Comparison Among EU-ADR, OMOP, Mini-Sentinel And MATRICE Strategies For Data Extraction And Management

Rosa Gini, MS, MS3, Patrick B Ryan, MEng3, Jeffrey S Brown, PhD4, Edoardo Vacchi, MS5, Massimo Coppola, MS, PhD6, Walter Cazzola, MS, PhD5, Preciosa M Coloma, MD, MS, PhD2, Roberto Berni, MS1, Gayo Diallo, PhD7, Paul Avillach, MD, PhD8,12, Gianluca Trifirò, MD, PhD9, José L Oliveira, PhD10, Peter R Rijnbeek, PhD2, Johan van der Lei, MD, PhD2, Miriam C JM Sturkenboom, PhD2 and Martijn J Schuemie, PhD2,3,11. 1Agenzia regionale di sanità della Toscana, Florence, Italy, 50141; 2Medical Informatics, Erasmus Medical Center, Rotterdam, Netherlands; 3Observational Medical Outcomes Partnership, Foundation for the National Institutes of Health, Bethesda, Maryland, United States; 4Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, Massachusetts, United States; 5Dipartimento di Informatica, Università degli Studi di Milano, Milan, Italy; 6Istituto di Scienze e Tecnologie dell’Informazione, Centro Nazionale delle Ricerche, Pisa, Italy; 7LESIM-ISPED Laboratory, University of Bordeaux Segalen, Bordeaux, France; 8Biomedical informatics and public health department, University Hospital HEGP, AP-HP, Paris, France; 9Università di Messina, Messina, Italy; 10Department of Electronics, Telecommunications and Informatics, Universidade de Aveiro, Aveiro, Portugal; 11Janssen Research and Development LLC, Titusville, New Jersey, United States and 12Information Sciences to support Personalized Medicine, Université Paris DescartesUniversité Paris Descartes, Paris, France.

Background: Data collection for epidemiologic studies is increasingly performed as data transformation from local organizations. Several research teams in Europe and US have recently tackled the additional challenge of pooling heterogeneous data sources. A conceptual framework describing the whole data transformation process
is lacking.

**Objectives:** Compare the strategies adopted by four projects: EU-ADR (EU), Mini-Sentinel (US), OMOP (US) and MATRICE (IT).

**Methods:** The data sources were compared. A conceptual framework describing the data transformation process was introduced splitting the process in three steps: metadata reorganization (T1), semantic transformation to produce datasets of specific events (T2), data linkage to produce datasets for analysis (T3). The strategies of the four groups were mapped onto this model and compared. Quality procedures and software tools were compared as well.

**Results:** EU-ADR managed the most heterogeneous pool of data sources with respect to country of origin, settings of data collection, coding systems and natural languages of clinical notes in free text. OMOP and Mini-Sentinel data sources had similar level of heterogeneity while MATRICE pooled data from homogeneous local databases. As for data transformation, step T3 was the most comparable across groups: common procedures were implemented in software tools. Step T1 was the most heterogeneous: from informal documentation (EU-ADR) to formally coded loading procedures (Mini-Sentinel and OMOP) to text files feeding a common software (MATRICE). Multiple definitions for the same clinical event were adopted in Step T2: in OMOP they all entered the study and impact of different definitions was evaluated on performance; in EU-ADR definitions were compared in terms of incidence rates within and between local data providers; Mini-Sentinel and EU-ADR conducted medical chart review, MATRICE and Mini-Sentinel performed validation studies. Ad-hoc programming languages were developed for T2 in OMOP, Mini-Sentinel and MATRICE.

**Conclusions:** The framework we introduced made comparison possible. EU-ADR pooled the most heterogeneous data with the least formally coded data transformation procedures.