



# 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

## Citation

Gini, R., M. Schuemie, J. Brown, P. Ryan, E. Vacchi, M. Coppola, W. Cazzola, et al. 2016. "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 4 (1): 1189. doi:10.13063/2327-9214.1189. <http://dx.doi.org/10.13063/2327-9214.1189>.

## Published Version

doi:10.13063/2327-9214.1189

## Permanent link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:26318777>

## Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

## Share Your Story

The Harvard community has made this article openly available. Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

2-8-2016

# Data Extraction And Management In Networks Of Observational Health Care Databases For Scientific Research: A Comparison Among EU-ADR, OMOP, Mini-Sentinel And MATRICE Strategies

Rosa Gini

*Agenzia regionale di sanità della Toscana, Florence, Italy; and Erasmus MC University Medical Center, Department of Medical Informatics, Rotterdam, Netherlands, rosa.gini@ars.toscana.it*

Martijn Schuemie

*Janssen Research & Development, Epidemiology, Titusville, New Jersey, United States; and Observational Health Data Sciences and Informatics (OHDSI) New York, New York, United States, mschuemi@its.jnj.com*

Jeffrey Brown

*Harvard Medical School, Department of Population Medicine, Boston, Massachusetts, United States, jeff\_brown@harvardpilgrim.org*

Patrick Ryan

*Janssen Research & Development, Epidemiology, Titusville, New Jersey, United States; and Observational Health Data Sciences and Informatics (OHDSI) New York, New York, United States, ryan@omop.org*

See next pages for additional authors

Follow this and additional works at: <http://repository.edm-forum.org/egems>



Part of the [Epidemiology Commons](#), and the [Health Services Research Commons](#)

## Recommended Citation

Gini, Rosa; Schuemie, Martijn; Brown, Jeffrey; Ryan, Patrick; Vacchi, Edoardo; Coppola, Massimo; Cazzola, Walter; Coloma, Preciosa; Berni, Roberto; Diallo, Gayo; Oliveira, José Luis; Avillach, Paul; Trifirò, Gianluca; Rijnbeek, Peter; Bellentani, Mariadonata; van Der Lei, Johan; Klazinga, Niek; and Sturkenboom, Miriam (2016) "Data Extraction And Management In Networks Of Observational Health Care Databases For Scientific Research: A Comparison Among EU-ADR, OMOP, Mini-Sentinel And MATRICE Strategies," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*: Vol. 4: Iss. 1, Article 2.

DOI: <http://dx.doi.org/10.13063/2327-9214.1189>

Available at: <http://repository.edm-forum.org/egems/vol4/iss1/2>

This Informatics Comparative Case Study is brought to you for free and open access by the the Publish at EDM Forum Community. It has been peer-reviewed and accepted for publication in eGEMs (Generating Evidence & Methods to improve patient outcomes).

The Electronic Data Methods (EDM) Forum is supported by the Agency for Healthcare Research and Quality (AHRQ), Grant 1U18HS022789-01. eGEMs publications do not reflect the official views of AHRQ or the United States Department of Health and Human Services.

---

# Data Extraction And Management In Networks Of Observational Health Care Databases For Scientific Research: A Comparison Among EU-ADR, OMOP, Mini-Sentinel And MATRICE Strategies

## Abstract

**Introduction:** We see increased use of existing observational data in order to achieve fast and transparent production of empirical evidence in health care research. Multiple databases are often used to increase power, to assess rare exposures or outcomes, or to study diverse populations. For privacy and sociological reasons, original data on individual subjects can't be shared, requiring a distributed network approach where data processing is performed prior to data sharing.

**Case Descriptions and Variation Among Sites:** We created a conceptual framework distinguishing three steps in local data processing: (1) data reorganization into a data structure common across the network; (2) derivation of study variables not present in original data; and (3) application of study design to transform longitudinal data into aggregated data sets for statistical analysis. We applied this framework to four case studies to identify similarities and differences in the United States and Europe: Exploring and Understanding Adverse Drug Reactions by Integrative Mining of Clinical Records and Biomedical Knowledge (EU-ADR), Observational Medical Outcomes Partnership (OMOP), the Food and Drug Administration's (FDA's) Mini-Sentinel, and the Italian network—the Integration of Content Management Information on the Territory of Patients with Complex Diseases or with Chronic Conditions (MATRICE).

**Findings:** National networks (OMOP, Mini-Sentinel, MATRICE) all adopted shared procedures for local data reorganization. The multinational EU-ADR network needed locally defined procedures to reorganize its heterogeneous data into a common structure. Derivation of new data elements was centrally defined in all networks but the procedure was not shared in EU-ADR. Application of study design was a common and shared procedure in all the case studies. Computer procedures were embodied in different programming languages, including SAS, R, SQL, Java, and C++.

**Conclusion:** Using our conceptual framework we found several areas that would benefit from research to identify optimal standards for production of empirical knowledge from existing databases.

## Acknowledgements

This study was funded by the project named "Integrazione dei contenuti informativi per la gestione sul territorio di pazienti con patologie complesse o con patologie croniche", acronym MATRICE, funded by the Italian Ministry of Health in the framework of the MATTONI Program.

## Keywords

Electronic Health Records, Health Services Research, Research networks, Data reuse, Pharmacoepidemiology, Data management

## Disciplines

Epidemiology | Health Services Research

---

## Creative Commons License



This work is licensed under a [Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 License](https://creativecommons.org/licenses/by-nc-nd/3.0/).

## Authors

Rosa Gini, *Agenzia regionale di sanità della Toscana, Florence, Italy; and Erasmus MC University Medical Center, Department of Medical Informatics, Rotterdam, Netherlands*; Martijn Schuemie, *Janssen Research & Development, Epidemiology, Titusville, New Jersey, United States; and Observational Health Data Sciences and Informatics (OHDSI) New York, New York, United States*; Jeffrey Brown, *Harvard Medical School, Department of Population Medicine, Boston, Massachusetts, United States*; Patrick Ryan, *Janssen Research & Development, Epidemiology, Titusville, New Jersey, United States; and Observational Health Data Sciences and Informatics (OHDSI) New York, New York, United States*; Edoardo Vacchi, *Università degli Studi di Milano, Dipartimento di Informatica, Milan, Italy*; Massimo Coppola, *Consiglio Nazionale delle Ricerche, Istituto di Scienza e Tecnologie dell'Informazione, Pisa, Italy*; Walter Cazzola, *Università degli Studi di Milano, Dipartimento di Informatica, Milan, Italy*; Preciosa Coloma, *Erasmus MC University Medical Center, Department of Medical Informatics, Rotterdam, Netherlands*; Roberto Berni, *Agenzia regionale di sanità della Toscana, Florence, Italy*; Gayo Diallo, *Université Bordeaux, LESIM - ISPED, Bordeaux, France*; José Luis Oliveira, *University of Aveiro, DETI/IEETA, Aveiro, Portugal*; Paul Avillach, *Harvard Medical School, Center for Biomedical Informatics, Boston, Massachusetts, United States*; Gianluca Trifirò, *Erasmus University Medical Center, Department of Medical Informatics, Rotterdam, Netherlands*; Peter Rijnbeek, *Erasmus University Medical Center, Department of Medical Informatics, Rotterdam, Netherlands*; Mariadonata Bellentani, *Agenzia nazionale per i servizi sanitari regionali, Rome, Italy*; Johan van Der Lei, *Erasmus University Medical Center, Department of Medical Informatics, Rotterdam, Netherlands*; Niek Klazinga, *University of Amsterdam, Academic Medical Center, Amsterdam, Netherlands*; Miriam Sturkenboom, *Erasmus University Medical Center, Department of Medical Informatics, Rotterdam, Netherlands*.



# 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

Rosa Gini;<sup>i,ii</sup> Martijn Schuemie;<sup>iii,iv</sup> Jeffrey Brown;<sup>v</sup> Patrick Ryan;<sup>iii,iv</sup> Edoardo Vacchi;<sup>v</sup> Massimo Coppola;<sup>vii</sup> Walter Cazzola;<sup>vi</sup> Preciosa Coloma;<sup>ii</sup> Roberto Berni;<sup>i</sup> Gayo Diallo;<sup>viii</sup> José Luis Oliveira;<sup>ix</sup> Paul Avillach;<sup>v</sup> Gianluca Trifirò;<sup>ii</sup> Peter Rijnbeek;<sup>ii</sup> Mariadonata Bellentani;<sup>x</sup> Johan van Der Lei;<sup>ii</sup> Niek Klazinga;<sup>xi</sup> Miriam Sturkenboom<sup>ii</sup>

## ABSTRACT

**Introduction:** We see increased use of existing observational data in order to achieve fast and transparent production of empirical evidence in health care research. Multiple databases are often used to increase power, to assess rare exposures or outcomes, or to study diverse populations. For privacy and sociological reasons, original data on individual subjects can't be shared, requiring a distributed network approach where data processing is performed prior to data sharing.

**Case Descriptions and Variation Among Sites:** We created a conceptual framework distinguishing three steps in local data processing: (1) data reorganization into a data structure common across the network; (2) derivation of study variables not present in original data; and (3) application of study design to transform longitudinal data into aggregated data sets for statistical analysis. We applied this framework to four case studies to identify similarities and differences in the United States and Europe: Exploring and Understanding Adverse Drug Reactions by Integrative Mining of Clinical Records and Biomedical Knowledge (EU-ADR), Observational Medical Outcomes Partnership (OMOP), the Food and Drug Administration's (FDA's) Mini-Sentinel, and the Italian network—the Integration of Content Management Information on the Territory of Patients with Complex Diseases or with Chronic Conditions (MATRICE).

**Findings:** National networks (OMOP, Mini-Sentinel, MATRICE) all adopted shared procedures for local data reorganization. The multinational EU-ADR network needed locally defined procedures to reorganize its heterogeneous data into a common structure. Derivation of new data elements was centrally defined in all networks but the procedure was not shared in EU-ADR. Application of study

<sup>i</sup>Agenzia Regionale di Sanità della Toscana, <sup>ii</sup>Erasmus MC University Medical Center, <sup>iii</sup>Janssen Research & Development, Epidemiology, <sup>iv</sup>Observational Health Data Sciences and Informatics (OHDSI), <sup>v</sup>Harvard Medical School, <sup>vi</sup>Università degli Studi di Milano, Dipartimento di Informatica, <sup>vii</sup>Consiglio Nazionale delle Ricerche, Istituto di Scienza e Tecnologie dell'Informazione, <sup>viii</sup>Université Bordeaux, LESIM - ISPED, <sup>ix</sup>José Luis Oliveira University of Aveiro, DETI/IEETA, <sup>x</sup>Agenzia Nazionale per i Servizi Sanitari Regionali, <sup>xi</sup>University of Amsterdam, Academic Medical Center

## CONTINUED

design was a common and shared procedure in all the case studies. Computer procedures were embodied in different programming languages, including SAS, R, SQL, Java, and C++.

**Conclusion:** Using our conceptual framework we found several areas that would benefit from research to identify optimal standards for production of empirical knowledge from existing databases. an opportunity to advance evidence-based care management. In addition, formalized CM outcomes assessment methodologies will enable us to compare CM effectiveness across health delivery settings.

### Introduction

Observational studies based on secondary use of existing data collected in the process of health care delivery have the potential to deliver sound evidence quickly enough to support health policy making, which it is often subject to time constraints [Salmon2012], thus complementing evidence generated by means of primary data collection. However, some epidemiological questions, especially those concerning rare events, rare exposures, and small groups of patients, require more data than is available in any single observational database.<sup>2,3,4</sup> Therefore a growing number of studies use data from *networks* of databases, sometimes from different countries. Although some of these networks were formed ad hoc for a particular study, several more permanent networks have now been established, where the partners have agreed on an infrastructure and workflow to be reused for different studies.

Privacy regulations and concerns about data ownership and interpretation prevent easy central pooling of original health care data that is now stored in different databases and can be used for

secondary purposes.<sup>5</sup> In spite of these barriers several approaches can be used to still employ this data for secondary purposes and pool the results. For example, investigators at each data source can independently create a protocol and execute the study, and estimates are only generated afterward through meta-analysis. A further step is to share the protocol across sites, but asking the local partners to adapt it to their local data and to implement it in their own usual software, to produce local estimates for meta-analysis that are compatible by design. However, most networks now go even further and adopt a *distributed analysis approach*: each database is locally transformed to a representation that is similar across the network, and one single computer program performing the analysis is shared and executed at each site.<sup>4,6</sup>

The need to pool data across different databases is most pronounced in the area of drug safety surveillance.<sup>7</sup> In Europe, the Exploring and Understanding Adverse Drug Reactions by Integrative Mining of Clinical Records and Biomedical Knowledge Project (EU-ADR)<sup>8,9</sup> was initiated in 2008 for investigating the feasibility of signal detection across multiple health care databases.



Meanwhile, the United States Food and Drug Administration's (FDA's) Mini-Sentinel Project<sup>10</sup> was developed to support medical product safety monitoring and now includes 18 data partners within a distributed network. Also in the United States, from 2010 to 2014 the Observational Medical Outcomes Partnership (OMOP)<sup>11</sup> performed methodological research on drug safety studies and developed tools and a database network for performing risk identification. Other networks have been developed in other countries, like the Canadian Network for Observational Drug Effect Studies (CNODES) project in Canada and the Asian Pharmacoepidemiology Network (ASPEN) network in Asia.<sup>4</sup> Pharmacoepidemiology is not the only field where the opportunities for combining multiple databases are increasing: in the context of public health or health services research, gathering data from different regions or countries has the added value that different policies can be compared. Mini-Sentinel and EU-ADR are also used to evaluate the impact of regulatory actions. And the Italian network—the Integration of Content Management Information on the Territory of Patients with Complex Diseases or with Chronic Conditions (*Integrazione dei Contenuti Informativi per la Gestione sul Territorio di Pazienti con Patologie Complesse o con Patologie Croniche*)(MATRICE) Project,<sup>12,13</sup> funded by the Italian Ministry of Health—created a distributed network to evaluate the impact of health policies on quality and equity of health care.

We developed a conceptual framework to analyze the process of data management in a network of databases adopting the distributed analysis approach to perform observational studies. We applied the framework to four case studies, and identified similarities and substantial differences.

### Purpose and Target of This Study

The purpose of this study was to compare processes that share the same aim but are presently described in separate scientific papers or other documents. Our intent was to find which choices were common among different networks and what the differences were. The comparison findings highlight topics for research. Research should be aimed to further explore if common choices are indeed optimal, and to assess which among the observed differences have an impact on the quality of the processes and on the generated evidence: as such, our findings may be of interest for researchers in medical informatics and methodologists of observational studies. Moreover, the framework and the findings from the comparison provide a unified presentation of strategic choices that are of interest to researchers who are setting or modifying their own networks.

### Methods

#### Sampling and Data Collection

Some of the paper's authors first conceived of the conceptual framework as an abstraction of the process in place in the European network EU-ADR and in the Italian network MATRICE. They reached out to the authors participating in the United States networks OMOP and Mini-Sentinel, to compare networks of different continents. Data collection was performed via document (scientific papers and websites) analysis and interviews with coauthors. The manuscript was reviewed by all the authors.

#### The Four Networks

The EU-ADR Project was funded by the European Commission under Framework Programme 7 (FP7) and ran from 2008 to 2012 with the aim of producing a computerized integrated system for the early detection of drug safety signals. The

project used data from eight databases from four European countries (Denmark, Italy, the Netherlands, and the United Kingdom) covering a population of about 20 million individuals overall with almost 60 million person-years (PYs) of follow-up.<sup>3</sup> Subsequently, the EU-ADR workflow has been further improved and applied in several collaborative drug-safety studies concerning NSAIDs (SOS),<sup>14,15</sup> pandemic influenza vaccine (VAESCO),<sup>16</sup> the arrhythmogenic potential of drugs (ARITMO),<sup>17</sup> and hypoglycemic drugs (SAFEGUARD).<sup>18</sup> The subjects of the studies performed in this network include methodology<sup>19,20,21,22</sup> drug utilization, disease incidence,<sup>23</sup> signal detection,<sup>24</sup> testing,<sup>25,26</sup> filtering,<sup>27</sup> and substantiation.<sup>28</sup> The workflow is currently being extended in the European Medical Information Framework (EMIF) project.<sup>29</sup>

The United States FDA's Mini-Sentinel program<sup>30</sup> began in 2008 and has created a distributed data network of 18 data partners covering a population of over 150 million persons and 380 million PYs in the United States [Curtis2012].<sup>6</sup> Mini-Sentinel was structured to produce both fast, standardized replies to specific queries (called Rapid Response queries [MiniSRRQ2013, MiniMP2013]) and studies based on ad hoc developed protocol (i.e., Protocol-based Assessments [MiniPBA2013]). Hundreds of Rapid Response queries are executed each year, and 14 Protocol-based Assessments have been completed or are underway. Network activities cover a broad range of topics including drug utilization, disease burden, the impact of regulatory policies, and the comparative safety of medical products.<sup>31</sup> At the same time, several studies focusing on methodology have been completed.<sup>32,33,34,35,36,37</sup> In 2015 the Mini-Sentinel pilot transitioned to the Sentinel system that is become part of the FDA's regulatory framework.

OMOP was a public-private partnership that ran from 2010 to 2014 and was part of the Innovation in Medical Evidence Development and Surveillance (IMEDS) program of the Reagan-Udall Foundation for the FDA. Its goal was to help determine best practices for use of observational health care data. OMOP currently maintained five commercial databases covering 164.9 million persons in its own central venue, and its data partner network included six other databases covering an additional 105 million persons [Stang2010].<sup>8,38</sup> The network was used to develop tools for performing observational studies in a database network, including the OMOP Common Data Model (CDM),<sup>39</sup> the OMOP Vocabulary,<sup>40</sup> and tools for assessing data quality,<sup>41</sup> as well as research into the development and evaluation of methods for drug-associated risk identification.<sup>42</sup> In 2014 the OMOP research team launched the Observational Health Data Sciences and Informatics (OHDSI) (pronounced "Odyssey") program<sup>43</sup> which is currently continuing the activity of OMOP.

The MATRICE project was funded by the Italian Ministry of Health and ran from 2011 to 2014 under the coordination of the Italian National Agency for Regional Health Services to measure quality of health care for chronic diseases. MATRICE developed a distributed network infrastructure specific to local and regional Italian administrative databases and is rapidly growing to include participants beyond the project. Currently, it covers a population of about 9 million subjects living in some of the Local Health Authorities in 9 of the 21 regional health care systems in the country. Studies completed so far using data from this network were aimed at evaluating the quality and equity of primary care, the impact of policies in this field [Visca2013,<sup>44,45,46,47</sup> and methodological challenges of such studies [Gini2014].<sup>48,49</sup> The network currently participates in several studies funded by the Italian Ministry of Health.



## Conceptual Framework

Figure 1 depicts our conceptual framework, showing a workflow consisting of data sets (D1, D2, D3, and D4) and transformation processes (T1, T2, and T3). The conceptual framework does not contain recommendations in itself: it is just a conceptual abstraction of the logical sequence of steps needed to perform studies in a network.

Figure 2 describes each step in detail. During a typical study, data transformation T2 and T3 might be performed iteratively: if additional analyses are

required to shed light on preliminary results, then T3 or both T2 and T3 can be repeated and new D4 can be produced to undergo statistical analysis. In some studies T2 (data derivation) may not be performed, if data needed for the study are all contained in the original data.

To ensure that T1-T3 are valid, both in terms of how well the transformation reflects the original data and of whether it achieves the aim of the transformation, quality control processes need to be in place. In Box 1 process and outcome verification steps are highlighted.

**Figure 1. Flowchart of the Data Transformation Process Occurring Locally in a Study Collecting Data from a Network of Databases**



D1, D2, D3, and D4 represent data sets; T1, T2, and T3 represent data transformations.

### Box 1. Definition of the Conceptual Framework

#### DATA SETS AND DATA TRANSFORMATIONS:

D1 (original databases: DBs) is a collection of data sources controlled by a single organization that has procedures in place to link them with each other at the individual level, thus creating a single data pool on the same subjects. The term “DB” refers to an organization that has access to the data.

T1 (data reorganization) is a data modeling step: transformation from the locally defined data repository into a global (common) schema with standardized variable and attribute names, without loss of information. Simple one-on-one recoding is performed as well, such as making data formats and coding of attributes (e.g., gender) identical. T1 is specific per DB but independent of the specific study.

D2 (global schema, GS) is a general database schema that contains all the attributes that are necessary to answer a realm of study questions (“use cases”) that are of general interest to the network, such as incidence of disease, drug utilization, or association studies. D2 has a defined set of table names, attribute names, and formats. D2 plays the same role as a GS of a data integration system.<sup>50</sup> Therefore, a set of correspondences are defined between this schema and the D1. Note that (1) these correspondences may not be complete for all databases: for instance, if a D1 does not have information about primary care diagnoses, these attributes will remain empty in the D2; and (2) some attributes (typically, diagnoses or drugs) might have different coding for different DBs in the network.

**Box 1. Definition of the Conceptual Framework (Cont'd)****DATA SETS AND DATA TRANSFORMATIONS (CONT'D):**

T2 (data derivation) is the step where novel meaning is obtained from D2 by means of an explicit manipulation and combination of D2 data. These manipulations are necessary when a study variable is not among those collected by one of the DBs in the network, and must therefore be represented, by proxy, as a combination of whatever pertinent information is available. When the study variable is a disease, this process is referred to in the literature as disease phenotyping.<sup>51</sup> T2 is often specific per DB, as it depends on the information that was originally collected, and is often specific per study, although conceivably past data derivations could be reused in new studies. As an example of T2, if the presence of diabetes in study subjects needs to be assessed, DBs collecting data from primary care can identify the information from a general practitioner's (GP's) diagnosis, whereas claims databases without clinical data from primary care may use dispensing of antidiabetic drugs as proxy, and combinations may also be possible.

D3 (derived data) are the data sets derived in T2, each containing one or more study-specific variables. Derived data may be occurrence of a disease, or other information like the duration of exposure to a specific drug. For instance a drug safety study has three basic types of derived data: the outcome of interest (often sudden occurrence of a condition), the exposure (a sequence of drug utilization episodes), and presence in the study cohort, with beginning and end dates of follow-up. While the tables for D2 contain multiple, longitudinal observations per subject, each generated during an encounter and each containing multiple codes, D3 contains as many observations per subject as requested by the study design (often one single observation). Original data (as modeled in D2) is therefore "rolled up" during T2 to create in D3 the best possible approximation of the variables needed in the specific study.

T3 (study design application) is data transformation for a specific analytic: based on the protocol of a study with specific design (application of inclusion and exclusion criteria, selection of exposure windows, propensity and disease score estimation, control selection, matching). T3 produces the data sets for statistical analysis. Within this transformation data may be de-identified and aggregated to various levels. T3 is specific to the study, but is the same across participant DBs.

D4 (data sets for analysis) is the result of T3. D4s from all the partners in the network are similar. Based on the level of sharing that is allowed, D4 may stay local at the database custodian or be pooled in a central repository. In both situations, statistical analysis on D4 follows and produces estimates to be interpreted.

**QUALITY:**

1. **Process verification:** assuring quality, transparency and reproducibility of the *stepwise* data extraction process, e.g., common standard process documentation, process automatization with common use of dedicated software, and parallel programming; and
2. **Outcome verification:** checking intermediate and final output against standards, including the following:
  - Benchmarking of D3 (derived data) against external data (e.g., determining whether observed disease rates are in line with those reported in literature);
  - Benchmarking of D3 within the network (comparison of DB-specific output to assess homogeneity);
  - Validation of D3 using a gold standard (e.g., chart review) to assess performance of data derivation (e.g., positive predictive value); and
  - Validation of D4 using expected results (i.e., using a reference set of known causal or noncausal associations).



To illustrate the steps of the workflow, an example from the MATRICE network is shown in Box 2.

### Box 2. An Example of Data Management in the MATRICE Network

The Italian National Agency for Regional Health Services promoted a study to assess whether regional Italian administrative databases can be used to measure whether patients with Chronic Obstructive Pulmonary Disease (COPD) are treated with recommended therapies. The study objective was to establish whether different cohorts, defined with different case-identification strategies, resulted in consistent estimates of therapy adherence. The MATRICE network was used for this study.

Five regions were involved in the study. In each Italian region several tables of administrative data are collected with content regulated by national law, in particular the following: the list of residents (citizens and regular migrants) entitled to receive health care; hospital discharge records, with six diagnosis codes; exemptions from copayments for health care; and drug prescriptions. In each region participating in the study, a copy of the four tables (D1) was stored, with different data models and format. The MATRICE network has established a specific data model for the above mentioned four tables (list of residents; hospital discharge records; exemptions from copayments for health care; and drug prescriptions), and the format is flat comma-separated files (D2).

Two of the regions had already participated in a previous study of the MATRICE network, so T1 had already been performed. In the other three regions, the format D2 was explained to a local expert by means of structured documents and a teleconference, a common software named TheMatrix was installed (see “T1 (reorganization)” in the “FINDINGS” Section below), and T1 was performed by the local expert and was checked with standard procedures embedded in the software.

The study protocol had defined several variables to be extracted or derived: gender, presence in the region at index date, age at index date, presence of a COPD diagnosis in the 1–5 years before index date, presence of some patterns of utilization of respiratory drugs in the 1–3 years before index date, and adherence to recommended therapies during follow-up. D3 was composed of a group of data sets, one per derived variable, each with a single observation per subject. Since in MATRICE all the participating data partners share the same data content (see “D1: original DBs” in the “FINDINGS” Section below), the transformation T2 was uniform across data partners. T2 was therefore embedded by the principal investigator in a single ad hoc procedure of the software TheMatrix, shared with the local partners and executed locally.

The data set D4 was designed in the protocol to be the aggregated data set that counted the frequency of each combination of the variables in D3. The transformation T3 was embedded by the principal investigator in another ad hoc procedure of the software TheMatrix, shared with the local partners and executed locally.

The D4s produced by the five regions were shared with the principal investigator, who executed the statistical analysis of the pooled data set using the statistical software Stata 13.1.

## Findings

We describe and compare T1-T3 and D1-D4 in the four networks.

### D1 (Original DBs)

We use “DB” to refer to an organization that has access to the data. Table 1 summarizes the DBs participating in the four networks. For each network a column represents a combination of data sources that are linked in at least one database. We classified data sources according to their provenance, and we indicated the data items available in the DB from that data source. If more than one DB in a network share the same combination, only one column is shown: the number of columns for a network in Table 1 is therefore a measure of heterogeneity of the DBs participating in the network. MATRICE has a single combination (M1), EU-ADR has seven (EU1-EU7), Mini-Sentinel has three (MS1-MS3), and OMOP has four (O1-O4).

### Differences and Similarities

First, in the two United States-based networks (OMOP and Mini-Sentinel) almost all databases (O1-O3 and MS1-MS3) obtain administrative information from primary, secondary, and inpatient care, while in both European networks (EU-ADR and MATRICE) each database lacks at least one setting. Second, EU-ADR pools data from the most heterogeneous databases: the eight databases showed seven different combinations. Third, in Italy, although administrative information from secondary care (such as specialty of the physician visiting the patient) is available, it does not contain diagnostic

codes (M1 and EU1-EU2). Fourth, access to laboratory test results is rare among databases in all networks. Fifth, in all but one United States database, enrollment of subjects in the data collection is due to the eligibility criteria for social insurance or an insurance company, while in Europe criteria include geographical residence or being listed with a GP. Sixth, only in EU-ADR and Mini-Sentinel are death and immunization registries available. Finally, only Mini-Sentinel involves partners collecting information from both clinical and administrative data sources. This is achieved by integrated delivery systems that operate medical facilities from which they collect electronic health care records data.

In addition, all the partners of Mini-Sentinel and some partners of the other networks can access full-text medical records for chart validation for their population.

Box 3 is a fictional example of the impact of the differences in D1 on the information captured from a patient history.

### T1 (Reorganization)

In Table 2, T1 is compared across case studies.

### Differences and Similarities

Besides local storage, in OMOP some databases also allow creating a central and cloud-based copy of the transformed data. In MATRICE and Mini-Sentinel, all original databases used the same coding systems, while in OMOP participating databases used different coding systems and even unstructured free text in different languages, in EU-ADR.



Table 1. Description of the D1 (Original DBs) Databases in Terms of Provenance and Data Items Collected from Each Data Source

PROVENANCE OF DATA SOURCE	COMBINATIONS OF DATA SOURCES AND DATA ITEMS AVAILABLE IN THE DBS OF THE NETWORK																
	IN MATRICE			IN EU-ADR			IN MINI-SENTINEL			IN OMOP							
	MA1	EU1	EU2	EU3	EU4	EU5	EU6	EU7	MS1	MS2	MS3	O1	O2	O3	O4		
Primary care	Administrative data																
	Clinical data					Dx	Dx Rx Text Refspec Refinpat Vac Res	Dx Rx Text Refspec Refinpat Vac Res								Dx Rx	
Secondary care	Administrative data	Spec Proc	Spec Proc	Spec Proc	Spec Proc	Spec Proc	Spec Proc	Spec Proc	Spec Proc	Spec Proc	Spec Proc						
	Clinical data																Dx Rx Text
Inpatient care	Administrative data	Dx Proc	Dx Proc	Dx Proc	Dx Proc	Dx Proc	Dx Proc	Dx Proc	Dx Proc	Dx Proc	Dx Proc						
	Clinical data																Dx Rx Text
Enrollment into the data collection	Pharmacies	Geo	Geo	Geo	Geo	Geo	Geo	Geo	Geo	Geo	Geo	Geo	Geo	Geo	Geo	Geo	Elig
	Registry of disease-specific exemptions from copayment of healthcare	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Elig
Death registry	Pharmacies	Rx	Rx	Rx	Rx	Rx	Rx	Rx	Rx	Rx	Rx	Rx	Rx	Rx	Rx	Rx	Rx
	Registry of disease-specific exemptions from copayment of healthcare	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Elig
Vaccination registry	Pharmacies																Rx
	Registry of disease-specific exemptions from copayment of healthcare	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Dx	Elig
Laboratory	Pharmacies	Lab	Lab	Lab	Lab	Lab	Lab	Lab	Lab	Lab	Lab	Lab	Lab	Lab	Lab	Lab	Res
	Registry of disease-specific exemptions from copayment of healthcare																Res

Notes: If more than one database in a network has access to the same combination of data, they are represented by a single column. Data items—Dx: diagnostic codes; Proc: procedure codes; Rx: prescriptions or dispensings of drugs; Spec: specialty of secondary care encounters; Refspec: Referrals from secondary care; Refinpat: Referrals from inpatient care; Text: Notes in free text; Lab: Labels of laboratory tests; Res: Laboratory test results; Vac: Vaccines; Geo: Presence in a geographical area; Charge: Being in assisted by a GP; Elig: Satisfying eligibility criteria for an insurance company or health plan.

### Box 3. A Short Illustration of the Differences in Original Data

In 2005, Irina, age 36, developed gestational diabetes during her second pregnancy, which was diagnosed by her gynecologist and treated with insulin prescribed by her GP. Irina gave birth to Louise in a hospital, and had her vaccinated against tetanus and diphtheria when the baby was six months old. The following year Irina's father Mario, age 67 and a smoker with a history of coronary heart disease, moved to the region where Irina lived. In 2007, Mario was diagnosed with diabetes by his GP, who was also his daughter's GP. After trying for a while to cope with his condition only through following a new diet, he started taking antidiabetic drugs in 2008. In 2010 he had severe angina and was admitted to the hospital for a few days. In 2013 Mario died in his sleep, and his death certificate indicated that the cause of death was myocardial infarction.

If Irina, Louise, and Mario were part of the database population of the four networks, the image of the story would be different. For databases lacking diagnosis from primary or secondary care, like M1 or EU1-EU4, Irina's beginning to take insulin could be misinterpreted as an occurrence of diabetes, even though a complex algorithm using hospital admittance for delivery or the ending of insulin prescriptions could effectively avoid misclassification. Louise's vaccine would be detected by MS1, MS2, and MS3. When Mario moved to Irina's region and entered the database population, only databases collecting clinical history from primary care—like EU6, EU7, MS1, and O4—could have detected that he was the father of Irina and was a smoker. While the history of coronary heart disease could also be deduced from the same databases or clinical notes of a cardiologist in MS1, the presence of the disease may be inferred from drug utilization data in all the databases, and angina precisely in 2010 in databases with diagnoses from inpatient care (MA1, EU1-EU4, all MS, and O1-O3). Diabetes would be detected in 2007 from primary care diagnosis in EU5-EU7 and all the United States databases, and in 2008 only from drug utilization in the others. Occurrence of myocardial infarction would be detected only by EU2, EU3, and all the MS databases.

Different strategies were adopted to transform the original data into a common data set: in EU-ADR, the transformation T1 was used only in internal discussions to define T2, and data sets in the common data model were never created.

In MATRICE, standard procedures for T1 are in place, and results are evaluated by local partners. In Mini-Sentinel, data is transformed to a general, common data model and is updated frequently; and checks for data completeness and consistency with the data model are Standard Operating Procedures (SOP) executed as part of each transformation and approval process.<sup>52</sup> OMOP recoded all data to a single system during T1, independently of a

study question, the transformation in T1 is evaluated by first generating descriptive statistics of all elements in D2 using a tool called Observational Source Characteristics Report (OSCAR), and by subsequently performing internal and external comparison of these statistics using a tool called Generalized Review of OSCAR Unified Checking (GROUCH). Both in OMOP and Mini-Sentinel, a formal Extraction, Transformation, and Loading (ETL) document is created as part of development and implementation of the data model. In MATRICE the transformation is executed via ad hoc software, called TheMatrix,<sup>53</sup> whose configuration is stored in a text file.



Table 2. Comparison with Respect to T1, D2, T2

T1 (DATA REORGANIZATION)				
NETWORK	RECODING	QUALITY: DATA COMPLETENESS	QUALITY: DOCUMENTATION	
<b>EU-ADR</b>	Does not require mapping to external standard: original coding and/or free text is maintained	Demanded to local partners, no formal procedure	No formal documentation	
<b>Mini-Sentinel</b>	Source data are homogeneous in coding systems	Local report on specific issues + feedback from standard programs checking for completeness and consistency	Data model, data elements and guiding principles approved by partners. ETL formal document, ad hoc per DB	
<b>OMOP</b>	Source data standardized to common vocabulary by domain: Drug (RxNorm), Condition (SNOMED), Labs (LOINC)	Formal procedures: OSCAR and GROUCH tools	ETL formal document, ad hoc per DB	
<b>MATRICE</b>	Source data are homogeneous in coding systems	Formal procedures checking data completeness	Local configuration of the TheMatrix software (text file)	
D2 (GLOBAL SCHEMA)				
NETWORK	NAMES OF TABLES CHOSEN ACCORDING TO	NAMES OF ATTRIBUTES CHOSEN ACCORDING TO	EVERY TABLE OF THE CDM HAS A VIEW IN EVERY DB	ATTRIBUTES ARE CODED UNIFORMLY ACROSS DBS
<b>EU-ADR</b>	Reason/setting of data recording	Clinical contents	N	N
<b>Mini-Sentinel</b>	Clinical content and data source (diagnosis, procedures, encounters, lab results) or reason/setting (outpatient pharmacy, death, enrollment)	Reason/setting of data recording for diagnosis and similar, clinical contents for pharmacy and death	N	Y
<b>OMOP</b>	Clinical content	Reason/setting of data recording	Y	Y
<b>MATRICE</b>	Reason/setting of data recording	Clinical contents	Y	Y

Table 2. Comparison with Respect to T1, D2, T2 (Cont'd)

T2 (DATA DERIVATION)				
NETWORK	LOGIC	SINGLE DEFINITION PER DERIVED DATA	QUALITY: PROCESS CONTROL	QUALITY: VALIDATION
<b>EU-ADR</b>	DB-specific algorithms, harmonized through a formal negotiation process	Y	No common procedures were implemented., although logic of local procedures was shared	Internal incidence rates comparison, comparison with literature, some validation with external gold standard (PPV)
<b>Mini-Sentinel</b>	The same algorithm was used across all DBs	Y	Shared SAS script	Systematic review of previously published validation studies, expert clinical, data, and epidemiologic guidance, medical chart review for PPV and assessment of difference in dates
<b>OMOP</b>	Multiple alternative algorithms were adopted to derive the same data, some were DB-specific	N	Shared parameterized SQL queries stored in common procedure (RICO)	Internal prevalence rates comparison, no external validation performed
<b>MATRICE</b>	Multiple algorithms were explored, decision was taken by means of a validation study	Y	Shared script in a scripting language developed ad hoc (TheMatrix)	Validation of algorithms with external gold standard: sensitivity, specificity, PPV, NPV

## D2 (Global Schema)

In Table 2, D2 is compared across case studies.

### Differences and Similarities

The main difference we observed in the evaluation of the data models was the way two main characteristics of an encounter were captured: the setting where the health care was administered (e.g., general practice, inpatient care, laboratory) and the medical content of the encounter (e.g., diagnosis, procedure, laboratory test). One possibility was that information

was grouped in tables according to the setting (e.g., a table for hospital admissions, another for laboratory tests) and facts were recorded as attributes. The alternative was that encounters were grouped in tables defined by medical content (e.g., a table for diagnoses, a table for procedures) and the care setting was recorded as an attribute. EU-ADR and MATRICE adopted the first approach, OMOP adopted the second, and Mini-Sentinel adopted a combination of the two approaches—death and pharmacy dispensations were organized in the first way and other information was organized in the second.



## T2 (Data Derivation for Specific Studies)

In Table 2, T2 is compared across case studies.

### Differences and Similarities

In EU-ADR each data custodian executed its algorithm with its own usual extraction tool to derive simple input files for a specific study, while execution was performed with common software on the GS in the other networks. OMOP and Mini-Sentinel adopted shared SQL and SAS code, respectively. In MATRICE an ad hoc scripting language was designed and a compiler (a computer program that transforms source code written in a programming language into another) from this language toward the Java virtual machine was developed; extraction in a shared code was then executed locally.

Since OMOP focused on methods development, it often used multiple algorithms for data derivation, to study the impact of the differences. In MATRICE, ongoing validation studies test several algorithms, but the plan is to use a single best definition per study in the end.

In EU-ADR, to overcome the heterogeneity across terminologies, a shared semantic foundation was built by using Unified Medical Language System (UMLS) concepts to define events<sup>5</sup>. Then, the definitive choice of algorithms was obtained through an iterative negotiation between databases: DBs with similar structures were invited to query the same tables and fields.<sup>54</sup> In Mini-Sentinel, algorithms are developed (or reused) for specific analyses and applied at the time of analysis; the result of those algorithms is not stored in the database, but analytic files for each assessment are retained locally.

As for validation of the event resulting from data derivation, all the networks compared incidence- or prevalence rates among databases as a tool to assess consistency. OMOP did not routinely compare with external standards nor with the literature. The

other networks performed either population-based external validation to estimate all validity indices (MATRICE) or external validation of a random sample of automatically detected events to estimate positive predictive value (EU-ADR, Mini-Sentinel).

## T3 (Study Design Application)

In Table 3, T3 is compared across case studies.

### Differences and Similarities

During steps T1 and T2, local partners in some of the networks were asked to implement the processes that had been agreed upon in their own local procedures; moreover the procedures were not shared. In step T3 (study design application), data transformation into analytical data sets was performed in all four networks using shared and common software. In Mini-Sentinel and OMOP, statistical analysis was needed in T3 to estimate propensity and disease scores, while in the studies implemented in the other networks only simpler tasks were needed: linkage between different tables, time splitting, random selection, matching, de-identification, and aggregation. The software Jerboa was developed and used by EU-ADR to execute T3. The software TheMatrix developed by MATRICE executes both T2 and T3: a Domain Specific Language (DSL) was designed and developed for this purpose. DSLs are computer programming languages whose features and expressiveness are restricted and designed ad hoc to fit a given field of application. They target a narrower set of programs than general-purpose languages like Java, but in exchange they provide a higher level of abstraction and can be programmed directly by domain experts rather than computer programmers.<sup>55</sup> In MATRICE, a DSL generating tool called Neverlang was used to develop the language,<sup>16,56</sup> and scripts in the language were generated by domain experts.

Mini-Sentinel and OMOP both used existing software (SQL, SAS, C, Java and R).

Table 3. Comparison with T3 and D4

T3 (APPLICATION OF STUDY DESIGN)					
NETWORK	LOCAL PARTNERS EXECUTE SHARED PROCEDURE	COMMON AMONG DBS	SCORES ESTIMATION	SPECIFIC SOFTWARE	PROGRAMMING LANGUAGE
EU-ADR	Y	Y	N	Jerboa	Java & Jerboa scripting language
Mini-Sentinel	Y	Y	Y	Modular programs and macros; PopMedNet	SQL, SAS, Java, R,
OMOP	Y	Y	Y	—	SQL, SAS, R, C, Java
MATRICE	Y	Y	N	TheMatrix	Java & TheMatrix scripting language

  

D4 (DATASETS FOR ANALYSIS)			
NETWORK	TYPE	FORMAT	QUALITY: STUDY RESULTS VALIDATION
EU-ADR	Intermediate files that can be shared among partners, analysis will follow	csv	Drug safety methodology: comparison of observed drug-event associations with previously classified true and false causal associations; impact on this of different definitions of the derived data
Mini-Sentinel	Level of granularity of dataset depends on study needs; always transfer minimum necessary. Some analyses transfer aggregate data, some use highly-summarized patient-level data Intermediate files saved locally by data partners	csv, SAS datafiles, HTML	To test code known associations are used. Rapid Response queries include data characterization and are reviewed manually by a data expert and an epidemiologist. Results are also reviewed by data partners. Protocol-based assessments might include chart reviews.
OMOP	Final estimates, intermediate files are discarded	csv, SAS datafiles, SQL tables	Drug safety methodology: comparison of observed drug-event associations with previously classified true and false causal associations; impact on this of different definitions of the derived data; estimate of residual bias per event by means of known non causal associations.
MATRICE	Intermediate files to be used for analysis or report generation	csv	Results are reviewed by data partners for comparison with similar analysis performed independently



## D4 (Data Sets for Analysis)

In Table 3, D4 is compared across case studies.

### Differences and Similarities

In OMOP only final estimates were shared, while in the other networks integrated data sets were shared to be pooled before statistical analysis.

EU-ADR and OMOP both adopted a similar validation strategy for their methodological studies in drug safety, which implicitly validated the whole sequence of data transformations at once: a set of positive controls (known adverse drug reactions) and negative controls (drug-outcome pairs that are believed to have no causal relationship) was created. The quality of each method of analysis was assessed by measuring its discriminating power, i.e., the ability of telling positive from negative controls.

## Discussion

In this paper we introduce a conceptual framework to analyze the data management process of a network performing distributed analyses. By applying the framework to four case studies we identify similarities and substantial differences. With this as the foundation, we highlight areas that need further research to identify optimal strategies.

### Differences in Original Databases (of DBs) Have Huge Consequences

The differences observed in the four networks when comparing the original databases (D1) are huge. Understanding such differences is a challenge in itself, as terminology describing health data sources is not shared across countries.<sup>57</sup> The three national networks (MATRICE, OMOP, and Mini-Sentinel) were much more homogeneous than EU-ADR. Since we expect that networks will continue to grow and new DBs will be different from existing DBs, the problems that EU-ADR encountered could indicate

challenges other networks will face in the future if the geographical area is extended. United States databases often have in- and outpatient diagnoses, whereas these are rarely all captured in European administrative databases. In contrast, in Europe general practice databases are very rich since in many countries GPs have a *gatekeeper* function, that is, nonemergency health care can be accessed free of charge only upon the prescription of a GP. Death registries are infrequently part of the data sources available to databases, and this hampers detection of conditions, like acute myocardial infarction or stroke, which may cause death before the patient can reach a health care facility. Due to the differences in available information in the different databases, various strategies need to be used in order to have a comprehensive data derivation of study variables, e.g., in the absence of outpatient diagnostic data, drug utilization or laboratory values may be used to identify certain conditions.

### Differences in Global Schemas (GS) Are Not Substantial

Differences in the GS (D2) between the networks exist but are not substantial, as each GS can be mapped into another, except for those data items that are specifically collected in a single network (for instance, exemptions from copayment, which are documented only in Italian DBs). It would be very valuable, however, to explicitly create such a mapping, as this would make it possible to run existing software procedures embodying T2 and T3 independently of the network: this happened, for instance, in a study replicating—in the EU-ADR network—results from the OMOP network.<sup>58</sup> One area of research should be the impact of different formats of GSs on study outcomes.

### Different Approaches to Terminology Mapping

In two networks (OMOP and EU-ADR), different disease and drug coding systems needed to be

managed. In OMOP the differences were addressed by mapping to homogeneous coding systems during T1, although the original codes were not discarded but were also included in D2. In EU-ADR, mapping was not conducted in T1, therefore all mapping was performed during T2 and only for study-specific conditions. Due to the large differences in the granularity and type of coding schemes, in European databases mapping was very time-consuming—yet this was necessary to obtain consensus across data custodians and investigators<sup>6</sup>—and is progressively growing a shared library. The impact of different mapping strategies, and whether mapping should be done at all versus addressed in the analytic phase, should be investigated.

### Sharing Aggregated Data Sets Versus Sharing Estimates

If network partners can share aggregated data sets in D4, the investigators maintain freedom to perform some subset and sensitivity analysis that were not strictly foreseen in the protocol without performing a new round of transformation. Sharing aggregated data would allow different levels of pooling and potentially more power with respect to meta-analysis, although previous research shows no improved performance of one approach over the other.<sup>14,59,60,61</sup> Given the privacy related issues around data sharing, it should be investigated when different levels of sharing may be indicated.

### Software Tools, Professional Skills, and Information Technology

Software tools used during the transformation process differed across case studies. This had implications for the type of professional skills needed to perform studies in the network as well as the readability of the programs for other investigators.

In principle, all data transformations must be documented to allow investigators to correctly

interpret study results and to understand study limitations and strengths. OMOP and Mini-Sentinel have complete websites where information is stored and can be openly accessed, while EU-ADR and MATRICE rely mainly on scientific papers and reports, a less efficient way of storing information.

How to develop transparent programs and how to store and share the corresponding complex body of information to make it easily available to investigators is also a relevant research topic.

### Validation

Validation of derived data is an imperative condition to produce good epidemiological estimates,<sup>62</sup> and this is even truer when heterogeneous databases participate in a network. Indeed, regularizing the process of creating research data sets from secondary data sets, although necessary, is not enough to ensure high data quality; and validation can quantify how much derived data fail in correctly identifying the study variables—failure that can differ across data partners.

In MATRICE—as data from primary care is lacking and information from secondary care is sparse—deriving chronic conditions, the primary focus of the network, is cumbersome. This is why MATRICE is leading a population-based validation study using diagnosis from a sample of GPs as a gold standard. In Mini-Sentinel a model for a typical validation study was developed<sup>13</sup> and implemented for some events, in particular acute myocardial infarction.<sup>14</sup> EU-ADR adopted a similar study design in some validation studies [Valkhoff2014].<sup>78</sup> Only positive predictive value could be estimated from the study design adopted in the two networks. A similar study was performed on an occasional basis in OMOP.<sup>63</sup> In order to estimate sensitivity, access to a population-based data source would be required, which is more complex than accessing clinical charts of selected candidate events. However, in the specific case of



acute myocardial infarction, death registries are estimated to add from 15 percent to 25 percent of cases to inpatient data where both data sources are available<sup>6</sup>. Therefore misclassification of non-cases, in principle, could have a relevant impact on study results, especially in older subpopulations. In EU-ADR it was observed that improving the positive predictive values of the outcome definition had a very small impact on estimates of additional risk of upper gastrointestinal bleeding in users of four drugs,<sup>7,64</sup> and in OMOP methodological studies varying the definition of several outcomes had little impact on system performance overall,<sup>65</sup> thus suggesting that outcome misclassification may not be a paramount concern when studying the safety of short exposure to drugs. This area has generated some research<sup>66</sup> and deserves further study.

The only attempt to automatically incorporate the result of a quality procedure in the interpretation of study results was performed in OMOP: the association with an outcome observed in a set of drugs that are a priori known not to cause the outcome was computed and applied as an estimate of overall bias in the association of any drug with the same outcome.<sup>67</sup>

Designing and developing a framework that allows for automatically incorporating validity indices in study design and analysis would be a useful follow-up for the effort invested in validation.

### Epistemological Framework of Reference

Unlike in the other steps, in T3 there was a very similar approach in the four networks: there is a uniform attempt to make study designs clearly specified and reusable across studies. This was achieved in all four networks by embedding this step into shared software, where the same procedure was executed across all data sites.

It could be argued that complexity arising from the network setting forces investigators to specify—right from the study design stage—every detail of data management and analysis, embedded in a sequence of computer instructions. A priori specification of the detail of the experiment is at the epistemological core of the experimental method, as it ensures falsifiability.<sup>68</sup> From this point of view, the intricacies of the network settings force investigators to do the right thing. Computer engineers have joined pharmacoepidemiologists and other population-based health scientists in supporting this effort, not just because computer programming is needed, but also and most of all, because a novel, more formal process must be streamlined and stabilized before investigators take control again of the new level of complexity.

### Limitations

The conceptual framework was useful to interpret similarities and differences among the four networks, which are heterogeneous for geographical coverage and purpose. However the choice of the sample of four was nonsystematic, therefore the framework may prove insufficient to include other networks in the comparison.

Data processing in networks of databases may suffer from subtle challenges: privacy laws may enable patients to opt out of sharing information based on some encounters only (for instance, for mental health issues); some databases may collect information from smaller health care providers, whose information is not effectively shared in digital form; regional or national differences in privacy regulations may affect differentially the partners of a network. We did not investigate how the four networks faced such challenges.

## Conclusion

We proposed a conceptual framework to analyze the data management process involved in observational studies taking place in distributed networks of databases. The framework was applied to four case studies to identify similarities and differences.

Several research questions were highlighted by this comparison, including interoperability among the available GSs, optimization of data harmonization, use of validity indices in study design and statistical analysis, development of an information infrastructure to support investigators in accessing details of data transformation, and optimal level of programming skills needed to manage the process.

Medical informatics is called on to support transparency, and quick and sound application of the experimental method to the production of empirical knowledge.

## Acknowledgements

This study was funded by the Integrazione dei Contenuti Informativi per la Gestione sul Territorio di Pazienti con Patologie Complesse o con Patologie Croniche (MATRICE) project, which is funded by the Italian Ministry of Health in the framework of the MATTONI Program.

## Conflict of Interest

Drs. Ryan and Schuemie are employees of Janssen Research and Development. The other authors declare no competing interests.

## Summary

### What was already known on this topic:

- Networks of data sources are being established to produce observational evidence from existing data about diverse, ample populations; and

- Standards for data management are not established, so each network is adopting different infrastructures and procedures.

### What this study added to our knowledge:

- Some steps in the data management process are very similar across existing networks, others are different; and
- Research is needed to identify optimal strategies and common standards.

## References

1. Salmon D, Yih WK, Lee G, Rosofsky R, Brown J, Vannice K, et al. Success of program linking data sources to monitor H1N1 vaccine safety points to potential for even broader safety surveillance. *Health Aff (Millwood)*. Nov 2012;31(11):2518–27.
2. Hernán MA, Savitz DA. From “Big Epidemiology” to “Colossal Epidemiology”. *Epidemiology*. May 2013;24(3):344–5.
3. Toh S, Platt R. Is Size the Next Big Thing in Epidemiology? *Epidemiology*. May 2013;24(3):349–51.
4. Trifirò G, Coloma PM, Rijnbeek PR, Romio S, Mosseveld B, Weibel D, et al. Combining multiple healthcare databases for post-marketing drug and vaccine safety surveillance: why and how? *J Intern Med*. 1 Mar 2014.
5. McGraw D, Rosati K, Evans B. A policy framework for public health uses of electronic health data. *Pharmacoepidemiology and Drug Safety*. Jan 2012;21:18–22.
6. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Med Care*. Jun 2010;48(6 Suppl):S45–51.
7. Valkhoff VE, Coloma PM, Masclee GMC, Gini R, Innocenti F, Lapi F, et al. Validation study in four health-care databases: upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk. *Journal of Clinical Epidemiology*. Aug 2014;67(8):921–31.
8. Avillach P, Mouglin F, Joubert M, Thiessard F, Pariente A, Dufour J-C, et al. A semantic approach for the homogeneous identification of events in eight patient databases: a contribution to the European eu-ADR project. *Stud Health Technol Inform*. 2009;150:190–4.
9. Coloma PM, Schuemie MJ, Trifirò G, Gini R, Herings R, Hippisley-Cox J, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiol Drug Saf*. Jan 2011;20(1):1–11.
10. Platt R, Carnahan RM, Brown JS, Chrischilles E, Curtis LH, Hennessy S, et al. The U.S. Food and Drug Administration’s Mini-Sentinel program: status and direction. *Pharmacoepidemiology and Drug Safety*. Jan 2012;21:1–8.



11. 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.* 11 Feb 2010;153(9):600–6.
12. Agenzia nazionale per i Servizi Sanitari Regionali. Programma Mattoni del SSN - Progetto MATRICE. [http://www.agenas.it/images/agenas/In%20primo%20piano/Matrice/Progetto\\_MATRICE\\_Scheda\\_informativa.pdf](http://www.agenas.it/images/agenas/In%20primo%20piano/Matrice/Progetto_MATRICE_Scheda_informativa.pdf). Accessed September 2015 [Italian]
13. Agenzia regionale di sanità della Toscana. Data integration for chronic diseases management in outpatient settings (MATRICE Project). <https://www.ars.toscana.it/en/project/chronic-diseases/2460-matrice-project.html>. Accessed September 2015
14. Cordis. Safety Of non-Steroidal anti-inflammatory drugs. [http://cordis.europa.eu/result/rcn/54210\\_en.html](http://cordis.europa.eu/result/rcn/54210_en.html). Accessed September 2015
15. Valkhoff VE, Schade R, 't Jong GW, Romio S, Schuemie MJ, Arfe A, et al. Population-based analysis of non-steroidal anti-inflammatory drug use among children in four European countries in the SOS project: what size of data platforms and which study designs do we need to assess safety issues? *BMC Pediatr.* 2013;13:192.
16. Vaccine Adverse Event Surveillance & Communication. <https://brightoncollaboration.org/vaesco.html>. Accessed September 2015
17. Cordis. Arrhythmogenic Potential of Drugs. [http://cordis.europa.eu/project/rcn/94061\\_en.html](http://cordis.europa.eu/project/rcn/94061_en.html). Accessed September 2015.
18. Safety Evaluation of Adverse Reactions in Diabetes. [www.safeguard-diabetes.org](http://www.safeguard-diabetes.org). Accessed September 2015
19. Avillach P, Joubert M, Thiessard F, Trifirò G, Dufour J-C, Pariente A, et al. Design and evaluation of a semantic approach for the homogeneous identification of events in eight patient databases: a contribution to the European EU-ADR project. *Stud Health Technol Inform.* 2010;160(Pt 2):1085–9.
20. Coloma PM, Valkhoff VE, Mazzaglia G, Nielsson MS, Pedersen L, Molokhia M, et al. Identification of acute myocardial infarction from electronic healthcare records using different disease coding systems: a validation study in three European countries. *BMJ Open.* Jan 2013;3(6).
21. Schuemie MJ, Coloma PM, Straatman H, Herings RMC, Trifirò G, Matthews JN, et al. Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods. *Med Care.* Oct 2012;50(10):890–7.
22. Trifirò G, Patadia V, Schuemie MJ, Coloma PM, Gini R, Herings R, et al. EU-ADR healthcare database network vs. spontaneous reporting system database: preliminary comparison of signal detection. *Stud Health Technol Inform.* 2011;166:25–30.
23. Wijnans L, Lecomte C, de Vries C, Weibel D, Sammon C, Hviid A, et al. The incidence of narcolepsy in Europe: before, during, and after the influenza A(H1N1)pdm09
24. Coloma PM, Schuemie MJ, Trifirò G, Furlong L, van Mulligen E, Bauer-Mehren A, et al. Drug-Induced Acute Myocardial Infarction: Identifying 'Prime Suspects' from Electronic Healthcare Records-Based Surveillance System. *PLoS ONE.* Aug 2013;8(8):e72148.
25. Dieleman J, Romio S, Johansen K, Weibel D, Bonhoeffer J, Sturkenboom M, et al. Guillain-Barre syndrome and adjuvanted pandemic influenza A (H1N1) 2009 vaccine: multinational case-control study in Europe. *BMJ.* 2011;343:d3908.
26. Romio S, Weibel D, Dieleman JP, Olberg HK, de Vries CS, Sammon C, et al. Guillain-Barré Syndrome and Adjuvanted Pandemic Influenza A (H1N1) 2009 Vaccines: A Multinational Self-Controlled Case Series in Europe. *PLoS ONE.* 3 Jan 2014;9(1):e82222.
27. Avillach P, Dufour J-C, Diallo G, Salvo F, Joubert M, Thiessard F, et al. Design and validation of an automated method to detect known adverse drug reactions in MEDLINE: a contribution from the EU-ADR project. *J Am Med Inform Assoc.* May 2013;20(3):446–52.
28. Bauer-Mehren A, van Mulligen EM, Avillach P, Carrascosa MDC, Garcia-Serna R, Piñero J, et al. Automatic filtering and substantiation of drug safety signals. *PLoS Comput Biol.* 2012;8(4):e1002457.
29. <http://www.imi.europa.eu/content/emif>. Accessed September 2015
30. Mini-Sentinel. [www.mini-sentinel.org](http://www.mini-sentinel.org). Accessed September 2015
31. Raebel MA, Penfold R, McMahon AW, Reichman M, Shetterly S, Goodrich G, et al. Adherence to guidelines for glucose assessment in starting second-generation antipsychotics. *Pediatrics.* 2014;134(5):e1308–14.
32. Carnahan RM, Moores KG. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative and claims data: methods and lessons learned. *Pharmacoepidemiology and Drug Safety.* Jan 2012;21:82–9.
33. Cutrona SL, Toh S, Iyer A, Foy S, Cavagnaro E, Forrow S, et al. Design for validation of acute myocardial infarction cases in Mini-Sentinel. *Pharmacoepidemiology and Drug Safety.* Jan 2012;21:274–81.
34. Cutrona SL, Toh S, Iyer A, Foy S, Daniel GW, Nair VP, et al. Validation of acute myocardial infarction in the Food and Drug Administration's Mini-Sentinel program. *Pharmacoepidemiol Drug Saf.* Jan 2013;22(1):40–54.
35. McClure DL, Raebel MA, Yih WK, Shoaibi A, Mullersman JE, Anderson-Smits C, et al. Mini-Sentinel methods: framework for assessment of positive results from signal refinement: FRAMEWORK FOR ASSESSMENT OF POSITIVE RESULTS FROM SIGNAL REFINEMENT. *Pharmacoepidemiology and Drug Safety.* 2014;23(1):3–8.
36. Raebel MA, Haynes K, Woodworth TS, Saylor G, Cavagnaro E, Coughlin KO, et al. Electronic clinical laboratory test results data tables: lessons from Mini-Sentinel: THE MINI-SENTINEL LABORATORY RESULTS TABLE. *Pharmacoepidemiology and Drug Safety.* 2014;23(6):609–18.

37. Rassen JA, Schneeweiss S. Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system. *Pharmacoepidemiology and Drug Safety*. Jan2012;21:41-9.
38. Ryan PB, Madigan D, Stang PE, Overhage JM, Racoosin JA, Hartzema AG. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership, *Stat Med*. 2012 Dec 30;31(30):4401-15.
39. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc*. 2012 Jan-Feb;19(1):54-60.
40. Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. *J Biomed Inform*. Aug2012;45(4):689-96.
41. Hartzema AG, Reich CG, Ryan PB, Stang PE, Madigan D, Welebob E, et al. Managing Data Quality for a Drug Safety Surveillance System. *Drug Safety*. 29 Oct2013;36(S1):49-58.
42. Ryan PB, Stang PE, Overhage JM, Suchard MA, Hartzema AG, DuMouchel W, et al. A Comparison of the Empirical Performance of Methods for a Risk Identification System. *Drug Safety*. 29 Oct 2013;36(S1):143-58.
43. Observational Health Data Sciences and Informatics. *ohdsi.org*. Accessed September 2015
44. Buja A, Gini R, Visca M, Damiani G, Federico B, Francesconi P, et al. Prevalence of chronic diseases by immigrant status and disparities in chronic disease management in immigrants: a population-based cohort study, Valore Project. *BMC Public Health*. 24 may 2013;13(1):504.
45. Buja A, Damiani G, Gini R, Visca M, Federico B, Donato D, et al. Systematic Age-Related Differences in Chronic Disease Management in a Population-Based Cohort Study: A New Paradigm of Primary Care Is Required. *PLoS ONE*. 14 Mar 2014;9(3):e91340.
46. Buja A, Gini R, Visca M, Damiani G, Federico B, Donato D, et al. Need and disparities in primary care management of patients with diabetes. *BMC Endocrine Disorders*. 10 Jul 2014;14(1):56.
47. Visca M, Donatini A, Gini R, Federico B, Damiani G, Francesconi P, et al. Group versus single handed primary care: a performance evaluation of the care delivered to chronic patients by Italian GPs. *Health Policy*. Nov 2013;113(1-2):188-98.
48. Gini R, Francesconi P, Mazzaglia G, Cricelli I, Pasqua A, Gallina P, 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. *BMC Public Health*. 9 Jan2013;13(1):15.
49. Gini R, Schuemie MJ, Lapi F, Cricelli I, Pasqua A, et al. Can Italian healthcare administrative databases be used to compare regions with respect to compliance with standards of care for chronic diseases? *PLoS ONE*. May 2014;9(5):e95419.
50. Cali A, Calvanese D, De Giacomo G, Lenzerini M. Data integration under integrity constraints. *Information Systems*, 29(2):147-163, 2004.
51. Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, et al. Defining Disease Phenotypes Using National Linked Electronic Health Records: A Case Study of Atrial Fibrillation. *PLoS ONE*. November 2014;9(11):e110900.
52. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Med Care*.2013 Aug;51(8 Suppl 3):S22-9.
53. *thematrix.isti.cnr.it*. Accessed September 2015.
54. Avillach P, Coloma PM, Gini R, Schuemie M, Mouglin F, Dufour J-C, 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 Inform Assoc*. 6 Sep2012
55. Cazzola W. Domain-Specific Languages in Few Steps: The Neverlang Approach. In *Proceedings of the 11th International Conference on Software Composition (SC'12)*, Prague, Czech Republic, May-June2012, LNCS 7306, pp. 162-177, Springer.
56. Cazzola W, Vacchi E. Neverlang 2: Componentised Language Development for the JVM. In *Proceedings of the 12th International Conference on Software Composition (SC'13)*, Budapest, Hungary, June 2013, LNCS 8088, pp. 17-32, Springer.
57. Adler-Milstein J, Ronchi E, Cohen GR, Winn LAP, Jha AK. Benchmarking health IT among OECD countries: better data for better policy. *J Am Med Inform Assoc*. 1 Jan 2014;21(1):111-6.
58. Schuemie MJ, Gini R, Coloma PM, Straatman H, Herings RMC, Pedersen L, et al. Replication of the OMOP Experiment in Europe: Evaluating Methods for Risk Identification in Electronic Health Record Databases. *Drug Safety*. Oct 2013;36(S1):159-69.
59. Toh SS, Gagne JJP, Rassen JAS, Fireman BH, Kulldorff M, Brown JS. Confounding Adjustment in Comparative Effectiveness Research Conducted Within Distributed Research Networks. *Medical Care*. 2013.
60. Toh S, Reichman ME, Houstoun M, Ding X, Fireman BH, Gravel E, et al. Multivariable confounding adjustment in distributed data networks without sharing of patient-level data. *Pharmacoepidemiol Drug Saf*. 2013;22(11):1171-7.
61. Toh SS, Shetterly SM, Powers JDM, Arterburn D. Privacy-preserving Analytic Methods for Multisite Comparative Effectiveness and Patient-centered Outcomes Research. *Medical Care*. 2014. 2014;52(7):664-8.
62. Hernán MA. With great data comes great responsibility: publishing comparative effectiveness research in EPIDEMIOLOGY. *Epidemiology*. May2011;22(3):290-1.
63. Hansen RA, Gray MD, Fox BI, Hollingsworth JC, Gao J, Zeng P. How Well Do Various Health Outcome Definitions Identify Appropriate Cases in Observational Studies? *Drug Safety*. 29 Oct2013;36(S1):27-32.
64. Valkhoff VE, Coloma PM, Lapi F, Gini R, Nielsson MS, Mosseveld M, Molokhia M, Schuemie MJ, Sturkenboom MCJM, Trifirò G. Positive predictive value for upper gastrointestinal bleeding in four health care databases using different coding systems in the EU-ADR project. Presented at the Digestive Disease Week, San Diego, California, May 19-22, 2012.
65. Reich CG, Ryan PB, Schuemie MJ. Alternative Outcome Definitions and Their Effect on the Performance of Methods for Observational Outcome Studies. *Drug Safety*. 29 ottobre2013;36(S1):181-93.



66. Maro JC, Brown JS, Dal Pan GJ, Kulldorff M. Minimizing signal detection time in postmarket sequential analysis: balancing positive predictive value and sensitivity. *Pharmacoepidemiol Drug Saf.* 2014
67. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Statistics in Medicine.* Jul 2013
68. Popper K. *Conjectures and refutations: the growth of scientific knowledge.* Routledge 1963. London.