>Germany</td>
<td>14 M</td>
</tr>
<tr>
<td>Danish National Health Service Register Database4</td>
<td>Denmark</td>
<td>5.5 M</td>
</tr>
<tr>
<td>Swedish Prescribed Drug Register5</td>
<td>Sweden</td>
<td>9 M</td>
</tr>
</tbody>
</table>

2. IMS® Disease Analyzer Tracking real-life patient care. IMS presentation, September 2010.
### Number of Children in different EU databases

<table>
<thead>
<tr>
<th>Database</th>
<th>Country</th>
<th>Number of Children</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPRD</td>
<td>UK</td>
<td>1.15 M</td>
</tr>
<tr>
<td>THIN</td>
<td>UK</td>
<td>502 K</td>
</tr>
<tr>
<td>Disease Analyzer UK</td>
<td>UK</td>
<td>460 K</td>
</tr>
<tr>
<td>Disease Analyzer DE</td>
<td>Germany</td>
<td>250 K</td>
</tr>
<tr>
<td>Disease Analyzer FR</td>
<td>France</td>
<td>190 K</td>
</tr>
<tr>
<td>PHARMO</td>
<td>Netherlands</td>
<td>&gt;360 K</td>
</tr>
<tr>
<td>Arno Observatory</td>
<td>Italy</td>
<td>1.5 M</td>
</tr>
<tr>
<td>Swedish Medical Birth Register</td>
<td>Sweden</td>
<td>3.23 M</td>
</tr>
</tbody>
</table>

Population-Based Health Registry Databases of Nordic Countries

Total population 25 mill

- Iceland: 0.3 mill
- Norway: 4.8 mill
- Sweden: 9.2 mill
- Denmark: 5.5 mill
- Finland: 5.2 mill

Source:
Furu K. et. al. The Nordic Countries as a Cohort for Pharmacoepidemiological Research. BCPT 2009; 106: 86-94
Consider using multiple databases, such as GPRD, THIN and IMS Disease Analyzer databases.

- THIN: 7.8 M
- GPRD: 11 M
- IMS Disease Analyzer UK: 5 M
- IMS Disease Analyzer Germany: 20 M
- IMS Disease Analyzer: 3.7 M
A hypothetical case study

• Use the 5 databases to increase sample size
  – Rare exposure
  – Rare outcome

• Need to identify and remove the practices that are common to the two databases: GPRD and THIN
  – Real Case Study follows
GPRD and THIN Databases at time of study

- **GPRD**
  - 487 practices.
  - 3-4 million currently active patients with research quality data.
  - Over 10 million patients with usable research data.

- **THIN**
  - 382 practices.
  - About 3 million currently active patients with research quality data.
  - About 7 million patients with usable research data.

- Data from both databases collected from the GP using Vision software
- GPRD and THIN databases are important in epidemiology research
“The objective of this research was to create algorithms to identify common practices to both the GPRD and THIN databases so that the two databases can be used together to maximize the number of patient records for analysis without duplicate records.”
Challenges

- Both database are large.
- Different database vendors have different data processing procedures.
- Different database vendors have different data updating schedule.
Methods

- Study period: 2001-2008
  - THIN practice: 495
  - GPRD practice: 613

- Two practice-level matching algorithms to identify common practices

- Patient-level data to verify: ≥80% identical patients
Algorithm 1

• Matched by geographic area.
• Total number of patients in the prescription dataset in a given practice +/-10%.
• Total number of patients in the clinical dataset +/-10%.
• Total number of users of specific drug matched by +/-3 users in each year from 2001 to 2006.
Algorithm 2

Calculate number of patients in 8 categories specified by gender and the following 4 age categories:

- Birth year less than 1900.
- Birth year between 1900-1905.
- Birth year between 1905-1910.
- Birth year less than the earliest birth year in the practice plus 10 years.

We looked for practices from GPRD that match those in THIN by the criteria that the difference in patient numbers was less than 5% for at least 4 out of 8 categories.
Results

THIN: 495

168 unmatched

327 matched:
• 312 by Algorithm 1&2
• 13 by Algorithm 1 only
• 2 by algorithm 2 only

GPRD: 613

286 unmatched

Redundant practices that were found by either algorithm were removed from a subsequent Epidemiological study.
Lessons learnt from study

• Combined the two databases to increase sample size
  – Rare exposure
  – Rare outcome
• Could identify and remove the practices that are common to the two databases
• This method could be extended to link different databases.
Comments

• First effort to match clinical practices
• Two algorithms work together to improve sensitivity
• Matching practice pairs are validate with patient level data.
• Limitations:
  – The results is time-sensitive, new overlapped practices occur when both database vendors continue recruiting new practices.
  – True set of matched pairs is unknown.

References:
Cai B et al. An algorithm to identify practices common to both the GPRD and THIN databases. ICPE Brighton 2010.
Cai B et al. An algorithm to identify medical practices common to both the General Practice Research Database and The Health Improvement Network database. PDS 2012; 21: 770-774.
Examples of International and National Initiatives addressing the use of multiple databases

- Innovative Medicines Initiative (IMI) project: PROTECT
  - European Community's Seventh Framework Programme (FP7/2007-2013) for the Innovative Medicine Initiative
- Observational Medicines Outcomes Partnership (OMOP)
- FDA Sentinel Initiative
- European Commission Seventh Framework Programme (FP-7) of the Research Directorate: EU_ADR
PROTECT Goal

To strengthen the monitoring of benefit-risk of medicines in Europe by developing innovative methods

to enhance early detection and assessment of adverse drug reactions from different data sources (clinical trials, spontaneous reporting and observational studies)
to enable the integration and presentation of data on benefits and risks

These methods will be tested in real-life situations.
Conclusions (1)

- There are now multiple rich heterogeneous and intricately constructed ‘real world’ data sets of observational databases
  - Need efforts for harmonized mappings between terminologies
    - Standard definitions of important concepts in multiple vocabularies
- Databases in the EU are rich and valuable, well established and widely used for Pharmacoepidemiology studies
- Even with large databases in EU, there is sometimes insufficient sample size for some desired pharmacoepidemiology research
- No single database is ideal for all studies and the suitability of a specific data set should be considered on a case by case basis
• Using multiple databases may be required to study some rare exposure or rare outcomes.

• Careful study specific consideration is needed before using multiple databases for pharmacoepidemiology research
  – There are public-private initiatives such as IMI PROTECT exploring methodological issues in such studies

• Analysis of ‘real world’ observational data is only one essential component of an overall continual assessment of benefit risk