8 April 2025
EMA/787647/2022
European Medicines Agency

# Good Practice Guide for the use of the HMA-EMA Catalogues of real-world data sources and studies

Version 2.0

| | |
|---|---|
| Start of public consultation | 26 September 2022 |
| End of consultation (deadline for comments) | 15 November 2022 |
| Agreed by MWP | 4 December 2024 |
| Adopted by CHMP | 17 February 2025 |
| Version 2 published | 8 April 2025 |

| | |
|---|---|
| Keywords | Data sources, metadata, study protocol, study report, data flows, data management, vocabulary, glossary, use cases, population, real-world data, observational studies |

Official address Domenico Scarlattilaan 6 ● 1083 HS Amsterdam ● The Netherlands
Address for visits and deliveries Refer to www.ema.europa.eu/how-to-find-us
Send us a question Go to www.ema.europa.eu/contact Telephone +31 (0)88 781 6000 An agency of the European Union

# Contents

# Abbreviations

| | |
|---|---|
| ATMPs | Advanced Therapy Medicinal Products |
| CDM | Common Data Model |
| DQF | Data quality framework |
| EMA | European Medicines Agency |
| EMRN | European Medicines Regulatory Network |
| ENCePP | European Network of Centres for Pharmacoepidemiology and Pharmacovigilance |
| ETL | Extract, Transform, Load |
| EU | European Union |
| EU PAS Register | European Union electronic register of post-authorisation studies |
| FAIR | Findable, Accessible, Interoperable, and Reusable |
| FDA | Food and Drug Administration |
| GDPR | General Data Protection Regulation 2016/679 (EU) |
| GP | General practitioner |
| HARPER | HARmonized Protocol template to Enhance Reproducibility |
| HMA | Heads of Medicines Agencies |
| ICU | Intensive care unit |
| ID | Identification |
| IMI | Innovative Medicines Initiative |
| ISO | Internal Organization for Standardization |
| MedDRA | Medical Dictionary for Regulatory Activities |
| MINERVA | Metadata for data dIscoverability aNd study rEplicability in obseRVAtional studies |
| OMOP | Observational Medical Outcomes Partnership |
| PICOT | Population, Intervention, Comparison, Outcome and Time horizon |
| RWD | Real-world data |
| RWE | Real-world evidence |
| SIFPD | Structured Process to Identify Fit-for-Purpose Data |
| TEHDAS | Towards the European Health Data Space |

# 1. Introduction

Real-world data (RWD) and real-world evidence (RWE) are increasingly used in the scientific evaluation of medicines to support regulatory decision-making. Conducting RWD research and interpreting study findings pose unique challenges and require a good understanding of associated methods, terminologies as well as deep knowledge of data source characteristics and of the healthcare system organisation in the respective countries [1, 2]. However, information on RWD sources is often lacking or not standardised. Such information, commonly provided as metadata[1] or "data about data", offers context about the data and includes details on a dataset's purpose, location, key-variables, generation, format, and ownership. As mentioned in the Data Quality Framework (DQF) for EU medicines regulation (hereafter referred to as EMRN DQF) [3], metadata are essential to understand the meaning of data and to assess the fitness for purpose (e.g. quality) of a dataset for a specific purpose. The HMA-EMA (Heads of Medicines Agencies-European Medicines Agency) Catalogues of RWD sources and studies along with the EMRN DQF and the Draft 'Data Quality Framework (DQF) for EU medicines regulation: application to Real-World Data' (hereafter referred to as RW-DQF) [1] will help to further address these aspects.

Having access to a standard and electronic set of complete and accurate metadata for data sources can support the identification of data sources suitable for a specific study, facilitate the description of the data sources planned to be used for a specific study or research, and help to properly assess the evidentiary value of the study results. Metadata are often published in data catalogues, which have the purpose of making data discoverable and assessable for fitness for purpose without revealing the raw data themselves.

The Heads of Medicines Agencies/European Medicines Agency (HMA/EMA) joint Big Data Task Force recommended "to promote data discoverability through the identification of metadata" as part of its priority recommendation III: "*Enable data discoverability. Identify key meta-data for regulatory decision making on the choice of data source, strengthen the current European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) resources database to signpost to the most appropriate data, and promote the use of the FAIR principles (Findable, Accessible, Interoperable and Reusable)*" (HMA/EMA, 2020). This goal was therefore included in the 2023-2025 Work Plan[2]of the HMA/EMA joint Big Data Steering Group.

To fulfil this mandate, EMA initiated in November 2020 the study "Strengthening Use of Real-World Data in Medicines Development: Metadata for Data Discoverability and Study Replicability" (MINERVA; EUPAS39322) [4]. The focus of the MINERVA study was to define a set of metadata for RWD sources, together with engagement of relevant stakeholders to reach broad agreement on this identified set, as well as to develop a good practice guide describing the metadata and recommendations.

Based on the results of MINERVA, multiple other studies[3], and consultation with the ENCePP community and other stakeholders, EMA has developed two electronic catalogues to provide metadata for RWD sources and real-world (observational) studies that can be found here: https://catalogues.ema.europa.eu/. These publicly available catalogues were launched on 15 February 2024 and have the following key objectives:

1) Facilitate the *discoverability* of adequate data sources to generate real-world evidence (RWE) for regulatory purposes (e.g. identification of RWD source suitable for investigating a specific research need);

[1]For a more comprehensive definition of metadata and its role, please see the EMRN DQF [1].
[2] HMA/EMA Workplan 2023-2025
[3] Studies: EUPAS39322, EUPAS49303, EUPAS104093

2) Aid in the suitability assessment of data sources by providing clear and easy access to information from the study protocol and study report;

3) Improve interoperability between studies and data sources;

4) Improve transparency.

# 2. Purpose of this document

The Good Practice Guide for the use of the HMA-EMA Catalogues of RWD sources and studies has been developed to provide regulators, researchers (including academia and pharmaceutical industry) and other interested stakeholders with recommendations on the use of the catalogues from the perspective of the data user. Specifically, the guide focuses on describing the steps and best practices on identifying a suitable data source, when planning a study.

For the users submitting data to the Catalogues (i.e.: data holders and investigators submitting study data) a user guide has also been published to help users navigate the Catalogues in this context. The user guide provides descriptions of the data fields and definitions, as well as guidance on how to submit and maintain a record in the Catalogues. The user guide is available on the Catalogues website (https://catalogues.ema.europa.eu/).

# 3. Considerations on the design of the Catalogues

The Catalogues collect information on the systems and processes behind data capture, as well as on descriptors of the data, and they are intended for capturing the extent of variety of existing data sources as well as facilitate data discoverability and fit-for-purpose assessments[4]. The information in the catalogues is organised in sections. For data sources, these sections include *Administrative details, Data elements collected (including Vocabularies used), Quantitative descriptors, Data flows and management*. For studies, these sections include *Administrative details, Methodological aspects*, and *Data management*. They are composed of qualitative information and quantitative metadata, e.g., counts and demographic distributions of the underlying population. The two Catalogues are supported with data from Networks and Institutions.

The Catalogues follow good practices for data management:

- FAIR (Findable, Accessible, Interoperable and Reusable) principles [5]: The Catalogues are supporting globally unique and persistent identifiers, ensuring seamless data integration and allowing users to reference specific entries consistently across different contexts and platforms. There is ongoing work to fully implement the FAIR principles to promote good data practices and allow integration with other catalogues.
- A controlled data entry process is run for the initial collection of metadata by the data holder and/or researchers; regular updates of metadata are foreseen at least once per year for data source records, and at each study milestone for study records.
- Change management and reproducibility are supported by enabling data source holders and/or researchers to edit the corresponding metadata while ensuring that the attribution of each data entry is traceable via appropriate version control, and by enabling the creation of a copy of the metadata and their update by the data holders and/or researcher.
- The General Data Protection Regulation (GDPR) is complied with by defining and communicating the purpose and use of the catalogues, establishing appropriate measures to protect the privacy of institutions and individuals, providing information on the process and approval needed for data availability and accessibility and ensuring a technical option to delete the metadata involving relevant personal data.

4 Assessing the fitness for purpose of a data source is strictly related to characterization of a source's data quality as noted in the EMA DQF. To this respect, we note how "systems and processes" and "descriptors of data" correspond to the concepts of "foundational determinants" and "intrinsic determinants" introduced in the framework.

A quality management process is in place, including an incident management system, a disaster recovery plan and a quality assurance office. It is also worth mentioning that the Catalogues do not provide access to primary records of data. The Catalogues only provide metadata about data sources and studies. Metadata do not contain individual patient data, and therefore informed consent is not applicable. Permission by the data holder to use the underlying data is always needed to access the data for analysis.

### *3.1. Scope of the data sources Catalogue*

The scope of the Catalogue of RWD sources (and this document) is on sources of secondary RWD, meaning data already collected for another purpose than the specific research question you would have, such as for patient monitoring, healthcare reimbursement, quality management or another administrative purpose. Primary data are in scope only when patient registries [6] are considered. Evidently, this document will also touch upon patient registries as sources of secondary data. The Catalogue provides information allowing for an initial evaluation of the suitability of data sources for a study.

### *3.2. Scope of the studies Catalogue*

The Catalogue of RWD studies, replacing the former EU PAS Register, aims to promote transparency of observational studies in the context of the medicine regulation activity. A central repository where observational research is registered benefits the scientific community and promotes the use of RWE [13]. For the purpose of this document, studies are described in relation to the selection of a suitable data source, as studies data is linked with its respective data source.

## 4. Use of the catalogues to identify and assess data sources

### *4.1. Categories of the RWD Catalogues metadata elements*

The following section describes the content of the Catalogues in summary terms with reference to the specific metadata elements (e.g., C2.7, D.1.2.1.1) in parentheses. The complete list of metadata elements is available in the document 'List of metadata for the HMA-EMA Catalogues of real-world data sources and studies' [7], which has been generated in line with the Big Data Steering Group workplan and the European medicines regulatory network Strategy. Both this Good Practice Guide and the metadata elements document [7] will be updated regularly as healthcare categories and data source requirements evolve over time.

Please note that herein "elements" is used to refer to metadata elements, while "variables" is used to refer to variables present within a data source.

#### *4.1.1. Metadata about systems, and processes and data quality metrics*

The following metadata elements provide information on the systems and processes that lead to the data. As noted in the EMRN DQF [3] and the RW-DQF [1], these are essential to assess the reliability and timeliness of data, but are also useful to understand other dimensions such as extensiveness and coherence (as defined in the EMRN DQF). Element identifiers are provided in parentheses and references in the 'List of metadata for the HMA-EMA Catalogue of real-world data sources and studies' [1].

- Data management, including the possibility of data validation (elements C2.7, C8.5 and C8.5.1) and the mapping to a common data model (CDM) (D1.2.1.1, D1.2, D1.2.1, D1.4 and D1.7). The

validation of data elements does not mean that access to the underlying data is granted to data sources in the Catalogue. The process of validation is the responsibility of the data holder.

- Mapping Extract, Transform, Load (ETL) to a CDM (B7.1 to B7.5). For more details, please see the EMRN DQF [3] and review maturity level-related sections (chapter 7) as well as the RW-DQF [1] section 1.4.
- Any qualification received by an institution/authority providing formal qualifications (e.g.: EMA, Internal Organization for Standardization (ISO) or other certifications) (C3.1, C3.1.1)
- Governance details such as data capture and management, data quality checks and validation of results of data quality checks (C2.3)
- The process of collecting and recording the data (C4.3), linkage information (B5.2, B.5.2.1, B5.3, B4.1)
- All vocabularies used in the data source[5]
- A link to the publications describing the data source (e.g.: validation, data elements, representativity) preferably by means of peer-reviewed publications that reflect the best description of the data[6]

Beyond the characterisation of systems and processes, the above elements touch on data quality metrics that measure intrinsic aspects of data quality (e.g.: completeness, coherence, etc.). Access to raw data and computational resources may be needed for a more in-depth assessment of intrinsic aspects of data quality, for example a verification of the records and values, data verification against reference or plausible values and other computations. Such assessment could be performed by the data holders and periodically updated, preferably by using automated tools (for an example, please consult [8]). Data holders are encouraged to make the methods and the results of these assessments publicly available for consultation to support the assessment and replication of studies.

The main goal of the data sources Catalogue is discoverability of data sources in the health domain and therefore the data sources' reliability should be assessed in light of a research question. EMA will not provide a benchmark or minimum level of reliability of data sources to be used for studies in order to be included in the Catalogues; however, the principles explained in the EMRN DQF and the RW-DQF will aid the assessment of data quality of each data source used to address a given research question.

### *4.1.2. Metadata describing the dataset content*

Beside a summary understanding of the data quality characteristics of a source, assessing the fitness for purpose of a data source requires an assessment of whether the source contains the appropriate data (relevance) and whether the data are suitable to generate valid evidence informing a specific research question based on the proposed study design. For example, this could inform the implementation of step 3 of the Structured Process to Identify Fit-for-Purpose Data (SPIFD)[9], or the key elements of a structured framework such as the target trial emulation (TTE) framework [10-12] or the Population, Intervention, Comparison, Outcome and Time horizon (PICOT) format [13].

[5] Note that the vocabularies in use can be also detected from an actual dataset.
[6] Such information provides a broad context on the why and how of a data source, that can help understand its biases and trade-offs.

The RWD Catalogues provide elements that cover information on type of data present, as well as their extensiveness. The Catalogues also provide the elements to be included in the table of data sources recommended by the HARmonized Protocol template to Enhance Reproducibility (HARPER) [14].

These data elements cover:

- Setting: country(-ies) (C1.5), region(s) (C1.5.1), type of data source (C5.1 and C5.1.1), care setting (C1.14). For sensitive variables, depending on country- and data holder-specific data privacy rules, not all items may be published in the catalogues due to privacy concerns.
- Population: total and active population size (C7.1), percentage of the population covered by the data source in the catchment areas (C1.11.2), description of the population for which data are not collected (C1.11.1), age groups (C1.8), sociodemographic information (C6.7), lifestyle factors (C6.8), family linkage (C6.6, C6.6.1), availability of data on pregnancy and neonates (C1.9), trigger for registration (C1.6, C1.6.1) and de-registration (C17.1, C1.7.1), median time between first and last records for all individuals (B6.3) and active individuals (B6.3.1).
- Exposure: availability of data on prescriptions and/or dispensing (C6.13), advanced therapy medicinal products (ATMPs) (C6.16), contraception (C6.17), vaccines (C6.19), other injectables (C6.19), medical devices (C6.20), procedures (C6.21), medicinal products (C6.15.1) and indication (C6.18), and biomarker data (C6.26).
- Outcomes: availability of data on hospital admission or discharge (C6.10), intensive care unit (ICU) admission (C6.10.1), death (mortality) and cause of death (C6.11), clinical measurements (C6.23), genetic data (C6.25), patient-generated data (C6.27), health care utilisation (C6.29), diagnostic codes (C6.9), specific diseases (C1.10) with disease information collected (C1.10.1).
- Time elements: date when the data source was established (C4.5), first collection date (C1.12) and last collection date (C1.13), median time between the first and the last available records for unique individuals captured in the data source (B6.3) and for unique active individuals (B6.3.1).

Link to the Catalogue of RWD studies also allows for identification of studies that have been performed with the same data source, and subsequently for an evaluation of its suitability to answer the research question of interest.

## *4.2. Data Quality considerations*

Assessing the suitability of data sources is inherently assessing the quality of a source overall and also in relation to a specific need. Data quality can be defined as “fitness for purpose for users’ needs in relation to health research, policy making and regulation, and the degree to which data reflect the reality which they aim to represent” [15]. Additionally, it is important to differentiate three broad aspects of data quality [1, 3]:

1. Quality of the processes and methods behind the generation of data e.g.: the detection and correction of errors, validation processes, time components and underlying calculations, the documentation of standardised processes leading to entry and exit of person, etc. Elements C9.5 and C2.3 inform on the data quality process employed by a data source, where applicable.
2. Quality of a dataset itself e.g.: in relevance to missing data and implausible values, coherency of formats, codes, the presence of unique identification numbers for each person, etc.
3. Quality as a characteristic of a data source in relation to a specific research question and method e.g.: the presence of the required data needed for a study, the availability and completeness of data

elements and the time span of such data, and whether the number of individuals included, and the population characteristics are adequate.

The Catalogues provide information on all three data quality aspects that could help the user assess the relevance of a dataset for their specific research question(s), and in some cases the relevance of data to their primary use / the original research question that leads to their capture (e.g.: existing publications based on a specific data source).

Furthermore, when considering the assessment of the suitability of data sources for a given study, it is important to consider that data quality and its management are different between studies with primary data collection and studies using secondary data [16]. In primary data collection studies, the study design and the study itself apply and control for all data quality management steps. In studies using secondary data, data quality assessment relies on generic data quality processes that may or may not be relevant to the research question at hand, e.g.: quality processes relevant to data generation, coding, curation, validation, processing and storage).

Several data quality frameworks have been proposed to help characterize these different aspects of data quality, aiming to reveal the strengths and limitations of a data source with regard to answering a given research question. These data quality frameworks differ as to how they categorise and define data quality dimensions, their levels of generality, the extent to which they address the relevance of data/dataset to a research question, as well as the overall goals what they are designed for.

The Towards European Health Data Space (TEHDAS) initiative has produced a data quality framework that defines six dimensions deemed most important at data source level: reliability, relevance, timeliness, coherence, coverage and completeness [15]. The release of the EMRN DQF [3] builds on TEHDAS and defines high-level principles and procedures that apply across EMA's regulatory mandate. This framework provides general considerations on data quality relevant for regulatory decision making, definitions for data quality dimensions and sub-dimensions, as well as their characterisation and related metrics. It also presents a strong distinction between the description of data quality (both foundational and intrinsic aspects) and its assessment with respect to a research question. The RW-DQF [1] describes RWD specific recommendations derived from the EMRN DQF.

As described above, the Data Quality section of the Catalogue will gradually include in its following iterations additional elements describing data quality metrics in line with the EMRN DQF[7] and the RW-DQF. This development will also be guided by the progress on development of the data quality and utility label as described in Art 56 of the European Health Data Space Regulation [12].

It should be noted that evidence does not only depend on data quality. Appropriate epidemiological and statistical methods must be applied to the study design and the analysis of data generated from a RWD source. These methods are not addressed by the Catalogue but are outlined in other guidance, e.g.: the ENCePP Guide on Methodological Standards in Pharmacoepidemiology, 11th Rev. (2023) [10]. General methodological principles of regulatory interest are also included in the Reflection paper on use of real-world data in non-interventional studies to generate real-world evidence, Draft (2024) [17].

## *4.3. Use cases*

The following section presents relevant use cases and illustrates how stakeholders can benefit from the Catalogues when planning a study and assessing data sources.

[7] The EMRN DQF is an umbrella framework applying to a variety of regulatory use cases and not specifically to real-world data. A draft document focusing on the application of the EMRN DQF to real-world data has been published.

### *4.3.1. Planning of a study*

*Use case: An investigator wants to identify suitable data sources for a planned study.*

The process for identification of suitable data sources may follow six successive steps (Figure 1):

1. In a first step, the investigator searches the catalogues to identify relevant data sources fulfilling the specifications of the research question of interest or, if there is a prior interest in using a specific data source, to access the record of this data source and consult the available information. The search may initially consist of the data elements deemed useful to assess pre-defined PICOT [13] criteria (see section 4.1.2) in order to identify possible suitable data sources.

2. In a second step, the investigator accesses the record of each potential data source and screens the information on the availability of data (incl. quantitative metadata) on the population, exposures, outcomes and confounding variables to confirm that the data source may be relevant to answer the research question.

3. In a third step, the investigator consults information on the governance, accessibility and availability of the selected data sources (C2.3) to determine whether they are accessible, as well as the conditions related to their use, and whether the investigator would be eligible to receive aggregated information or get access to raw data.

4. In a fourth step, the investigator screens the metadata allowing to perform a preliminary assessment of the reliability of each potential data source based on important quality aspects of the data source that are relevant for the specific study (see section 4.2). Publications describing the data source and its validation can be extracted and consulted. Missing information for some of these variables may raise doubts about the presence of an adequate quality management process or may question whether the data holder gives sufficient attention to quality management.

   At this stage, the investigator should establish a first list of candidate data sources (if there is no *a priori* choice of a specific data source).

5. In a fifth step, the investigator uses the links providing access to the studies indexed in the Catalogue of RWD studies (via EU PAS numbers) that have been performed with the same data sources to address research questions similar to the current one (in any). After restricting to studies with a similar design as for the planned study (if relevant), the investigator accesses the study information to:

   - Confirm the suitability of the data source as regards to the PICOT criteria; if the study protocol and/or the study report have been uploaded, more granular information can be extracted on the time frame for the use of the database, the number of active study participants originating from the data source (providing useful information for the sample size calculation for the proposed study), the data elements used for the study (e.g. exposure and outcome variables, confounding factors), variable definitions and vocabularies (and any need for mapping of terms), the transformation of data into categories and the analyses that could be performed with the data;

   - Check in the study protocol or study report (if available) the codes and algorithms (if any) that have been used to identify diseases or outcomes of interest as well as their characteristics (if relevant, such as their severity), and which prompts and contents were used in such algorithm(s);

   - Learn about the data sources' strengths and weaknesses (limitations) encountered during the study conduct; in case a limitation is acknowledged showing that the data source is not optimal

to identify all the variables of interest (e.g. diagnosis of the disease, levels of severity, treatments, confounding variables), use of the data source should be reconsidered or a strategy could be devised to complement the information obtained from the data source with that from another, possibly by data linkage;

- Search for use of the data source in studies published in peer-reviewed journals and review comments made on study limitations.

If there are remaining uncertainties regarding the reliability and relevance of the data source for the proposed study, the investigators of similar studies in terms of research question (and study design if relevant) can be contacted to gather additional information.

If past studies using the same data source cannot be found, it may still be valuable to leverage the data source if its quality and appropriateness for the specific research question seems adequate based on the assessment done. Alternatively, investigating the information available for another relevant data source may be a valid option.

6. If the previous steps have been successful, the data holders of the data sources of interest can be contacted to discuss the feasibility of using the data sources for the specific study and the conditions of this use. Also, in case of additional questions, the investigator can contact the data holder directly.

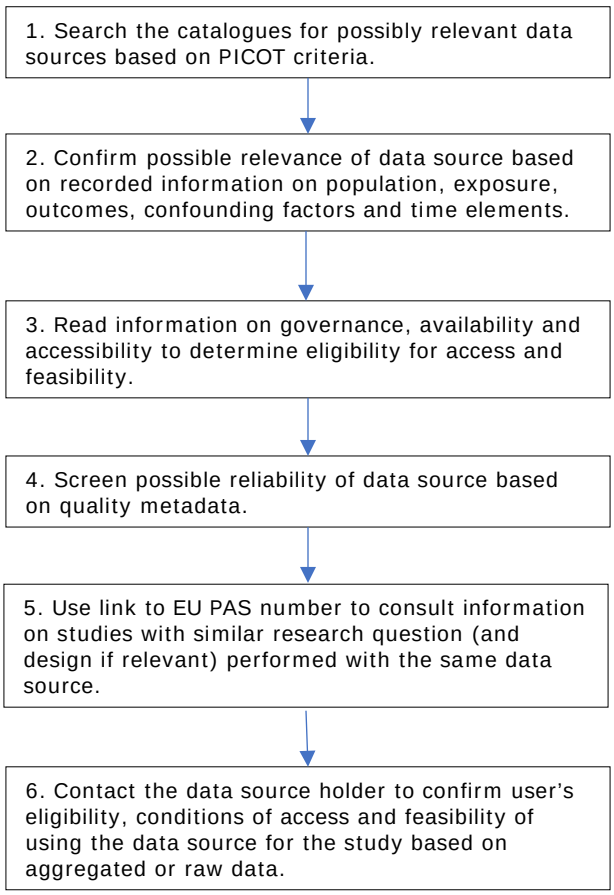


Figure 1. Steps for using the metadata catalogues when planning a real-world study

The Catalogues can also be a useful tool for regulators to plan studies and for assessing data sources. An example is the DARWIN EU® initiative (https://www.darwin-eu.org/). The European Medicines Regulatory Network (EMRN) is the principal user of DARWIN EU® by requesting studies to support its

scientific evaluations and regulatory decision-making process. DARWIN EU® collaborates with data partners who help generate RWE that can be used in scientific evaluations and regulatory decision-making. Via DARWIN EU®, the Catalogue of RWD sources for use in medicines regulation will also be extended, providing high-quality, validated RWD on the use, safety and efficacy of medicines thanks to onboarding of healthcare databases enabling distributed data access via DARWIN EU® that will expand over time.

### *4.3.2. Assessment of a study protocol*

*Use case: A data source is mentioned in the study protocol submitted for a study and the assessor needs to understand in detail the suitability of the data source proposed to be used.*

The assessor may verify if the data source has been registered in the Catalogues.[8] Depending on the information that is already available in the protocol, the assessor accesses different sections of the Catalogue to verify or complement the provided information. For example, in order to examine the fit-for-purpose and potential representativeness of the study population described in the protocol relative to the target population of the study, the assessor may access qualitative information, such as the geographical coverage, the type of data source, the care setting and the trigger for registering a person in the data source, as well as quantitative metadata on the percentage of the population covered by the data source in the catchment area and the estimated sample size of active patients per age category.

In the Data elements section, the assessor may find information on exposure, outcomes, covariates, and subgroups (which may be derived from covariates) collected in the data source and identify those that have not been proposed to be extracted but could be useful to include in the study.

The assessor can also explore technical information supporting the evaluation of the protocol such as the vocabularies used to define variables, the process of data collection, the CDM, the ETL specifications and any linkage strategy.

The extent of the validation of the data source and the possibility to contact patients provides regulatory assessors of studies required to pharmaceutical companies' information about the need and the possibility to request additional data validation. The link to studies using the same data source and registered in the Catalogue of RWD studies will allow to further document use cases where the data source was used with its strengths and limitations.

### *4.3.3. Assessment of a study report or publication*

*Use case: A data source is mentioned in the study report or publication and the reader needs to understand the suitability of the data source used to interpret the study results.*

The process is similar to the process described above for the assessment of a study protocol. The main difference resides in the fact that the study report (or publication) contains results and generally quantitative information on the characteristics of the study population originating from the data source. The reader may therefore identify, and investigate if needed, differences between the information provided in the study report and in the Catalogues. Some verification may be applied to the description of the study population, the sample size originating from the data source included in the report (publication), the sampling technique used to obtain the study population, the nature and categories of

[8] Except of specific circumstances, there is no legal obligation to register a data source into the Catalogue of RWD sources. It is however expected that data source holders will register their data source, and update the records, whenever it will be used for public health or regulatory purpose, as absence of information on the data source in the Catalogue may affect the scientific credibility and public confidence on study results. In case where a data source user has got access to a data source based on a contractual agreement, the contract may include a provision that the data source is registered, or the records updated, in the Catalogue as part of the agreement.

variables included in the analysis, the methods and key variables used to take into account potential biases, the coding system provided, limitations stated in the study report (publication) regarding the reliability and relevance of the data, and whether the analysis followed the protocol or were unable to follow it due to data limitations.

Insight into the characteristics of the data source also helps interpret the study results and understand the strengths and limitations of the study independently from the investigator's own interpretation.

### *4.3.4. Writing of a study protocol or study report*

*Use case: An investigator writes a study protocol or a study report for which they need to describe the data source(s) proposed to be used or used in the study. The information on the data source they find in other publications or other documentation is heterogeneous, and a comparison between the characteristics of several data sources used or to be used in the study is difficult to perform*.

The investigator can extract from the Catalogues standardised information on each data source and provide a reference to public information for the registered data sources. They can provide in the Methods section of the protocol or report the identification number and the link of the data source in the Catalogues.

If a data source is not registered in the Catalogues, this registration can be made simultaneously to the writing of the protocol or report by the data holder.

### *4.3.5. Benchmarking of several data sources*

*Use case: A data holder or data user may wish to compare the characteristics of a specific data source with other ones covering fully or partially the same population.*

The different data sources may have different primary purposes, contain different data elements, and cover different population groups. It is nevertheless important to be able to perform comparisons to help understand the heterogeneity of results obtained in some analyses conducted in the same country or region, or to perform a validation of a data source in comparison to another one considered a gold standard. For this purpose, the Catalogues provide:

- A harmonised description of the characteristics of each data source that allow to compare differences, e.g., in demographics (such as age, etc.) covered.
- Information on common variables and variable categories by which analyses can be stratified to map sources of heterogeneity.
- Information on possible linkages with other data sources, including availability of linkages to the same data sources (or cross-linkage between data sources) allowing to harmonise data on the same individuals and provide additional information, e.g., on confounding factors.

### *4.3.6. Analysis of a data source used in a study*

*Use case: An investigator, statistician or analyst wants to benefit from the experience of others for the programming of the data transformation and statistical analysis.*

If a CDM is used to conduct a study, the analyst will find in the catalogues the specifications of the ETL procedure from the data source to the CDM, along with the CDM version. Irrespective of whether the data holder has converted the entire data source to the CDM, or only an extraction thereof, this information supports the programmer in developing the study script. Using the link to the studies, the

analyst can also access detailed information on the studies conducted using the same data source and select the studies that investigated the same research question with a similar study design. The study protocol or statistical analysis plan of these studies may contain information on how to operationalise the variables of the study in their respective data sources. The detailed programming script may also be available in a public repository, e.g., a GitHub repository, along with debugging records.

At the end of the analysis, the analyst is encouraged to publish the script in a public repository, along with debugging records, and upload the link to the Catalogues as part of the study record, thus enabling transparency, quality control, and facilitating reproducibility.

## Glossary

This glossary addresses the main terms and definitions that have been used in this Good Practice Guide.

| Definitions | Explanation |
|---|---|
| Active population | Individuals alive and currently registered with active records in a data source. For example, an active population for administrative healthcare data refers to the collection of patients for which there is an active record, i.e. the record was created and not closed because the patient moved or died. |
| Catalogue | A collection of dataset descriptions, which is arranged in a systematic manner and consists of a user-oriented public part, where information concerning individual dataset parameters is accessible by electronic means through an online portal. |
| Common data model (CDM) | Common structure and format for data that allows for interoperability, e.g. the efficient execution of the same analysis code against different local database for an efficient execution of programs against local data. |
| Contributor | An institution that contributes content to the Catalogues. |
| Data characterisation | The summarisation of features of a data source, including quantitative measures. |
| Data holder | Any natural or legal person, which is an entity or a body in the health or care sector, or performing research in relation to these sectors, as well as Union institutions, bodies, offices and agencies who has the right or obligation, in accordance with the Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space [18], applicable Union law or national legislation implementing Union law, or in the case of non-personal data, through control of the technical design of a product and related services, the ability to make available, including to register, provide, restrict access or exchange certain data. |
| Data quality | Data quality is defined as fitness for purpose for users' needs in relation to health research, policy making, and regulation and that the data reflect the reality, which they aim to represent. Data quality is relative to the research question and does not address the question on what level is the quality measured e.g., variable, data source or institutional level. These aspects are addressed in the data quality determinants and dimensions of data quality. |

| | |
|---|---|
| Data Quality Framework | A Data Quality Framework provides a set of definitions, guidelines, and recommendation to assess and govern data quality. The framework here presented addresses a wide range of data sources for the purpose of characterising, assessing, and assuring data quality for regulatory decision making. |
| Dataset | A structured collection of electronic health data |
| Data source | Dataset (or a set of linked datasets) sustained by a specified organisation, which is the data holder. The data source is characterised by the underlying population that can potentially contribute records to it, the event occurring that triggers the creation of a record in the data source, and the data model used in the data source. |
| Data user | A natural or legal person who has lawful access to personal or non-personal electronic health data for secondary use. |
| Elements | “Elements” is used to refer to metadata elements, while “variables” is used to refer to the variables present within a data source. |
| Extract, transform, load (ETL) | A data integration process that involves combining and/or ingesting data from the original source(s) (extraction), integrating into a format suitable for analysis (transformation), and making available on a target system(s) for final analysis and computation (loading). In the context of the EMA-HMA catalogues of data sources and non-interventional studies, ETL refers in particular to a repeatable process for converting data from one format to another, such as from a source native format to a common data model format. In this process, mappings to the standardised dictionary may be added. It is typically implemented as a set of automated scripts. |
| FAIR principles | Findable, accessible, interoperable, and reusable principles [5]. |
| Institution | An organisation connected to one or more data sources - such as a data holder, or a research organisation running a study. |
| Metadata | Metadata are defined as “data about data” providing context about their purpose and generation. It’s a set of data that describes and gives information on other data providing context about their purpose, location, key-variables, generation, format, and ownership of a dataset. Metadata are often published in data catalogues, which have the purpose of allowing data to be discoverable and checked for fitness for purpose, without revealing the data themselves. |
| Underlying population | The population of individuals in one or more geographical locations who can potentially contribute information to a data source. This is a population defined by an administrative characteristic, a disease, a medical condition or any other relevant characteristic. |
| Vocabulary | Standardised medical terminologies; may be an international standard (e.g., International Classification of Diseases, Anatomical Therapeutic Chemical) or a country/region-specific system or modification. |

# References


1. *Data Quality Framework for EU medicines regulation: application to Real-World Data. [Draft] EMA/503781/2024. Available at: https://www.ema.europa.eu/en/documents/other/draft-data-quality-framework-eu-medicines-regulation-application-real-world-data_en.pdf.*
2. *Real-world evidence framework to support EU regulatory decision-making. Report on the experience gained with regulator-led studies from September 2021 to February 2023. Available at: https://www.ema.europa.eu/en/documents/report/real-world-evidence-framework-support-eu-regulatory-decision-making-report-experience-gained-regulator-led-studies-september-2021-february-2023_en.pdf.*
3. *Data Quality Framework for EU medicines regulation. EMA/326985/2023. Available at: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en.pdf.*
4. *MINERVA: Strengthening Use of Real-World Data in Medicines Development: Metadata for Data Discoverability and Study Replicability (2022). https://www.encepp.eu/encepp/viewResource.htm?id=45375.*
5. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Sci Data, 2016. 3: p. 160018.
6. *Guideline on registry-based studies. EMA/426390/2021. Available at: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-registry-based-studies_en.pdf.*
7. *HMA/EMA. List of metadata for Real World Data catalogues (2022). Available at: https://www.ema.europa.eu/en/documents/other/list-metadata-real-world-data-catalogues_en.pdf.*
8. Oh, S.W., et al., *Data Quality Assessment for Observational Medical Outcomes Partnership Common Data Model of Multi-Center.* Stud Health Technol Inform, 2023. 302: p. 322-326.
9. Gatto, N.M., et al., *The Structured Process to Identify Fit-For-Purpose Data: A Data Feasibility Assessment Framework.* Clin Pharmacol Ther, 2022. 111(1): p. 122-134.
10. *The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). Guide on Methodological Standards in Pharmacoepidemiology (Revision 11). EMA/95098/2010. https://www.encepp.eu/standards_and_guidances/documents/01.ENCePPMethodsGuideRev.11.pdf.*
11. Hernán, M.A., et al., *Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease.* Epidemiology, 2008. 19(6): p. 766-79.
12. Hernán, M.A. and J.M. Robins, *Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available.* Am J Epidemiol, 2016. 183(8): p. 758-64.
13. Brown, P., et al., *How to formulate research recommendations.* BMJ, 2006. 333(7572): p. 804-6.
14. Wang, S.V., et al., *HARmonized Protocol Template to Enhance Reproducibility of hypothesis evaluating real-world evidence studies on treatment effects: A good practices report of a joint ISPE/ISPOR task force.* Pharmacoepidemiol Drug Saf, 2023. 32(1): p. 44-55.
15. *TEHDAS. European Health Data Space Data Quality Framework (2022). Available at: https://tehdas.eu/app/uploads/2022/05/tehdas-european-health-data-space-data-quality-framework-2022-05-18.pdf.*
16. Wang, S.V. and S. Schneeweiss, *Assessing and Interpreting Real-World Evidence Studies: Introductory Points for New Reviewers.* Clin Pharmacol Ther, 2022. 111(1): p. 145-149.
17. *Reflection paper on use of real-world data in non-interventional studies to generate real-world evidence. [Draft] EMA/CHMP/150527/2024. Available at: https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-use-real-world-data-non-interventional-studies-generate-real-world-evidence_en.pdf.*
18. *Proposal for a Regulation of the European Parliment and of the council on the European Health Data Space, COM(2022) 197 final, 2022/0140 (COD). Available at: https://eur-lex.europa.eu/resource.html?uri=cellar:dbfd8974-cb79-11ec-b6f4-01aa75ed71a1.0001.02/DOC_1&format=PDF.*