Challenges faced by Europe in implementing a CDM

Olaf Klungel

Utrecht Institute for Pharmaceutical Sciences (UIPS), Utrecht University & Julius Center for Health Sciences and Primary Care, UMCU.

Pharmacoepidemiology & Clinical Pharmacology

Utrecht University

Disclosure

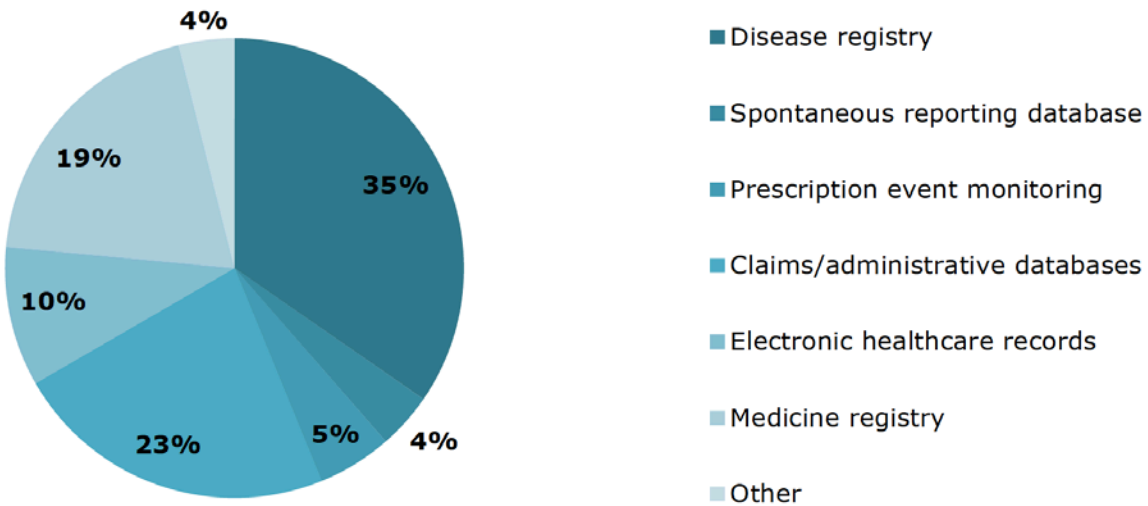
- The division of pharmacoepidemiology has received research grants from the Innovative Medicines Initiative (IMI-PROTECT, IMI-EU2P), GSK (HTA methodological research), Lygature (public private partnership with EFPIA/EBE).
- Educational lecture fee from Roche

Outline

- Overview and evolution of multi-database studies in EU
- Scientific challenges distributed data networks and CDM
  - Design
    - Selection bias
    - ***Information bias***
    - ***Confounding bias***
  - Analysis
    - Effect estimation
    - Control for Confounding
  - Reporting

ENCePP inventory of data sources

104 Data sources (Sep 2017)

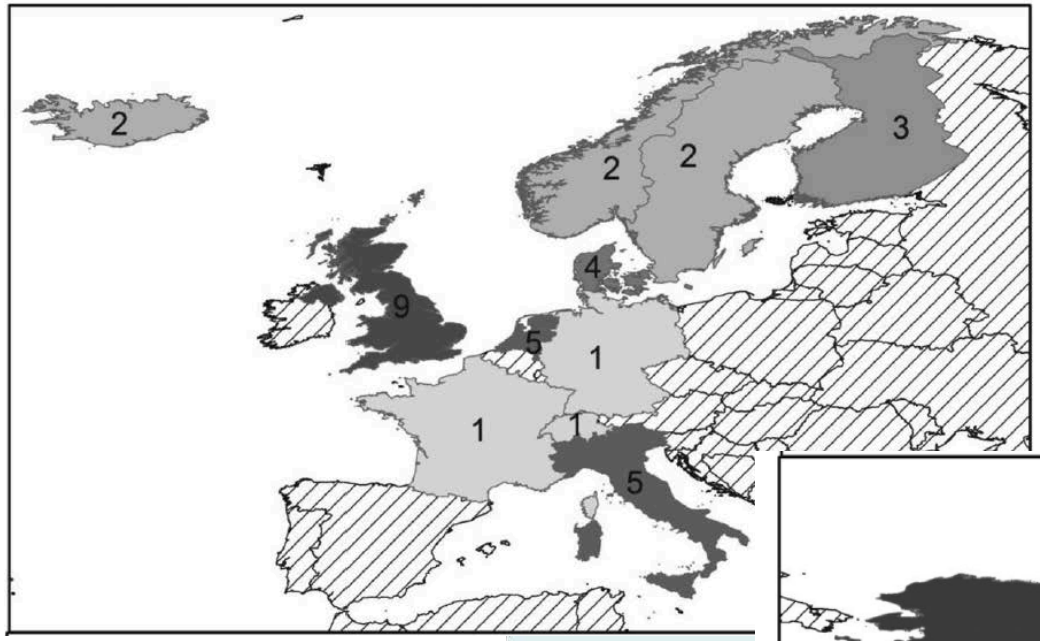


Presented by S. Perez-Gutthan at 10th Anniversary of ENCePP

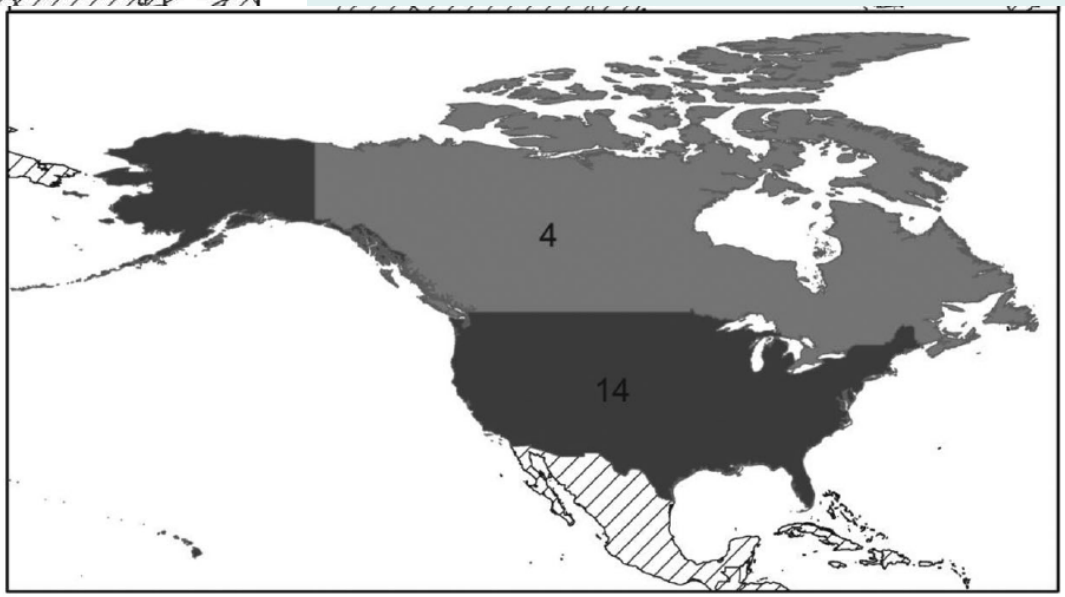
Characteristics of selected EU healthcare databases

Database	Country	Cumulative population (2008)	Data source	Coding diagnoses	Free text	Coding drugs	Coding product	Recording of drug use
BIFAP	ES	7.5 M	GP	ICPC-2, ICD-9	Spanish	ATC	CNF	Prescribing
SIDIAP	ES	7.0 M	GP	ICD-10	No	ATC	-	Prescribing
ARS	IT	4.0 M	Hospital claims/death	ICD-9-CM	No	ATC		Dispensing
Health Search Italy	IT	1.0 M	GP	ICD-9-CM	Italian	ATC	Brand names	Prescribing
CPRD	UK	12.5 M	GP	READ	English	BNF	Prod code	Prescribing
THIN	UK	7.8 M	GP	READ	English	BNF	Prod code	Prescribing
IPCI AHC	NL NL	0.75 M 0.26 M	GP GP/Pharmacy	ICPC ICPC	Dutch Dutch	ATC ATC	HPK HPK	Prescribing Prescribing + dispensing
PHARMO	NL	3 M	Pharmacy/Hospital/Laboratory/GP	ICD-9-CM, ICPC	Dutch	ATC	HPK	Prescribing /dispensing
The Danish national registries	DK	5.2 M	Hospital/ Pharmacy/death	ICD-8/9/10	No	ATC	Varenr	Dispensing
Bavarian Claims	DE	10.5 M	Claims	ICD-10-GM	No	ATC	PZN	Dispensing
AOK Nordwest	DE	2.7 M	Claims	ICD-10-GM	No	ATC	PZN	Dispensing
EGB	FR	0,7/60 M	Claims	ICD-10	No	ATC	CIP13	Dispensing

MULTI-DATABASE STUDIES: A SYSTEMATIC LITERATURE REVIEW



8 studies



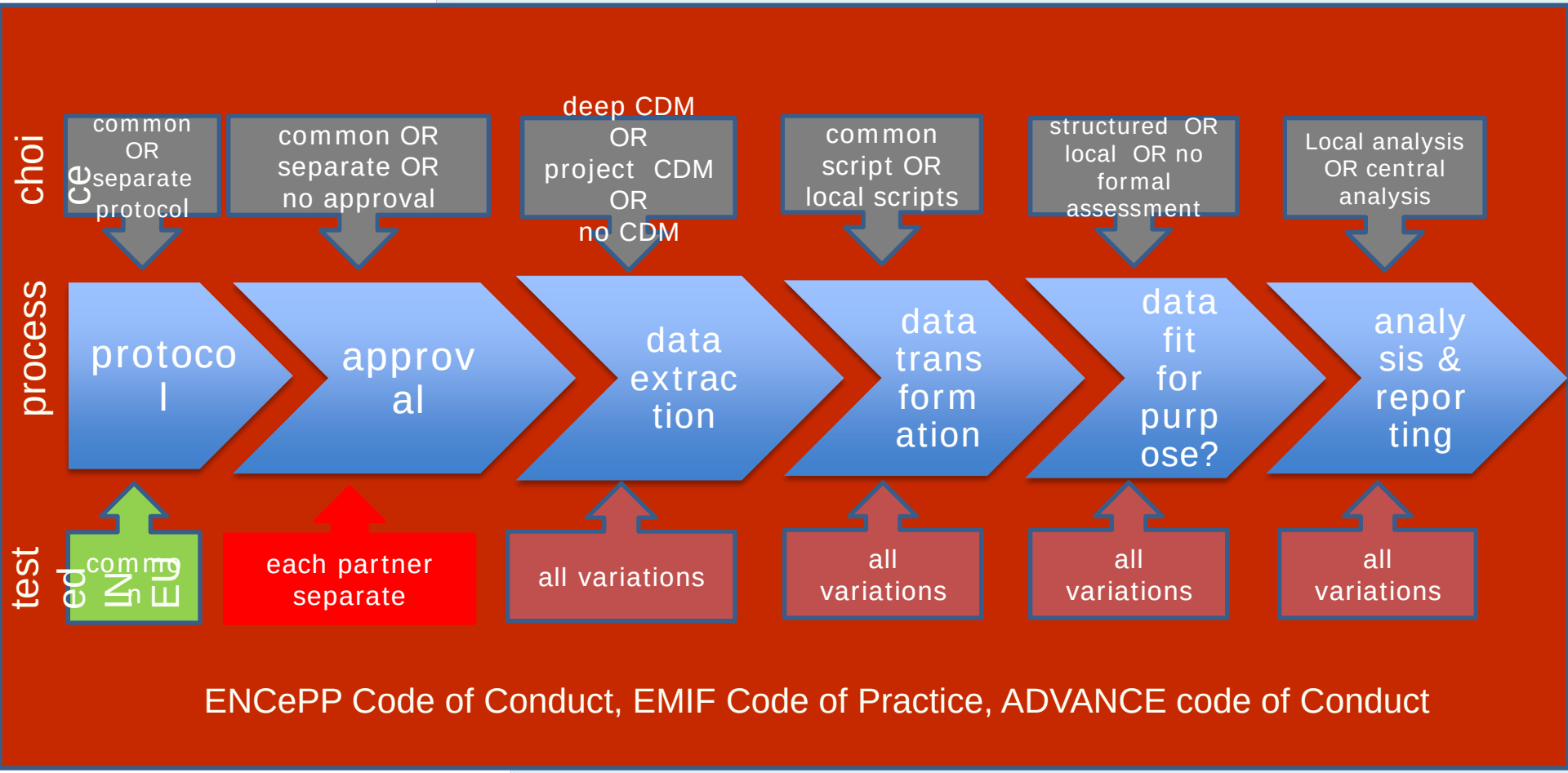
14 studies

Pharm

Bazelier MT, et al. Pharmacoepidemiol Drug Saf 2015;24:897-905

Utrecht University

Process flow for multi-site drug safety studies in EU



Courtesy: M.c.j.Sturkenboom@umcutrecht.nl

‘Increasing harmonization’: the evolution

across FP-7 & IMI EU-funded drug safety projects

Common protocol, no CDM

- PROTECT, GetReal

Project based CDM

- EU-ADR (2008), SOS (2009), ARITMO (2010), VAESCO (2010), SAFEGUARD (2011).... EMIF (2012), ADVANCE (2013), CARING (2011)

Common scripts: Jerboa, SAS, R

- EU-ADR (2008), SOS (2009), ARITMO (2010), VAESCO (2010), SAFEGUARD (2011).... EMIF (2012), ADVANCE (2013), CARING (2011)

datawarehouse & pooling

- SOS project (2009) Datawarehouse at Milano-Bicocca, CARING (2011) pooling and centralised analysis at Statistics Denmark

ADVANCE Codemapper

- ADVANCE project (2013)

OMOP CDM test

- EMIF project (2017), IMI Bigdata@heart (2017)

Courtesy: m.c.j.sturkenboom@umcutrecht.nl

Information bias

- Misclassification of outcomes and exposures due to loss of information in mapping to a CDM
  - No mapping possible to standard vocabulary CDM
  - Different granularity source codes
  - Free text source
- Non-differential => bias towards null
- Example of acute liver injury
  - Sentinel CDM: ICD-9-CM codes
  - OMOP CDM: ICD-9-CM, LOINC codes, laboratory tests
  - PROTECT: CPRD (Read codes, laboratory tests), BIFAP (ICPC codes, laboratory tests, free text)

Classification of ALI in PROTECT

Eur J Clin Pharmacol (2014) 70:1227–1235 1229

Table 1 Computer search algorithms to ascertain acute liver injury. Operational case definition

Case status	Ia. Diagnosis of liver injury or symptoms recorded by specific codes or text[a] for liver injury	Ib. Diagnosis of liver injury or symptoms recorded by unspecific codes or text[a] indicating only positive results for liver tests	II. Complete laboratory criteria: an increase of more than two times ULN in ALT or a combined increase in AST, AP and total bilirubin provided one of them was two times ULN within 2 months of the event	III. A referral to a specialist orhospital within 2 weeks of a recorded diagnosis of liver injury
Definite	Yes	No	Yes	Yes
Probable A	Yes	No	Yes	No
Probable B	No	Yes	Yes	Yes
Possible	No	Yes	Yes	No
No case	Yes	No	No (normal LFTs or just increased values not with complete criteria)	Yes
	No	Yes	No (normal LFTs or just increased values not with complete criteria	No

ULN upper limit of normal, *ALT* alanine aminotransferase, *AP* alkaline phosphatase, *AST* aspartate aminotransferase

[a] In BIFAP, database ICPC codes were used along with computer search of keywords in text

Validation of ALI in PROTECT

Table 2 Computer case ascertainment and manual review process in BIFAP database

Pre-review computer case status[a]	Patients with a ICPD code of ALI (N=19,074)	Status after manual review of free text					
		Definite-confirmed	%	Probable-confirmed	%	No-case confirmed[b]	%
1. Definite	179	43	24.02	19	10.61	117	65.36
2. A-Probable A	119	14	11.76	22	18.49	83	69.75
2. B-Probable B	1,038	51	4.91	122	11.75	865	83.33
3. Possible	1,537	16	1.04	149	9.69	1372	89.26
4. No case	16,201	Manually reviewed a sample n=120, 100 % no case					

[a] As in Table 1

[b] Reason for exclusion during manual review: other liver disease (691), cancer (23), alcohol-related problems (186), gallbladder and pancreatic disease (120), routine testing (1,322) or not confirmed cases (95)

Eur J Clin Pharmacol (2014) 70:1227–1235

Table 3 Computer case ascertainment and manual review process in CPRD database

Pre-review computer case definition[a]	Sample cases to review	Status after manual review	
		Confirmed	%
1. Definite	101 (47 with free text)	64	63.4
2. Definite+probable	208 (59 with free text)	122	58.6

[a] As in Table 1

Outcome definition and rates of ALI

Fig. 1 Incidence rate of acute liver injury (ALI) in BIFAP and in CPRD (*dashed line*) by age and sex. Using the ALI narrow definition (**a**) and broad definition (**b**)

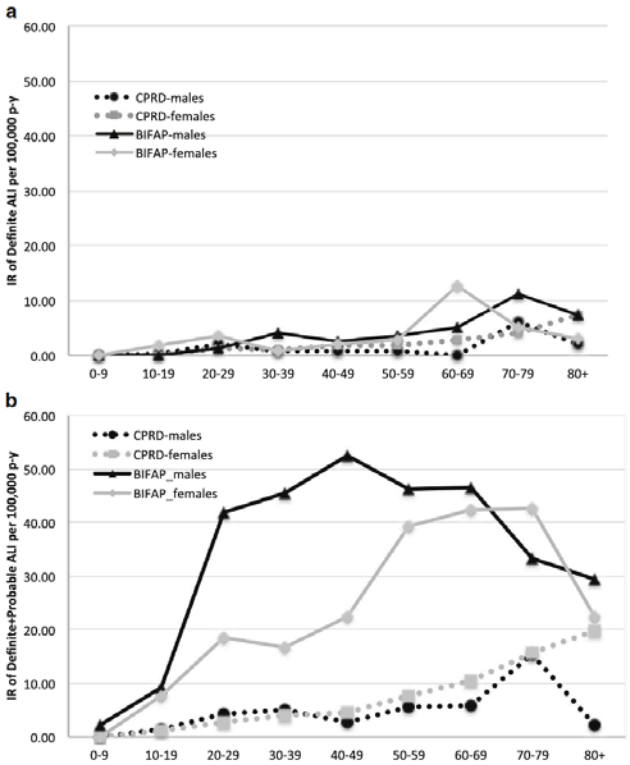
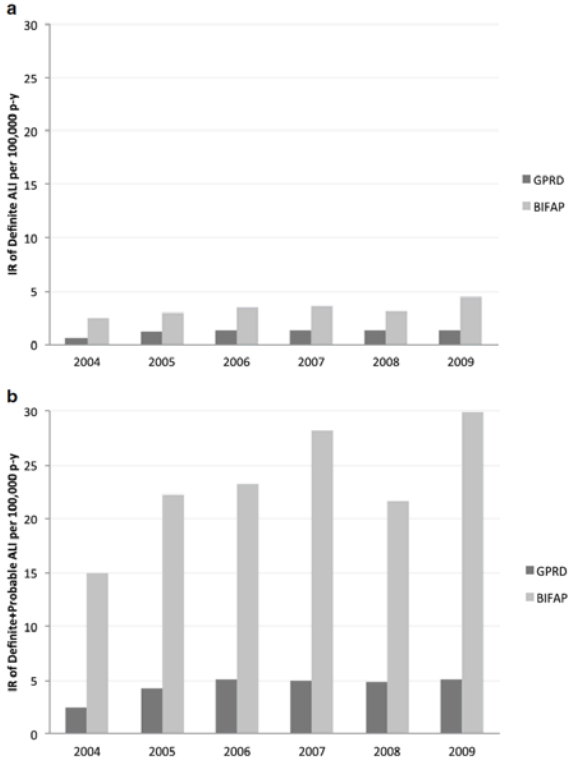


Fig. 2 Incidence rate of acute liver injury in BIFAP and in CPRD by year (standardised by age and sex EURO weights 2008). Using the ALI narrow definition (**a**) and broad definition (**b**)



Outcome definition and RR of ALI associated with antibiotic use

	Cohort		Case-control	
	CPRD	BIFAP	CPRD	BIFAP
Definite	10.0 [7.0-14.0]	5.8 [3.5-9.6]	5.7 [3.5-9.4]	2.6 [1.3-5.4]
Definite+ probable	8.3 [6.8-10.1]	5.1 [3.8-6.8]	3.6 [2.8-4.6]	3.1 [2.1-4.6]

Brauer R, et al. Pharmacoepidemiol Drug Saf 2016;25 (Suppl 1):29-38

Impact of exposure misclassification

- Incomplete mapping to OMOP CDM
  - 10,3% of drug exposure records in CPRD[1]
  - 7% of drug exposure records (55% of exposure terms) in THIN[2]
- Complex exposure definitions require adaptation to specific study/database

1 Matcho M, et al. Drug Saf 2014;37:945-959
2 Zhou X, et al. Drug Saf 2013;36:119-34.

Impact of confounder misclassification

- Incomplete mapping to OMOP CDM
  - 0,15% of condition records, 2,3% of procedure records in CPRD[1]
  - 6% of condition records (25% of condition terms), 4% of procedures in THIN[2]
- Residual confounding due to incomplete measurement of confounding factors

1 Matcho M, et al. Drug Saf 2014;37:945-959
2 Zhou X, et al. Drug Saf 2013;36:119-34.

Impact of confounder misclassification

- Impact depends on:
  - strength of association between confounder-outcome and confounder-exposure
  - Type B vs Type A adverse drug reaction, intended effects

- Multilevel multiple imputation before transformation to CDM?[1]

1 Jolani S, et al. Stat Med 2015;34:1841-63.

Data collection and analytical options

1. Aggregate level approach (e.g. PROTECT, CNODES)
   - No sharing of individual patient data
   - Overall results are collected for meta-analysis
   - Allows optimization for individual database

2. Semi-aggregate level approach (e.g. EU-ADR, CARING, SENTINEL)
   - Stratified datasets collected from all databases
   - Outcomes, Exposure, Covariate patterns
   - One common analysis

3. Individual level approach (e.g. NORPEN)
   - Individual patient data collected from all databases for one common analysis

1. Aggregate level analysis

- Decentral analysis
- Control for confounding
  - Conventional Multivariable Regression
    - Common set of confounders
    - Additional adjustment in individual databases with maximum amount of information
  - High dimensional Propensity Score
  - Disease Risk Scores
  - Distributed regression

Collaboration EMA-Health Canada

- Framework contract EU PE&PV (former PROTECT consortium)
  - 8 EU databases, ~47 M patients
- "Characterising the risk of major bleeding in patients with Non-Valvular Atrial Fibrillation: non-interventional study of patients *taking Direct Oral Anticoagulants in the EU"*
- Common protocol, statistical analysis plan/programmning instructions, no CDM
- Replicate findings in Canadian Network of Observational Drug Effect Studies (CNODES)
- Which CDM if replication is needed?

2. Semi-aggregate level analysis

- Datasets collected from each database stratified on
  - Outcome
  - Exposure
  - Confounders

- Central privacy preserving analysis on semi-aggregated dataset
  - Control for confounding limited by number of confounders (e.g. propensity score) stratified on
  - Case-centered logistic regression

3. Individual patient level analysis

- Individual patient data collected from each database on
  - Outcome
  - Exposure
  - Confounders

- Central analysis on individual patient dataset
  - Control for confounding limited by number of confounders that are common to each database
  - Can be complemented by meta-analysis utilizing site-optimized estimates

Reporting of (multi-)database studies

- **The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement for Pharmacoepidemiology (RECORD-PE)**
- Developed as an extension of the existing STROBE guidelines (**ST**rengthening the **R**eporting of **OB**servational studies in **E**pidemiology), with the overall goal to enhance transparency by providing researchers with the minimum reporting requirements needed to adequately convey the methods and results of their research.

http://www.record-statement.org

Reproducability and replicability

Received: 21 July 2017 | Revised: 25 July 2017 | Accepted: 25 July 2017
DOI: 10.1002/pds.4295

WILEY

ORIGINAL REPORT

Reporting to Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies V1.0

Shirley V. Wang[1,2] | Sebastian Schneeweiss[1,2] | Marc L. Berger[3] | Jeffrey Brown[4] | Frank de Vries[5] | Ian Douglas[6] | Joshua J. Gagne[1,2] | Rosa Gini[7] | Olaf Klungel[8] | C. Daniel Mullins[9] | Michael D. Nguyen[10] | Jeremy A. Rassen[11] | Liam Smeeth[6] | Miriam Sturkenboom[12] |
on behalf of the joint ISPE-ISPOR Special Task Force on Real World Evidence in Health Care Decision Making

Value in Health 2017;20:1009-22
Pharmacoepidemiol Drug Saf 2017;26:1018-32

Reproducability and replicability

TABLE 1 Reproducibility and replicability

		Data	Methods
Reproducibility	Direct replication *Reproduction of a specific study*	**Same**	**Same**
	Conceptual replication *Reproduction of a finding for the exposure (and comparator), outcome and estimand of interest*	**Different** **Same** **Different**	**Same** **Different** **Different**

Value in Health 2017;20:1009-22
Pharmacoepidemiol Drug Saf 2017;26:1018-32

TABLE 2 Reporting specific parameters to increase reproducibility of database studies*

	Description	Example	Synonyms
A. Reporting on data source should include:			
A.1 Data provider	Data source name and name of organization that provided data.	Medicaid Analytic Extracts data covering 50 states from the Centers for Medicare and Medicaid Services.	
A.2 Data extraction date (DED)	The date (or version number) when data were extracted from the dynamic raw transactional data stream (e.g. date that the data were cut for research use by the vendor).	The source data for this research study was cut by [data vendor] on January 1st, 2017. The study included administrative claims from Jan 1st 2005 to Dec 31st 2015.	Data version, data pull
A.3 Data sampling	The search/extraction criteria applied if the source data accessible to the researcher is a subset of the data available from the vendor.		
A.4 Source data range (SDR)	The calendar time range of data used for the study. Note that the implemented study may use only a subset of the available data.		Study period, query period
A.5 Type of data	The domains of information available in the source data, e.g. administrative, electronic health records, inpatient versus outpatient capture, primary vs secondary care, pharmacy, lab, registry.	The administrative claims data include enrollment inforn outpatient diagnc procedure (ICD9/ well as outpatient for 60 million live The electronic he diagnosis and prc records, problem prescription and l results, inpatient well as unstructur notes and reports encounters at AB system.	
A.6 Data linkage, other supplemental data	Data linkage or supplemental data such as chart reviews or survey data not typically available with license for healthcare database.	We used Surveillance, Epidemiology, and End Results (SEER) data on prostate cancer cases from 1990 through 2013 linked to Medicare and a 5% sample of Medicare enrollees living in the same regions as the identified cases of prostate cancer over the same period of time. The linkage was created through a collaborative effort from the National Cancer Institute (NCI), and the Centers for Medicare and Medicaid Services (CMS).	
A.7 Data cleaning	Transformations to the data fields to handle missing, out of range values or logical inconsistencies. This may be at the data source level or the decisions can be made on a project specific basis.	Global cleaning: The data source was cleaned to exclude all individuals who had more than one gender reported. All dispensing claims that were missing day's supply or had 0 days' supply were removed from the source data tables. Project specific cleaning: When calculating duration of exposure for our study population, we ignored dispensation claims that were missing or had 0 days' supply. We used the most recently reported birth date if there was more than one birth date reported.	
A.8 Data model conversion	Format of the data, including description of decisions used to convert data to fit a Common Data Model (CDM).	The source data were converted to fit the Sentinel Common Data Model (CDM) version 5.0. Data conversion decisions can be found on our website (http://ourwebsite). Observations with missing or out of range values were not removed from the CDM tables.	

Conclusions

- Characterise loss-of-information when different EU databases are transformed into CDM
- Assess impact of transformation into CDM on effect estimates from analytic studies
  - Empyrical studies comparing original database studies vs CDM based studies
- Complete CDM (eg OMOP) for all EU databases versus basic CDM for EU databases enhanced with study/database specific variables
- Further development and assessment of analytic methods for distributed data networks/multi-database studies

Key publications regarding methods & tools

- Trifirò G, Coloma PM, Rijnbeek PR, Romio S, Mosseveld B, Weibel D, Bonhoeffer J, Schuemie M, van der Lei J, Sturkenboom M. **Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how?** J Intern Med. 2014 Jun;275(6):551-61.
- Gini R, Schuemie M, Brown J, Ryan P, Vacchi E, Coppola M, Cazzola W, Coloma P,Berni R, Diallo G, Oliveira JL, Avillach P, Trifirò G, Rijnbeek P, Bellentani M, van Der Lei J, Klazinga N, Sturkenboom M. **Data Extraction and Management in Networks of Observational Health Care Databases for Scientific Research: A Comparison of EU-ADR, OMOP, Mini-Sentinel and MATRICE Strategies**. EGEMS (WashDC). 2016 Feb 8;4(1):1189.
- Klungel OH, Kurz X, de Groot MC, Schlienger RG, Tcherny-Lessenot S, GrimaldiL, Ibáñez L, Groenwold RH, Reynolds RF. **Multi-centre, multi-database studies with common protocols: lessons learnt from the IMI PROTECT project.** Pharmacoepidemiol Drug Saf. 2016 Mar;25 Suppl 1:156-65.
- Bazelier MT, Eriksson I, de Vries F, Schmidt MK, Raitanen J, Haukka J,Starup-Linde J, De Bruin ML, Andersen M. **Data management and data analysis techniques in pharmacoepidemiological studies using a pre-planned multi-database approach: a systematic literature review**. Pharmacoepidemiol Drug Saf. 2015Sep;24(9):897-905.
- But A, de Bruin ML, Bazelier MT,
- Becker BFH, Avillach P, Romio S, van Mulligen EM, Weibel D, Sturkenboom MCJM, Kors JA; ADVANCE consortium. **CodeMapper: semiautomatic coding of case definitions. A contribution from the ADVANCE project**. Pharmacoepidemiol Drug Saf.2017 Aug;26(8):998-1005.
- Kurz X, Bauchau V, Mahy P, Glismann S, van der Aa LM, Simondon F; ADVANCE consortium. **The ADVANCE Code of Conduct for collaborative vaccine studies**. Vaccine. 2017 Apr 4;35(15):1844-1855.