MINERVA: Metadata for Data Discoverability and Study Replicability in Observational Studies

Romin Pajouheshnia,[1] Rosa Gini,[2] Eleanor Hyde,[3] Morris Swertz,[3] Miriam Sturkenboom,[4] Andrea Margulis,[5] Carla Franzoni,[5] Alejandro Arana,[5] Vera Ehrenstein,[6] Karin Gembert,[7] Ella Jansen,[8] Ron Herings,[8] Nicolas Thurin,[9] Igor Locatelli,[10] Janja Jazbar,[10] Špela Žerovnik,[10] Mitja Kos,[10] Steven Smit,[11] Sirje Lind,[11] Andres Metspalu,[11] Silvia Zaccagnino,[12] Maria Paula Busto,[12] Bas Middelkoop,[12] Manuel Barreiro-de Acosta,[13] Francisco Sanchez-Saez,[14] Clara Rodriguez-Bernal,[14] Gabriel Sanfélix-Gimeno,[14] Beatriz Poblador-Plou,[15] Jonás Carmona-Pírez,[15] Antonio Gimeno-Miguel,[15] Miguel Gil,[16] Wiebke Schäfer,[17] Ulrike Haug,[17,18] Stefania Simou,[19] Karin Hedenmalm,[19] Ana Cochino,[19] Paolo Alcini,[19] Xavier Kurz,[19] Lia Gutierrez,[5] Susana Perez-Gutthann[5] (on behalf of the MINERVA project consortium)

[1] Employee of Utrecht University, Netherlands, at the time this project was performed; [2] Agenzia Regionale di Sanità (ARS) della Toscana, Florence, Italy; [3] University Medical Center Groningen, Groningen, Netherlands; [4] Julius Center for Health Sciences and Primary Care, Department of Data Science & Biostatistics, University Medical Center Utrecht, Utrecht, Netherlands; [5] RTI Health Solutions, Pharmacoepidemiology and Risk Management, Barcelona, Spain; [6] Department of Clinical Epidemiology, Aarhus University and Aarhus University Hospital, Aarhus, Denmark; [7] Centre for Pharmacoepidemiology, Karolinska Institutet, Stockholm, Sweden; [8] PHARMO Institute for Drug Outcomes Research, Utrecht, Netherlands; [9] Bordeaux PharmacoEpi, INSERM CIC-P 1401, University of Bordeaux, Bordeaux, France; [10] University of Ljubljana, Faculty of Pharmacy, Ljubljana, Slovenia; [11] Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia; [12] European Society for Blood & Marrow Transplantation (EBMT), Europe; [13] Spanish Working Group on Crohn's Disease and Ulcerative Colitis–GETECCU, Spain; [14] Health Services Research Unit. Foundation for the Promotion of Health and Biomedical Research of the Valencia Region (FISABIO), Valencia, Spain; [15] EpiChron Research Group, Instituto Aragonés de Ciencias de la Salud (IACS), IIS Aragon, Zaragoza, Spain; [16] Agencia Española de Medicamentos y Productos Sanitarios, Madrid, Spain; [17] Department of Clinical Epidemiology, Leibniz-Institute for Prevention Research and Epidemiology–BIPS, Bremen, Germany; [18] Faculty of Human and Health Sciences, University of Bremen, Germany; [19] European Medicines Agency, Amsterdam, Netherlands

DISCLOSURES

The content of this poster relates to the MINERVA project funded by the European Medicines Agency (EMA) through the framework contract No EMA/2017/09/PE/16. The views expressed in this poster are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the EMA or one of its committees or working parties. The project was implemented collaboratively by members of the SIGMA Consortium and Collaborators and the EU PE&PV Research Network. All coauthors affiliated with the institutions listed above have no conflicts of interest to disclose related to the research conducted and their contributions to this study.

OBJECTIVES

- Define a set of metadata and provide detailed information on source, spectrum, and quality of data sets and pilot metadata collection in a proof-of-concept (POC) catalogue.
- Provide recommendations on a sustainable metadata collection process and use of metadata for identifying real-world (RW) data sources for specific regulatory use cases.

BACKGROUND

- Identification of RW data sources for valid and relevant pharmacoepidemiologic research requires comprehensive assessment of their characteristics and contents.
- Identifying appropriate RW data sources and defining a set of metadata information are increasingly needed for regulatory decision making. This European Medicines Agency (EMA)-commissioned project (EUPAS39322) stemmed from the Heads of Medicines Agencies (HMA)–EMA Joint Big Data Task Force recommendations.[1]

METHODS

- MINERVA was a partnership/consortium of 18 ENCePP research centers and collaborators in 12 European countries covering 15 heterogeneous RW data sources, including electronic health records and patient registers.
- A list of candidate metadata was derived from information gathered from 57 publicly available documents and 8 structured interviews with external experts from European and international real world evidence (RWE) research networks. The list of metadata was finalized after EMA's feedback and input of stakeholders during a public technical workshop in April 2021.[2]
- A POC catalogue was built based on the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles using the open-source software MOLGENIS.
- The POC catalogue aimed to provide an interactive environment for exploring and interacting with recommended metadata for several use cases and applied examples of use.
- The POC catalogue population was piloted following two processes: (1) import of metadata from a preexisting catalogue and (2) collection using an interview tool. Pilot results informed a set of recommendations for future metadata collection, use, and sustainability.
- Retrieval of quantitative metadata (age and sex distribution of the underlying population of a data source) was tested using a mock-up dataset mapped to four different common data models (CDMs) (OMOP, IMI- ConcePTION, Nordic and TheShinISS) through a programming script that could run on multiple CDMs.

RESULTS

- The final proposed metadata list included 436 variables; 241 variables labelled as priority for regulatory purposes were collected in the pilot.[3]
- The POC metadata catalogue comprised 6 highly interconnected domains: Institutions, Data Sources, Data Banks, Common Data Models, Networks, and Studies (Figure 1).
- 11 data access partners (DAPs) collected a subset of metadata in a compatible way in previously used catalogues; the data models of those tools were mapped to the MINERVA metadata list and transferred into the MINERVA POC catalogue. Missing values could be added manually. The remaining 4 DAPs had no preexisting metadata available. Metadata were collected using a process developed initially for the IMI-ConcePTION project. The process for retrieving preexisting metadata and entering new metadata is displayed in Figure 2.
- Metadata use cases included: (1) a DAP provides expertise throughout the cycle of a study; (2) an investigator uses the catalogue throughout the life cycle of a study; (3) a programmer programs a study; (4) a consumer of evidence assesses reproducibility and quality of evidence generated by a study; (5) a FAIR process launched by an external organisation, including a data originator, populates the catalogue; and (6) an institution with research capabilities becomes a DAP for a data bank and/or a data source.
- Considerable resources and pharmacoepidemiologic expert knowledge were required for entry and review of qualitative metadata to ensure that metadata concepts and terminology were interpreted consistently across contributors to the catalogue entries.
- For the quantitative data, the 4 output result data sets were proven to be the same for age and sex distributions. This exercise is available on GitHub (https://github.com/ARS-toscana/MINERVA_samplescript).
- The 15 RW data sources included a variable number of data banks ranging from 1 to 16. Completeness of qualitative metadata varied across data sources. Recommendations were compiled in a guidance document available in the EU PAS register.[4]

Figure 1. Structure of the MINERVA POC Catalogue

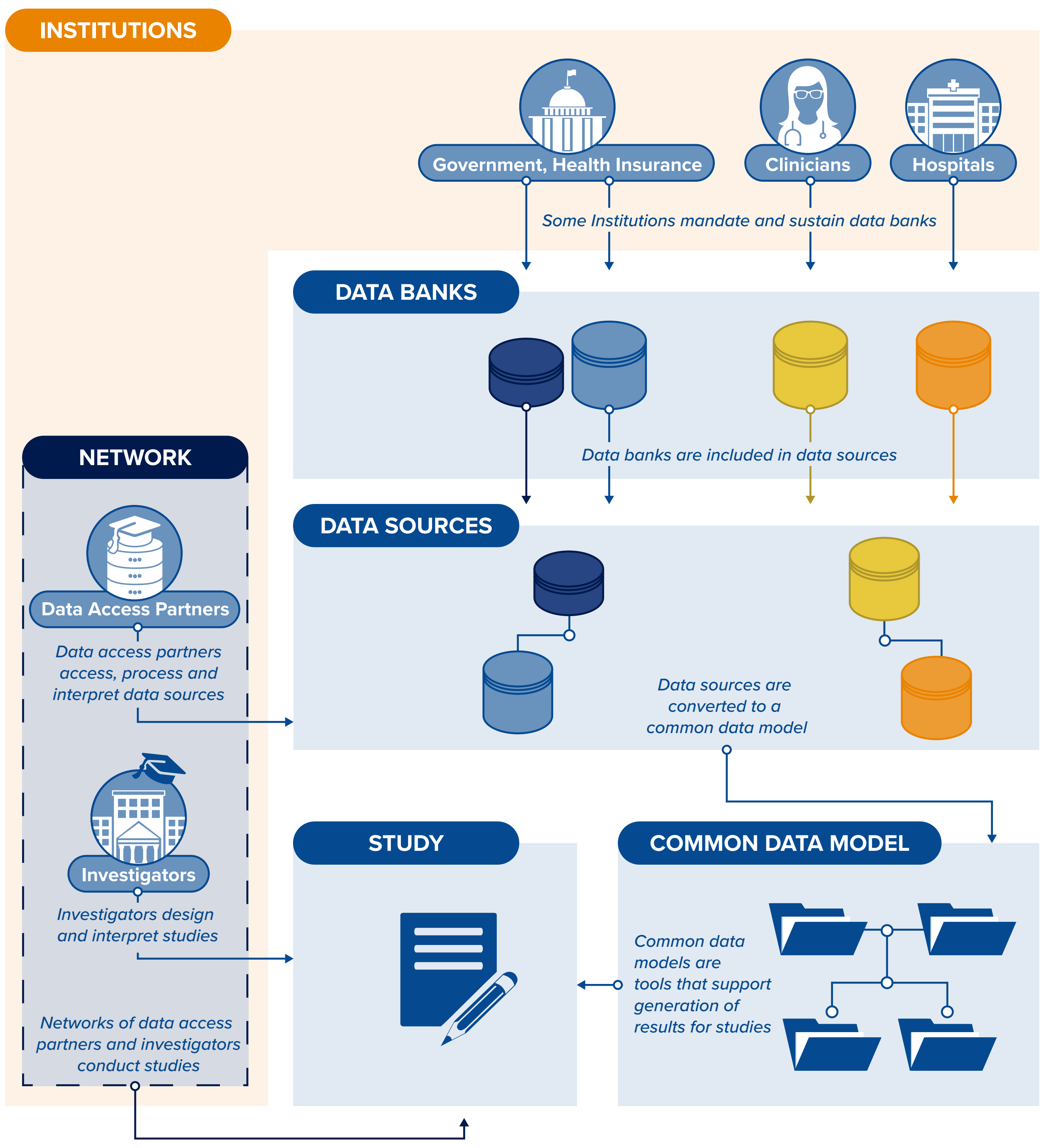
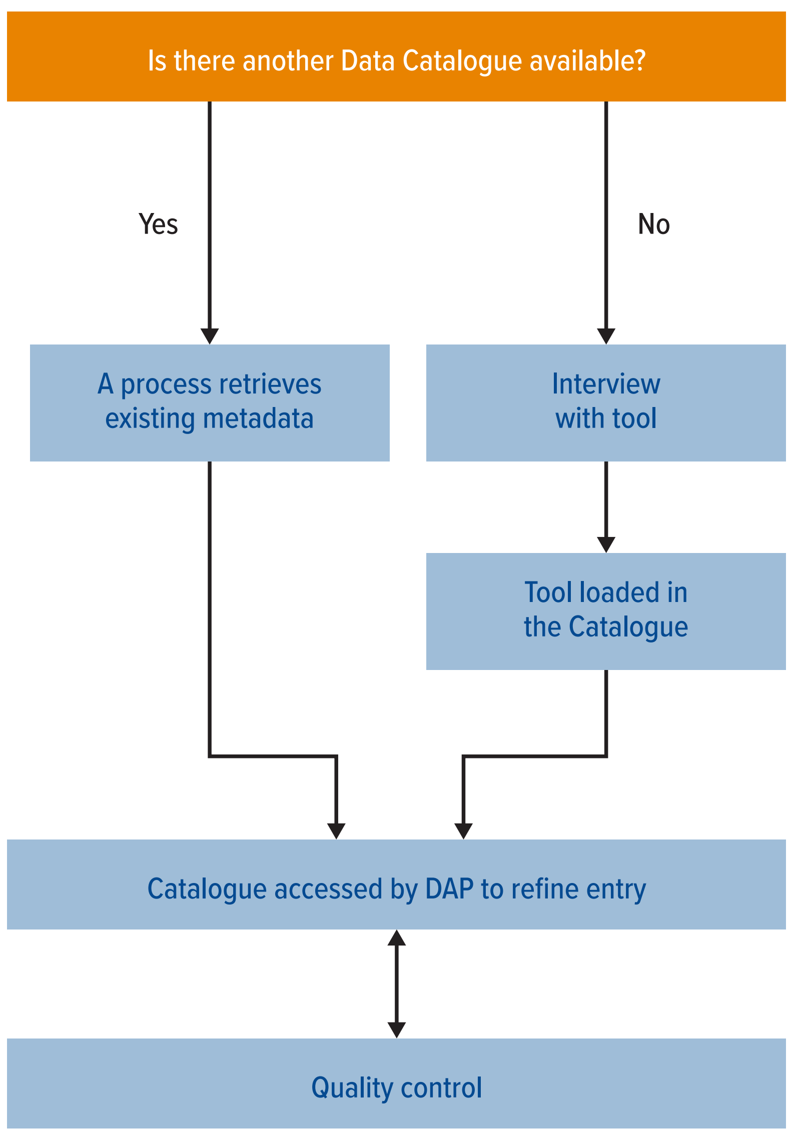

INSTITUTIONS
Government, Health Insurance
Clinicians
Hospitals
Some Institutions mandate and sustain data banks
DATA BANKS
Data banks are included in data sources
NETWORK
Data Access Partners
Data access partners access, process and interpret data sources
Investigators
Investigators design and interpret studies
Networks of data access partners and investigators conduct studies
DATA SOURCES
Data sources are converted to a common data model
STUDY
COMMON DATA MODEL
Common data models are tools that support generation of results for studies

Figure 2. Retrieval of Preexisting Metadata and Entry of New Metadata


Is there another Data Catalogue available?
Yes
No
A process retrieves existing metadata
Interview with tool
Tool loaded in the Catalogue
Catalogue accessed by DAP to refine entry
Quality control

CONCLUSIONS

- The MINERVA pilot showed the value of piloting major metadata catalogue processes and a need for data curation.
- Based on the MINERVA pilot, setting up and maintaining an operating metadata catalogue on real-world data sources require substantial effort to implement FAIR principles, adhere to data protection rules, and effectively support discoverability of data sources and reproducibility of studies in Europe.
- The MINERVA pilot proposed list of metadata is publicly available through the EU PAS Register (see QR code).[3]
- Recommendations for future implementation are publicly available in a guidance document.[4]

REFERENCES

1. HMA-EMA. 15 Dec 2020. https://www.ema.europa.eu/en/documents/other/priority-recommendations-hma-ema-joint-big-data-task-force_en.pdf.
2. HMA-EMA. 12 Apr 2021. https://www.ema.europa.eu/en/documents/other/summary-report-technical-workshop-real-world-metadata-regulatory-purposes_en.pdf.
3. MINERVA. 10 January 2022. https://www.encepp.eu/encepp/openAttachment/documents.otherDocument-2/45372.
4. MINERVA. https://www.encepp.eu/encepp/openAttachment/studyResult/45315.

OTHER PRESENTATIONS ON MINERVA AT ICPE

Data source heterogeneity in multidatabase pharmacoepidemiologic studies: an ISPE-sponsored scoping review. DIVERSE Symposium, August 26, 2022.

Rosa Gini, Olga Paoletti, Romin Pajouheshnia, Patrick Souverein, Nicolas Thurin, Vera Ehrenstein, et al. (on behalf of the MINERVA project Consortium). Study scripts supporting multiple common data models. Poster no. 137, Publication 1182, Poster Session C, Sunday, 28 August 2022.

CONTACT INFORMATION

Lia Gutiérrez, BScN, MPH
Senior Director, Epidemiology

RTI Health Solutions

Av. Diagonal, 605, 9-1
08028 Barcelona Spain

Phone: +34.93.241.7764
Email: lgutierrez@rti.org

Presented at the 38th International Conference on Pharmacoepidemiology & Therapeutic Risk Management; 24-28 August 2022; Copenhagen, Denmark