MINERVA: Metadata for data dIscoverability aNd study rEplicability in obseRVAtional studies

Strengthening Use of Real-World Data in Medicines Development: Metadata for Data Discoverability and Study Replicability (EMA/2017/09/PE/16)

Selection of Data Sources and Justification: Final

A Partnership of the EU PE&PV Research Network, the SIGMA Consortium, and Collaborators

EUPAS 39322

26 March 2021

Prepared for:

European Medicines Agency

Katerina-Christina Deli, Stefania Simou
Healthcare Data, Data Analytics and Methods Task Force

Lead authors:

Lia Gutierrez, Andrea Margulis, Romin Pajouheshnia, Vera Ehrenstein, Morris Swertz, Karin Gembert, Susana Perez-Gutthann
On behalf of the MINERVA Project Consortium
Contact: lgutierrez@rti.org

MINERVA Project Consortium

Names	Key Roles Organisation	Organisation, Country
Susana Perez-Gutthann, Lia Gutierrez, Carla Franzoni, Andrea Margulis, Alejandro Arana Joan Fortuny, Estel Plana	**Proposal and Project Coordination** **Lead Tasks 1-3, 5, 10** Co-lead SIGMA office	RTI-HS, Barcelona, Spain
Miriam Sturkenboom, Daniel Weibel, Vjola Hoxhaj	**Lead Task 7** President VAC4EU PI ConcePTION	UMCU, Netherlands
Olaf Klungel, Satu Johanna Siiskonen, Romin Pajouheshnia, Helga Gardarsdottir, Patrick Souverein, Marloes Bazelier	**Lead Task 4** EU PE&PV research network lead and management PM EU **DAP, CPRD Aurum, UK**	UU, Netherlands
Rosa Gini, Giuseppe Roberto	**Lead Tasks 6+9, co-lead Task 4** **DAP, Tuscany, Italy**	ARS Toscana (ARS), Italy
Morris Swertz, Eleanor Hyde	**Lead Task 8, Co-lead Task 4**	UMCG, Netherlands
Cécile Droz-Perroteau, Nicolas Thurin, Régis Lassalle	**DAP, SNDS, France**	BPE, University of Bordeaux, France
Vera Ehrenstein	**DAP, Denmark**	DCE-AU, Aarhus, Denmark
Ron Herings, Josine Kuiper	**DAP, PHARMO; Netherlands** Co-lead SIGMA office	PHARMO, Netherlands
Ulrike Haug, Bianca Kollhorst	**DAP, GePaRD, Germany**	BIPS, Bremen, Germany
Helle Kieler, Karin Gembert	**DAP, Swedish registers**	CPE KI, Stockholm, Sweden
Ian Douglas, Anna Schultze	**DAP, CPRD GOLD, UK**	LSHTM, London, United Kingdom
Miguel Gil García, Miguel Ángel Maciá	**DAP, BIFAP, Spain**	AEMPS, Madrid, Spain
Gabriel Sanfélix-Gimeno, Aníbal Garcia-Sempere, Isabel Hurtado, Clara Rodríguez-Bernal, Francisco Sanchez-Saez	**DAP, Valencia, Spain**	FISABIO, Valencia, Spain
Beatriz Poblador-Plou, Jonás Carmona-Pírez, Alexandra Prados-Torres, Antonio Gimeno-Miguel, Antonio Poncel-Falcó	**DAP, EpiChron, Aragon, Spain**	IACS, Zaragoza, Spain
Mitja Kos, Igor Locatelli	**DAP, Slovenia**	UL FFA, Ljubljana, Slovenia
Manuel Barreiro, Eugeni Domènech, Yamile Zabana	**DAP, ENEIDA patient registry, Spain**	GETECCU, Spain
Bas Middelkoop, Eoin McGrath, Andreu Gusi, Silvia Zaccagnino, Maria Paula Busto	**DAP, EBMT patient registry, Europe**	EBMT, Europe
Andres Metspalu, Steven Smit, Merit Kreitsberg, Kadri Raav	**DAP, Estonian Biobank**	University of Tartu, Institute of Genomics, Estonia

DAP = data access provider.

Note: Full organisation names are included in the abbreviation section.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AEMPS	*Agencia Española de Medicamentos y Productos Sanitarios* (Spain)
ARS	*Agenzia Regionale di Sanità della Toscana* (Italy)
ATMP	advanced therapy medicinal product
AU-DCE	Aarhus University, Department of Clinical Epidemiology (Denmark)
BIFAP	*Base de datos para la Investigación Farmacoepidemiológica en Atención Primaria* (Spain)
BIPS	Leibniz Institute for Prevention Research and Epidemiology (Germany)
BPE	Bordeaux PharmacoEpi (France)
CBV	Centraal Beheer Verrichtingenbestand (central management transaction file)
CCAM	Classification commune des actes médicaux (medical acts)
CDM	common data model
CET	Central European Time
CNODES	Canadian Network for Observational Drug Effect Studies
COVID-19	coronavirus disease 2019
CPE KI	Centre for Pharmacoepidemiology, Karolinska Institutet
CPRD	Clinical Practice Research Datalink
CVV	Classificatie van Verrichtingen (classification of transactions)
DAP	data access provider
EBB	Estonian Biobank
EBMT	The European Society for Blood and Marrow Transplantation
EHDEN	European Health Data & Evidence Network
EHR	electronic health record
EpiChron cohort	a data source for studying chronic diseases (Aragón, Spain)
EMA	European Medicines Agency
ENEIDA	*Estudio Nacional en Enfermedad Inflamatoria intestinal sobre Determinantes genéticos y Ambientales*
EOB	end of the business day
EU	European Union
FAIR	Findable, Accessible, Interoperable, and Reusable
FDA	Food and Drug Administration
FISABIO	Foundation for the Promotion of Health and Biomedical Research of the Valencia Region (Spain)
GePaRD	German Pharmacoepidemiological Research Database
GETECCU	Spanish Working Group on Crohn's Disease and Ulcerative Colitis
GOLD	general practitioner online database of the CPRD
IACS	*Instituto Aragonés de Ciencias de la Salud* (Spain)
ISPE	International Society for Pharmacoepidemiology
KI	Karolinska Institutet
LSHTM	London School of Hygiene and Tropical Medicine
MINERVA	Metadata for data dIscoverability aNd study rEplicability in obseRVAtional studies
NABM	Nomenclature des actes de biologie médicale (lab tests)
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
PHARMO	PHARMO Institute for Drug Outcomes Research (Netherlands)
RTI-HS	RTI Health Solutions

SNDS	*Système National des Données de Santé* (France)
SNOMED	Systematized Nomenclature of Medicine
UK	United Kingdom
UL FFA	University of Ljubljana, Faculty of Pharmacy (Slovenia)
UMCG	University Medical Centre Groningen (Netherlands)
UMCU	University Medical Centre Utrecht (Netherlands)
US	United States
UU	Utrecht University (Netherlands)
VID	Valencia Health System Integrated Database (Spain)
ZA	ZorgActiviteit (care act)

GLOSSARY

- Biological medicinal product: Products intended for medicinal use in humans, such as biological products manufactured using biotechnology in a living system or advanced therapy medicinal products (ATMPs) or naturally derived biologicals. Biological products include a wide range of products, such as vaccines, blood and blood components, allergenics, somatic cells, gene therapy, tissues, and recombinant therapeutic protein.
- Catalogue: The set of standardised tools used to access, search, and visualise metadata to illustrate each data bank and data source.
- Common data model (CDM): Common structure and format for data that allows for an efficient execution of programmes against local data.
- Data access provider (DAP): An organisation that can gain protocol-based access to one or multiple data banks (e.g., for the purpose of research or surveillance).
- Data bank: Data collections mandated and sustained by a specified organisation. A data bank is defined by the data originator that sustains the collection of records in the data bank, the underlying population that can potentially contribute records to a data bank, and the prompt that leads to creation of a record in the data bank. For example: data banks of the UK Clinical Practice Research Datalink (CPRD) data source include the CPRD GOLD primary care data (CPRD GOLD) and the Hospital Episode Statistics (HES); data banks of the Danish National Registries data source include Danish National Patient Registry, Danish National Prescription Registry, Medical Birth Registry. The data bank terminology has been used in international collaborations in the genomics research field (NCBI-NIH, 2020) and in the Clinical Interchange Standards Consortium (CDISC) (Clinical Data Interchange Standards Consortium, 2020). Further examples of data banks and their relation to data sources and DAPs are provided in deliverable 3 and deliverable 5, Section 2.1.
- Data source: All the data banks referring to the same underlying source population that a given data access provider can access and link to one another at an individual level. For example: if a data access provider gains protocol-based access to the CPRD GOLD data source including HES, the resulting data source will be composed of those two data banks. If an institution gains protocol-based access to an extraction from the Danish National Registry including the Danish National Patient Registry, Danish National Prescription Registry, and Medical Birth Registry, the resulting data source will be composed of those three data banks.
- FAIR (Findable, Accessible, Interoperable, and Reusable) principles:
  - *Findability*: Any healthcare database that is used for analysis should, from a scientific perspective, persist for future reference and reproducibility. A comprehensive record of the database in terms of purpose, sources, vocabularies and terms, access-control mechanisms, licence, consents, etc., should be available.
  - *Accessibility*: Data should be accessible through a standardised and well-documented method, with suitable interfaces for humans and programmes.
  - *Interoperability*: The use of a CDM, standardised dictionary, and common statistical approaches should allow healthcare data from multiple data sources to be leveraged for evidence generation. The use of standard formats and interfaces supports automated data ingest, maintenance, and exchange.
  - *Reusability*: For data to be reusable, the data licences should explicitly allow the data to be used by others, and the data provenance (understanding how the data came into existence) needs to be specified and updated as needed.
- Medical products: Products intended for medical use in humans, including medicinal drugs, biological medicinal products, medical procedures, and medical devices or equipment.

- Metadata: A set of data that describes and gives information about other data. More specifically, information describing the generation, location, and ownership of a data set; key variables; and the format (coding, structured versus not) in which the data are collected is needed to enable accurate identification and qualification of the exposure and outcome information available. In addition, metadata also include the provenance and timespan of the data, clearly documenting the input, systems, and processes that define data of interest. Finally, metadata include details on the storage, handling processes, access, and governance of data.
- Standardised dictionary: A tool containing a list of variables and their definitions to enable transparent and consistent content across disparate observational databases and allows efficient and reproducible observational research.

1 BACKGROUND

1.1 Data Source Selection Criteria From Existing Research Networks and Initiatives

Enabling discoverability of data is among the ten priority recommendations of the European Medicines Agency (EMA) Big Data Taskforce (HMA-EMA, 2019). In this context, finding suitable data sources to deliver data of sufficient depth and details in several European countries to allow addressing specific research questions is critical for regulatory decision making. Information on the general principles for the selection of data sources from existing collaborative research networks in North America, Europe, and internationally, as well as published guidelines, is of interest to the MINERVA project.

The Sentinel System, sponsored by the US Food and Drug Administration (FDA), is a collaboration between the FDA and both data and academic partners that provide access to healthcare data and/or scientific, technical, methodologic, and organisational expertise, as needed. The Harvard Pilgrim Health Care Institute leads the Sentinel System Coordinating Center. Currently, 16 data partners, including organisations such as academic medical centres and healthcare systems with electronic health record systems and health insurance companies with administrative claims data, have data in the Sentinel Common Data Model (CDM) (Sentinel, 2021). At its earlier stages of development and to inform decisions about potential data partners for Sentinel, the following aspects were assessed: (1) size of the population, degree of capture (e.g., number of records for the previous complete year for variables such as unique encounters, admissions, physicians, and laboratory results), and duration of longitudinal follow-up across different patient care settings and payment systems; (2) structure and coding, including consistency with widely recognised standards; (3) completeness, timeliness, and accessibility of data, including estimated times from service delivery to the accessibility of data via queries; (4) level of detail that would be available to examine temporal relationships between product administration and associated adverse events/outcomes; and (5) potential limitations if used for postmarketing product safety surveillance. Additional detail is provided in Annex 1, Table A1.

The Canadian Network for Observational Drug Effect Studies (CNODES), a joint initiative of Health Canada and the Canadian Institutes of Health Research, is an academically based distributed network of Canadian researchers and data centres that work together to respond to research questions about the safety and effectiveness of drugs marketed in Canada. CNODES currently uses population databases from eight Canadian provinces (British Columbia, Alberta, Saskatchewan, Manitoba, Ontario, Quebec, Nova Scotia, and Newfoundland and Labrador), the Clinical Practice Research Datalink (CPRD) in the UK, and MarketScan in the US and includes more than 100 million patients (Clarkson, 2020; Suissa et al., 2012). No specific criteria for the selection of provinces or data sources were applied.

The European Health Data and Evidence Network (EHDEN) is a public-private consortium with 22 partners from academia, patient associations, regulatory authorities, and pharmaceutical and other companies (additional detail is provided in Annex 1, Table A2). Led by the Erasmus Medical Centre, EHDEN receives funding from the European Commission. In EHDEN's September 2020 call for data partners, prioritisation criteria for data partners with access to person-level observational health data from electronic health records, claims, hospitals, or registries were data impact, network impact, and metadata sharing. These three areas would weight equally towards the total score. Data impact includes size, coverage, uniqueness, and perceived quality. Specifically, preference would be given to data sources with a larger size, more complete coverage from various care settings, more complete representation of the underlying population, availability of more data domains when size is comparable, and longer follow-up within each data type. Network impact included considerations on the willingness and ability to participate in network studies, including feasibility assessments; availability of information on ethical and governance mechanisms for study participation; and experience in projects using the proposed data source. The evaluation of metadata sharing included the ability and willingness to share information on data provenance, available data domains, size of the database, univariate statistics, and others (EHDEN, 2020b).

The Observational Health Data Sciences and Informatics (OHDSI) research network is an international open network consisting of over 100 institutions with access to patient-level healthcare data that converted their data to the Observational Medical Outcomes Partnership (OMOP) CDM. Information provided by sites is collected in the Data Network census maintained by OHDSI. Each site decides to participate in any given study; final inclusion in any given study depends on the suitability of the data for the study (OHDSI, 2021). OHDSI and EHDEN have collaborated; for example, EHDEN hosted an OHDSI two-day research meeting in March 2020 to organise and launch COVID-19–related research (EHDEN, 2020a).

The International Society of Pharmacoepidemiology's (ISPE) special interest group on database research provided recommendations on relevant aspects to consider for an adequate selection of databases to address specific research questions in pharmacoepidemiological research. Aspects highlighted of relevance to our project were database coverage related to population size and representativeness, the capture of study variables and accessibility, continuity and consistent data capture, record duration and latency, and database expertise. When the use of multiple data sources is considered, data linkage capabilities and reliability of individual-level linkage and data storage and analyses were also highlighted. Additional aspects related to data extraction, coding, data retrieval and analysis, availability of validated algorithms, data privacy, and security and quality (Hall et al., 2012).

1.2 Selection and Characterisation of Data Sources

The EMA Big Data Initiative acknowledges the increasing complexity of data being captured across multiple settings and highlights the importance of *"understanding the quality and representativeness of Big data to allow regulators to select the optimal data set to study an important question impacting the benefit-risk of a medicine"* (HMA-EMA, 2019).

This document describes the proposed criteria for selecting appropriate real-world data sources in the MINERVA project and summarises the characteristics of the selected real-world data sources that will contribute a defined set of metadata to the proof-of-concept MINERVA metadata catalogue and how they fit the proposed criteria for participating data sources. We also provide details on the justification for the selection of the listed data sources.

2 OBJECTIVES

Objective 1 EMA Specification. "To define a list of criteria to identify relevant real-world data sources from which the data sources to be included in this study will be selected. Real-world data sources that should be included are:

- **"Databases allowing to link drug utilisation data to subsequent and existing clinical events and demographic variables for individual patients: i) primary care, ii) specialist care, iii) hospital care data from EHRs, iv) claims databases, and v) disease registries;**
- **"Databases allowing to measure for each individual patient the duration of use and the cumulative doses of medicines prescribed/delivered: vi) longitudinal drug prescription, dispensing or other drug utilisation."**

Objective 2. EMA Specifications. "To identify a list of minimum 10 databases to use in the study. The selected databases should include databases of different scope and format (e.g., having both databases in source format and databases converted in an OMOP Common Data Model format) covering all the different types of real-world data sources mentioned in objective 1."

The following are the objectives of these subtasks:

- Develop a list of criteria that will guide the decisions for the selection of relevant real-world data sources for inclusion in the proof-of-concept catalogue in the MINERVA project.
- Select at least ten European data sources of real-world health data that will contribute metadata to the proof-of-concept catalogue in the MINERVA project, ensuring diversity in geographic representativeness, types of data sources, and healthcare settings.
- Evaluate the characteristics of data sources in relation to the criteria proposed for selecting data sources in the MINERVA project and provide the justification for their selection.

The MINERVA Consortium is submitting to the EMA the list of criteria according to the timelines in Annex 2, Table B1 and the characteristics and justification of the selected data sources according to the timelines in Table B2 in Annex 2.

3 CRITERIA FOR SELECTION OF RELEVANT REAL-WORLD DATA SOURCES

The selection criteria for inclusion in the proof-of-concept MINERVA metadata catalogue, listed below, take into account the characteristics listed in the EMA specifications for objective 1 and relevant data source selection criteria from existing initiatives and collaborative networks. In addition, the criteria follow the general principles highlighted by researchers who have assessed electronic health databases for potential use in regulatory drug decision making related to accessibility, longitudinal dimension, recording of exposure and outcomes, and generalisability of databases (Hall et al., 2012; Pacurariu et al., 2018).

The proposed criteria should be fulfilled by the data sources to be included in a metadata catalogue and do not include specific metadata. In Section 5, Other Considerations, we list some relevant items related to data access providers (DAPs)—i.e., research

institutions with access to the data sources—that are instrumental to accessing data and conducting research that are also of interest for the metadata proof-of-concept catalogue.

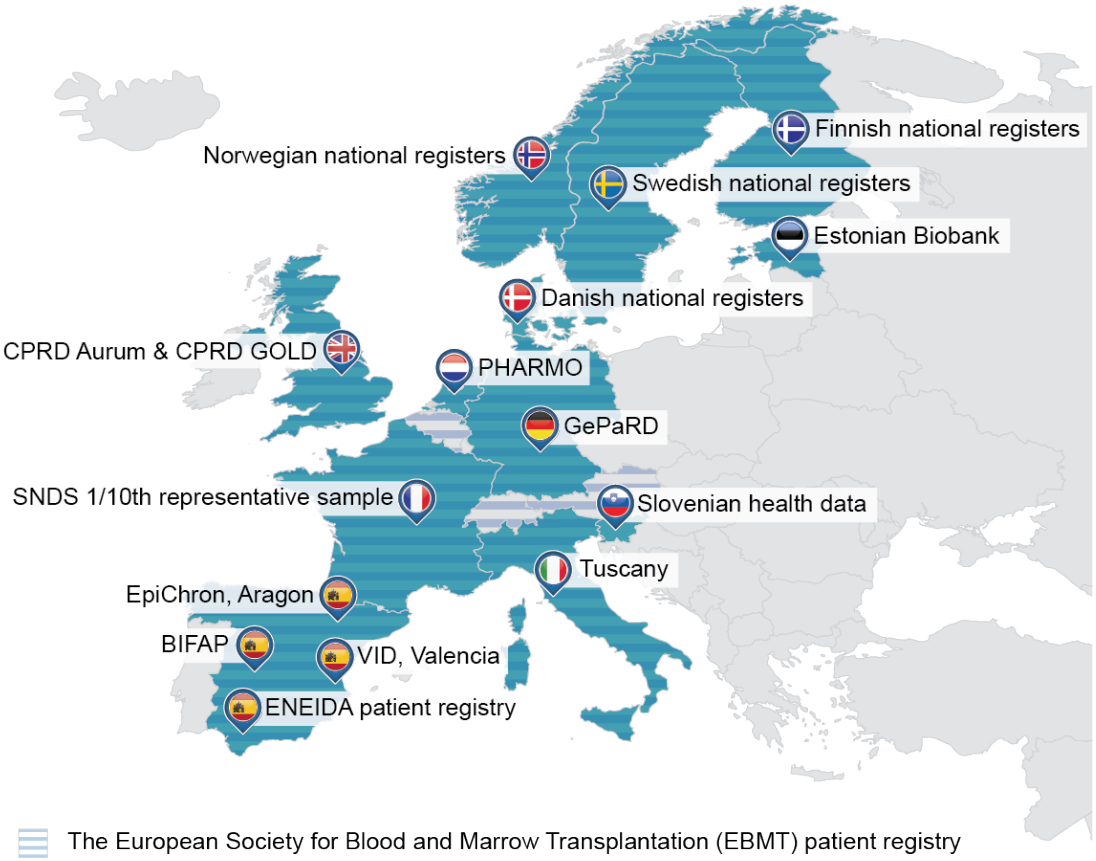
The following criteria are proposed for the selection of real-world data sources for inclusion in the MINERVA proof-of-concept metadata catalogue and for future expansion of this or similar catalogues. Selected data sources must fulfil all of the proposed criteria:

1. Data sources collecting health data routinely. These can be electronic healthcare databases, claims databases, disease registries, or genomics data sets covering defined populations at a local or national level in countries in Europe. Other types of data sources will not be excluded.
2. Data sources collecting patient-level data. Patient-level data collected in a data source, covered through a single data bank or through linkages of multiple data banks, are medicinal products utilisation data and demographic variables including at least one of the following settings:
   - Primary care: data from primary care/general practice
   - Specialist care: data from outpatient hospital clinics or specialised outpatient centres
   - Hospital care: data from inpatient hospital settings, which may or may not include emergency care data
3. Data sources collecting detailed patient-level measures of medicinal product use. Data on the timing of prescription/dispensing/administration/delivery of medicinal drugs are the critical elements. Longitudinal capture of medicinal drugs prescription/dispensing/administration, as appropriate, is required. In addition, the information should allow for estimation of duration of use and the cumulative doses. Capture of use of biologics, medical devices and other medical products is of value.
4. Data sources with continuous and consistent data capture. This refers to the absence of systematic or prolonged temporal gaps in data collection; for example, a data source that skipped collecting pharmacy data in a given year would not be eligible. Data sources where the data collection ceased will not be considered for this project.
5. Data sources where structured data is collected. For example, data sources collecting data only as free text will not meet this criterion. Data sources that have not converted to an existing CDM will be eligible for inclusion in the proof-of-concept catalogue. Data sources that have converted to an existing CDM will be described to facilitate inclusion in the proof-of-concept catalogue.
6. Data sources with procedures in place to comply with applicable patient data privacy and confidentiality rules, allowing use of data for public health research.

4 SELECTION OF DATA SOURCES

As part of the work performed for this project during its initial proposal planning phase, the MINERVA Consortium identified a list of real-world data sources considered of interest for this project. An additional patient registry suggested by the EMA, the EBMT registry, has recently agreed to join the Consortium. These data sources are the target for the data source selection component of the project. Research centres that maintain or access these data sources are integrated into the Consortium. The Consortium established contact with one additional data source suggested by the EMA, the Estonian Biobank (EBB), which on 08 March 2021 communicated its willingness to join the project. Figure 1 shows the geographic distribution of the participating data sources.

Figure 1. Participating Data Sources and Countries



4.1 Characteristics of Data Sources Related to Selection Criteria

The data sources identified in the planning phase of the MINERVA project were assessed in relation to their characteristics related to the proposed criteria listed in Section 3.

Table 1 and Table 2 summarise the characteristics of the data sources focusing on those aspects related to the specific criteria defined for selecting the data sources that will participate in the MINERVA project. The information compiled was based on publicly available information and completed by the respective DAPs. For the Estonian Biobank, which agreed to join MINERVA on 08 March 2021, the information provided is based on publicly available information, which was augmented by the EBB researchers.

Table 1. Characteristics of Data Sources: Criteria 1, 2, and 3

		Criterion 1		Criterion 2 and Criterion 3						
Country: Data Source (DAP)	**Type**	**Routine Data Collection/ Since**	**Active Population (Million)/ Country Coverage**	**Primary Care**	**Specialist Care**	**Hospital Care**	**Age and/or Date of Birth**	**Sex**	**Diagnoses**	**Longitudinal Patient-Level Capture of Medicinal Drugs/ Estimated Duration of Use and Cumulative Dosage**
Denmark: Danish national registers (AU-DCE)	Record linkage system	+/1994	5.8/100%	–	(partial)	+	+ (all ages)	+	+	+/+
Europe: EBMT patient registry	Patient registry	+/1974	~640K/Multiple countries (~60 countries and ~700 centres, including centres that are no longer active)	–	–	+	+ (all ages)	+	+	+/+[a]
Estonia: Estonian Biobank	Population-based cohort and record linkage system	+/2002	202K/15%	+	+	+	+ (≥ 18 years)	+	+	+/–[b]
Finland: Finnish national registers (local institutions)	Record linkage system	+/1993	5.5/100%	+	(partial)	(partial)	+ (all ages)	+	+	+/+
France: SNDS 1/10th representative sample (BPE, University of Bordeaux)	Claims	+/2006	6.6/10%	–	–	+	+ (all ages)	+	+ (inpatient)	+/+
Germany: GePaRD (BIPS)	Claims	+/2004	15.0/20%	+	+	+	+ (all ages)	+	+	+/+
Italy: Tuscany (ARS)	Claims	+/1996	3.6/6%	–	–	+	+ (all ages)	+	+	+/+

		Criterion 1		Criterion 2 and Criterion 3						
Country: Data Source (DAP)	**Type**	**Routine Data Collection/ Since**	**Active Population (Million)/ Country Coverage**	**Primary Care**	**Specialist Care**	**Hospital Care**	**Age and/or Date of Birth**	**Sex**	**Diagnoses**	**Longitudinal Patient-Level Capture of Medicinal Drugs/ Estimated Duration of Use and Cumulative Dosage**
Netherlands: PHARMO (PHARMO Institute)	Record linkage system	+/1990	7/25%	(partial)	+	+	+ (all ages)	+	+	+/+
Norway: Norwegian national registers (local institutions)	Record linkage system	+/2004	5.3/100%	–	(partial)	+	+ (all ages)	+	+	+/+
Slovenia: Slovenian health data (UL FFA)	Claims	+/2000	2.1/100%	(partial)	(partial)	+	+ (all ages)	+	+	+/+
Spain: BIFAP (AEMPS) (Macia-Martinez et al., 2020)	Electronic medical records	+/2001	8.0/17%	+	–	(partial)	+ (all ages)	+	+	+/+
Spain: ENEIDA patient registry (GETECCU)	Patient registry IBD	+/2005	70,000 patients	–	IBD	IBD	+	+	+	+/+[c]
Spain: VID, Valencia (FISABIO) (Garcia-Sempere et al., 2020)	Record linkage system	+/2008	5.0/11%	+	+	+	+ (all ages)	+	+	+/+
Spain: EpiChron, Aragon (IACS) (Prados-Torres et al., 2018)	Record linkage system	+/2011	1.2/2.6%	+	+	+	+	+	+	+/+
Sweden: Swedish national registers (CPE KI)	Record linkage system	+/2005	10.2/100%	–	(partial)	+	+ (all ages)	+	+	+/+
United Kingdom: CPRD Aurum (UU)(CPRD, 2021)	Electronic medical records	+/1995	13.3/19.9%	+	–	(partial ~90%)	+ (all ages)	+	+	+/+

		Criterion 1		Criterion 2 and Criterion 3						
Country: Data Source (DAP)	**Type**	**Routine Data Collection/ Since**	**Active Population (Million)/ Country Coverage**	**Primary Care**	**Specialist Care**	**Hospital Care**	**Age and/or Date of Birth**	**Sex**	**Diagnoses**	**Longitudinal Patient-Level Capture of Medicinal Drugs/ Estimated Duration of Use and Cumulative Dosage**
United Kingdom: CPRD GOLD (LSHTM)	Electronic medical records	+/1987	3.0 /4.6%	+	–	(partial (~47%)	+ (all ages)	+	+	+/+

Key: + = yes; – = no.

DAP = data access provider; IBD = inflammatory bowel disease.

[a] Data on haematopoietic cellular therapies only.

[b] Data on duration and dosage of medications currently not available; work is in progress to obtain access to this information.

[c] Data on treatments for ulcerative colitis only.

Note: Full organisation names are included in the abbreviation section.

Table 2. Characteristics of Data Sources: Criteria 4, 5, and 6

	Criterion 4		Criterion 5		Criterion 6
Country: Data Source (DAP)	**Data Capture Is Continuous and Consistent**	**Data Are Being Currently Collected**	**Collected Data Are Structured (Coding Systems or Description of Structure)**	**Data Transformed Into a CDM[a]**	**Data Privacy Procedures in Place, Use for Public Health Research Allowed**
Denmark: Danish national registers (AU-DCE)	+	+	+ ICD-8, ICD-10, ATC, and other coding systems	+ ConcePTION, Nordic [CARING project; (Andersen et al., 2015; But A, 2017)]	+
Europe: EBMT patient registry	+	+	Internal coding system	–	+
Estonia: Estonian Biobank	+	+	ICD-10, ATC	OMOP	+
Finland: Finnish national registers (local institutions)	+	+	+ ICD-10, ATC	+ ConcePTION, Nordic (CARING project)	+
France: SNDS 1/10th representative sample (BPE, University of Bordeaux)	+	+	+ ICD-10 (French modification), French procedural codes (CCAM, NABM), ATC	+ ConcePTION, OMOP	+
Germany: GePaRD (BIPS)	+	+	+ ICD-10 (German modification), OPC, ATC	+ ConcePTION	+
Italy: Tuscany (ARS)	+	+	+ Data are stored in 8 linkable tables, each with a number of predefined fields	+ ConcePTION, OMOP, TheShinISS (Trifirò, 2020)	+
Netherlands: PHARMO (PHARMO Institute)	+	+	+ ICD-10, Dutch procedural codes (CBV/CVV/ZA), ATC	+ ConcePTION, OMOP	+
Norway: Norwegian national registers (Local institutions)	+	+	+ ICD-10, ATC	+ ConcePTION, Nordic (CARING project)	+

	Criterion 4		Criterion 5		Criterion 6
Country: Data Source (DAP)	**Data Capture Is Continuous and Consistent**	**Data Are Being Currently Collected**	**Collected Data Are Structured (Coding Systems or Description of Structure)**	**Data Transformed Into a CDM[a]**	**Data Privacy Procedures in Place, Use for Public Health Research Allowed**
Slovenia: Slovenian health data (UL FFA)	+	+	+ ICD-10 (Australian modification), Slovenian procedural codes, ATC	–	+
Spain: BIFAP (AEMPS)	+	+	+ ICD-9, ICPC-BIFAP, ATC	+ ConcePTION	+
Spain: ENEIDA patient registry (GETECCU)	+	+	+ Locally defined set of variables	–	+
Spain: VID, Valencia (FISABIO)	+	+	+ ICD-9-CM, ICD-10 (Spanish modification), ATC	+ ConcePTION	+
Spain: EpiChron, Aragon (IACS)	+	+	+ ICPC, ICD-9 (clinical modification), DRG, ATC	–	+
Sweden: Swedish national registers (CPE KI)	+	+	+ ICD-8, ICD-9, ICD-10 (Swedish modification), ATC codes	+ Nordic (CARING project), ConcePTION	+
United Kingdom: CPRD Aurum (UU)	+	+	+ SNOMED, Read and EMIS codes	+ ConcePTION	+
United Kingdom: CPRD GOLD (LSHTM)	+	+	+ Read codes	+ ConcePTION	+

Key: + = yes; – = no.

ATC = Anatomical Therapeutic Chemical (ATC) Classification System; CBV = Centraal Beheer Verrichtingenbestand (central management transaction file); CCAM = Classification commune des actes médicaux (medical acts); CDM = common data model; CVV = Classificatie van Verrichtingen (classification of transactions); DAP = data access provider; DRG = diagnosis related groups; ICD = International Classification of Diseases (number indicates revision, CM means "clinical modification") ICPC = International Classification of Primary Care; NABM = Nomenclature des actes de biologie médicale (lab tests); OMOP = Observational Medical Outcomes Partnership; OPC = Operative Procedure Code; SNDS = Système National des Données de Santé; SNOMED = Systematized Nomenclature of Medicine; ZA = ZorgActiviteit (care act).

[a] CDMs into which data have been mapped by the participating MINERVA DAP.

Note: Full organisation names are included in the abbreviation section.

5 OTHER CONSIDERATIONS

The criteria listed in Section 3 do not take into consideration the possible overlap of populations included in different data sources. This aspect should be assessed when selecting data sources for future studies using this type of catalogue.

As part of Task 4, Defining Metadata, the MINERVA Consortium will explore the process and requirements for data accessibility and timeliness for each data source and whether metadata can be made available to regulatory authorities or to third parties for research purposes. These requirements may vary over time.

We aim for a diverse and representative selection of data sources covering different European geographic regions and settings of health care services (primary care, specialist care, hospital care) and types of data (claims data, medical care data, disease registries, and others). Experience in pharmacoepidemiology and the availability of informatics tools and human resources with knowledge and expertise to handle the data source and provide the requested information in a timely manner will be instrumental to the success of the project.

6 JUSTIFICATION FOR SELECTION OF DATA SOURCES

- The data sources proposed to be included in the MINERVA proof-of-concept catalogue meet the selection criteria.
- The selected data sources provide wide geographic representativeness and distribution across European regions, different healthcare settings and health systems, as well as different types of data sources, including two patient registries.
- The selected data sources have been included in numerous regulatory-driven post-authorisation safety studies, which highlights the importance of the use of observational data to support regulatory decision making.
- MINERVA DAPs can gain protocol-based access to one or multiple data banks in the selected data sources and have extensive experience in research studies, often regulatory-driven multi-database studies.
- MINERVA DAPs have agreed to contribute to the metadata task.

7 REFERENCES

Andersen M, Thinsz Z, Citarella A, Bazelier MT, Hjellvik V, Haukka J, et al. Implementing a Nordic common data model for register-based pharmacoepidemiological research. Nor Epidemiol. 2015 Sep;25(Suppl 1):P27.

But A DBM, Bazelier MT, Hjellvik V, Andersen M et al. Cancer risk among insulin users: comparing analogues with human insulin in the CARING five-country cohort study. Diabetologia. 2017;2017(60):1691-703.

Clarkson M. What is CNODES and what does it do? Saskatchewan Health Quality Council; 28 January 2020. Available at: https://www.hqc.sk.ca/news-and-events/hqc-blog/what-is-cnodes-and-what-does-it-do#What%20exactly%20is%20CNODES? Accessed 21 January 2021.

Clinical Data Interchange Standards Consortium C. Glossary. 18 December 2020. Available at: https://www.cdisc.org/standards/glossary. Accessed 10 March 2021.

CPRD. CPRD Aurum February 2021 Dataset. 2021. Available at: https://cprd.com/cprd-aurum-february-2021. Accessed 23 February 2021.

EHDEN. COVID19 Study-a-thon. 2020a. Available at: https://www.ehden.eu/covid19-study-a-thon/. Accessed 20 January 2021.

EHDEN. Data partner call description. 14 September 2020b. Available at: https://www.ehden.eu/wp-content/uploads/2020/09/Data-Partner-Call-Description-v4.0.pdf. Accessed 19 January 2021.

FDA. Evaluation of potential data sources for the FDA Sentinel Initiative – final report. US Food and Drug Administration; 1 October 2010. Available at: https://www.regulations.gov/document?D=FDA-2009-N-0192-0017. Accessed 21 January 2021.

Garcia-Sempere A, Orrico-Sanchez A, Munoz-Quiles C, Hurtado I, Peiro S, Sanfelix-Gimeno G, et al. Data Resource Profile: The Valencia Health System Integrated Database (VID). Int J Epidemiol. 2020 Jun 1;49(3):740-1e.

Hall GC, Sauer B, Bourke A, Brown JS, Reynolds MW, LoCasale R. Guidelines for good database selection and use in pharmacoepidemiology research. Pharmacoepidemiol Drug Saf. 2012 Jan;21(1):1-10.

HMA-EMA. Joint Big Data Taskforce phase II report: evolving data-driven regulation. 2019. Available at: https://www.ema.europa.eu/en/documents/other/hma-ema-joint-big-data-taskforce-phase-ii-report-evolving-data-driven-regulation_en.pdf. Accessed 20 January 2021.

Macia-Martinez MA, Gil M, Huerta C, Martin-Merino E, Alvarez A, Bryant V, et al. Base de Datos para la Investigacion Farmacoepidemiologica en Atencion Primaria (BIFAP): A data resource for pharmacoepidemiology in Spain. Pharmacoepidemiol Drug Saf. 2020 Oct;29(10):1236-45.

NCBI-NIH. GenBank Overview. 2020. Available at: https://www.ncbi.nlm.nih.gov/genbank/. Accessed 11 March 2021.

OHDSI. The book of OHDSI. Observational Health Data Sciences and Informatics; 11 January 2021. Available at: https://ohdsi.github.io/TheBookOfOhdsi/. Accessed 20 January 2021.

Pacurariu A, Plueschke K, McGettigan P, Morales DR, Slattery J, Vogl D, et al. Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation. BMJ Open. 2018;8(9):e023090.

Prados-Torres A, Poblador-Plou B, Gimeno-Miguel A, Calderon-Larranaga A, Poncel-Falco A, Gimeno-Feliu LA, et al. Cohort Profile: The Epidemiology of Chronic Diseases and Multimorbidity. The EpiChron Cohort Study. Int J Epidemiol. 2018 Apr 1;47(2):382-4f.

Sentinel. Who is involved. 2021. Available at: https://www.sentinelinitiative.org/about/who-involved. Accessed 18 January 2021.

Suissa S, Henry D, Caetano P, Dormuth CR, Ernst P, Hemmelgarn B, et al. CNODES: the Canadian Network for Observational Drug Effect Studies. Open Med. 2012;6(4):e134-40.

Trifirò GM, M; Da Cas, R; Menniti Ippolito, F; Sultana J, et al. Renin–Angiotensin–Aldosterone System Inhibitors and Risk of Death in Patients Hospitalised with COVID-19: A Retrospective Italian Cohort Study of 43,000 Patients. Drug Saf. 2020;43(August 2020):1297-308.

Annex 1. Data Source Eligibility Criteria and Related Evaluations in Existing Research Networks

Table A1. Data Elements Included in the Assessment of Data Sources: Sentinel System

Item	Assessed Aspects
General questions	1. Organisation type 2. Types of electronic health data 3. Time period covered 4. Settings 5. Types of insured population(s) 6. Geographic coverage
Data attribute: population coverage	7. Length of time (median and maximum) for any one patient 8. Frequency of data source updates 9. Total number of unique patients and encounters for the most recent full year 10. Total number of prescriptions for the most recent full year 11. Total number of laboratory test results for the most recent full year 12. Number of hospitals included in the most recent full year 13. Number of hospital admissions for the most recent full year 14. Number of physicians for the most recent full year 15. Number of group practices for the most recent full year 16. Percentage of the population with health insurance that includes prescription benefits 17. Identification of pregnancies
Data attribute: structure and coding	18. Data source structured in flat files or a relational database
Data attribute: data linkage capabilities	19. Unique or more than one identification number (ID) for an individual patient 20. Interfaces with the following types of systems: ▪ Registries ▪ Biospecimen management and tracking systems ▪ Medical examiner/coroner's office ▪ Medical image management and tracking systems ▪ Electronic prescribing systems ▪ Personal health record systems ▪ Vital statistics 21. Standards for storage and exchange of clinical data
Data attribute: medical reconciliation capabilities	22. Clinical decision support software or homegrown logic in place to identify drug-drug interactions
Data attribute: devices	23. Are devices included? 24. Device information linked or potentially linkable to medical records 25. Device information linked or potentially linkable to claims data
Data attribute: clinical trials	26. Existence of field that identifies if the patient is/was enrolled in a clinical trial and reference to the trial 27. Is the organisation a member of or does it contribute data to any clinical research networks?
Source	28. FDA Evaluation of Potential Data Sources for the FDA Sentinel Initiative – Final Report (FDA, 2010)

Table A2. Data Source Eligibility/Prioritisation Criteria: EHDEN (September 2020 Data Partner Call)

Item	Criteria
Data types	Person-level observational health data from these types of sources:[a] • Electronic health records • Claims • Hospital • Registry
Standardised data should contain…	• Exposures; e.g., drugs, devices • Procedures • Outcomes: e.g., conditions and measurements
Prioritisation of data sources based on…	A score that is given to each data source based on • Data impact: size, coverage, uniqueness, and perceived quality • Network impact: expected impact for the EHDEN ecosystem • Metadata sharing: data provenance, size, domains; univariate statistics; etc.
Source	EHDEN September 2020 call for data partners (EHDEN, 2020b)

[a] EHDEN mentions that data sources that were not included in this call because they could not be accommodated into the common data model in use at the time are Biobanks, Research Networks, Patient Reported Outcomes, and Genetic data.

Annex 2. Deliverable 2 Timelines

Table B1. Organisation and Timelines Deliverable 2a: Criteria for Selection of Data Sources

Subtask	Who	Date (EOB CET)
2a. Draft list of criteria to identify data sources based on EMA requirements	**RTI-HS, UU, UMCU, ARS, BPE, AU-DCE, PHARMO, FISABIO**	26 Jan 2021 (complete)
2b. Review and provide input on draft criteria	**Consortium**	03 Feb 2021 (complete)
2c. Consolidate input from reviewers	**RTI-HS**	08 Feb 2021 (complete)
2d. Submit draft list of criteria to EMA (Deliverable 2a)	**RTI-HS**	10 Feb 2021 (complete)
2e. EMA provides feedback on draft list of criteria	**EMA**	19 Feb 2021 (complete)
2f. EMA comments are integrated and reviewed, and final Deliverable 2a is integrated with draft Deliverable 2b	**RTI-HS, UU, UMCU, ARS, BPE, AU-DCE, PHARMO, FISABIO**	25 Feb 2021 (complete)

CET = central European time; EMA = European Medicines Agency; EOB = end of the business day.

Table B2. Organisation and Timelines Deliverable 2b: Data Sources Selection and Justification

Subtasks	Who	Date (EOB CET)
3a. Draft report mapping the list of criteria to the list of participating data sources	**RTI-HS, UU, UMCU, ARS, BPE, AU-DCE, PHARMO, FISABIO**	10 Feb 2021 (complete)
3b. Review of draft report with mapping of criteria	**Consortium**	12 Feb 2021 (complete)
3c. Review and finalise draft report	**RTI-HS, UU, UMCU, ARS, BPE, AU-DCE, PHARMO, FISABIO, UMCG, DAPs**	23 Feb 2021 (complete)
3d. Submit draft report combining the list of criteria for selection of data sources (Del. 2a) and selection of data sources and justification to EMA (Del. 2b) **(Deliverable 2)**	**RTI-HS**	25 Feb 2021 (complete)
3e. EMA provides feedback on draft report	**EMA**	04 Mar 2021 (complete)
3f. Address feedback and distribute updated draft report to Consortium for review	**RTI-HS, UU, UMCU, ARS, BPE, AU-DCE, PHARMO, FISABIO**	08 Mar 2021 (complete)
3g. Review and provide input to updated draft report	**Consortium**	10 Mar 2021 (complete)
3e. Consolidate input from reviewers	**RTI-HS**	12 Mar 2021 (complete)
3f. Submit final report combining the list of criteria for selection of data sources and selection of data sources and justification to EMA (**Deliverable 2**)	**RTI-HS**	15 Mar 2021 (complete)

CET = central European time; EMA = European Medicines Agency; EOB = end of business day.