Pharmacoepidemiology and Drug Safety

WILEY

ORIGINAL ARTICLE OPEN ACCESS

Metadata for Data dIscoverability aNd Study rEplicability in obseRVAtional Studies (MINERVA): Development and Pilot of a Metadata List and Catalogue in Europe

Romin Pajouheshnia[1,2] | Rosa Gini[3] | Lia Gutierrez[2] | Morris A. Swertz[4] | Eleanor Hyde[4] | Miriam Sturkenboom[5] | Alejandro Arana[2] | Carla Franzoni[2] | Vera Ehrenstein[6] | Giuseppe Roberto[3] | Miguel Gil[7] | Miguel Angel Maciá[7] | Wiebke Schäfer[8] | Ulrike Haug[8,9] | Nicolas H. Thurin[10] | Régis Lassalle[10] | Cécile Droz-Perroteau[10] | Silvia Zaccagnino[11] | Maria Paula Busto[11] | Bas Middelkoop[11] | Karin Gembert[12] | Francisco Sanchez-Saez[13] | Clara Rodriguez-Bernal[13] | Gabriel Sanfélix-Gimeno[13] | Isabel Hurtado[13] | Manuel Barreiro-de Acosta[14] | Beatriz Poblador-Plou[15] | Jonás Carmona-Pírez[15,16] | Antonio Gimeno-Miguel[15] | Alexandra Prados-Torres[15] | Anna Schultze[17] | Ella Jansen[18] | Ron Herings[18] | Josine Kuiper[18] | Igor Locatelli[19] | Janja Jazbar[19] | Špela Žerovnik[19] | Mitja Kos[19] | Steven Smit[20] | Sirje Lind[20] | Andres Metspalu[20] | Stefania Simou[21] | Karin Hedenmalm[21] | Ana Cochino[21] | Paolo Alcini[21] | Xavier Kurz[21] | Susana Perez-Gutthann[2]

[1]Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht University, Utrecht, The Netherlands | [2]Department of Epidemiology, RTI Health Solutions, Barcelona, Spain | [3]Agenzia Regionale di Sanità della Toscana, Florence, Italy | [4]Department of Genetics, University Medical Centre Groningen, Groningen, The Netherlands | [5]University Medical Centre Utrecht, Utrecht, The Netherlands | [6]Department of Clinical Epidemiology, Aarhus University and Aarhus University Hospital, Aarhus, Denmark | [7]Base de datos para la Investigación Farmacoepidemiológica en Atención Primaria, Agencia Española de Medicamentos y Productos Sanitarios, Madrid, Spain | [8]Leibniz Institute for Prevention Research and Epidemiology, Bremen, Germany | [9]Faculty of Human and Health Science, University of Bremen, Bremen, Germany | [10]Bordeaux PharmacoEpi, INSERM CIC-P 1401, Université de Bordeaux, Bordeaux, France | [11]European Society for Blood & Marrow Transplantation, Leiden, The Netherlands | [12]Department of Epidemiology, Karolinska Institutet, Stockholm, Sweden | [13]Foundation for the Promotion of Health and Biomedical Research of the Valencia Region, Valencia, Spain | [14]Spanish Working Group on Crohn's Disease and Ulcerative Colitis, Bilbao, Spain | [15]EpiChron Research Group, Aragon Health Sciences Institute (IACS), IIS Aragón, Miguel Servet University Hospital, Zaragoza, Spain | [16]Technical Advisory Subdirectorate of Information Management (STAGI), Andalusian Health Service (SAS), Seville, Spain | [17]London School of Hygiene and Tropical Medicine, London, UK | [18]PHARMO Institute for Drug Outcomes Research, Utrecht, The Netherlands | [19]Faculty of Pharmacy, University of Ljubljana, Ljubljana, Slovenia | [20]Institute of Genomics, University of Tartu, Tartu, Estonia | [21]European Medicines Agency, Amsterdam, The Netherlands

Correspondence: Romin Pajouheshnia (rpajouheshnia@rti.org)

Received: 16 June 2023 | **Revised:** 30 May 2024 | **Accepted:** 24 June 2024

Funding: This project was funded by the European Medicines Agency (EMA) through the framework contract No. EMA/2017/09/PE/16.

Keywords: catalogue | FAIR | metadata | observational studies | real-world data sources | reproducibility

ABSTRACT

Purpose: Metadata for data dIscoverability aNd study rEplicability in obseRVAtional studies (MINERVA), a European Medicines Agency–funded project (EUPAS39322), defined a set of metadata to describe real-world data sources (RWDSs) and piloted metadata collection in a *prototype* catalogue to assist investigators from data source discoverability through study conduct.

Research from this project was presented at the 2022 International Conference of Pharmacoepidemiology (ICPE); August 2022. Copenhagen, Denmark: (1) MINERVA: Metadata for data dIscoverability aNd study rEplicability in obseRVAtional studies; Pajouheshnia R. et al. Poster no. 163, Publication No. 1297, August 28, 2022; (2) MINERVA: Study Scripts Supporting Multiple Common Data Models; Gini R. et al. Poster no. 137, Publication 1182, August 28, 2022; and (3) Data source heterogeneity in multidatabase pharmacoepidemiologic studies: An ISPE-sponsored scoping review. DIVERSE Symposium, August 26, 2022.

The views expressed in this article are the personal views of the author(s) and may not be understood or quoted as being made on behalf of or reflecting the position of the regulatory agency/agencies or organizations with which the author(s) is/are employed/affiliated.

Methods: A list of metadata was created from a review of existing metadata catalogues and recommendations, structured interviews, a stakeholder survey, and a technical workshop. The prototype was designed to comply with the FAIR principles (findable, accessible, interoperable, reusable), using MOLGENIS software. Metadata collection was piloted by 15 data access partners (DAPs) from across Europe.
Results: A total of 442 metadata variables were defined in six domains: institutions (organizations connected to a data source); data banks (data collections sustained by an organization); data sources (collections of linkable data banks covering a common underlying population); studies; networks (of institutions); and common data models (CDMs). A total of 26 institutions were recorded in the prototype. Each DAP populated the metadata of one data source and its selected data banks. The number of data banks varied by data source; the most common data banks were hospital administrative records and pharmacy dispensation records (10 data sources each). Quantitative metadata were successfully extracted from three data sources conforming to different CDMs and entered into the prototype.
Conclusions: A metadata list was finalized, a prototype was successfully populated, and a good practice guide was developed. Setting up and maintaining a metadata catalogue on RWDSs will require substantial effort to support discoverability of data sources and reproducibility of studies in Europe.

Summary

- We created a list of 442 metadata elements to describe real-world healthcare data sources.
- Metadata were organized within six interconnected domains: institutions, data banks, data sources, studies, networks, and common data models.
- Collection of selected metadata was piloted for 26 institutions and 15 data sources in a catalogue prototype using the MOLGENIS software.
- Qualitative metadata, describing how and why records are collected in a data source, are essential to provide context to quantitative metadata, such as event rates.
- Implementation of the full metadata list will require substantial effort, and prioritization of metadata is recommended.

1 | Introduction

Pharmacoepidemiologic researchers have for decades used real-world data and generated real-world evidence (RWE) to support regulatory decision-making on medicines. For many stakeholders, increased data discoverability can have a clear, positive impact; it allows more efficient and higher-quality studies and enhances the transparency and reproducibility of data [1]. The Heads of Medicines Agencies–European Medicines Agency (HMA–EMA) joint Big Data Taskforce Phase II report identified the creation of sustainable and FAIR (findable, accessible, interoperable, and reusable) [2] data sources and metadata catalogues as a challenge [3].

In November 2020, the MINERVA (Metadata for data dIscoverability aNd study rEplicability in obseRVAtional studies) project (EUPAS39322) was initiated in response to the HMA–EMA joint Big Data Task Force recommendation on "the identification of metadata" for regulatory decision-making on the choice of data source [2–5]. The MINERVA Consortium included 18 institutions and worked in collaboration with the EMA.

The project primarily aimed at defining a set of metadata to describe data sources that could be used to support investigators throughout the phases of a study: from the identification of suitable data sources to the interpretation and transparent reporting of results. The second objective was to pilot the list of metadata in a proof-of-concept metadata catalogue (hereinafter called *prototype*) to identify opportunities and challenges for future implementations of a metadata catalogue. We report on the methods and principal results of the MINERVA project.

2 | Methods

2.1 | Conceptual Framework

To acknowledge the heterogeneous landscape of European data sources [6], we based the design of MINERVA metadata catalogue on a conceptual framework (Figure 1), which arose from a prior qualitative study [7]. Based on the study, six domains of the metadata list were chosen: institutions, data banks, data sources, studies, networks, and common data models (CDMs). A glossary of terms is provided in Supporting Information S1.

Some *institutions* (e.g., insurance companies, governments) acting as *data originators* mandate and sustain regular collections of information on defined sets of healthcare-related services/events for purposes normally unrelated to research (e.g., reimbursement) and store such data in electronic format. These collections of data are herein called *data banks*. Each data bank comprises tables, variables, values, and metadata labels that can be annotated with the following information to support interpretation:

- *Prompts*: record-prompting events (also called triggers), for example, access to an emergency department, dispensation of a reimbursable medication.
- *Underlying population*: population whose events prompt the record creation, for example, the persons entitled to healthcare services funded by a specific payer, legal inhabitants of a geographic area.

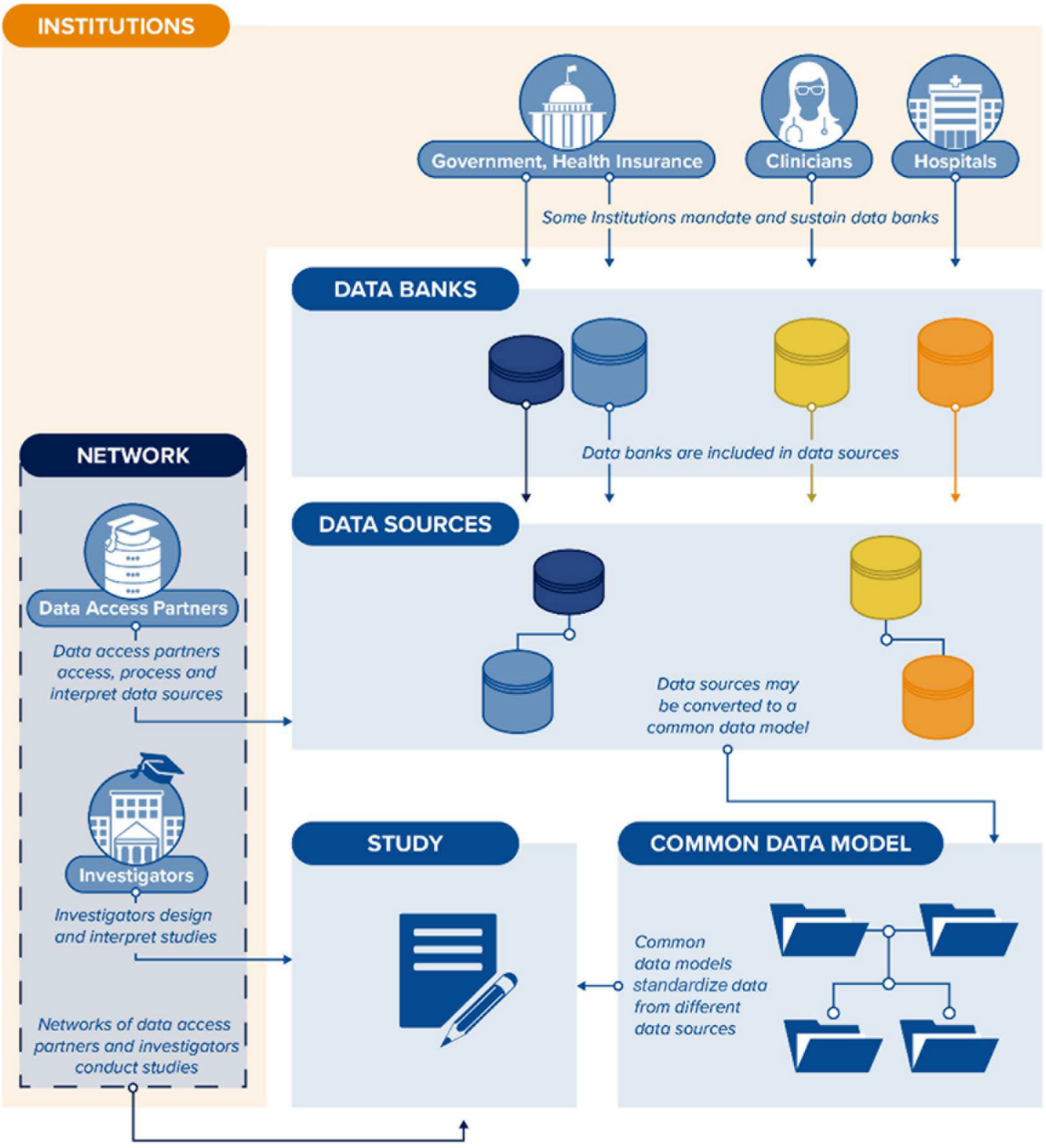

FIGURE 1 | Illustration of the conceptual framework underlying the metadata list.

Data banks can be grouped into *families* with similar record prompts and originators.

Data sources are collections of data banks that relate to the same or partially overlapping underlying populations, all linkable to one another at the person level, either probabilistically or deterministically. A data source could include one or more data banks.

Research institutions are institutions with expertise in conducting pharmacoepidemiologic *studies*, including interpreting their results. A data access partner (DAP) is a research institution that can obtain access to data, typically subject to local authorization requirements, such as a protocol and/or ethical approval. DAPs have experience and capabilities in generating RWE to support health research, including regulatory decision-making. Data originators themselves may have data access for research purposes and act as DAPs, provided they have research expertise.

Research institutions (including DAPs) may form *networks* that conduct studies.

To conduct multi-database studies in networks, it may be useful to transform the data from participating data sources from their original native format into a *CDM* through data standardization, for an efficient execution of programming code against local data [8].

Figure 1 illustrates the relationships among the six domains. In Figure 2, the conceptual framework is applied to the Danish National Registers, whose selected data banks are all mandated by government institutions and to the data source of ARS Toscana, Italy, whose data banks have heterogeneous originators.

2.2 | Identification of the Metadata List

A search of the websites of key organizations and consortia with experience in cataloguing health databases and their metadata was conducted in January 2021. Recommendations for metadata collection and relevant metadata fields were extracted from publications and catalogue tools identified by the search (the full list of materials is available online [9]). An initial list of unique metadata fields was derived from the extracted information.

Next, structured interviews were conducted with representatives from eight organizations or networks with experience in pharmacoepidemiological studies involving multiple data sources or with expertise in metadata in the health domain:

- FDA Sentinel Initiative [10]
- CNODES [11]
- IMI-EHDEN [12]
- IMI-ConcePTION [13]
- AsPEN [14]
- FAIRplus [15]
- Maelstrom [16]
- Aetion [17]

The interviews helped to refine the scope of the metadata list, identify additional key metadata variables to collect, gather additional resources to inform the metadata list, identify existing tools to access or visualize metadata, and identify challenges or barriers for implementing the prototype. Experts consistently underscored the need of central quality checks on the metadata entered in the visualization tools to ensure they remain functional and that the updates are consistent with the definitions.

Feedback on the preliminary metadata list was gathered through a stakeholder survey and an EMA Technical Workshop [18] in April 2021. The initial metadata list was updated and a prioritization of key metadata for collection was made based on feedback from the workshop.

2.3 | Metadata Collection in the Prototype

2.3.1 | Identification and Representativeness of Data Sources

Implementation of the metadata list was piloted together with DAPs by collecting metadata in a prototype catalogue (described below). During the proposal phase, we identified 13

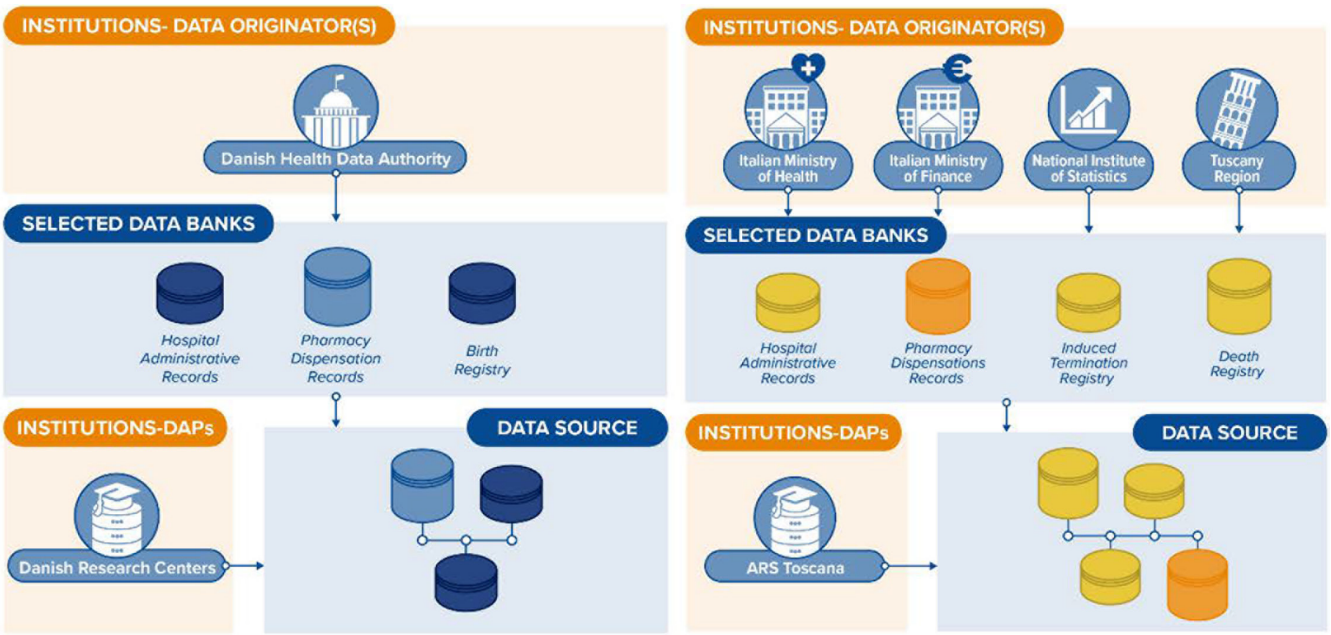

FIGURE 2 | Streamlined examples of the conceptual framework inspired by the Danish National Registers and by the ARS Toscana data source. The figure shows families of selected data banks included in the data sources and institutions involved as DAPs or as data originators. The specific data banks may not exist in a linked state unless used for studies. In this example, the names of the data banks were replaced by the name of the families as classified in the ontology referred to in the results section. ARS Toscana, *Agenzia Regionale di Sanità della Toscana* (Regional Health Agency of Toscana), Italy; DAP, data access partner.

real-world data sources (RWDSs) of interest and a corresponding DAP with expert knowledge of the data source. Following project initiation, the EMA proposed two additional RWDSs and DAPs to provide an additional representation of registries and biobanks. We mapped the data sources for inclusion in the prototype to the general principles on suitability of healthcare data sources for potential use in regulatory medicines decision-making: accessibility, longitudinal dimension, recording of exposure and outcomes, and generalizability [6, 19]. The data sources comprised data collected continuously and consistently; at the person level; timing of prescription, dispensation, administration of medicinal products; in a structured manner (i.e., no free-text only data sources); and compliant with data privacy and confidentiality rules. The 15 data sources with mapping to these criteria are described in tables 1 and 2 of MINERVA deliverable 2 [20]. They provided wide geographic representativeness and distribution across European regions in different healthcare settings. From each data source, selected data banks were included.

2.3.2 | Development of the Prototype

The prototype was implemented using MOLGENIS EMX2, an open-source framework for FAIR scientific data available online [21]. The catalogue tool is also used for multicenter cohort studies in H2020 projects such as EUCAN-Connect [22], LifeCycle [23], LongITools [24], Athlete [25], the European Human Exposome Network [26], and IMI-ConcePTION [13]. This made the prototype natively interoperable with such catalogues. Based on a file defining our catalogue structure, MOLGENIS enables creation of FAIR catalogues via auto-generated user interfaces to help users navigate metadata entered in the catalogue and programmatic interfaces to promote interoperability and reuse, such as DCAT [27] or FAIR Data Point and semantic web Resource Description Framework (RDF) export. We configured the MINERVA data model in MOLGENIS and used these forms for data entry. Finally, we developed a custom user interface on top of MOLGENIS building blocks that enabled users to query and search the metadata.

2.3.3 | Piloting Qualitative Metadata Collection and Quality Assurance

Two processes to collect qualitative metadata were enacted (Figure 3). For data sources that had previously submitted metadata to the ConcePTION Catalogue [13], metadata were automatically retrieved, allowing us to examine whether the prototype adhered to the FAIR principles of interoperability and reusability (see Supporting Information S1 for an operational definition). For the other data sources, DAPs were interviewed following a standard process and answers were collected in a spreadsheet. The resulting information was uploaded to the prototype. All DAPs subsequently accessed the prototype and refined the content. Finally, a round of quality checks was performed by experts for completeness, correct understanding of metadata items, and typographical errors.

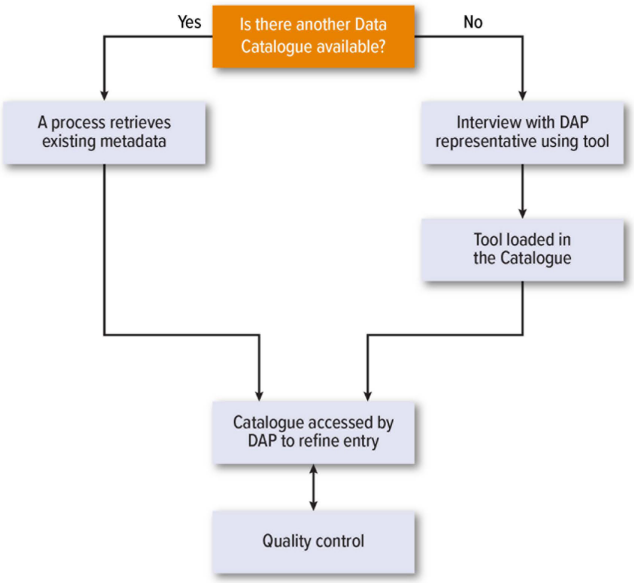

FIGURE 3 | Flow diagram of the two processes enacted to populate the qualitative metadata of the prototype. DAP, data access partner.

2.3.4 | Piloting Quantitative Metadata Collection

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 CDMs (Observational Medical Outcomes Partnership [OMOP] [28], ConcePTION [7], Nordic [29], and The ShinISS [30]) through an R programming script that could run on multiple CDMs, based on a data processing analysis [8]. The four output result datasets were proven to be the same (exercise available on GitHub https://github.com/ARS-toscana/MINERVA_samplescript). Four DAPs with mappings to at least one of the four CDMs were selected to run the script on existing instances of their data sources to retrieve the quantitative metadata.

3 | Results

3.1 | Metadata List

Supporting Information S2 (available at https://zenodo.org/records/10422428) contains the final list of 442 proposed metadata variables and definitions; 260 were labeled as priority variables for regulatory purposes, among which 202 were collected during the pilot. The metadata variables are organized in tables within six domains—institution, data source, data bank, study, network, and CDM—each including ≥1 tables plus 11 technical metadata (Table 1). Each table contains the name of the metadata variables (with a hierarchical code), a description of the content, the standard allowable values, and an explanation of how metadata should be entered in the catalogue for each variable (e.g., by manual entry or automized process), including variables that link and cross-populate tables in the catalogue, and the unit of observation of the metadata.

TABLE 1 | Metadata list domains and content.

Domain	Number of metadata variables[a]	Content of metadata variables
Institution	39	Describe contributors to the prototype and other institutions connected to ≥1 data sources and/or data banks—such as a DAP, data originator—or with other roles, for example, investigators.
Data source	128	Describe the underlying population, a list of institutions with access to it, the data banks that make up the data source, quantitative descriptors, ETL specifications for mapping of the data source to a CDM, and a list of studies and publications related to the data source.
Data bank	177	Describe the data bank's originator, the data bank's family and content, records-creation prompts, the underlying population, the tables' data model, regularity of updates and time lags, qualitative descriptions of quality, and qualitative and quantitative descriptors of completeness. The data sources including the data bank and the institutions that have access to it.
CDM	15	CDMs to which data sources described in the prototype had been previously converted.
Network	8	Describe research networks that the institutions are part of. Provide information on how data sources and institutions' expertise may be brought together for the purpose of a study and could potentially interface with the ENCePP Resources Database.
Study	64	Designed to interface with the EU PAS Register to link information on studies with information on use, content, and quality of data sources, and data banks. This domain is key to reproducibility. Metadata of data sources and data banks used in the study are duplicated in each study (additionally to the 64 metadata variables) to create a snapshot of the metadata specifications at the time of the study. Software used for extraction and processing of data (including ETL to a CDM, if any) are part of the metadata. It also captures expertise and lessons learnt specific to the study, specifically for study variables.

Abbreviations: CDM, common data model; DAP, data access partner; ENCePP, European Network of Centres for Pharmacoepidemiology and Pharmacovigilance; ETL, extract, transform, load; EU PAS Register, European Union electronic Register of Post-Authorization Studies.
[a]An additional 11 technical metadata to support catalogue entry and maintenance are included in the list.

For the metadata "data bank family" (C5.1), a draft ontology was created and stored in BioPortal (https://bioportal.bioontology.org/ontologies/DFO).

3.2 | Summary of Metadata Collected in the Prototype

Data from 26 institutions were collected in the prototype, including all institutions in the Consortium plus other institutions listed as DAPs and/or data originator.

Each of the 15 DAPs in the Consortium populated the metadata of 1 data source and at least 1 data bank from the data source. Table 2 lists each DAP with its data source, the number of data banks entered into the prototype and the number of mappings to different CDMs available for each data source. The number of data banks per family of data bank entered in the prototype is summarized in Supporting Information S3. Five data sources included one data bank, 6 included 3–10 data banks, and 4 included >10 data banks. The most common families of data banks were hospital administrative records and the pharmacy dispensation records (both available in 10 data sources each). Selected data banks from 12 data sources had been converted into ≥1 and 7 into ≥2 of the following CDMs: ConcePTION CDM, OMOP CDM, Nordic CDM, and ShinISS.

Collecting all the metadata of each data source and data bank was unfeasible due to time and effort constraints: data source metadata completeness ranged from 21% to 67% and data bank metadata completeness ranged from 3% to 78%. Three DAPs successfully ran the script to compute quantitative metadata, and results were uploaded into the prototype; one DAP was unable to obtain authorization to run the script in time and instead provided aggregated counts from a public source. To demonstrate the functionality of the other catalogue domains in the prototype, mock-up metadata were entered for four studies, three networks, and one CDM.

3.3 | Recommendations

Based on piloting the prototype, instructions for how to create an entry in the catalogue de novo were developed and can be found in Supporting Information S4. A detailed discussion of the full results of the pilot and lessons learned are presented in the final deliverable of the project: Final Good Practice Guide for Metadata Collection for Real-World Data Sources [31].

TABLE 2 | List of data sources included in the catalogue prototype.

DAP	Country	Data source	Underlying population of the data banks in the data source	No. of data banks in prototype	No. of mappings to a CDM
Aarhus University, Department of Clinical Epidemiology	Denmark	Danish National Registers	Legal inhabitants of Denmark	4	2
University of Tartu	Estonia	Estonian Biobank	Voluntary participants agreeing to donate to the biobank, must be legal residents of Estonia and >18 years of age	1	1
European Society for Blood and Marrow Transplantation (EBMT)		EBMT patient registry	Network of centers; the population in the registry is composed of people attending such centers for selected treatments; EBMT covers majority of centers offering such treatments in Europe, as well as some centers across the world	1	0
BPE, University of Bordeaux	France	Système National des Données de Santé (SNDS)	Persons insured by the French national insurance: The Health Insurance Institute of France (>99% of French population, including legal residents who are not citizens)	14	2
BIPS, Leibniz Institute	Germany	German Pharmaco epidemiological Research Database (GePaRD)	Persons insured with four statutory health insurance providers, ~20% of the general population (90% of overall German population is insured with statutory insurances; civil servants and high-income groups are not or underrepresented)	5	1
Agenzia regionale di sanità della Toscana (ARS Toscana)	Italy	ARS	Inhabitants of the Tuscany region of Italy registered with the healthcare system (all legal inhabitants can register)	16	2
PHARMO Institute	Netherlands	PHARMO Data Network	The underlying populations of the data banks only partially overlap. These populations are nested within the population of the Netherlands	8	2
Utrecht University	Netherlands	Clinical Practice Research Datalink Aurum (CPRD Aurum)	Individuals registered with UK primary care practices using EMIS Web general practice patient management software who have opted into providing data to CPRD	1	2

(Continues)

TABLE 2 | (Continued)

DAP	Country	Data source	Underlying population of the data banks in the data source	No. of data banks in prototype	No. of mappings to a CDM
Faculty of Pharmacy, University of Ljubljana (UL FFA)	Slovenia	Slovenian health data—NIJZ	Persons insured by the Slovenian national insurance. >99% of Slovenian population, including legal residents who are not citizens	3	0
Agencia Española de Medicamentos y Productos Sanitarios (AEMPS)	Spain	Base de datos para la Investigación Farmacoepidemiológica en el Ámbito Público (BIFAP)	Inhabitants of nine regions of Spain registered with a GP or pediatrician	5	1
Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO)	Spain	The Valencia Health System Integrated Database (VID)	Individuals in the Valencia Region insured by the Spanish National Health system, including legal residents who are not citizens (97% of the ~5 million inhabitants)	12	2
Grupo Español de Trabajo en Enfermedad de Crohn y Colitis Ulcerosa (GETECCU)	Spain	ENEIDA patient registry	Patients with inflammatory bowel diseases treated by a physician belonging to the Spanish network ENEIDA	1	0
Instituto Aragonés de Ciencias de la Salud (IACS)	Spain	EpiChron	Legal inhabitants of the Aragon region in Spain	12	1
Centre for Pharmacoepidemiology Karolinska Institute (CPE KI)	Sweden	Swedish national registers	Legal inhabitants of Sweden	4	2
London School of Hygiene and Tropical Medicines	UK	Clinical Practice Research Datalink Gold (CPRD Gold)	People in any of the four nations in the United Kingdom who are registered with a GP practice that uses the Vision software and has opted into providing data to CPRD	1	2

Abbreviations: BIPS, Leibniz Institute for Prevention Research and Epidemiology; BPE, Bordeaux PharmacoEpi; CDM, common data model; DAP, data access partner; ENEIDA, *Estudio Nacional en Enfermedad Inflamatoria intestinal sobre Determinantes genéticos y Ambientales*; GP, general practitioner; NIJZ, National Institute of Public Health.

4 | Discussion

The MINERVA Consortium, in collaboration with the EMA and experts from international RWE research networks, along with the input of stakeholders during a public workshop, developed a metadata list and tested it within a catalogue prototype.

The MINERVA metadata list included 442 variables. A selection of variables was collected from 26 institutions, 15 data sources and selected data banks, and dummy data were generated for four studies, three networks, and one CDM. Two different methods to populate the metadata list were piloted following the FAIR principles. For most DAPs and data sources, metadata could be retrieved from the catalogue of another project and transferred to the MINERVA prototype, demonstrating the potential efficiency if metadata catalogues are designed to be interoperable.

4.1 | Metadata to Support Data Discovery and Interpretation of RWE

The prototype was designed to assist investigators throughout the phases of a study. For example, if an investigator wanted to investigate the safety of a specific medicinal product, they could first search the *Study* section of the catalogue for protocols of other studies investigating the medicinal product (Supporting Information S2, Section F2), which data sources were used, and their strengths and limitations (Section F3). They could then search the corresponding *Data Source* and *Data Bank* sections for details on the coverage and contents of the data source (Sections B1, B4–B6 and C1, C5–C8), data quality (Section C9), as well as governance and access (Sections B2 and C2). Finally, they could identify *Institutions* with expertise in working with the data source for collaboration (Sections A1–A4). If the investigator could not find a suitable data source, they could search for a data bank that contains information on the medicinal product of interest. They might then search for other data banks with a similar underlying population (e.g., geographic region and eligibility) to identify data banks that could potentially be linked to create a new data source. In this case, epidemiological expertise is essential to understand whether the data source would contain the necessary components for a study.

Qualitative metadata can help investigators conduct analyses and interpret the results. In most data sources included in the prototype, the dates of entry and exit of each individual from the underlying population are recorded in one of the data banks. Whether these dates are captured in a data bank can be inferred from the data dictionary of the data bank (Supporting Information S2, Section C6). In such cases, the underlying population can be seen as an epidemiological cohort, where persons can be observed over a defined time period, and the linked data banks allow investigators to derive study variables that can be analyzed to provide quantitative metadata at the data source level. Simple examples include distributions of age and sex, both of which can be captured in data source Section B6 and data bank Section C7 of the prototype.

When the underlying population of a data source represents a general population (e.g., it includes all the inhabitants of a geographic area), quantitative metadata are comparable across data sources. However, when this is not the case, metadata on the underlying population and prompts will support investigators to draw comparisons across data sources. For example, differences in the prompts of the data sources in Table 3 will cause the occurrence of diabetes in data source A to be higher than in B, and lower than C, even though the true population prevalence may be the same. If the population is not representative of a general population, this has an important impact: in data source D, occurrence is higher than E and lower than F. Quantitative metadata providing estimates of occurrence of a disease in a population must be designed, implemented, and interpreted using research expertise.

4.2 | Challenges Identified in the Pilot

We encountered several challenges when piloting the prototype. This included a need to find ways to authenticate contributions

TABLE 3 | Comparability of the occurrence of diabetes across data sources.

Data source	Prompt for diagnoses records	Population	Context required for interpretation
A	Primary care visits		Diagnoses must be interpreted as an occurrence of access to *primary care* for diabetes
B	Inpatient visits		Diagnoses must be interpreted as an occurrence of *hospitalization* for diabetes
C	Primary care visits and/or inpatient visits		Diagnoses must be interpreted as the occurrence of either *hospitalization* or access *to primary care* for diabetes
D		General population	Diagnoses must be interpreted as occurrence in the general population
E		Pediatric population	Diagnoses must be interpreted as occurrence in the pediatric population
F		Population in a diabetes registry	Diagnoses must be interpreted as occurrence in the diabetes population

to the catalogue and apply version control (metadata for this are described in Supporting Information S2, Section M1), a need to bring together the expertise of DAPs on data sources and expertise on the conceptual framework of the catalogue, and a lack of existing persistent identifiers and ontologies, which are needed for the catalogue to be interoperable with other catalogues.

The metadata list was designed to be extensive to support the entire cycle of a study, as well as to provide a candidate master data model to which future metadata lists could be mapped. However, a substantial amount of time and effort was needed to collect and enter metadata during the pilot. It was unfeasible to implement the full set of 442 metadata variables in the catalogue prototype and to populate the metadata fields for each DAP in the consortium within the timeframe of the study. This motivated the selection of a set of 260 priority metadata variables for regulatory purposes, which were the focus of the pilot and catalogue prototype. It was determined by the research team that this priority set of metadata would be sufficient to support the most important regulatory use cases. Despite this selection, it was still not feasible for DAPs to complete all entries for all data banks of each data source within the study time frame, for example, in the case of quantitative metadata, where four DAPs were selected to pilot these metadata fields. In addition, we did not implement some technical metadata specifically to describe catalogue entries, such as contact details of the person making or editing a catalogue entry, as this would have required links between parts of the catalogue data model that were complex to add to the prototype. The largest investment was required when making a catalogue entry de novo. We found that fewer resources were required to update existing entries, which were imported from the IMI-ConcePTION catalogue [13], and we could allocate more time for quality checks.

During the pilot, we considered reusing data previously approved in the context of a different study to compute simple quantitative metadata (age/sex distributions). However, this proved more challenging than expected because reusing the data to produce the same numbers in a different way for a different project was perceived as "data repurposing" and not authorized by the governing authority. This highlights the limitations of data use outside of a specific study and is an important consideration for the European Health Data Space [32].

Recommendations for how to overcome these challenges for future catalogue implementations are presented in a commentary article in this issue *Pharmacoepidemiology and Drug Safety* [33].

4.3 | Strengths and Limitations

We provide a comprehensive metadata list based on a conceptual framework that can describe heterogeneous data sources. The catalogue prototype was piloted by a broad range of European data sources/DAPs, including two registries and a biobank. The definitions, standards, and rules for data entry (Supporting Information S2), and manual for creating metadata entries de novo (Supporting Information S3), as well as further publicly available guidance and recommendations [34], will help future adoption of the metadata list.

In this study, we piloted only a subset of the total metadata list in the prototype. The pilot did not cover all conceivable families of data banks, such as collections of data from wearable devices or social media. However, we anticipate the metadata list will accommodate these types of data. Further testing will help to identify areas for improvement for future catalogue implementations.

4.4 | Implementation and Follow-Up

The MINERVA metadata list is currently in use in the catalogue of the IMI-ConcePTION Project and in the catalogue of the international nonprofit association Vaccine Monitoring Collaboration for Europe (VAC4EU) [13, 35]. Both catalogues use MOLGENIS software [36]. A comparison between the work developed in MINERVA and other real-world data catalogues and catalogues representing cohorts has been published [37], and the MINERVA metadata list is compatible with recommendations to describe diversity across data sources generating RWE identified in a recent scoping review [38].

On the basis of the results of the MINERVA study and the consultation of the ENCePP community and other stakeholders, the HMA–EMA Catalogues of real world-data sources and studies have been developed:

- The catalogue of studies covers studies performed on the data sources, enhancing and replacing the EU PAS Register [39].
- The catalogue of data sources covers information on real-world databases, enhancing and replacing the ENCePP resources database [40].
- The HMA/EMA draft Good Practice Guide for the use of the Metadata Catalogue of Real-World Data Sources V 1.0 is publicly available [41].

5 | Conclusion

The MINERVA project's key publicly available deliverables are the list of metadata and the related guidance document. Based on the experience of the MINERVA pilot, setting up and maintaining an operating metadata catalogue on RWDSs will require a substantial effort to implement the FAIR principles, adhere to data protection rules, and effectively support discoverability of data sources and reproducibility of studies in Europe. The research community will value a transparent presentation of metadata related to the strengths and limitations of data sources.

5.1 | Plain Language Summary

Collections of data relating to patient health status and/or the delivery of health care routinely generated from a variety of sources (so-called "real-world data sources") are used to generate scientific evidence. The MINERVA project defined a list of 442 characteristics of data sources (metadata) that should be collected to describe a data source comprehensively. The list

includes many dimensions that describe not only the content of the data source, but also the context that defines the data, the institutions that are involved in generating and using the data source, the network of such institutions, the studies they conduct, and the technical aspects of the conduct of the studies. To pilot the list, we collected a selection of the metadata in a sample of 15 data sources and displayed them in a *prototype* using an open-source software called MOLGENIS (www.molgenis.org/). The investigators identified features of the metadata list and the prototype that will help support the use of real-world data sources to study the safety and effectiveness of medicines, as well as important considerations for the future collection of metadata.

Author Contributions

Authors worked collaboratively and contributed to the work performed throughout the duration of the project and were also involved in the preparation or review of the manuscript.

Ethics Statement

The research conducted did not involve human subjects and as such ethics reviews or approvals were not required.

Consent

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The information that supports the findings of this project, including main deliverables and the metadata list, is available through the HMA-EMA Catalogues of Real-World Data Sources and Studies at https://catalogues.ema.europa.eu/node/3409/administrative-details.

References

1. X. Kurz, "Darwin EU—First Experience and Regulatory Use Cases," Data Analytics and Methods Taskforce, European Medicines Agency. EMA/HMA Big Data Stakeholder Forum 2022, accessed December 15, 2022, https://www.ema.europa.eu/en/documents/presentation/session-2-1-darwin-eu-r-first-experiences-regulatory-use-case-xavier-kurz_en.pdf.

2. M. Wilkinson, M. Dumontier, I. Aalbersberg, et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 15, no. 3 (2016): 160018, https://doi.org/10.1038/sdata.2016.18.

3. Heads of Medicines Agencies, European Medicines Agency, "HMA–EMA Joint Big Data Taskforce Phase II Report: Evolving Data-Driven Regulation," accessed October 11, 2022, https://www.ema.europa.eu/en/documents/other/hma-ema-joint-big-data-taskforce-phase-ii-report-evolving-data-driven-regulation_en.pdf.

4. Heads of Medicines Agencies, European Medicines Agency, "Priority Recommendations of the HMA–EMA Joint Big Data Task Force," HMA–EMA Big Data Steering Committee Group, accessed January 9, 2023, https://www.ema.europa.eu/en/documents/other/priority-recommendations-hma-ema-joint-big-data-task-force_en.pdf.

5. Heads of Medicines Agencies, European Medicines Agency, "Big Data Steering Group," accessed January 9, 2023, https://www.ema.europa.eu/system/files/documents/work-programme/workplan-2023-2025-hmaema-joint-big-data-steering-group_en.pdf.

6. A. Pacurariu, K. Plueschke, P. McGettigan, et al., "Electronic Healthcare Databases in Europe: Descriptive Analysis of Characteristics and Potential for use in Medicines Regulation," *BMJ Open* 8, no. 9 (2018): e023090.

7. N. H. Thurin, R. Pajouheshnia, G. Roberto, et al., "From Inception to ConcePTION: Genesis of a Network to Support Better Monitoring and Communication of Medication Safety During Pregnancy and Breastfeeding," *Clinical Pharmacology and Therapeutics* 111, no. 1 (2022): 321–331, https://doi.org/10.1002/cpt.2476.

8. R. Gini, M. C. J. Sturkenboom, J. Sultana, et al., "Different Strategies to Execute Multi-Database Studies for Medicines Surveillance in Real-World Setting: A Reflection on the European Model," *Clinical Pharmacology and Therapeutics* 108, no. 2 (2020): 228–235.

9. European Medicines Agency, "Catalogue of RWD Studies," MINERVA, Final Set of Metadata and Definitions, Process, and Catalogue Tool, accessed May 2, 2024, https://catalogues.ema.europa.eu/sites/default/files/document_files/MINERVA_D5_Report_v1.1_28June2021.pdf.

10. US Food and Drug Administration, "FDA's Sentinel Initiative," accessed October 11, 2022, https://www.fda.gov/safety/fdas-sentinel-initiative.

11. Cnodes.ca, "CNODES | Canadian Network for Observational Drug Effects Studies," accessed October 11, 2022, https://www.cnodes.ca.

12. Ehden.eu, "European Health Data Evidence Network – ehden.eu," accessed October 11, 2022, https://www.ehden.eu/.

13. Imi-conception.eu, "ConcePTION," accessed October 11, 2022, https://www.imi-conception.eu/.

14. Aspennet.asia, "Asian Pharmacoepidemiology Network 2021," accessed October 11, 2022, https://aspennet.asia/.

15. Fairplus-project.eu, "FAIRplus | Home Page," accessed October 11, 2022, https://fairplus-project.eu.

16. Maelstrom-research.org, "Home Page | Maelstrom Research," accessed 11 October 2022, https://www.maelstrom-research.org.

17. Aetion, "Real-World Evidence Solution | RWE Analytics | Aetion," accessed October 11, 2022, https://aetion.com.

18. European Medicines Agency, "Technical Workshop on Real-World Metadata for Regulatory Purposes," Virtual Meeting, European Medicines Agency, accessed January 16, 2023, https://www.ema.europa.eu/en/documents/other/summary-report-technical-workshop-real-world-metadata-regulatory-purposes_en.pdf.

19. G. C. Hall, B. Sauer, A. Bourke, J. S. Brown, M. W. Reynolds, and R. LoCasale, "Guidelines for Good Database Selection and Use in Pharmacoepidemiology Research," *Pharmacoepidemiology and Drug Safety* 21, no. 1 (2012): 1–10.

20. European Medicines Agency, "Catalogue of RWD Studies. MINERVA Deliverable 2, Selection of Data Sources and Justification: Final," accessed May 2, 2024, https://catalogues.ema.europa.eu/system/files/2024-02/MINERVA_DataSourceSelection_Final_26March2021.pdf.

21. K. J. van der Velde, F. Imhann, B. Charbon, et al., "MOLGENIS Research: Advanced Bioinformatics Data Software for Non-Bioinformaticians," *Bioinformatics* 35, no. 6 (2019): 1076–1078.

22. EUCAN-connect, "Catalogue," accessed April 24, 2023, https://catalogue.eucanconnect.eu/#/.

23. LifeCycle, "Project Summary," accessed April 24, 2023, https://lifecycle-project.eu/about-lifecycle/project-summary/.

24. J. Ronkainen, R. Nedelec, A. Atehortua, et al., "LongITools: Dynamic Longitudinal Exposome Trajectories in Cardiovascular and Metabolic Noncommunicable Diseases," *Environmental Epidemiology* 6, no. 1 (2021): e184.

25. M. Vrijheid, X. Basagaña, J. R. Gonzalez, et al., "Advancing Tools for Human Early Lifecourse Exposome Research and Translation (ATHLETE): Project Overview," *Environmental Epidemiology* 5, no. 5 (2021): e166.

26. The European Human Exposome Network, "Human Exposome," accessed April 24, 2023, www.humanexposome.eu.

27. Data Catalogue Vocabulary (DCAT), "Data Catalogue Vocabulary – Version 2," accessed April 24, 2023, https://www.w3.org/TR/vocab-dcat-2/.

28. GitHub, "Observational Medical Outcomes Partnership Common Data Model," accessed January 26, 2023, https://ohdsi.github.io/CommonDataModel/.

29. A. But, M. L. De Bruin, M. T. Bazelier, et al., "Cancer Risk Among Insulin Users: Comparing Analogues With Human Insudalin in the CARING Five-Country Cohort Study," *Diabetologia* 60, no. 9 (2017): 1691–1703.

30. G. Trifirò, V. Isgrò, Y. Ingrasciotta, et al., "Large-Scale Postmarketing Surveillance of Biological Drugs for Immune-Mediated Inflammatory Diseases Through an Italian Distributed Multi-Database Healthcare Network: The VALORE Project," *BioDrugs* 35, no. 6 (2021): 749–764.

31. European Medicines Agency, "Catalogue of RWD Studies," MINERVA Deliverable 9, Final Good Practice Guide for Metadata Collection for Real-World Data Sources, accessed May 2, 2024, https://catalogues.ema.europa.eu/sites/default/files/document_files/MINERVA_GoodPracticeGuide_10Jan2022.pdf.

32. European Commission, "European Health Data Space," accessed December 5, 2023, https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en.

33. R. Gini, R. Pajouheshnia, L. Gutierrez, et al., "Metadata for Data dIscoverability aNd Study rEplicability in obseRVAtional Studies: Lessons Learnt From the MINERVA Project in Europe: Lessons Learnt on Metadata Catalogues," *Pharmacoepidemiology and Drug Safety* (2024), https://doi.org/10.1002/pds.5884.

34. European Medicines Agency, "List of Metadata for Real World Data Catalogues," accessed November 23, 2023, https://www.ema.europa.eu/en/documents/other/list-metadata-real-world-data-catalogues_en.pdf.

35. VAC4EU, "Toolbox Catalogue," accessed December 19, 2023, https://vac4eu.org/catalogue/.

36. M. A. Swertz, M. Dijkstra, T. Adamusiak, et al., "The MOLGENIS Toolkit: Rapid Prototyping of Biosoftware at the Push of a Button," *BMC Bioinformatics* 11, no. 12 (2010): S12, https://doi.org/10.1186/1471-2105-11-S12-S12.

37. M. Swertz, E. van Enckevort, J. L. Oliveira, et al., "Towards an Interoperable Ecosystem of Research Cohort and Real-World Data Catalogues Enabling Multi-Center Studies," *Yearbook of Medical Informatics* 31, no. 1 (2022): 262–272.

38. R. Gini, R. Pajouheshnia, H. Gardarsdottir, et al., "Describing Diversity of Real World Data Sources in Pharmacoepidemiologic Studies: The DIVERSE Scoping Review," *Pharmacoepidemiology and Drug Safety* 33, no. 5 (2024): e5787, https://doi.org/10.1002/pds.5787.

39. European Medicines Agency, "Catalogue of RWD Studies | HMA–EMA Catalogues of Real-World Data Sources and Studies (europa.eu)," accessed May 2, 2024, https://catalogues.ema.europa.eu/catalogue-rwd-studies.

40. European Medicines Agency, "Catalogue of RWD Sources | HMA–EMA Catalogues of Real-World Data Sources and Studies (europa.eu)," accessed May 2, 2024, https://catalogues.ema.europa.eu/catalogue-rwd-sources.

41. European Medicines Agency, "Good Practice Guide for the use of the Metadata Catalogue of Real-World Data Sources," V 1.0, accessed November 23, 2023, https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/good-practice-guide-use-metadata-catalogue-real-world-data-sources_en.pdf.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.