MINERVA: Study Scripts Supporting Multiple Common Data Models

Rosa Gini,[1] Olga Paoletti,[1] Romin Pajouheshnia,[2] Patrick Souverein,[3] Nicolas Thurin,[4] Vera Ehrenstein,[5] Manuel Barreiro-de Acosta,[6] Miriam Sturkenboom,[7] Lia Gutierrez,[8] Susana Perez-Gutthann[8]; on behalf of the MINERVA project Consortium

[1] Agenzia Regionale di Sanità (ARS) della Toscana, Florence, Italy; [2] Employee of Utrecht University, Netherlands, at the time this project was performed; [3] Utrecht University, Netherlands; University Medical Center Groningen, Groningen, Netherlands; [4] Bordeaux PharmacoEpi, INSERM CIC-P 1401, University of Bordeaux, Bordeaux, France; [5] Department of Clinical Epidemiology, Aarhus University and Aarhus University Hospital, Aarhus, Denmark; [6] Spanish Working Group on Crohn's Disease and Ulcerative Colitis–GETECCU, Spain; [7] Julius Center for Health Sciences and Primary Care, Department of Data Science & Biostatistics, University Medical Center Utrecht, Utrecht, Netherlands; [8] RTI Health Solutions, Pharmacoepidemiology and Risk Management, Barcelona, Spain

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, its Collaborators, and the EU PE&PV Research Network.

Rosa Gini, Olga Paoletti, Patrick Souverein, Nicolas Thurin, Vera Ehrenstein, Miriam Sturkenboom, Lia Gutierrez, and Susana Perez-Gutthann work in institutions partially funded by studies using the ConcePTION data pipeline. The rest of the coauthors have no conflicts of interest to disclose related to the research conducted and their contributions to this study.

OBJECTIVE

- To evaluate the feasibility of retrieving quantitative metadata (age and sex distribution) from real-world (RW) data sources mapped to different CDMs.

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 EMA-commissioned project (EUPAS39322) stemmed from the Heads of Medicines Agencies (HMA)–EMA Joint Big Data Task Force recommendations.[1,2]
- Multi-data-source RW studies can be performed across a distributed network of data sources by converting the original data into a common data model (CDM) and then running a common analysis script at each site.[3]
- Multiple CDMs have been successfully deployed.[4-7] However, analysis scripts designed to run against one CDM cannot run against another CDM. Therefore, it is important to develop scripts that can support multiple CDMs, to effectively utilize existing instances of CDMs or existing mappings.
- In a previous study,[8] the analytic pipelines of multiple CDMs, including OMOP and Sentinel, were conceptually mapped to a sequence of transformation steps: (T1) conversion to the CDM, resulting in a CDM instance; (T2) study variable creation, resulting in data sets of observations on the study population (this is the first step of a study script); (T3) application of study design, resulting in analytic data sets; and (T4) statistical analysis, resulting in data sets of study results.
- In practice, scripts can be structured according to this sequence of steps. In such scripts, only step T2 needs adaptation to the different CDMs.

METHODS

- In the European project MINERVA, we simulated a data source with sex and dates of birth and data source entry and exit.
- We piloted the conversion of quantitative metadata from data sources to 4 commonly used CDMs: OMOP, ConcePTION, Nordic, and TheShinISS and developed an R analysis script to calculate annual sex and age distributions of the population.
- 4 versions of step T2 were programmed, one per CDM, to generate the same output.
- Then steps T3 and T4 were designed to run on the output of T2 and were programmed just once.
- The script was run against the 4 conversions of the simulated data source, and the resulting 4 outputs were merged to test whether they were the same.
- Finally, the script was run against 2 real instances of the Clinical Practice Research Datalink (CPRD) and the ARS Toscana (ARS) data sources converted, respectively, to the ConcePTION and TheShinISS CDM.

RESULTS

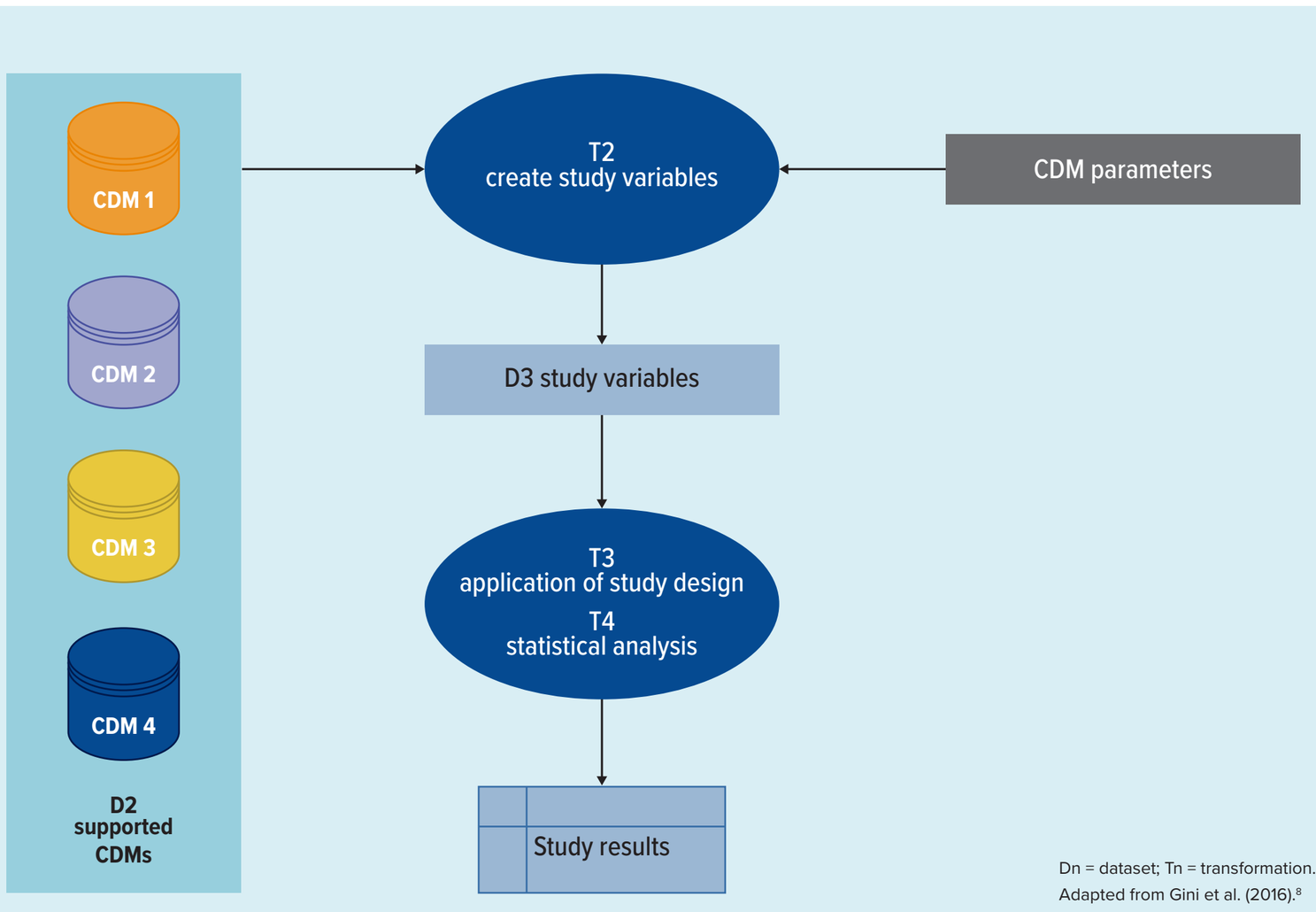
- The script took a few hours to develop and is loaded in a GitHub repository (https://github.com/ARS-toscana/MINERVA_samplescript).
- After running the script against the 4 CDMs to retrieve quantitative data, the 4 output result data sets were proven to be the same for age and sex distributions.
- The script ran successfully against the 2 data sources (CPRD and ARS) and correctly calculated their annual age and sex distribution.
- Figure 1 illustrates the steps of a programming script to generate quantitative metadata from multiple CDMs.

WHAT IS A D3 DATASET?

- A D3 dataset contains individual-level data
- It lists **study variables** observed on the **study subjects**, including:
  - Exclusion criteria
  - Variables enacting the study design (start and end of observation period, matching variables, censoring variables, etc.)
  - Exposure
  - Covariates
  - Outcome
- The **codebook of each D3 is specified** in the statistical analysis plan, including the logical rules (also known as measurements, or phenotypes) to populate it based on the data available in the data sources participating in the study.
- During the script execution, **each D3 is populated**, and the program must support all CDMs.
- Once the D3 data sets are populated, **the next steps are independent of the CDM**.

Adapted from Gini et al. (2016).[8]

Figure 1. Steps of a Script to Generate Study Results From Multiple Common Data Models



Dn = dataset; Tn = transformation.
Adapted from Gini et al. (2016).[8]

CONCLUSIONS

- It is possible to structure study scripts in a common sequence of steps. This minimizes the effort to adapt them to multiple CDMs, because only one step (T2) requires adaptation.
- Structuring scripts this way has the potential to support collaboration in studies, as well as data source characterization, by enabling the use of multiple CDMs.

REFERENCES

1. MINERVA. 2022. Available at https://www.encepp.eu/encepp/openAttachment/studyResult/45315.
2. MINERVA. 10 January 2022. Available at https://www.encepp.eu/encepp/openAttachment/documents.otherDocument-2/45372.
3. Gini R, et al. Clin Pharmacol Therapeut. 2020;108(2):228-35.
4. Thurin NH, et al. 2021. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpt.2476.
5. Observational Health Data Sciences and Informatics. 2021. Available at: https://ohdsi.github.io/TheBookOfOhdsi/.
6. But A, et al. Diabetologia. 2017 Sep 1;60(9):1691-703.
7. Alegiani SS, et al. Rheumatology (Oxford). 2021 Apr 15.
8. Gini R, et al. EGEMS (Wash DC). 2016 Feb 8;4(1):1189.

OTHER PRESENTATIONS ON MINERVA AT ICPE

MINERVA: Metadata for data dIscoverability aNd study rEplicability in obseRVAtional studies. Pajouheshnia R. et al. Poster No. 163, Publication No. 1297. Poster Session C, Sunday, 28 August 2022.

Representing and Leveraging Heterogeneity Between Data Sources in Multi-Database Pharmacoepidemiologic Studies. Symposia & Workshops: Session 1. 26 August 2022; 10:30 am-12:00 pm CEST.

CONTACT INFORMATION

Rosa Gini, PhD
Head, Pharmacoepidemiology Unit

ARS Toscana
Via Dazzi 1
55100 Florence, Italy

Phone: +39.335.77.57.388
Email: rosa.gini@ars.toscana.it

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