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

Rotterdam, 22 November 2012
Acknowledgements

This presentation is elaborated from a workshop with the same title given last summer in Pisa at MIE2012. Thanks to coauthors, to José Luis Oliveira and to Peter Rijnbeek.
Contents

Introduction

Conceptual framework

The three projects compared on the conceptual framework

Final considerations
Contents

Introduction

Conceptual framework

The three projects compared on the conceptual framework

Final considerations
Data use and reuse in epidemiology

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

Traditional data reuse  Secondary use of available information, such as routinarily 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
Focus

Compare three recent projects: EUADR (EU), OMOP (USA) and MATRICE (I) to

Detect challenges
Understand differences
Highlight best practices
Identify weaknesses
Contents

Introduction

Conceptual framework

The three projects compared on the conceptual framework

Final considerations
Conceptual framework: general process

D1 Original DBs
T1 Reorganization
D2 Common Data Model
T2 Semantic transformation
D3 Events
T3 Implementation of study design
D4 Datasets for analysis

Q Quality cycle
D1: Original DBs, data sources

Administrative data sources (aka claims data sources) Collected by a healthcare provider to account for its activity

Electronic medical records (EMR) Collected by clinicians to document clinical activity

Disease or birth or death registry Collected by a public health authority to perform disease/cause of death surveillance
D2: Common data model

Not really common! Indeed

Different codings  Eg diagnosis refer to different coding systems or might be in free text

Partial views of the CDM  Eg is a DB does not have lab results, another does not have causes of death

Still: more common than before transformation  “Hospital discharge records” or “GP clinical database” are called in the same way across DBs

Lossless?  Documentation is needed!
D3: Events

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

Study-specific Specific clinical definition per 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

Think traditional! The traditional dataset for analysis after traditional data collection & data management for statistical analysis

Data that can be shared  Deidentified or even aggregated data

Tables or datasets  Both tables of results ready for interpretation and/or datasets that need further statistical analysis

Not large  It is expected that they are not large datasets
T1: metadata reorganization and data recoding

Metadata reorganization  Tables and attributes with similar content across DBs are mapped to the same names

Data simple recoding  Only in case 1:1 mappings are possible (eg male becomes ”M”)

Keep original codings if complex  Otherwise the original coding is maintained, eg coding systems/free text for diagnoses

Permanent data transformation  Study-independent, performed when the DBs first gather

Why is this transformation conceptually useful? To emphasize difference between metadata and semantics!
T2: imposing semantics

Think traditional! Same role as “asking questions” or “collecting clinical information” in traditional data collection.
**T2: imposing semantics**

**Think traditional!** Same role as “asking questions” or “collecting clinical information” in traditional data collection

**Algorithms** Answer to the “question”: sequences of recordings are mapped to the event
T2: imposing semantics

Think traditional! Same role as “asking questions” or “collecting clinical information” in traditional data collection

Algorithms Answer to the “question”: sequences of recordings are mapped to the event
T2: imposing semantics

Think traditional! Same role as “asking questions” or “collecting clinical information” in traditional data collection

Algorithms Answer to the “question”: sequences of recordings are mapped to the event
T2: imposing semantics

Think traditional! Same role as “asking questions” or “collecting clinical information” in traditional data collection

Algorithms Answer to the “question”: sequences of recordings are mapped to the event
T2: imposing semantics

Think traditional! Same role as “asking questions” or “collecting clinical information” in traditional data collection

Algorithms  Answer to the “question”: sequences of recordings are mapped to the event
T2: imposing semantics

Think traditional! Same role as “asking questions” or “collecting clinical information” in traditional data collection

Algorithms Answer to the “question”: sequences of recordings are mapped to the event

L G S P Diabetes!
T2: imposing semantics

Think traditional! Same role as “asking questions” or “collecting clinical information” in traditional data collection

Algorithms Answer to the “question”: sequences of recordings are mapped to the event

Multiple alternative definitions In DBs with different information

Diabetes!
T2: imposing semantics

Think traditional! Same role as “asking questions” or “collecting clinical information” in traditional data collection

Algorithms Answer to the “question”: sequences of recordings are mapped to the event

Multiple alternative definitions In DBs with different information

Diabetes!
T2: imposing semantics

Think traditional! Same role as “asking questions” or “collecting clinical information” in traditional data collection

**Algorithms** Answer to the “question”: sequences of recordings are mapped to the event

**Multiple alternative definitions** In DBs with different information

**Dates** When did the event happen?

![Diagram](image)
T2: imposing semantics

Think traditional! Same role as “asking questions” or “collecting clinical information” in traditional data collection

**Algorithms** Answer to the “question”: sequences of recordings are mapped to the event

**Multiple alternative definitions** In DBs with different information

**Dates** When did the event happen?

**Uncertainty in quality of transformation** . . . might happen to be really dramatic!
T2: imposing semantics

Think traditional! Same role as “asking questions” or “collecting clinical information” in traditional data collection

Algorithms Answer to the “question”: sequences of recordings are mapped to the event

Multiple alternative definitions In DBs with different information

Dates When did the event happen?

Uncertainty in quality of transformation ... might happen to be really dramatic!

Same question, more possible answers ... might try several algorithms
T2: imposing semantics

Think traditional! Same role as “asking questions” or “collecting clinical information” in traditional data collection

Algorithms Answer to the “question”: sequences of recordings are mapped to the event

Multiple alternative definitions In DBs with different information

Dates When did the event happen?

Uncertainty in quality of transformation . . . might happen to be really dramatic!

Same question, more possible answers . . . might try several algorithms

Differences in coding systems Not the worst problem, resources as UMLS are useful
T2: imposing semantics

Think traditional! Same role as “asking questions” or “collecting clinical information” in traditional data collection

**Algorithms** Answer to the “question”: sequences of recordings are mapped to the event

**Multiple alternative definitions** In DBs with different information

**Dates** When did the event happen?

**Uncertainty in quality of transformation** ... might happen to be really dramatic!

**Same question, more possible answers** ... might try several algorithms

**Differences in coding systems** Not the worst problem, resources as UMLS are useful

**What if more data sources or text mining techniques become available?** Conceptually the same
T3: implement study design

Record linkage  Create individual-level dataset by linking events
Create study population  Apply inclusion/exclusion time and criteria
Time splitting  If necessary split person-time
Privacy  Deidentify, aggregate
Last local data transformation  After this step data can be shared and pooled
Q: quality cycle

**Documentation**  Improve transparency and reproducibility

**Automatization**  Prevent mistakes, improve speed of process

**Double programming**  Detect bugs

**Feedback on semantic transformation (T2)**  Via validation, internal rate comparison, external rate comparison

**Feedback on best methods (T3)**  Via comparison with expected results

**Management**  Allow parallel tasks
Contents

Introduction

Conceptual framework

The three projects compared on the conceptual framework

Final considerations
## Conceptual framework: details and comparisons

<table>
<thead>
<tr>
<th></th>
<th>D1</th>
<th>T1</th>
<th>D2</th>
<th>T2</th>
<th>D3</th>
<th>T3</th>
<th>D4</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUADR</td>
<td>Original DBs</td>
<td>Metadata reorganization and data recoding</td>
<td>Common Data Model</td>
<td>Semantic transformation</td>
<td>Events</td>
<td>Implementation of study design</td>
<td>Datasets for analysis</td>
<td>Quality</td>
</tr>
<tr>
<td>OMOP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATRICE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparison</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

**Goal**
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
Observational Medical Outcomes Partnership (OMOP)

**Setting**  US

**What**  Public-private partnership managed by Foundation for the National Institutes of Health, chaired by the Food and Drug Administration

**Timeframe**  Initiated in 2008

**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
MATRICE

**Setting**  Italy

**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

**Timeframe**  2011-2014

**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
### D1 in EUADR: original DBs

<table>
<thead>
<tr>
<th>Characteristics of DB</th>
<th>Pedianet (ITA)</th>
<th>SIMG (ITA)</th>
<th>Lombardy Regional DB (ITA)</th>
<th>ARS (ITA)</th>
<th>IPCI (NL)</th>
<th>PHARMO (NL)</th>
<th>Aarhus Hospital DB (DK)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type of data</strong></td>
<td>FP</td>
<td>GP</td>
<td>Claims</td>
<td>Claims/Reg</td>
<td>GP</td>
<td>Claims/GP</td>
<td>Claims/Reg</td>
</tr>
<tr>
<td><strong>Drug code</strong></td>
<td>ATC</td>
<td>ATC</td>
<td>ATC</td>
<td>ATC</td>
<td>ATC</td>
<td>ATC</td>
<td>ATC</td>
</tr>
<tr>
<td><strong>Event code</strong></td>
<td>ICD9-CM</td>
<td>ICD9-CM</td>
<td>ICD9-CM</td>
<td>ICD9-CM</td>
<td>ICPC</td>
<td>ICD9-CM</td>
<td>ICD10</td>
</tr>
<tr>
<td><strong>Free text</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
D2 in EUADR: common data model

- **PERSONS**
  - PERSON_ID
  - GENDER_CONCEPT_ID
  - DATE_OF_BIRTH
  - STARTDATE
  - ENDDATE
  - GP_CODE

- **GP**
  - PERSON_ID
  - DIAGNOSIS
  - SPEC_DIAG
  - DATE
  - DRUG
  - DRUG_INDICATION

- **SPEC**
  - PERSON_ID
  - SPEC_CODE
  - SPEC_DATE
  - SPEC_DIAG

- **HOSP**
  - PERSON_ID
  - START_DATE
  - MAIN_DIAGNOSIS
  - SECONDARY_DIAGNOSIS_1-5
  - PROCEDURE_CODE_1-6
  - PROCEDURE_DATE_1-6

- **LAB**
  - PERSON_ID
  - PROC_CODE
  - PROC_DATE
  - PROC_RESULT

- **DRUGS**
  - PERSON_ID
  - DRUG_EXPOSURE_START_DATE
  - DDD
  - ATC

- **DEATH**
  - PERSON_ID
  - CAUSE_OF_DEATH
  - DEATH_DATE
# T2 in EU-ADR: harmonization

<table>
<thead>
<tr>
<th>Event</th>
<th>UMLS Concept</th>
<th>Preferred term</th>
<th>ICD9CM</th>
<th>ICD10</th>
<th>RCD</th>
<th>ICPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMI</td>
<td>C0155626</td>
<td>Acute myocardial infarction</td>
<td>410.x</td>
<td>I21.x</td>
<td>G30z., XE0Uu</td>
<td>K75, K75002</td>
</tr>
<tr>
<td></td>
<td>C0428953</td>
<td>ECG: myocardial infarction</td>
<td></td>
<td></td>
<td>323.., 323Z.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C0232320</td>
<td>ECG: antero-septal infarct.</td>
<td></td>
<td></td>
<td>3233</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C0428956</td>
<td>ECG: posterior/inferior infarct</td>
<td></td>
<td></td>
<td>3234</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C0428955</td>
<td>ECG: subendocardial infarct</td>
<td></td>
<td></td>
<td>3235</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C0232325</td>
<td>ECG: lateral infarction</td>
<td></td>
<td></td>
<td>3236</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C0428953</td>
<td>ECG: myocardial infarction</td>
<td></td>
<td></td>
<td>323.., 323Z.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C0340324</td>
<td>Silent myocardial infarction</td>
<td>X200a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C0340283</td>
<td>Only for refinement use</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other acute and subacute ischemic heart disease NOS</td>
<td>411</td>
<td></td>
<td>G31.., G31yz</td>
<td></td>
</tr>
<tr>
<td>AS</td>
<td>C0002792</td>
<td>Anaphylactic shock</td>
<td></td>
<td></td>
<td>T78.2</td>
<td>SN50</td>
</tr>
<tr>
<td></td>
<td>C0375697</td>
<td>Other anaphylactic shock</td>
<td></td>
<td></td>
<td>995</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C0685898</td>
<td>Anaphylactic shock due to adverse food reaction</td>
<td>995.6</td>
<td>T78.0</td>
<td>X70vm, X70w1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C0161840</td>
<td>Anaphylactic shock due to serum</td>
<td>999.4</td>
<td>T80.5</td>
<td>SP34, X70vl</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C0274304</td>
<td>Anaphylactic shock, due to adverse effect of correct medicinal substance properly administered</td>
<td></td>
<td></td>
<td>T88.6</td>
<td>SN501</td>
</tr>
</tbody>
</table>
Q in T2 in EU-ADR: evaluation of different event definitions through comparison of incidence rates

<table>
<thead>
<tr>
<th>Event</th>
<th>DB</th>
<th>IR for basic query</th>
<th>Additional data: IR and increment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>HOSP-main</td>
<td>GP</td>
</tr>
<tr>
<td>AMI</td>
<td>AARHUS</td>
<td>101.4</td>
<td>126.5 (+25%)</td>
</tr>
<tr>
<td></td>
<td>ARS</td>
<td>77.8</td>
<td>90.2 (+15%)</td>
</tr>
<tr>
<td></td>
<td>HSD</td>
<td>58.7</td>
<td>59.1 (+0.5%)</td>
</tr>
<tr>
<td></td>
<td>IPCI</td>
<td>148.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PHARMO</td>
<td>93.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LOMBARDY</td>
<td>82.5</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Event</th>
<th>DB</th>
<th>IR for basic Query</th>
<th>Additional data: IR and increment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>HOSP-main</td>
<td>GP</td>
</tr>
<tr>
<td>AS</td>
<td>AARHUS</td>
<td>5.7</td>
<td>6.4 (+12%)</td>
</tr>
<tr>
<td></td>
<td>ARS</td>
<td>12.0</td>
<td>12.7 (+6%)</td>
</tr>
<tr>
<td></td>
<td>HSD</td>
<td>5.2</td>
<td>12.8 (+0%)</td>
</tr>
<tr>
<td></td>
<td>IPCI</td>
<td>7.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PHARMO</td>
<td>1.9</td>
<td>2.4 (+26%)</td>
</tr>
<tr>
<td></td>
<td>LOMBARDY</td>
<td>2.2</td>
<td>2.8 (+27%)</td>
</tr>
</tbody>
</table>
Q in T2 in EU-ADR: evaluation through comparison of incidence rates
## Q in T2 in EU-ADR: validation

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Coding system</th>
<th>No. of cases confirmed (%)</th>
<th>Overall PPV (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GP/specialist diagnoses (IPCI, NL)</td>
<td>ICPC, free text</td>
<td>119 (29.8)*</td>
<td>46.5 (40.4 - 52.6)</td>
</tr>
<tr>
<td>GP/specialist diagnoses (HSD, Italy)</td>
<td>ICD9-CM, free text</td>
<td>115 (57.5)*</td>
<td>95 (91.2 - 98.9)</td>
</tr>
<tr>
<td>Primary hospital discharge diagnoses (Aarhus, DK)</td>
<td>ICD-10</td>
<td>148 (100)</td>
<td>100 (100 - 100)</td>
</tr>
</tbody>
</table>
Q in T3 in EU-ADR: methods evaluation

![Graph showing AUC of methods comparison](image-url)
D3 in EU-ADR: events deemed to be important for drug safety

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

- 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

Prescriptions
- Date
- PatientID
- Duration
- ATC

Events
- Date
- PatientID
- Event type

Population
- PatientID
- Birthdate
- Gender
- System entry date
- System exit date

Table 1: exclusive time /events
Table 2: Non-exclusive time /events
Table 3: general time /events
Table 4: Patient start profile
Table n..
## D1 in OMOP: original DBs

<table>
<thead>
<tr>
<th>Name</th>
<th>General Database Description</th>
<th>Pop Size (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GE Centricity Electronic Health Record (GE)</td>
<td>Derived from data pooled by providers using GE Centricity Office (an ambulatory electronic health record) into a data warehouse in a HIPAA-compliant manner.</td>
<td>11.2</td>
</tr>
<tr>
<td>MarketScan® Research Databases from Thomson Reuters</td>
<td></td>
<td>1.5</td>
</tr>
<tr>
<td><em>MarketScan Lab Database (MSLR)</em> - Represents privately insured population, with administrative claims from inpatient, outpatient, and pharmacy services supplemented by laboratory results.</td>
<td>11.1</td>
<td></td>
</tr>
<tr>
<td><em>MarketScan Medicaid Multi-State Database (MDCD)</em> - Contains administrative claims data for Medicaid enrollees from multiple states, containing inpatient, outpatient, and pharmacy services.</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td><em>MarketScan Medicare Supplemental and Coordination of Benefits Database (MDCR)</em> - 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.</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td><em>MarketScan Commercial Claims and Encounters (CCAE)</em> - 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.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
D2 in OMOP: Common Data Model
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.
Q in 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* 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 <=1.5 mg/dl, >=1.0 mg/dl for patients with a baseline level of 1.6–4.9 mg/dl, and >=1.5 mg/dl for patients with a baseline level >= 5.0 mg/dl
Q in T3 in OMOP: methods evaluation

- Tailoring method increased AUC for all outcomes
- Self-controlled cohort designs were optimal in MDCR across all outcomes, but parameters were different in each outcome
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
- ...
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.
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
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

```xml
<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>
  </dataset>
</iadMapping>
```

- 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)
T2 in MATRICE: scripting the change

**Scripting language** Data transformation expressed via a scripting language
- filter
- summarize
- classify
- ... 

From Jerboa The syntax of the scripting language was derived from Jerboa

Extend Jerboa? Cumbersome as flexible data model is needed
T2 in MATRICE: implementing a novel DSL

Domain Specific Languages  DSLs are programming languages that target a *specific problem area* (e.g., SAS language, Stata language, SQL, ...)

Interpreter  TheMatrix must include an interpreter for the DSL that we defined

Developing a DSL  DSLs are usually developed using specific frameworks:

- as compared to hand-writing, frameworks are generally easier to use and understand;
- also, employing a frameworks simplifies the maintenance activities.

We are planning to use one such framework

http://cazzola.di.unimi.it/neverlang.html
The Neverlang framework eases the development of DSLs and focuses on *modularity*. Componentisation makes maintenance and extensibility of the DSL implementation easier. Each basic syntactic construct of the DSL is defined separately (e.g. Script Modules). Each component defines the syntax of a part of the language (as a portion of a formal grammar) and its semantics. Neverlang uses the well-known *syntax-directed translation* technique. In this technique semantic actions are *attached* to nonterminals of the rules occurring in the formal grammar.
Example: Filter patients alive on a date $DATE$

**Script**

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
Example: Filter patients alive on a date $\$DATE$

FilterModule

```plaintext
module matrix.parser.FilterModule {
  role(syntax) {
    // main syntax of the FilterModule; has 2 sections "inputs" and "parameters"
    Module ← ModuleId ""FilterModule"
    "inputs" FilterInputDef
    "parameters" FilterParameterList
    "end";
    FilterInputDef ← InputName "" ModuleId
    FilterParameterList ← ConditionList BooleanExpr
    ConditionList ← Condition ConditionList;
    ConditionList ← Condition;
    // e.g. conditions = [{DATE_OF_DEATH, >, $DATE};{DATE_OF_BIRTH, <=, $DATE}]
    Condition ← "conditions" ""[
    ConditionSpecifierList ""
    ConditionSpecifierList ← ConditionSpecifier
    ConditionSpecifierList ← ConditionSpecifier;
    ConditionSpecifier ← "ColumnId", "ComparisonOp", "Variable"
    BooleanExpr ← "boolExpr" "" BooleanExprType;
  }
  role(evaluation) {
    $0.moduleObject = new FilterModule(
      $1.moduleId, // "Filter Alive Patients"
      $2.inputDef, // dataset = Patients File
      $2.parameterList // [{DATE_OF_DEATH, >, $DATE};{DATE_OF_BIRTH, <=, ecc.}
    );
    // other rules omitted here */
  }
}
```
Example: Filter patients alive on a date $DATE

Output Java: Syntax

```java
public class FilterModule$role$syntax extends Syntax {
    public FilterModule$role$syntax() {
        declareProductions(
            p(nt("Module"), nt("ModuleId"), "(" + "FilterModule" + ")", "inputs", nt("InputDef"), "parameters", nt("FilterParameterList"), "end"),
            p(nt("InputDef"), nt("InputId"), ",", nt("ModuleId")),
            p(nt("FilterParameterList"), nt("ConditionList"), nt("BooleanExpr")),
            p(nt("ConditionList"), nt("Condition"), nt("ConditionList")),
            p(nt("ConditionList"), nt("Condition")),
            p(nt("Condition"), "conditions", ",", "]"),
            p(nt("ConditionSpecifierList"), nt("ConditionSpecifier"), ",", nt("ConditionSpecifierList")),
            p(nt("ConditionSpecifierList"), nt("ConditionSpecifier")),
            p(nt("ConditionSpecifier"), ",", nt("ColumnId"), ",", nt("ComparisonOp"), ",", nt("Variable"), "]"),
            p(nt("BooleanExpr"), "boolExpr", "="),
        );
    }
}
```
Example: Filter patients alive on a date $DATE$

Output Java: Semantic Action attached to nonterminal n.0 (Module)

```java
public class FilterModule$role$evaluation$0 implements SemanticAction {
    public apply(ASTNode n) {
        n.moduleObject = new FilterModule(
            n.ntchild(0).moduleId, // "Filter Alive Patients"
            n.ntchild(1).inputDef, // dataset = Patients File
            n.ntchild(2).parameterList
                // [{DATE_OF_DEATH, >, $DATE}; {DATE_OF_BIRTH, <=, $DATE}] ecc.
        );
    }
}
```
T2 in MATRICE: programming TheMatrix via visual interface

Script generation  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

Usability  Usability tests will be performed for the VI to be usable by epidemiologies – no programming skills
Q in T2 in MATRICE: chronic diseases definition validation

ABC has IHD according to both algorithm 1 and 2
CBA has IHD according to 1 but not according to 2
BAC has IHD according to no algorithm
CAB has IHD according to no algorithm

ABC has IHD
CBA has not IHD
BAC has IHD
CAB has IHD

P₁ has IHD according to both algorithm 1 and 2, and has a diagnosis
P₂ has IHD according to 1 but not according to 2, and has no diagnosis
P₃ has IHD according to no algorithm, but has a diagnosis
P₄ has IHD according to no algorithm, and has no diagnosis
## D3 in MATRICE: stages of chronic diseases

<table>
<thead>
<tr>
<th>Diabetes</th>
<th>Hypertension</th>
<th>Heart failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncomplicated</td>
<td>Uncomplicated</td>
<td></td>
</tr>
<tr>
<td>Subclinical organ damage</td>
<td>Complicated, no heart failure</td>
<td></td>
</tr>
<tr>
<td>Complicated</td>
<td>Heart failure</td>
<td></td>
</tr>
<tr>
<td>Ischaemic heart disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coronaropathy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left ventricular dysfunction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heart failure</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
## Comparison wrt D1: original DBs

<table>
<thead>
<tr>
<th>Setting</th>
<th>attributes</th>
<th>EUADR</th>
<th>OMOP</th>
<th>MATR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>HOSP</td>
<td>main diagnosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>secondary diagnosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>procedures</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEATH</td>
<td>cause of death</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GP</td>
<td>diagnosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>specialist</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>lab</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>drug</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAB</td>
<td>classification</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>result</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPEC</td>
<td>classification</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>diagnosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EXE</td>
<td>diagnosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRUGS</td>
<td>classification</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Project</th>
<th>Where data transformation</th>
<th>Where data storage</th>
<th>Data format</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUADR</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>OMOP</td>
<td>Local</td>
<td>5 DBs in central/cloud database and 10 DBs local copies</td>
<td>Database</td>
</tr>
<tr>
<td>MATRICE</td>
<td>Local</td>
<td>Local</td>
<td>csv</td>
</tr>
</tbody>
</table>
Comparison wrt D2: common data models

<table>
<thead>
<tr>
<th>Project</th>
<th>Table classified according to</th>
<th>Attributes</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUADR</td>
<td>Setting of data recording</td>
<td>Clinical content (⁺.⁺...)</td>
<td>Data capture mechanism clearer</td>
</tr>
<tr>
<td>OMOP</td>
<td>Clinical content</td>
<td>Setting of data recording (⁺.⁺...)</td>
<td>Optimize performance of next data transformations</td>
</tr>
<tr>
<td>MATRICE</td>
<td>Setting of data recording</td>
<td>Clinical content (⁺.⁺...)</td>
<td>Easiest solution (it was an easy problem anyway)</td>
</tr>
</tbody>
</table>
## Comparison wrt T2: semantic transformations

<table>
<thead>
<tr>
<th>Project</th>
<th>Multiple vs unique within DB</th>
<th>Across DBs</th>
<th>Final choice</th>
<th>Estimates of algorithm validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUADR</td>
<td>Multiple</td>
<td>“Harmonization”</td>
<td>Y</td>
<td>Informal (but PPV for some)</td>
</tr>
<tr>
<td>OMOP</td>
<td>Multiple</td>
<td>Centralised</td>
<td>N</td>
<td>Informal</td>
</tr>
<tr>
<td>MATRICE</td>
<td>Multiple</td>
<td>Distributed</td>
<td>Y</td>
<td>Sensitivity, specificity, PPV, PPN</td>
</tr>
</tbody>
</table>
### Comparison wrt D3: events

<table>
<thead>
<tr>
<th>Project</th>
<th>Acute events</th>
<th>Conditions lasting for some time</th>
<th>Chronic conditions</th>
<th>Drug use as such</th>
<th>Use of other health services as such</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUADR</td>
<td>O</td>
<td>-</td>
<td>-</td>
<td>E, C</td>
<td>-</td>
</tr>
<tr>
<td>OMOP</td>
<td>O,C</td>
<td>-</td>
<td>C</td>
<td>E, C</td>
<td>-</td>
</tr>
<tr>
<td>MATRICE</td>
<td>O</td>
<td>-</td>
<td>X, C, O</td>
<td>O, C</td>
<td>O, C</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Project</th>
<th>Automatic</th>
<th>Common among DBs</th>
<th>Specific software</th>
<th>Programming language</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUADR</td>
<td>Y</td>
<td>Y</td>
<td>Jerboa</td>
<td>Java &amp; Jerboa scripting language</td>
</tr>
<tr>
<td>OMOP</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td>SQL, SAS, R, C, Java</td>
</tr>
<tr>
<td>MATRICE</td>
<td>Y</td>
<td>Y</td>
<td>TheMatrix</td>
<td>Java &amp; TheMatrix scripting language</td>
</tr>
</tbody>
</table>
## Comparison wrt D4: datasets for analysis

<table>
<thead>
<tr>
<th>Project</th>
<th>Type</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUADR</td>
<td>Intermediate files, analysis will follow</td>
<td>csv</td>
</tr>
<tr>
<td>OMOP</td>
<td>Final estimates, intermediate files discarded</td>
<td>csv/SAS datafiles/-database tables...</td>
</tr>
<tr>
<td>MATRICE</td>
<td>Intermediate files, analysis or report generation will follow</td>
<td>csv</td>
</tr>
</tbody>
</table>
Comparison wrt Q: quality cycle – documentation & automatization

<table>
<thead>
<tr>
<th>Project</th>
<th>T1 (reorganization)</th>
<th>T2 (semantics)</th>
<th>T3 (aggregation)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Documentation</td>
<td>Automatization</td>
<td>Documentation</td>
</tr>
<tr>
<td>EUADR</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>OMOP</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>MATRICE</td>
<td>Y</td>
<td>TheMatrix</td>
<td>Y</td>
</tr>
</tbody>
</table>
## Comparison wrt Q: quality cycle – feedback

<table>
<thead>
<tr>
<th>Project</th>
<th>T2: internal comparison</th>
<th>T2: external comparison</th>
<th>T2: validation</th>
<th>T3: methods comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUADR</td>
<td>Y</td>
<td>Y</td>
<td>some</td>
<td>Y</td>
</tr>
<tr>
<td>OMOP</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>MATRICE</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
</tbody>
</table>
Contents

Introduction

Conceptual framework

The three projects compared on the conceptual framework

Final considerations
Wrap up

Emerging need  The need for multi-centre epidemiology frameworks is emerging across continents
Wrap up

**Emerging need**  The need for multi-centre epidemiology frameworks is emerging across continents

**Challenges**  Extracting evidence from heterogeneous data sources presents multiple and novel challenges
Wrap up

Emerging need  The need for multi-centre epidemiology frameworks is emerging across continents

Challenges  Extracting evidence from heterogeneous data sources presents multiple and novel challenges

Conceptual framework  The conceptual framework is useful to highlight strengths and weaknesses of the frameworks that are being put in place
Wrap up

Emerging need  The need for multi-centre epidemiology frameworks is emerging across continents

Challenges  Extracting evidence from heterogeneous data sources presents multiple and novel challenges

Conceptual framework  The conceptual framework is useful to highlight strengths and weaknesses of the frameworks that are being put in place

Recommendations  Document (possibly automatize) T1; formally document, validate and automatize T2; document T3.
Wrap up

Emerging need The need for multi-centre epidemiology frameworks is emerging across continents

Challenges Extracting evidence from heterogeneous data sources presents multiple and novel challenges

Conceptual framework The conceptual framework is useful to highlight strengths and weaknesses of the frameworks that are being put in place

Recommendations Document (possibly automatize) T1; formally document, validate and automatize T2; document T3.

Goal Make the process faster, more transparent, more valid and more reliable
Wrap up

Emerging need  The need for multi-centre epidemiology frameworks is emerging across continents

Challenges  Extracting evidence from heterogeneous data sources presents multiple and novel challenges

Conceptual framework  The conceptual framework is useful to highlight strengths and weaknesses of the frameworks that are being put in place

Recommendations  Document (possibly automatize) T1; formally document, validate and automatize T2; document T3.

Goal  Make the process faster, more transparent, more valid and more reliable

Final goal  ...to better assist production of epidemiological evidence
Thanks for your attention!
Validation through comparison with a gold standard

<table>
<thead>
<tr>
<th>Classification</th>
<th>Gold standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Y</td>
<td>TP</td>
</tr>
<tr>
<td>N</td>
<td>FN</td>
</tr>
<tr>
<td>N</td>
<td>TN</td>
</tr>
</tbody>
</table>

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** 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 with a unique identifier (Concept Unique Identifier - CUI).

**Semantic network** Concepts are interlinked by relationships generally inherited from the original terminologies.