

Frameworks for Data Extraction and Management from Electronic Healthcare Databases for Multi-Center Epidemiologic Studies: a Comparison among EU-ADR, MATRICE, and OMOP Strategies

Rosa GINI, Massimo COPPOLA, Patrick B RYAN, Giacomo RIGHETTI, Iacopo PERI, Roberto BERNI, Paul AVILLACH, Preciosa M COLOMA, Gianluca TRIFIRÒ, Gayo DIALLO, Johan VAN DER LEI, Miriam CJM STURKENBOOM and Martijn J SCHUEMIE

Pisa, MIE2012, August 2012

Speakers

Speaker	Institution	Presents
Rosa Gini	Agenzia regionale di sanità della Toscana (ARS) & Erasmus University Medical Center (EMC)	Moderator, comparison
Gayo Diallo	Université Bordeaux	General process
Preciosa M Coloma (Patrick B Ryan)	EMC Observational Medical Outcomes Partnership (OMOP)	EUADR Project OMOP (RG)
(Martijn J Schuemie)	EMC & OMOP	OMOP (RG)
Massimo Coppola (Roberto Berni)	Consiglio Nazionale Ricerche ARS	MATRICE MATRICE (RG)

Outline

Background

Conceptual framework

General process

The three projects

Specific aspects

Comparison on steps of the process

Final considerations

Contents

Background

Conceptual framework

General process

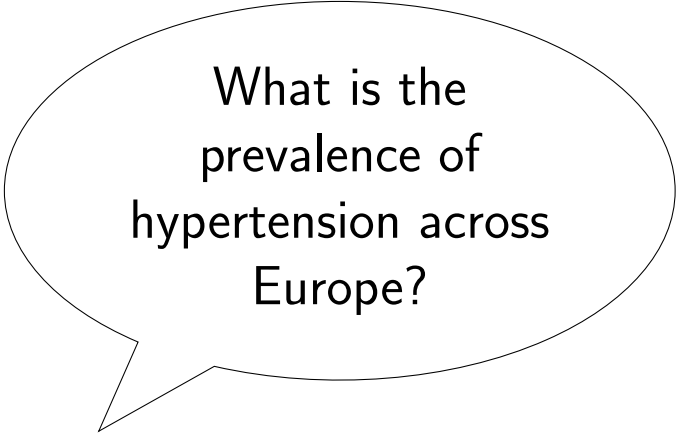
The three projects

Specific aspects

Comparison on steps of the process

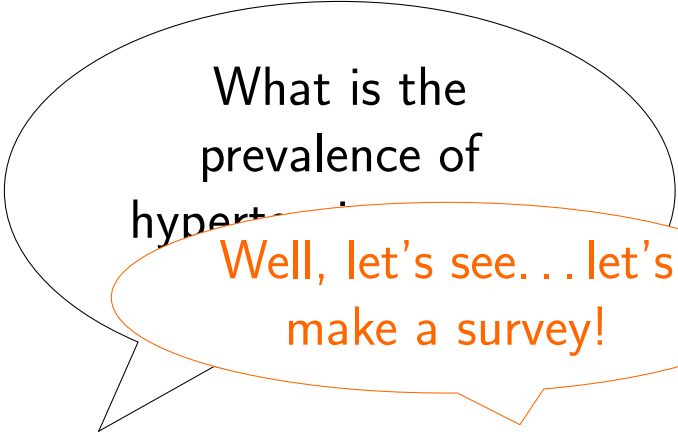
Final considerations

Questions




What is the prevalence of hypertension across Europe?

Questions



What is the
prevalence of
hypertension?



Well, let's see... let's
make a survey!

Questions

What if...

It will take years, by the way – when was it that you needed the information exactly?

Questions

Ooops... and if I
wanted to follow up
the patients to
compare the quality
of care they receive?

exactly?

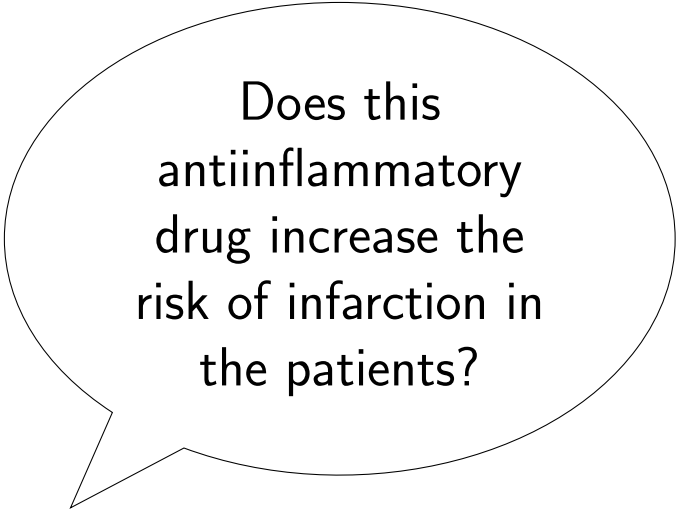
Questions

Ooops... and if I
wanted to follow up
the...

Mmmmh, then I
guess you are in
trouble...

exactly

Questions



Does this
antiinflammatory
drug increase the
risk of infarction in
the patients?

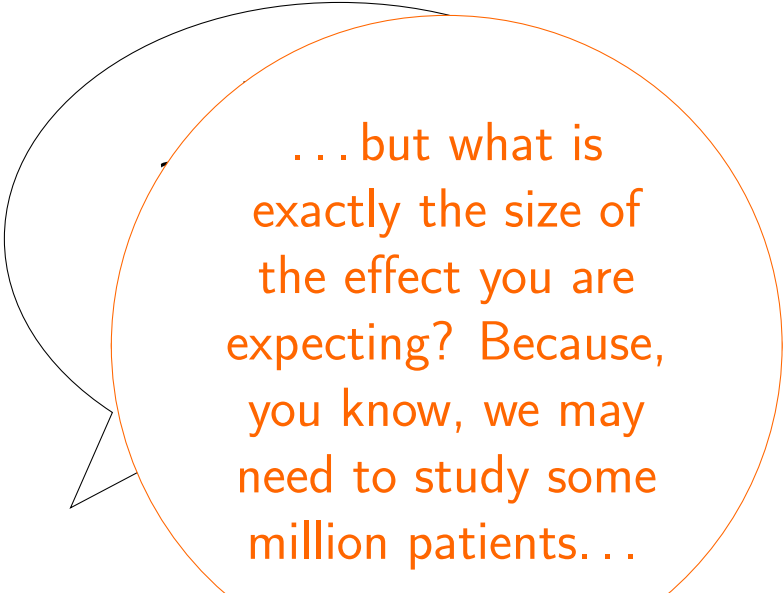
Questions

Does this
antiinflammatory
drug increase the

risk?

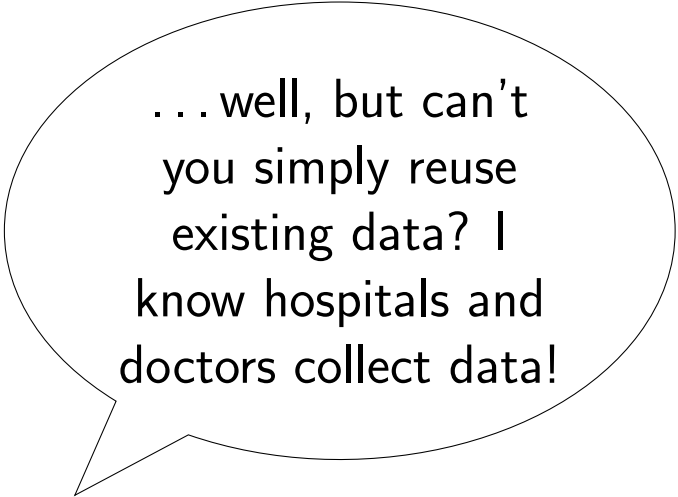
Well, let's perform a
drug safety study!

Questions



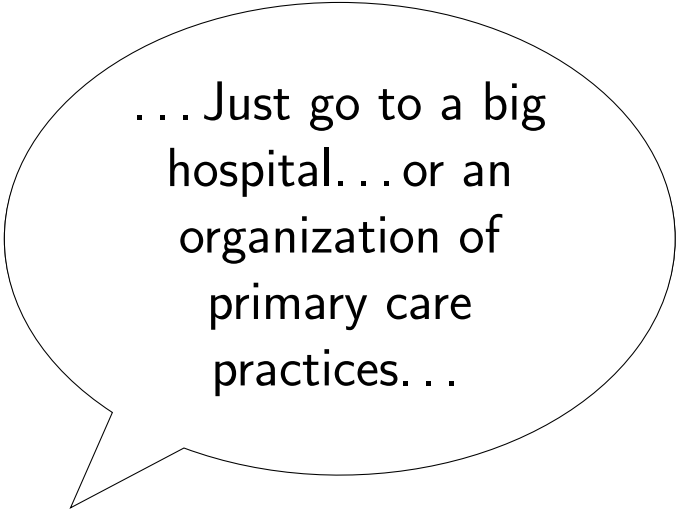
... but what is exactly the size of the effect you are expecting? Because, you know, we may need to study some million patients...

Opportunity



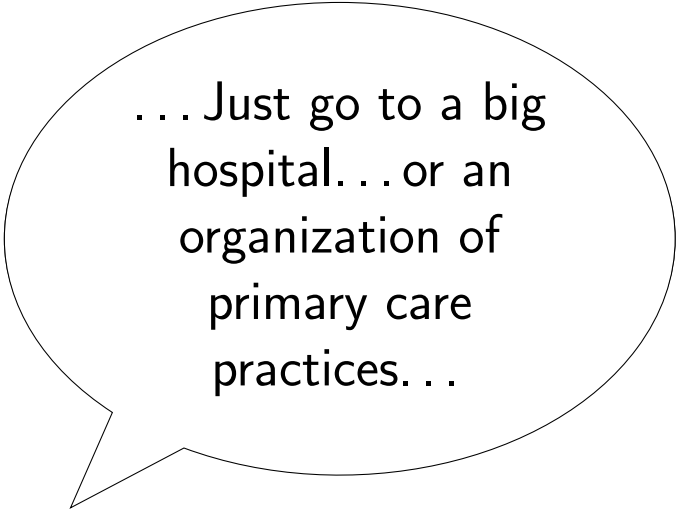
... well, but can't
you simply reuse
existing data? I
know hospitals and
doctors collect data!

Opportunity



... Just go to a big hospital... or an organization of primary care practices...

Opportunity



... Just go to a big hospital... or an organization of primary care practices...

Challenge

Ok, but, beware that
we will need
hundreds of hospitals
and thousands of
primary care
practices to obtain
the data you need...

Challenge

Ok, but what
... and they need to
be scattered around
countries, and they
record data in
different ways...

the

...

Challenge

Ok, but what

... recall that hospitals will not see patients just because they are hypertensive and primary care practices might fail to record emergency admissions. . .

the

. . .

Challenge

... moreover the
data were not
collected for us by
those guys, so we
cannot impose
anything on the way
they collect them!

Challenge

We will need to use
a flexible data
collection approach
then!

we
cannot impose
anything on the way
they collect them!

Background

- Epidemiology** empirical science that draws from data information on
- health phenomena geographic/temporal distribution
 - health phenomena causation
 - broadly, health services functioning

Causal inference Support for an *exposure* (ie previous event) causing an *outcome* (ie subsequent event) comes if observed distribution of outcome is different across exposure strata, under specific conditions upon data generation mechanism called *study designs*

Observational Unless exposure is assigned by the investigator the study is called *observational*, and to infer causality exposure and outcome must be observed across presence/absence of other relevant events (*covariates*)

Data reuse

Traditional data collection Data are freshly collected from direct clinical observation, questionnaires, environmental measurements. . .

Traditional data reuse Secondary use of available information, such as routinely collected demographic information, has always been performed as well.

Electronic data reuse Recently, regular electronic storage of information has made available huge amount of possibly useful data

Data reuse from heterogeneous data sources The challenge of using heterogeneous electronic data sources has been faced in the last few years

Heterogeneous Data stored in different places, collected for different reasons, in different formats and coding systems

System heterogeneous Different programming or querying languages or different models or data management systems

General strategies for data integration

Federated database Retrieve and store in the same place all necessary data. Feeding the central repository is generally done in batch mode through an Extraction, Transformation and Load (ETL) module. The datawarehouse needs to be regularly refreshed as the local data can be updated.

Mediator based integration A global schema is defined, linked to data stored into the local databases through a set of mappings. Adapters: dedicated components which play a role close to ETL: translate a global query into a specific local language. Mediator: component dedicated to querying the system and the global schema as well as managing information about each local schema.

Different kinds of mappings arity can be considered: 1:1, n:1, 1:n, n:n

Adapting the general strategy to the case multi-center epidemiologic studies: specific issues

Events Some crucial data are not collected as such and need to be inferred: eg a person in the study population having hypertension or an infarction – eg a person failing to be inferred as being hypertensive will be implicitly considered to be free from hypertension

Static vs dynamic data transformation While the information recorded in the DB can be statically projected to a common data model/schema, events require a dynamic transformation

Sophisticated mappings Some event definition rely on sophisticated algorithms that map **sequences of longitudinal recordings** to the event: for instance in an administrative database a person being prescribed antidiabetic drugs for some time is mapped to "Diabetes"

Validation Validation is needed

Ethical/legal issues In some cases data cannot be shared unless aggregated, so the whole data management necessary to

Focus

This talk compares strategies of data integration that were adopted by three recent projects: EUADR (EU), OMOP (USA) and MATRICE (I)

Contents

Background

Conceptual framework

General process

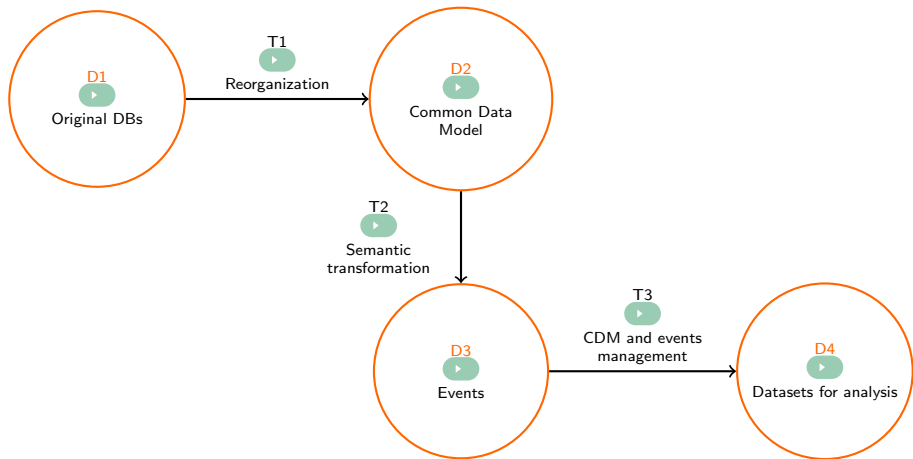
The three projects

Specific aspects

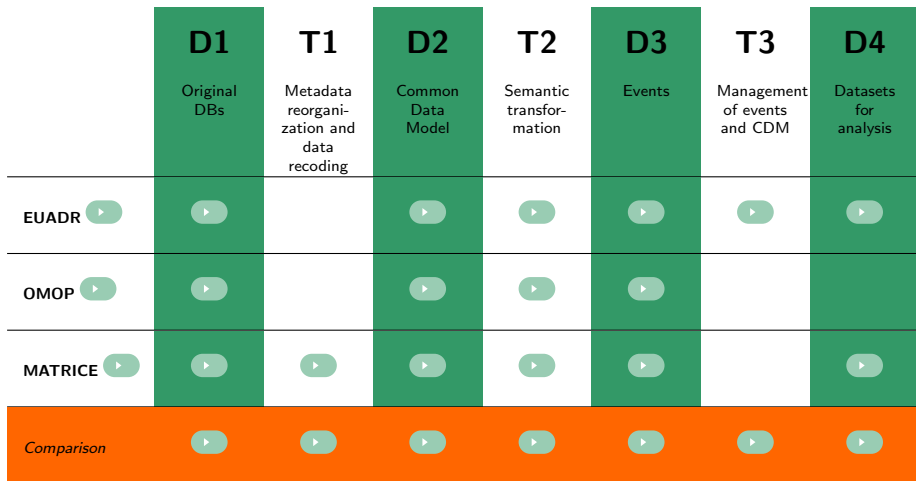
Comparison on steps of the process

Final considerations

Conceptual framework: general process



Conceptual framework: comparisons and details



▶ Wrap up

Contents

Background

Conceptual framework

General process

The three projects

Specific aspects

Comparison on steps of the process

Final considerations

D1: Original DBs, definition

What a DB is

In this talk a database (DB) is an organization that has access to a collection of data sources referring to an identifiable population, meaning that record linkage among the records referring to the same individual is possible, and time windows when a person's data are collected is recorded as well

D1: Original DBs, data sources

Administrative data sources (aka claims data sources) Information is collected by a healthcare provider (hospital, primary or secondary care practice, laboratory, pharmacy) to be submitted to a payer or a health manager in order to account for its activity

Electronic medical records (EMR) Information is collected by a clinician (such as a general practitioner or a specialist physician) or organization of clinicians (such as a hospital ward) to document clinical activity

Disease or birth or death registry Information is collected by a public health authority to perform disease/cause of death surveillance



D2: Common data model

- Statically transformed data** Not a *traditional* CDM: here CDM only contains the static part of the data
- Possibly partial views of the CDM** Some local view of the CDM might be partial (eg if laboratory values are not recorded in one DB this local view will miss corresponding structures)
- Possibly differently coded local views of the CDM** Only simple 1:1 recordings have been performed in T1, therefore some attributes (eg diagnosis) refer to different coding systems, according to the local view



D3: Events

For instance Hypertension, Acute Myocardial Infarction, Diabetes, . . .

Study-specific In a study an event has its specific clinical definition and event datasets are created ad hoc wrt to the aim of the study

Simple data model Person identifier, event identifier, start date, possibly: end date

Use events might be used as inclusion/exclusion criteria (X), exposure (E), outcome (O) or covariate (C)



D4: Datasets for analysis

Data that can be shared Even when a mediator-based integration was chosen, D4 must be in a format that can be shared: deidentified or even aggregated data

Tables or datasets D4 might contain both tables of results ready for interpretation and/or datasets that need further statistical analysis after pooling

Not large It is expected that they are not large datasets



T1: metadata reorganization and data recoding

Static part The first transformation maps the local databases into a common data model/schema that represents the core of the shared information, and is just performed once (up to local data updating)

Metadata reorganization Metadata are reorganized (in case of relational databases: data are reorganized into possibly new tables)

Data simple recoding Data are recoded using common vocabulary, but only in case 1:1 mappings are possible

Diagnosis In case 1:1 mappings are not possible **the original coding is maintained**. This is the case when DBs use different coding systems for diagnoses



T2: imposing semantics – issues

Events At this stage events that are meaningful for the specific study can be defined against the CDM

Algorithms An event definition is an algorithm that maps sequences of longitudinal recordings to the event (eg having a hospital admission with a diagnostic code AND having some laboratory analysis within 60 days from discharge will be mapped to 'AMI' meaning Acute Myocardial Infarction)

Compulsory multiple alternative definitions of the same event The same event must be defined in different ways in DBs with different CDM partial views

Resources to merge diagnostic coding systems In case different coding systems must be handled we need to identify the relevant codes in all of these coding systems and establish mappings: resources such as **UMLS** or ontologies are useful for that

Dates The algorithm must identify dates as well: date of start and (possibly) date of end

T2: imposing semantics – validation

If you did not conduct your own validation study, be prepared to cite others who did. Validation studies increase cost and take time, but they may be the difference between cranking out analyses and sound epidemiologic research.

Hernán M. With great data comes great responsibility. *Epidemiology* 2011 May.

- T2 not only detects people who incurred in a condition, but (implicitly) implies that the others did not incur in it
- Validation *through external gold standard*: computes **specificity, sensitivity, PPV and NPV**, and distribution of covariates (severity, ...) among detected events wrt undetected
- Evaluation through *comparison* of (incidence, prevalence) rates
 - with literature
 - internally: comparison among DBs referring to similar populations but different CDM partial views (eg admin vs GP) might provide information on distribution of covariates among events detected in one DB wrt the other and about precision of terminology mapping

T3: managing

- in case a physical central copy of the data does not exist, T3 might be executed locally, and in this case the transformation produces data that the partners are allowed/willing to share
- the process executing T3 being automatic allows for
 - if executed locally: avoid mistakes or misunderstandings/ambiguity among DBs
 - several assumptions on data semantics being implemented at this stage and tested sequentially: eg several event definitions, assumptions on actual drug use
 - several study designs being tested sequentially (eg different strategies of controls sampling)
 - if new inputs come from data analysis T3 might be changed and repeated



Contents

Background

Conceptual framework

General process

The three projects

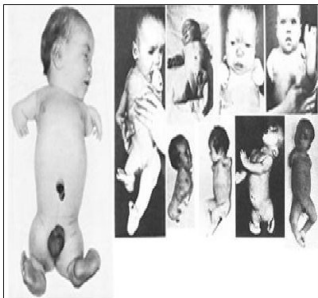
Specific aspects

Comparison on steps of the process

Final considerations

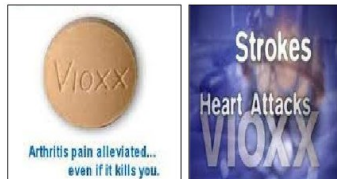
EUADR: motivation

Then...



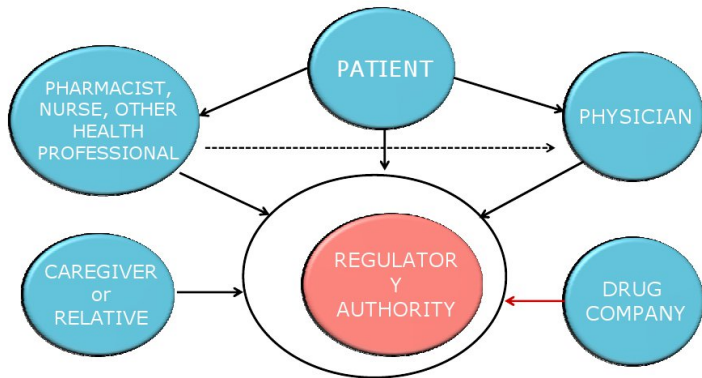
Thalidomide and babies
born with limb
abnormalities, 1960s

and now.



Rofecoxib and heart
attack, 2004

EUADR: motivation



Limitations: underreporting, biased reporting, lack of information on population of medication users and patterns of use

Exploring and Understanding Adverse Drug Reactions by Integrative Mining of Clinical Records and Biomedical Knowledge (EU-ADR) Project

Objective Design, development, and validation of a computerized system that exploits data from electronic healthcare records and biomedical databases for the early detection of adverse drug reactions

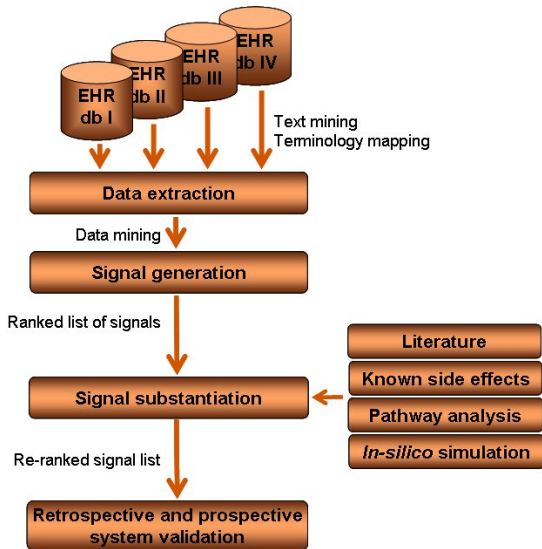
Funding Information and Communication Technologies (ICT) area of the European Commission under the VII Framework Programme

Partners Aarhus University Hospital, Aarhus Sygehus, Denmark; Agenzia regionale di Sanità, Italy; AstraZeneca AB, Sweden; Erasmus University Medical Center, Netherlands; Fundació IMIM, Spain; Health Search - Italian College of General Practitioners, Italy; London School of Hygiene & Tropical Medicine, UK; PHARMO Coöperatie UA, Netherlands; Società Servizi Telematici SRL, Italy; Tel-Aviv University, Israel; Università di Milano-Bicocca, Italy; Université Victor-Segalen Bordeaux II, France; University of Aveiro IEETA, Portugal; University of Nottingham, UK; University of Santiago de Compostela, Spain; University Pompeu Fabra, Spain

EUADR: challenges

- ① Federation of different databases of electronic healthcare records in order to create a large-scale resource for monitoring adverse events.
- ② **Different coding systems for both drugs and diagnoses**
- ③ In EU-ADR eight DBs containing healthcare data of more than 20 million European citizens are involved.
- ④ Exploitation of European diversity for routine drug monitoring.
- ⑤ Automated exploitation of heterogeneous sources of information to reduce the number of spurious signals.

EUADR: workflow



EUADR: current status

- The original EU-ADR project is finished
- The offspring that uses the infrastructure: SOS, Aritmo, Vaesco (ECDC), Safeguard
- Formation of the EU-ADR Alliance: maintain infrastructure
- EU-ADR Alliance: conducting EMA safety studies
- US: OMOP
- Website <http://www.euadr-project.org/>
- Web platform
<https://bioinformatics.ua.pt/euadr/Welcome.jsp>

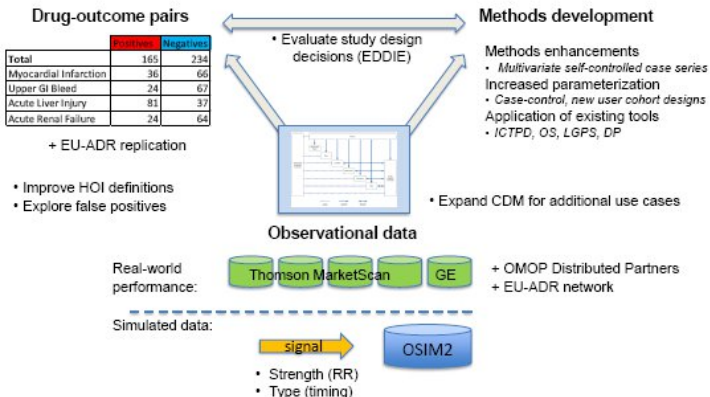
EU-ADR selected list of publications

- Trifirò G, et al. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoepidemiol Drug Saf.* 2009 Dec; 18(12):1176-84.
- Coloma PM, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiol Drug Saf.* 2011 Jan;20(1):1-11.
- Coloma PM, et al. Electronic healthcare databases for active drug safety surveillance: is there enough leverage?. *Pharmacoepidemiol Drug Saf.* 2012 Jun;21(6):611-21. doi: 10.1002/pds.3197. Epub 2012 Feb 8.
- Avillach P, Coloma PM, et al. Harmonization process for the identification of medical events in eight European healthcare databases: the experience from the EU-ADR project. *J Am Med Informatics Assoc* 2012 (in press)

OMOP

- What** The Observational Medical Outcomes Partnership (OMOP) is a public-private partnership initiated in 2008, managed by Foundation for the National Institutes of Health, chaired by the Food and Drug Administration
- Support** Pharmaceutical industry with active engagement from academia, industry, healthcare providers in US and internationally.
- Goal** Methodological research about use of electronic healthcare data to explore the real-world effects of medical products

OMOP 2011-2012 agenda



OMOP website

http://omop.fnih.org

The screenshot shows the OMOP website homepage. The browser address bar displays 'omop.fnih.org'. The main header is a dark blue bar with the text 'Observational Medical Outcomes Partnership' on the left and navigation links 'contact us', 'OMOP charter', 'privacy policy', 'terms of use', and 'login' on the right. Below the header, there is a search box on the left with the text 'Search this site:' and a 'Search' button. A central white box contains the title 'Observational Medical Outcomes Partnership' and a paragraph: 'In partnership with PhRMA and the FDA, the Foundation for the National Institutes of Health launched the Observational Medical Outcomes Project (OMOP), a public-private partnership. This interdisciplinary research group has tackled a surprisingly difficult task that is critical to the research community's broader aims: identifying the most reliable methods for analyzing huge volumes of data drawn from heterogeneous sources.' Below this box is a section titled 'Call for Public Comment: Proposed OMOP Research Priorities, Infrastructure tasks, and Research Tasks for 2013 - 2016' with a sub-link 'Draft Report for Public Comment'. On the right side, there are sections for 'Syndicate' with a RSS icon and 'Quick Links' listing various resources like '2012 Research Findings Summary', 'Evidence-based Learning from Medical Observations CDM & Vocabularies', 'Health Outcomes of Interest Investigators', 'Methods Library', 'Publications', 'Simulated Datasets', 'Symposium Presentations', and 'White Papers'. A left sidebar contains a 'Navigation' menu with links: 'About Us', 'Call for Public Comment', 'Research', 'OMOP Implementation', 'Simulated Data', 'Resources', and 'Events & Presentations'.

OMOP selected list of publications

- Stang PE, Ryan PB, Dusetzina SB, Hartzema AG, Reich C, Overhage JM, Racoosin JA. Health Outcomes of Interest in Observational Data: Issues in Identifying Definitions in the Literature. Health Outcomes Research in Medicine (2011). doi: 10.1016/j.ehrm.2011.11.003
- Overhage JM, Ryan PB, Reich CG et al. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19(1):54-60. Epub 2011 Oct 28.
- Madigan D, Ryan P. What can we really learn from observational studies? The need for empirical assessment of methodology for active drug safety surveillance and comparative effectiveness research. Epidemiology. 2011;22:629631.
- Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010 Nov 2;153(9):600-6.

MATRICE: outline of the project

Partners National Agency for Regional Health Services, Italian Ministry of Health, Regional Agency for Public Health of Tuscany, National Research Council, 5 Local Health Units, Medical Informatics Department of Erasmus Medical Center University

Timeframe 2011-2013

Goal Design and develop an automatic system to support local clinical governance of chronic disease management quality assessment and regional/national chronic disease quality of care surveillance

MATRICE: workplan

Chronic disease algorithm validation Some chronic diseases and clinical staging definitions were identified, algorithms for event definition will be validated against GP data, others will be generated through machine learning techniques

Quality of care indicators QOF indicators that can be computed on IAD were chosen and their measure will be validated as well

Data integration tools A specific software tool called TheMatrix is being developed to support data integration

System testing The system is being tested in the 5 Local Health Units, in 2 regions at a central level

Management and communication

MATRICE: publications

- Gini R, et al. Chronic disease prevalence from Italian administrative databases in the VALORE project: a validation through comparison of population estimates with general practice databases and national survey. Submitted for publication



Contents

Background

Conceptual framework

General process

The three projects

Specific aspects

Comparison on steps of the process

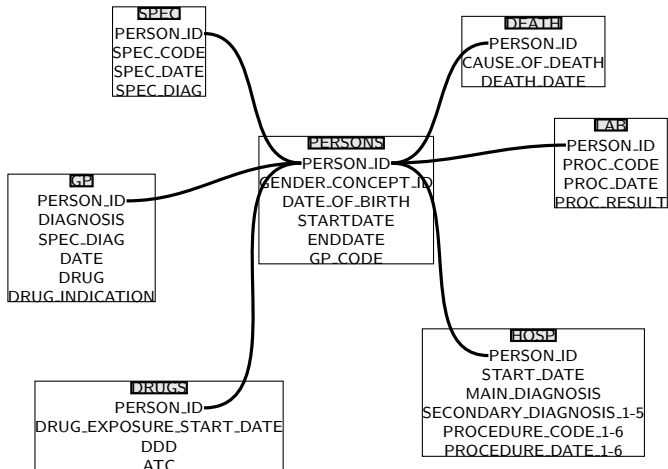
Final considerations

D1 in EUADR: original DBs

Characteristics of DB	Pedinet (ITA)	SIMG (ITA)	Lombardy Regional DB (ITA)	ARS (ITA)	IPCI (NL)	PHARMO (NL)	Aarhus Hospital DB (DK)
Type of data	FP	GP	Claims	Claims/Reg	GP	Claims/GP	Claims/Reg
Observation years	2003-2010	2003-2010	2003-2008	2003-2010	2000-2010	2004-2010	2002-2010
Drug code	ATC	ATC	ATC	ATC	ATC	ATC	ATC
Event code	ICD9-CM	ICD9-CM	ICD9-CM	ICD9-CM	ICPC	ICD9-CM	ICD10
Free text	Yes	Yes	No	No	Yes	No	No



D2 in EUADR: common data model



D2 in EUADR: common data model

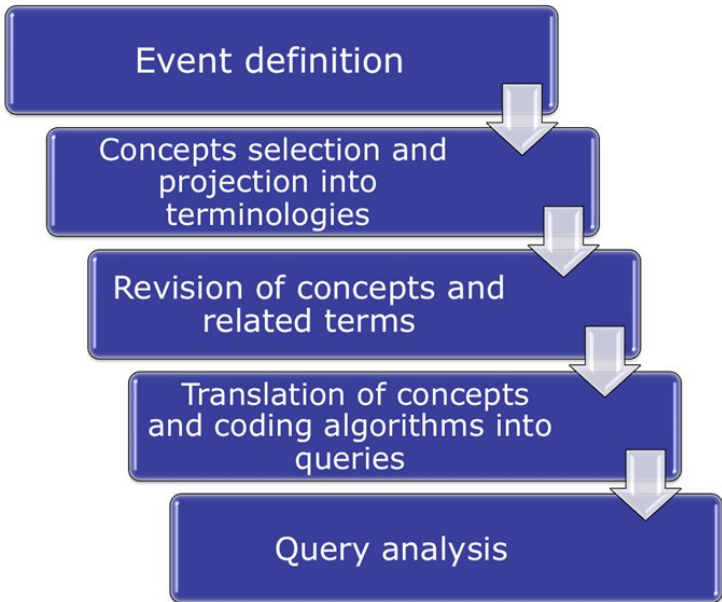
Setting	attributes	A	B	C	D	E	F	G	H
HOSP	main diagnosis	■	■	■	■				
	secondary diagnosis								
	procedures								
DEATH	cause of death	■	■						
GP	diagnosis					■	■	■	■
	specialist								■
	drug					■	■	■	■
LAB	classification	■	■	■	■	■			
	result	■			■	■			
SPEC	classification	■	■	■					
	diagnosis								
DRUGS	classification	■	■	■	■				

HOSP: hospitals; DEATH: death registry; GP: general practitioners practices; LAB: laboratories; SPEC: specialist physicians' practices; DRUGS: pharmacies

Attributes recorded in GP - specialist: specialist diagnosis; lab: diagnostic tests results; drug: drug prescriptions



T2 in EU-ADR: harmonization



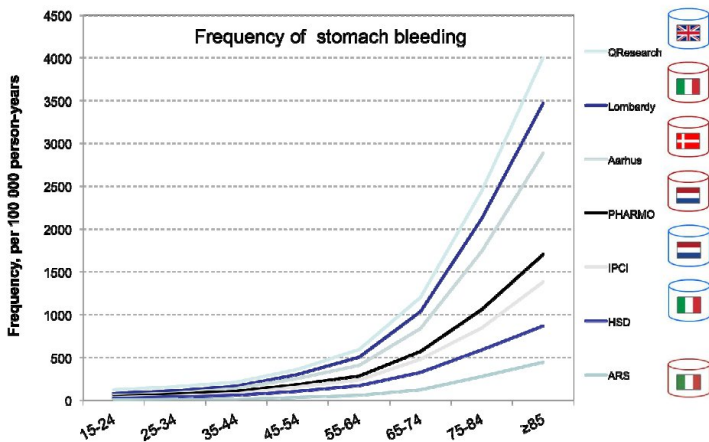
T2 in EU-ADR: harmonization

Event	UMLS Concept	Preferred term	ICD9CM	ICD10	RCD	ICPC
AMI	C0155626	Acute myocardial infarction	410.x	I21.x	G30z.,XE0Uh	K75, K75002
	C0428953	ECG: myocardial infarction			323., 323Z.	
	C0232320	ECG: antero-septal infarct.			3233	
	C0428956	ECG: posterior/inferior infarct			3234	
	C0428955	ECG: subendocardial infarct			3235	
	C0232325	ECG: lateral infarction			3236	
	C0428953	ECG: myocardial infarction			323., 323Z.	
	C0340324	Silent myocardial infarction			X200a	
C0340283	Other acute and subacute ischemic heart disease NOS <i>Only for refinement use</i>	411		G31.,G31yz		
AS	C0002792	Anaphylactic shock		T78.2	SN50	A12004, A92005
	C0375697	Other anaphylactic shock	995			
	C0685898	Anaphylactic shock due to adverse food reaction	995.6	T78.0	X70vm, X70w1	
	C0161840	Anaphylactic shock due to serum	999.4	T80.5	SP34, X70vl	
	C0274304	Anaphylactic shock, due to adverse effect of correct medicinal substance properly administered		T88.6	SN501	

T2 in EU-ADR: evaluation of different event definitions through comparison of incidence rates

Event	DB	IR for basic query		Additional data: IR and increment	
		HOSP-main	GP	add DEATH	add concept with refinement
AMI	AARHUS	101.4		126.5 (+25%)	
	ARS	77.8		90.2 (+15%)	
	HSD		58.7		59.1 (+0.5%)
	IPCI		148.4		
	PHARMO	93.4			
	LOMBARDY	82.5			
		IR for basic Query		Additional data: IR and increment	
Event	DB	HOSP-main	GP	add HOSP-sec	add DEATH
AS	AARHUS	5.7		6.4 (+12%)	6.4 (+0%)
	ARS	12.0		12.7 (+6%)	12.8 (+0%)
	HSD		5.2		
	IPCI		7.9		
	PHARMO	1.9			2.4 (+26%)
	LOMBARDY	2.2			2.8 (+27%)

T2 in EU-ADR: evaluation through comparison of incidence rates



T2 in EU-ADR: validation

Data Source	Coding system	No. of cases confirmed (%)	Overall PPV (95% CI)
GP/specialist diagnoses (<i>IPCI, NL</i>)	ICPC, free text	119 (29.8)*	46.5 (40.4 - 52.6)
GP/specialist diagnoses (<i>HSD, Italy</i>)	ICD9-CM, free text	115 (57.5)*	95 (91.2 - 98.9)
Primary hospital discharge diagnoses (<i>Aarhus, DK</i>)	ICD-10	148 (100)	100 (100 - 100)



D3 in EU-ADR: events deemed to be important for drug safety

System/organ	Event
Hematologic	Hemolytic anemia Aplastic anemia/pancytopenia Neutropenia Thrombocytopenia
Cutaneous	Maculo-papular erythematous eruptions Bullous eruptions (Stevens Johnson Syndrome, Lyell's Syndrome)
Liver and gastrointestinal	Acute liver injury Acute pancreatitis Upper gastrointestinal bleeding
Cardiac and vascular	Acute myocardial infarction QT prolongation Cardiac valve fibrosis Venous thrombosis
Neurologic	Convulsions Peripheral neuropathy Extrapyramidal disorders Rhabdomyolysis
Psychiatric	Confusional state Mood changes: depression and mania Amnesias
Renal	Suicidal behavior/attempt Acute renal failure
Multi-systemic	Anaphylactic shock

- Trigger for drug withdrawal
- Trigger for black box warning
- Emergency Room visit, hospitalization
- Likelihood of being drug-induced
- Leads to death



T3 in EU-ADR: Jerboa

Jerboa EU-ADR developed a Java software tool called Jerboa to aggregate data from an EHR database in its local environment.

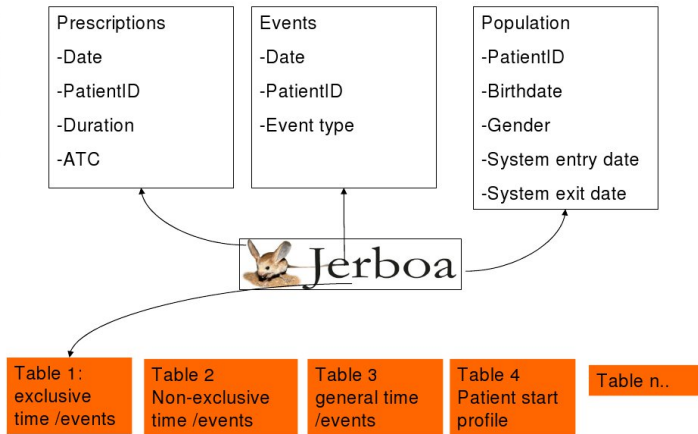
Input files Contain drug exposure, occurrence of adverse events, and patient information, as flat csv files

Output Number of events and exposure time, stratified according to ATC code, age category, and gender, as csv files

Scripting language Data processing and aggregation parameters of Jerboa are specified in a script file. This makes it easy to test different parameter settings and facilitates uniform data aggregation across the databases.



D4 in EU-ADR: resulting datasets

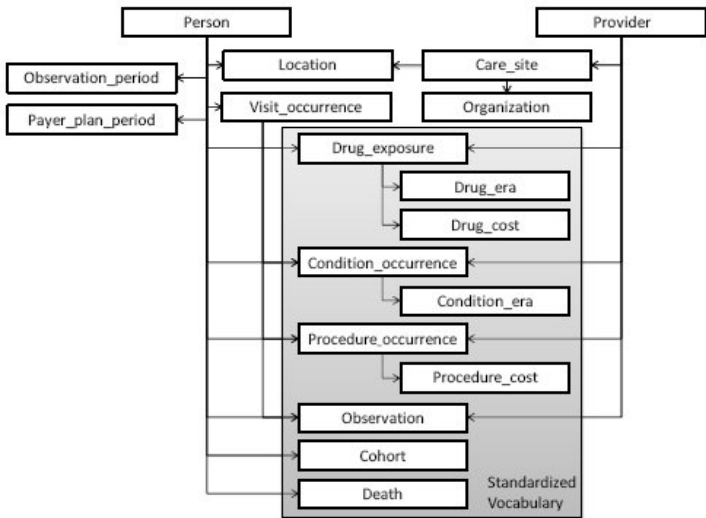


D1 in OMOP: original DBs

Name	General Database Description	Pop Size (M)
GE Centricity Electronic Health Record (GE)	Derived from data pooled by providers using GE Centricity Office (an ambulatory electronic health record) into a data warehouse in a HIPAA-compliant manner.	11.2
MarketScan® Research Databases from Thomson Reuters	<i>MarketScan Lab Database (MSLR)</i> - Represents privately insured population, with administrative claims from inpatient, outpatient, and pharmacy services supplemented by laboratory results.	1.5
	<i>MarketScan Medicaid Multi-State Database (MDCD)</i> - Contains administrative claims data for Medicaid enrollees from multiple states, containing inpatient, outpatient, and pharmacy services.	11.1
	<i>MarketScan Medicare Supplemental and Coordination of Benefits Database (MDCR)</i> - Captures administrative claims for retirees with Medicare supplemental insurance paid by employers, including services provided under Medicare-covered payment, employer-paid portion, and any out-of-pocket expenses.	4.4
	<i>MarketScan Commercial Claims and Encounters (CCAE)</i> - Represents privately insured population and captures administrative claims with patient-level de-identified data from inpatient and outpatient visits and pharmacy claims of multiple insurance plans.	58



D2 in OMOP: Common Data Model



T2 in OMOP: Regularized Identification of Cohorts (RICO)

Parameterized extraction strategy against CDM Under the assumption that the definition of an event is a variable combination of fixed operations, a corresponding extraction strategy is defined

Either in SQL or SAS Two versions of RICO were developed, one for CDMs in Oracle environment and one for CDMs in a SAS environment

Parameters Primary selection Either an episode of drug use or a diagnosis in any setting

Restriction Date/age range, gender. . .

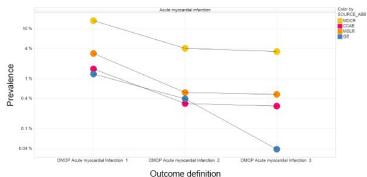
Inclusion or exclusion other drugs, conditions, procedures, visits, observations temporally related to primary selection criteria (within x days. . .)

Type of relationships AND, OR, AT LEAST

T2 in OMOP: evaluation of alternative HOI definitions through prevalence rate comparison

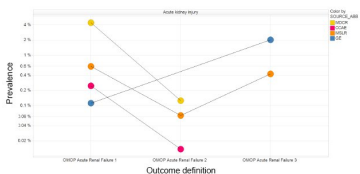
Acute Myocardial Infarction (AMI)

1. Occurrence of at least one broad diagnosis code ICD9 410*, 411.1, 411.8* OR angina (413.9) during hospitalization
2. Occurrence of at least one narrow diagnosis code ICD9 410*
3. Occurrence of at least one narrow diagnosis code AND (diagnostic procedure <=30d before diagnosis OR (treatment procedure >=60d after diagnosis))



Acute Kidney Injury (AKI)

1. Occurrence of at least one diagnosis code ICD9 584*
2. Occurrence of at least one diagnosis code AND treatment procedure for acute dialysis >= 60d after EXCLUDING diagnosis code for chronic dialysis status
3. An increase in serum creatinine level (LOINC 2160-0) of >=0.5 mg/dl for patients with a baseline serum creatinine level of <=1.9 mg/dl, >=1.0 mg/dl for patients with a baseline level of 2.0-4.9 mg/dl, and >=1.5 mg/dl for patients with a baseline level >= 5.0 mg/dl

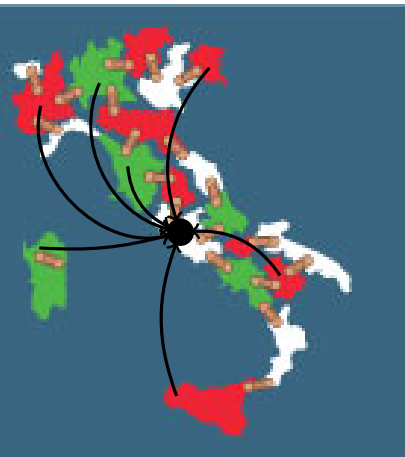


D3 in OMOP: Health Outcomes of Interest

- Aplastic anaemia
- Liver disorder
- Renal failure acute
- Acute myocardial infarction
- Gastrointestinal haemorrhage
- Anaphylactic shock
- Erythema multiforme
- Neutropenia
- Rhabdomyolysis
- Mitral valve disease
- Progressive multifocal leukoencephalopathy
- Embolism venous
- ...



D1 in MATRICE: Italian Administrative Databases



Italy has a universal, tax-based national health system, organized at regional level. Some patient-level, deidentified datafiles must be transmitted yearly by every Local Health Unit or Region to the central government according to a fixed data model

Image from the XXXV conference of the Italian Epidemiology Association

D1 in MATRICE: opportunities and challenges of Italian Administrative Databases

Technically homogeneous Same data model, same content: technically it might become an actual distributed database

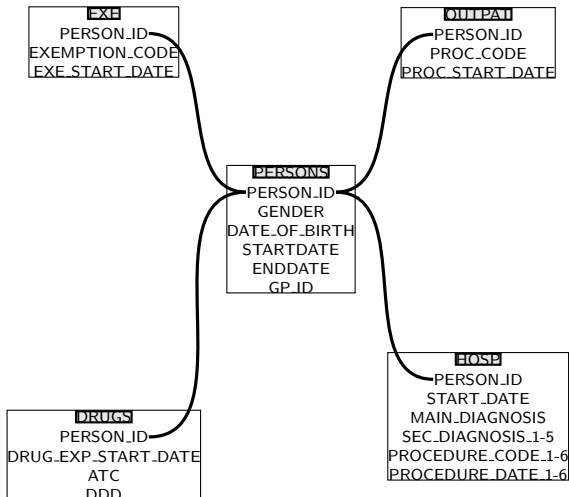
Profoundly heterogeneous In fact, data are stored in organizations dedicated to healthcare organizations that have often little interest and resources dedicated to data integration, are geographically sparse are diverse

Semantic problem Laboratory analyses don't record results, specialistic visits don't record diagnosis, hence sophisticated algorithms need to be developed to replace the missing information

Validation Validation is crucial



D2 in MATRICE: IAD



T1 in MATRICE: installing TheMatrix

- Java application, database access via JDBC

Production Settings portability over many OS and DBMS platforms

Experimental Settings can also read plain CSV data

- **IAD administrative model** – normative data model
 - semantically clear
 - meant for unfrequent, anonymized data interchange
 - different concrete implementation
- **XML file descriptor** `mapping.xml`
 - map the “*physical*” input data model onto IAD
 - reconcile implementation choices:
field names, split fields, different data formats . . .

Deal with low-level DBMS transformations just once,
BEFORE the real analysis task

T1 in MATRICE: mapping example

```
<iadMapping>                                <!-- mapping of PERSON -->
  <dataset name="PERSON">
    <joinName>PERSON</joinName>
    <joinClause>FLUSSI.U_ATA_EXT</joinClause>
    <simpleMapping name="BIRTH_LOCATION_CONCEPT_ID">
      <sourceTable>FLUSSI.U_ATA_EXT</sourceTable>
      <sourceAttribute>COMNASC</sourceAttribute>
    </simpleMapping>
    <simpleMapping name="DATE_OF_BIRTH">
      <sourceTable>FLUSSI.U_ATA_EXT</sourceTable>
      <sourceAttribute>DATANASC</sourceAttribute>
    </simpleMapping>
```

- Automatic access to DBMS tables, including JOIN, data caching
- Automatic field renaming, concatenation, table lookup
- **Extendable Approach**
 - future : other models and IAD extensions
 - experiment with diverse data sources (MATRICE Gold standard)

T1 in MATRICE: scripting the change

- Data transformation expressed via a scripting language
 - all ordinary analysis tasks *(see step T2)*
 - data transformation, summarization, algorithm application
- Scripting language derived from Jerboa
- can deal with extendable data model
 - ease experiments in the research phase *... essential in T2!*
 - ease script distribution to data sources
- Over a common model, script can be executed on **separate sources** and their results gathered back

```
Filter Alive Patients (FilterModule)
inputs
    dataset = Patients File
parameters
    conditions = [{DATE_OF_DEATH,>,$DATE}];{DATE_OF_BIRTH,<=,$DATE}]
    conditions = [{STARTDATE,<=,$DATE}];{ENDDATE,>,$DATE}]
    boolExpr = AND
end
```

T1 in MATRICE: scripting the change

- scripts exploited in a modular way
 - automatic on-demand invocation
 - results can be cached
 - results reuse **only if** all parameters match

A data extraction related to year 2010 but performed today can exploit available consolidated data and modern criteria

```
Fill CI11ECG_betabloc (ScriptInputModule)
parameters
//sub script name
    scriptFilename = computeCI11ECG_betabloc_rev4
//script parameters
    scriptParams = [$DATE]
//file generated by the sub-script
    inputName = Fill CI11ECG_betabloc
end
```



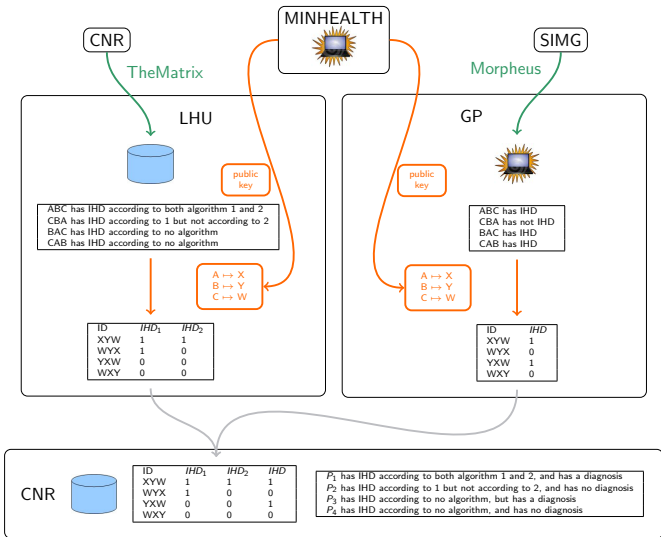
T2 in MATRICE: programming TheMatrix via visual interface

Script generation A script is generated through a sequence of simple choices on a visual interface

Recursive Script generation might be recursive: a previously generated script can be loaded and presented to the user as a new table of the global schema, the final script will encapsulate previous scripts

Usability Usability tests will be performed to identify strategy for the VI to be usable by epidemiologists without programming skills

T2 in MATRICE: chronic diseases definition validation



D3 in MATRICE: stages of chronic diseases

Diabetes <ul style="list-style-type: none">● Uncomplicated● Subclinical organ damage● Complicated	Hypertension <ul style="list-style-type: none">● Uncomplicated● Complicated, no heart failure● Heart failure
Ischaemic heart disease <ul style="list-style-type: none">● Coronaropathy● AMI● Left ventricular dysfunction● Heart failure	Heart failure



D4 in MATRICE: output of TheMatrix

- TheMatrix produces data files (CSV) and **additional metadata**
- Data layout : standard (IAD, Jerboa...) or customized by scripts
- Metadata : CSV comments *plus* an XML file

- Data **accessibility and reuse**
 - data field names, types as CSV comment lines
- **Integrity**
 - MD5 signatures of the data file
 - prevent data corruption and file misuse/tampering
- **Privacy**
 - restricted diffusion of data files / fields
- **Traceability and reproducibility**
 - what script produced the data (id, version), when (date), what source data (dataset origins, reference dates).
 - improve productivity and reliability of the validation process
 - key feature for safe modular reuse of scripts



Contents

Background

Conceptual framework

General process

The three projects

Specific aspects

Comparison on steps of the process

Final considerations

Comparison wrt D1: original DBs

Project	Type	Countries	Subjects maintaining the DB	Level of heterogeneity
EUADR	administrative databases, GP electronic health records, diseases/death registries	UK, DK, NL, I	Academics, bodies of national/regional health systems, scientific associations, private research companies	High
OMOP	administrative databases, primary care ("GP") electronic health records	US	FNIH maintains a central lab with 5 de-identified databases; distributed partners maintain patient-level data and share only summary results to the OMOP cloud-based research lab	Medium
MATRICE	administrative databases	I	Bodies of national/regional health systems	Low

Comparison wrt D1: original DBs – distinct combination of data settings

Setting	attributes	EUADR							OMOP				MATR
		A	B	C	D	E	F	G	A	B	C	D	all
HOSP	main diagnosis	orange	orange	orange	orange					orange	orange	orange	orange
	secondary diagnosis	orange	orange	orange	orange					orange	orange	orange	orange
	procedures	orange	orange	orange	orange					orange	orange	orange	orange
	cause of death	green	green								green		
GP	diagnosis					orange	orange	orange	orange		orange		
	specialist					orange	orange						
	lab					orange	orange	orange	orange				
	drug					orange	orange	orange	orange				
LAB	classification	green				green				green		green	green
	result	green				green				green		green	green
SPEC	classification	orange	orange	orange						orange	orange		orange
	diagnosis	orange								orange	orange		orange
EXE	diagnosis		green	green									green
DRUGS	classification	orange	orange	orange	orange					orange	orange	orange	orange

HOSP: hospitals; DEATH: death registry; GP: general practitioners practices; LAB: laboratories; SPEC: specialist physicians' practices; EXE: disease-specific exemptions from copayment; DRUGS: pharmacies

Attributes recorded in GP - specialist: specialist diagnosis; lab: diagnostic tests results; drug: drug prescriptions



Comparison wrt T1: data reorganization

Project	Storage	Documentation	Mechanism	Recoding
EUADR	Virtual (local)	No formal documentation	Demanded to single DB	Clinical data original coding or free text
OMOP	5 DBs in federated central/cloud database and 10 DBs in federated local copies of the database	ETL formal document, ad hoc per DB	Demanded to single DB	Source data standardized to common vocabulary by domain: Drug (RxNorm), Condition (SNOMED), Labs (LOINC)
MATRICE	Virtual (local)	Local configuration of the TheMatrix software (XML file)	The Matrix installation	No meaningful recoding except ATC and DDD classification of drug boxes



Comparison wrt D2: common data models

Project	Table classification	Every table of the CDM is instantiated in every DB	Features
EUADR	According to reason/setting of data recording, clinical contents are recorded as attributes	False	The set of tables which a DB instantiates gives information about its data capture mechanism, hence applicability, sensitivity and specificity of semantic algorithms (T2)
OMOP	According to data clinical content, setting is recorded as attribute	True	Optimize performance of next data transformations (T2 and T3)
MATRICE	According to setting of data recording, as embodied in technical documents from national regulations: clinical contents are attributes	True	Easiest solution (it was an easy problem anyway)



Comparison wrt T2: semantic transformations

Project	Logic	Documentation	Validation	Mechanism
EUADR	DB-specific algorithms for event definition harmonized through a formal process	Formal documentation	Internal incidence rates comparison, comparison with literature, some validation with external gold standard	Demanded to single DB
OMOP	HOI: Multiple alternative algorithms for HOI definition, not all are applicable to all DBs	RICO parameters	Internal prevalence rates comparison, no external validation performed	RICO is a tool to generate SQL or SAS code which runs on the CDM
MATRICE	Multiple algorithms	Scripting language of TheMatrix, script generated through visual interface and shared among DBs	Validation for events in course	TheMatrix run



Comparison wrt D3: events

Project	Acute events	Conditions lasting for some time	Chronic conditions	Drug use as such	Use of other health services as such
EUADR	O	-	-	E, C	-
OMOP	O,C	-	C	E, C	-
MATRICE	O	-	X, C, O	O, C	O, C

X: event used as inclusion or exclusion criterion to select cohorts, E: event used as exposure, O: event used as outcome, C: event used as covariate



Comparison wrt T3: management of events

Project	Automatic	Common among DBs	Specific software	Programming language
EUADR	Y	Y	Jerboa	Java & Jerboa scripting language
OMOP	Y	Y	-	SQL, SAS, R, C, Java
MATRICE	Y	Y	TheMatrix	Java & TheMatrix scripting language



Comparison wrt D4: datasets for analysis

Project	Type	Format	Documentation
EUADR	Intermediate files that can be shared among partners, analysis will follow	csv	Jerboa script (only documenting T3)
OMOP	Final estimates, intermediate files are discarded	csv/SAS datafiles/SQL table...	ETL + RICO params+ SQL/Java/R... code
MATRICE	Intermediate files to be used for analysis or report generation	csv	TheMatrix specific XML format



Contents

Background

Conceptual framework

General process

The three projects

Specific aspects

Comparison on steps of the process

Final considerations

Wrap up

- Emerging need** The need for multi-centre epidemiology frameworks is emerging across continents
- New type of data integration** Traditional strategies do not address completely the issues raised by this challenge
- Tentative solutions** Tentative solutions have been introduced in the first projects, with differences but a general common conceptual framework can be identified
- Room for improvement** There is obvious room for improvement

Discussion...

Validation through comparison with a gold standard

	Gold standard		
		Y	N
Classification	Y	TP	FP
	N	FN	TN

Sensitivity $\frac{TP}{TP+FN}$ Probability that a case is detected

Specificity $\frac{TN}{TP+FN}$ Probability that a non case is classified as such

Negative predictive value (NPV) $\frac{TN}{TN+FN}$ Probability that subject classified as non case is actually such

Positive predictive value (PPV) $\frac{TP}{TP+FP}$ Probability that subject classified as case is actually such

Unified Medical Language System

UMLS A major resource in the biomedical domain maintained by the National Library of Medicine. Two semantic elements: the Metathesaurus and the semantic network.

UMLS Metathesaurus A large graph constituted about more than 2 million concepts (2.6 million in the 2011AB version), defined by integrating more than 150 (161 in the 2011AB) of biomedical terminologies (MeSH, SNOMED CT, ICD10, etc.) Each Metathesaurus concept is a cluster of synonym terms, possibly defined in different languages, which has a unique identifier (Concept Unique Identifier - CUI) and to which is associated sometimes a definition.

Semantic network Concepts are interlinked by relationships generally inherited from the original terminologies. They are hierarchical or associative (ie horizontal) and some of them are defined from a logical point of view (ie they are typed; for instance *isa*, *finding_site_of*, *treats*). The semantic network of the UMLS has a hierarchy of 133 semantic types aggregated in 15 semantic groups which correspond to different sub-domains of biomedicine (e.g. Anatomy, Disorders). Each Metathesaurus concept is categorized by one or several semantic types (for example Acute Myocardial Infarction is categorized by the semantic type Disease or Syndrom).