USE OF PRIMARY CARE DATABASES IN EPIDEMIOLOGIC RESEARCH

Luis Alberto García Rodríguez

Spanish Centre for Pharmacoepidemiologic Research (CEIFE), Madrid, Spain
The structure to successfully perform a pharmacoepidemiology study with automated primary care databases can be divided into the following items:

- **Hypothesis**: correct research question
- **Study design**: best suited design
- **Rich data source**: numbers and detail (quantity and quality)
- **Correct definition and classification** of the outcome/exposure
- **Analysis**: adequate analysis plan
- Statement about the relationship between a drug and a disease that can be tested

- **Hypothesis definition requires** the researcher to have knowledge on:
  
  • **Biological mechanism**: detailed knowledge about the study drug and disease; mechanism of action (i.e. whether the effect is acute or chronic, local or systemic, modified by other factors…)
  
  • **Data sources**: detailed knowledge about the source of information (i.e. population included, type of health system, strengths and limitations…)
Study Hypothesis

The **study design and methodological approaches** will be subject to the **study question**

- **When there are prior data available:**
  - this must be considered when conducting a new study
  - especially when a new study aims to replicate what was already observed in a previous one

- **When no prior data are available:**
  - defining the hypothesis might be more challenging
  - the marginal contribution of a single study is the greatest
How to decide the most appropriate design under the observational approach:

- **Cohort**: e.g. follow-up study, long-term outcomes, survival
- **Case-control**: e.g. specific dose/duration effect
- **Case-crossover**: e.g. transient exposures/acute/prompt outcomes
- **Drug utilization**: e.g. treatment patterns, switching patterns, comparison with guidelines recommendations
- **Meta-analysis**: e.g. pooled estimates to reach conclusions on hypothesized association performed by different studies
Traditional “differences” between cohort and nested case-control studies

**COHORT STUDIES**

- Temporal sequence between exposure and outcome can be established
- Useful to study rare exposures, i.e. a specific chemical product
- Multiple outcomes associated with the exposure can be studied
- Information on confounding factors can be obtained
- Useful for estimating the risk of disease, the incidence rate and/or relative risks. Time-to-event analysis is possible as well
- Less prone to selection and information biases as compared to other designs

**NESTED CASE-CONTROL STUDIES**

- They may be less expensive and time consuming than cohort studies
- Rare diseases may be explored
- Diseases with long latency period can be studied
- It is possible to investigate multiple exposures
- When a risk set sampling is used to select the controls, the outcome estimate is similar to the risk ratio
- In "nested" case-control designs, information on exposures have been collected before cases had been diagnosed, and may be less prone to bias.
Clinical studies: where do pharmacoepidemiology (PE) studies fit in?

Clinical studies

- Interventional
  - Open-label
  - RCT

- Observational
  - Database study
  - Registry study
  - Field study
    - Claims/ administrative databases
    - Saskatchewan database
    - Primary care databases
Sources of information in pharmacoepidemiology

• **Field-based studies**: Researcher captures the information directly from the patient
  • Interview
  • Survey
  • Nurses

• **Registries**: Systematic collection of data about specific conditions. There can be population-based as well as hospitalized-based
  • Cancer
  • Pregnancy
  • Autoimmune diseases (Multiple Sclerosis)

• **Automated/computerized database**: Digital version of a paper chart that contains collections of clinical records with a defined structure and purpose. These databases include information at patient-level, on demographics, drug prescriptions, specialist referrals, hospital admissions, hospital discharge diagnoses, operations, ambulatory care
Automated databases in pharmacoepidemiology

- **Claims/administrative databases**: Captured claims data for individuals or multiple insurers, describing all transactions that results in a claim for reimbursement.
  - **Medicaid**: Federal state program for financing medical care to low income population.
  - **Commercial insurance databases**: HealthCore, PharMetrics

- **Saskatchewan databases**: Provincial health system (government held database), universal coverage for most residents, information accumulated on computer. Possibility of linking to other files (death records, cancer registry, mental health services)

- **Primary care databases**: Electronic medical records
Revolution of primary care database started in the ‘80s

Some characteristics converged together to make the first EMR in the UK a possibility

• Universal Health Care system in the UK

• Replacement of medical charts with medical records (Alan Dean, VAMP, eventually EMIS, Vision, etc)

• Dictaphones to record diagnostic and treatment information

• Active counseling and training to primary care physicians willing to participate

• Validation exercises for recorded information to ensure quality of information (Boston Collaborative Drug Surveillance Program (BCDSP))
Pioneers sources of information in pharmacoepidemiology...

- CPRD and THIN (two primary care databases in the UK) are gold-standard sources of information to perform pharmaco-epidemiologic studies

  - Strengths of these resources include:

    i) Highly granulated dictionaries (Read, drug Gemscript)

    ii) Drugs and devices prescriptions automatically recorded

    iii) Completeness of information: primary care practitioners (PCPs) are gatekeepers in the health care system

    iv) Feasibility of validation processes

Use of large databases with access to millions of patients enables the study of rare diseases or outcomes difficult to capture
The Health Improvement Network

- Medical research database of systematically recorded anonymous patient records
- Validated for use in pharmacoepidemiology\(^1\)
- Contains details on close to 4 million patients currently registered with primary care physicians (PCPs) in the UK
- Prospectively recorded information on patients’ demographics, medical history (symptoms, diagnoses), prescriptions, additional health data (laboratory results), details of outpatient visits and hospitalisations
- Free text comments are available
- Subset of practices linked to HES files

Choosing a good data source alone does not guarantee the validity of the study results.

- Transforming data into relevant information regarding the study hypothesis will depend on researcher’s ability to:
  - Extract
  - Process
  - View
  - Analyze

Importance of using specific tools to support efficient data collection and analysis.
Outcome and exposure validation

How to start?

1) **Outcome of interest**

- Relevant *operational definition*: constructing specific *diagnostic algorithms*
- *Validation* of diagnostic algorithms

This validation process should include:

I) **Manual review** of patients’ profiles with anonymized free text comments

II) **Questionnaires** sent to Primary Care Physicians requesting them to send anonymized copies of notes (hospital summaries, discharge/referral letters, reports of diagnostic procedures, etc.)
Outcome and exposure validation

- The full validation process, including both steps, should always be performed when:
  - There is no prior experience with the particular disease in the database
  - Concerns exist regarding our ability to validly identify cases using only coded information

- When validation of the entire case set might not be feasible for logistic reasons:
  - Random subset(s) of all computer-detected cases is required to undergo validation
Validation in primary care databases

The task of transforming these data into relevant information addressing the study hypothesis depends on researchers’ expertise in extracting, processing and analysing the data.

The process of case ascertainment and validation of the outcome of interest is summarised below.

A. Relevant operational definition for case ascertainment

- Feasible operational definition
- Constructing specific diagnostic algorithms
- Objective eligibility criteria

B. Manual review of patient profiles for case validation

- B (i) Manual review of patients’ records (without free-text comments) identified by the algorithms. This is useful for assessing the validity of the initial computer search strategy and algorithms.
- B (ii) Manual review of the patients’ records including anonymized free-text comments related to the outcome of interest. (e.g. PCP notes, hospital summaries, discharge/referral letters, reports of diagnostic procedures)

C. Questionnaires to PCPs for case confirmation

PCP, primary care practitioner
2) Drug exposure

- Classification of individuals or person-time according to recency, dose, and duration of exposure

Extracting information from the database:

- Not always a straightforward process

- Often requires fairly complex programming algorithms, considering:
  - time period elapsed between repeat prescriptions, prescribed daily dose and number of pills in each prescription
  - should be able to deal with partial missing data and to assume some degree of non-compliance
Classification of individuals or person-time according to recency, dose, and duration of exposure:

- might vary between studies depending on the strategy followed
- these differences of drug exposure definition could have an impact on the results

Invalid exposure ascertainment might prevent documenting dose or duration response when inferring causality

Errors in exposure measurement will compromise interpretation of the results
Outcome and exposure validation

Our ability to correctly ascertain drug exposure might be limited by quality of the information in the data source.

For example:

- Drug exposures obtained from biennial self-reported questionnaires (e.g., Health Professionals or Nurses’ Health Study), or
- Drug exposure ascertained at initial discharge from hospital and assumed to remain constant over a long follow-up period.

Primary care databases based on prospective recording avoid the likelihood of recall biases present in sources of information with either retrospective recording or recording based on memory of participants.
Outcome validation: True Positive

Follow up contribution

New Episode

Low dose ASA initiation CVD indication
Outcome validation: False Positive

50088 20/02/2007  Telephone encounter FROM HUSBAND, GETTING [OF] <9N31.00>
ON WELL, SPEECH IS COMING BACK, WANTS TO COME HOME? FIT ENOUGH, HE WILL SEE SPECIALIST ON THURSDAY

50091 23/02/2007  CAT Scan <567..13> computerised tomography
50091 23/02/2007  Echocardiogram <5853.11> Echocardiogram

50091 23/02/2007  Letter from specialist CLINICAL LETTER OTHER [O] <9N16.00>
HOSPITAL GASTROENTEROLOGY

50091 23/02/2007  Letter from specialist CLINICAL LETTER OTHER [O] <9N16.00>
HOSPITAL GASTROENTEROLOGY

50092 24/02/2007  Discharged from hospital DISCHARGE [O] <8H8..00>
LETTER/SUMMARY OTHER HOSPITAL

50092 24/02/2007  Discharged from hospital DISCHARGE [O] <8H8..00>
LETTER/SUMMARY OTHER HOSPITAL

50097 1/03/2007  Urine test Urine test <46..11> Urine exam - general
50097 1/03/2007  Urine test Urine protein test <46H..00> Urine protein test
50097 1/03/2007  Urine test Culture sensitivity <4J15..11> Culture sensitivity
50097 1/03/2007  Urine test Urine dipstick for nitrite <46X2..00> Urine dipstick for nitrite

50097 1/03/2007  Blood pressure DIASTOLIC 70 SYSTOLIC 130 <246..00> O/E = blood pressure reading

50097 1/03/2007  Urinalysis - general Urinalysis - general [OF] <461..00>
LEUKS ETC, TO LAB

50097 1/03/2007  Stroke and cerebrovascular accident unsp REFER CAUSE CLINIC, [OF] <666..00>
AND DAY HOSPITAL

50097 1/03/2007  <93619997> SIMVASTIN 1 EVERY NIGHT D=0 N=56
50097 1/03/2007  <97217998> BENDROFLUMETHIAZIDE 1 EVERY MORNING D=0 N=56
50097 1/03/2007  <96550997> ARIDARONE 1 EVERY MORNING D=0 N=56
50097 1/03/2007  <98762997> ASPIRIN 1 EVERY MORNING D=0 N=56
50097 1/03/2007  <89659997> BISOPROLOL FUMARATE 1 EVERY MORNING D=0 N=56

50098 2/03/2007  Other rehabilitation [6 S] <8F...00>

50102 5/03/2007  Other rehabilitation

50103 7/03/2007  Stroke and cerebrovascular accident unsp [R S] <966..00>

50104 8/03/2007  Seen in speech and language clinic CLINICAL LETTER [O] <9N0Q..00>
[DELETED] FT SPEECH AND LANGUAGE THEREBY

50105 9/03/2007  Urine tests [O] <46..11>
PRESCRIPTION/[DELETED]

50108 12/03/2007  Encounter administration [O] <9N11..00>
PATIENT HAS APPT WITH SPEECH THERAPY AT 9 AM AND APPT IN DR [DELETED] CLINIC AT 11 AM ON WEDNESDAY. [DELETED] SECRETARY, WANTS TO KNOW WHICH APPT SHOULD BE GIVEN PRIORITY AS IT WON'T BE POSSIBLE FOR HER TO ATTEND BOTH. [DELETED] GIVEN MESSAGE TO CALL [DELETED]

50109 13/03/2007  Blood pressure DIASTOLIC 60 SYSTOLIC 110 <246..00> O/E = blood pressure reading
# Over-the-counter low-dose aspirin use
## Questionnaire

<table>
<thead>
<tr>
<th>Practice ID</th>
<th>Patient ID</th>
<th>Sex</th>
<th>Age on 1/1/2013</th>
</tr>
</thead>
</table>

Please answer all the questions according to the information held in the patients written records by ticking the appropriate box or writing in the space provided.

1. Is the patient currently taking low-dose aspirin or in the last three months?  
   - Yes  
   - No  
   - Unknown

   **IF YES:**  
   - Are they taking prescription aspirin or over-the-counter aspirin?  
   - How long have they been taking aspirin?  
   - What is the indication for low-dose aspirin use?  
     - Primary Prevention  
     - Myocardial Infarction  
     - Unstable Angina  
     - Revascularization  
     - Cerebrovascular disease (including stroke or transient ischaemic attack)  
     - Other (please specify)  

   **IF NOT:**  
   - Did the patient ever take low-dose aspirin?  
     - Yes  
     - No  
     - Unknown

   **If yes:**  
   - Were they taking prescription aspirin or over-the-counter aspirin?  
   - When did they stop taking low-dose aspirin?  
   - What was the reason to stop taking low-dose aspirin?  
     - Adverse event (list which)  
     - Change antiplatelet  
     - Other

   - What is the indication for low-dose aspirin use?  
     - Primary Prevention  
     - Myocardial Infarction  
     - Unstable Angina  
     - Revascularization  
     - Cerebrovascular disease (including stroke or transient ischaemic attack)  
     - Other (please specify)

2. Is the patient currently or in the last three months taking over-the-counter one of the following drugs?  
   a. Proton Pump Inhibitors  
      - Yes  
      - No  
      - Unknown
   
      Please specify name of individual PPI:  

   b. H2 receptor antagonists  
      - Yes  
      - No  
      - Unknown
   
      Please specify name of individual H2RA:  

   c. NSAIDs  
      - Yes  
      - No  
      - Unknown
   
      Please specify name of individual NSAID:  

Impact of case validation in THIN

Incidence of major bleeds among combined study cohorts

Including all bleeds identified by Read codes leads to misclassification of outcome and corresponding over-estimation of incidence of major bleeds

IC, GI and UG bleeds identified by Read codes (incidence per 1,000 person-years)

- IC bleeds: 3.16
- GI bleeds: 20.51
- UG bleeds: 24.49

Manual review of patients’ complete records with free-text comments

- Major\(^a\) IC bleeds: 3.06
- Major\(^a\) GI bleeds: 9.35
- Major\(^a\) UG bleeds: 11.13

- Hospitalized IC bleeds: 3.06
- Hospitalized GI bleeds: 4.51
- Hospitalized UG bleeds: 4.09

\(^a\)Major bleed = requiring referral or hospitalization

GI, gastrointestinal; IC, intracranial; UG, urogenital
Use of large databases with access to millions of patients enables the study of rare diseases or outcomes difficult to capture

Yet caution is needed:

- This phenomenal statistical power ensures precision on study results irrespective of their validity; this “amplifying” effect is particularly pernicious when erroneous data management or study design led to invalid estimates

- Requires careful planning and data analysis by experienced researchers

- Facilitate methodological mistakes translate into flawed conclusions
Good hypothesis, good data source, and a good design might not be enough... *we might spoil it in the analyses*

- **Issues to take into account:**
  - **Discuss an analysis plan**
  - **Minimizing confounding**
    - Identify all potential confounders
    - Use the most adequate statistical test and other analyses such as stratification
  - **Sensitivity analysis**

In general, the use of complicated analytical methods is not recommended, unless they are well understood and *add* documented relevant contributions to conventional analyses.
• Observational studies that use electronic medical records provide valuable data and can be just as important as RCTs

• UK primary care data are of global significance

• The quality/experience of the researchers, as well as the quality of the data source, is key