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 GINIa,b,1, Massimo COPPOLAc, Patrick B. RYANd, Giacomo RIGHETTIc, Iacopo PERIc, Roberto BERNId, Paul AVILLACHc, Preciosa M. COLOMAb, Gianluca TRIFIROc,f, Gayo DIALLOg, Johan VAN DER LEIb, Miriam C.J.M. STURKENBOOMB and Martijn J. SCHUEMIEb

a Agenzia regionale di sanità della Toscana, Florence, Italy
b Department of Medical Informatics, Erasmus Medical Center, Rotterdam, The Netherlands
c Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo”, National Research Council, Pisa, Italy.
d Observational Medical Outcomes Partnership, USA
e HEGP Hospital, University Paris Descartes, Paris, France
f Department of Clinical and Experimental Medicine and Pharmacology, Section of Pharmacology, University of Messina, Messina, Italy.
g Univ. of Bordeaux, ISPED, LESIM, F-33000, Bordeaux, France

Abstract. Reuse of healthcare databases offers great potential for producing timely epidemiological evidence reflecting ‘real-world’ circumstances. The advantage of such databases is maximized when data from different countries and organizations are pooled. Several ongoing initiatives are facing the technical and scientific challenges of pooling data from substantially diverse sources. In this workshop some notable experiences from USA and Europe will be presented and compared.

Keywords. Data reuse, observational studies, pharmacoepidemiology, health services research

Introduction

Several ongoing initiatives are aimed at exploiting existing administrative claims and electronic medical record databases from different institutions/countries to automatically produce datasets for the conduct of epidemiological studies. Although all the healthcare databases involved in these projects share a basic semantic similarity, challenges arise in intelligent, semantic interoperability from the heterogeneity of the

1 Corresponding Author. Rosa Gini, PhD; Agenzia regionale di sanità della Toscana, Via Pietro Dazzi, 50141 Florence, Italy; Email: rosa.gini@ars.toscana.it
Data collection and management strategies from three projects will be compared with focus on (1) data model adopted by each project (2) strategies for data storage (3) development of innovative software tools (4) adaptability and flexibility to various study designs (5) level of heterogeneity across databases participating in each system (6) transparency and efficiency of the process from study design to data elaboration.

2. Contributions: Outline of the Initiatives

2.1. The EU-ADR Project

Exploring and Understanding Adverse Drug Reactions by integrative mining of clinical records and biomedical knowledge (EU-ADR) was a FP7 project funded by the European Commission from 2008 to 2012 [3]. The aim of the project was to develop an automatic system to detect potential adverse drug reactions in healthcare databases. It involved eight databases from four European countries, highly heterogeneous in original aim, structure and coding systems. In some databases, free text fields were available as well, in two different natural languages. Data sources were logically mapped into a common, person-centered data model including demographic data, outpatient drug dispensing/prescriptions, inpatient diagnoses, general practitioner diagnoses and clinical observations, laboratory results, causes of death [4]. The eight local views of the common data model were all semantically different. A common identification strategy of outcome variables was first defined in the Unified Medical Language System (UMLS) to account for different coding systems and languages, then logically projected to the local data model of every single database. Extraction of cohort, exposure and outcome datasets from each database was performed locally through own software procedures under the responsibility of the data owner. A Java-based tool, called Jerboa [5], was developed centrally and applied locally to transform the cohort, exposure and outcome datasets into several integrated datasets, each corresponding to a different study design. The datasets were then centrally pooled.

2.2. The Observational Medical Outcomes Partnership (OMOP)

The Observational Medical Outcomes Partnership [6] (OMOP) is a public-private partnership initiated in 2008, managed by the Foundation for the National Institutes of Health, chaired by the Food and Drug Administration, and supported by pharmaceutical industry with active engagement from academia, industry, healthcare providers in US and internationally. Its goal was methodological research about use of electronic healthcare data to explore the real-world effects of medical products. OMOP has created a network of observational database comprising patient-level data for over 150m lives. OMOP developed a common data model (CDM) that encompasses both a single relational data structure (with associated specifications for tables, fields,
datatypes) and common set of vocabularies to standardize the structured content of drugs, conditions, procedures, and other clinical observations. It defines table structures for each of the data domain in a Person and Provider-centric model. The OMOP CDM is not intended to be an integration point for multiple source data sets into a large pool; instead a separate CDM instance is expected to be generated for each source data set. To date, OMOP has been technology platform agnostic, with data partners creating CDM instances in SAS, Oracle, SQL Server, mySQL. Aggregate summary analysis results from disparate sources are being brought together within the OMOP central coordinating center’s cloud-based research lab.

2.3. The MATRICE Project

The MATRICE Project is an initiative of the Italian National Agency for Regional Health Services which was started in 2011 and will end in 2013. It is aimed at developing tools to exploit Italian administrative databases to produce information on the prevalence of chronic disease and on standards of care across the country. Italy has a regional-based, tax-funded universal healthcare system. Regions must collect patient-centered administrative information on healthcare activities according to a national common data model (IAD). Data cover demographic information, disease-specific exemptions from copayment, outpatient drug prescriptions, inpatient diagnosis and procedures and a simple description of outpatient activity. When data are sent to the central government, due to confidentiality rules the personal identifier is discarded, making data integration at the national level impossible. The MATRICE Project developed a Java tool, called TheMatrix, that interacts with the regional DBMS and maps the local data model onto the IAD. Higher-level conversions and processing are expressed in a scripting language. A complementary visual interface was developed to assist the final user in defining the data transformation needed for each specific study. Each extraction is fully reproducible and traceable: script, parameters and dates are saved along with the resulting XML file. The same script runs in different local installations and produces datasets that can be pooled for national-level analysis.

3. Expected Outcome

Identifying possible avenues for further development, improvement, and synergy with similar ongoing activities.

References