1 1 September 2022
2 EMA/787647/2022
3 European Medicines Agency

4 Good Practice Guide for the use of the Metadata 5 Catalogue of Real-World Data Sources

6 V 1.0

Start of public consultation	27 September 2022
End of consultation	16 November 2022

7

Comments should be provided using this template. The completed comments form should be sent to metadata@ema.europa.eu

8

Keywords	Data sources, studies, metadata, study protocol, study report, data flows, data management, vocabulary, glossary, use cases, population

9

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

CDM	Common Data Model
EMA	European Medicines Agency
ENCePP	European Network of Centres for Pharmacoepidemiology and Pharmacovigilance
ETL	Extract, Transform, Load
EU	European Union
EUDPR	Regulation (EU) 2018/1725 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data
EU PAS Register	European Union electronic register of post-authorisation studies
FAIR	Findable, Accessible, Interoperable, and Reusable
FDA	Food and Drug Administration
GDPR	Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)
HARPER	HARmonized Protocol template to Enhance Reproducibility
HMA	Heads of Medicines Agencies
ID	identification
IMI	Innovative Medicines Initiative
MINERVA	Metadata for data dIscoverability aNd study rEplicability in obseRVAtional studies
OMOP	Observational Medical Outcomes Partnership
RWE	Real-world evidence
SIFPD	Structured Process to Identify Fit-for-Purpose Data
TEHDAS	Towards the European Health Data Space

Glossary

- 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 metadata catalogue.
- Data quality: Set of attributes of a data source that define its fitness for purpose for users' needs in relation to health research, policy making and regulation.
- Data source: Data set 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 trigger that leads to the creation of a record in the data source, and the data model used in the data source.
- Dataset: a structured collection of electronic health data.
- 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 this Regulation, 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.
- Extract, transform, load (ETL): 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 are added. It is typically implemented as a set of automated scripts.
- 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.
    - Interoperability: The ability of organisations as well as software applications or devices from the same manufacturer or different manufacturers to interact towards mutually beneficial goals, involving the exchange of information and knowledge without changing the content of the data between these organisations, software applications or devices, through the processes they support.
    - 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.
- Institution: An organisation connected to one or more data sources—such as a Data Holder, or a research organisation running a study.
- Metadata: A set of data that describes and gives information about a dataset. More specifically, information describing the generation, location, and ownership of the 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. Metadata also include the provenance and time span 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.
- Underlying population: The population of individuals in a geographical location 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.

1. Introduction

Identification of appropriate data sources is becoming an increasing need for regulatory decision making. While data needs are becoming more complex, standardised information and statistics on real-world data sources is lacking. Metadata are descriptive data that characterise other data to create a clearer understanding of their meaning and to achieve greater reliability and quality when using the data for a specific purpose. Access to a standard and electronic set of complete and accurate metadata information can contribute to identifying the data sources suitable for a specific study, facilitate description of the data sources planned to be used in a study protocol or research proposal, and contribute to assessing the evidentiary value of the results of studies.

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 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 is therefore included in the 2020-2021 Work Plan of the HMA-EMA joint Big Data Steering Group (HMA-EMA Big Data Steering Group, 2022).

To fulfil this mandate, EMA in November 2020 the study "Strengthening Use of Real-World Data in Medicines Development: Metadata for Data Discoverability and Study Replicability" (MINERVA; EU PAS Register number EUPAS39322). The main focus of the study was the definition of a set of metadata on real-world data sources, including engagement with stakeholders to reach broad agreement and the development of a good practice guide describing the metadata and recommendations based on a pilot.

Based on the results of the MINERVA study and the consultation of the ENCePP community and other stakeholders, the EMA is developing an electronic catalogue that will provide metadata for real-world data sources. This catalogue has two objectives: 1) to facilitate the *discoverability* of data sources to generate adequate evidence for regulatory purpose, i.e., the initial identification of data sources suitable to investigate a specific research question, and 2) to support the assessment of study protocols and study results by providing quick access to information on the suitability of data source(s) proposed to be used in the study protocol or referred to in the study report.

The *Good Practice Guide for the use of the Metadata Catalogue of Real-World Data Sources* has been developed to provide regulators, researchers and other interested stakeholders with recommendations on the use of the EU metadata catalogue of real-world data sources.

2. Purpose of this document

The Good Practice Guide aims to provide recommendations for the use of the EU metadata catalogue to identify real-world data sources suitable for specific research questions and to assess the suitability of data sources proposed to be used in a study protocol or referred to in a study report.

It also provides a detailed description of all the metadata elements as envisaged to be used in the EMA catalogue, which have been published by HMA/EMA in the List of metadata for Real World Data catalogues[1], and it guides the user for the insertion and maintenance of data in the catalogue.

The catalogue is targeted for release in late 2023.

[1] HMA/EMA. List of metadata for Real World Data catalogues (2022).

3. Format of the catalogue

The structure of the catalogue is based on the MINERVA catalogue pilot project.[2] A ***data source*** is a data collection (or a set of linked data collections) sustained by a specified organisation, which is the data holder. It is characterised by the underlying population that can potentially contribute records, the event triggering the creation of a record in the data source and the data model. The mechanisms that put data into existence are heterogeneous across data sources. The catalogue is therefore divided into the following sections allowing to capture the variety of existing data sources and facilitate data discoverability: *Characteristics*, *Population*, *Data elements*, *Data flows and management* and *Vocabularies*. It is composed of qualitative information and quantitative metadata, e.g. counts and demographic distributions of the underlying population.

The catalogue follows good practices for data management:

- FAIR principles are complied with: the data are Findable, Accessible, Interoperable and Reusable,[3] and there is interoperability with the EU PAS register for studies conducted with the data sources and with other catalogues to be developed in the future.
- A controlled data entry process is run for the initial collection of metadata by the data holder, regular updates of metadata are foreseen with trusted relationship between the data holder and the EMA.
- Change management and reproducibility are supported by enabling data holders of a data source 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.
- Quantitative metadata for data sources are provided at the level of the total and active populations.
- Personal data will be processed in compliance with European data protection legislation and, in particular, Regulation (EU) 2018/1725 (EUDPR) and Regulation (EU) 2016/679 (GDPR) as applicable. In this regard, EMA will publish a record of processing activity and a data protection notice as required. A quality management process is in place, including an incident management system, a disaster recovery plan and a quality assurance office.

4. Use of the catalogue to assess the suitability of data sources

4.1. Reliability and relevance of data sources

The assessment of the suitability of data sources for studies needs to consider the differences between studies with primary data collection and studies based on secondary use of data already collected for another purpose, such as patient monitoring, healthcare reimbursement, quality management or another administrative purpose. In primary data collection, the study itself applies and controls all the quality management steps related to the data collected. In secondary data collection, use of already collected data relies on existing processes for data quality, i.e., which data have been collected for the initial purpose and how they were generated, and many aspects of the data processes, i.e., how the data were coded, curated, validated and stored.

[2] MINERVA: Strengthening Use of Real-World Data in Medicines Development: Metadata for Data Discoverability and Study Replicability (2022). EUPAS39322
[3] FAIR Principles. https://www.go-fair.org/fair-principles/

The assessment of the suitability of data sources should therefore differentiate between two broad aspects of data quality[4,5]:

- quality in relation to the *reliability* of the primary data, based on e.g. the detection and correction of errors, missing data and implausible values, the verification and validation of formats, codes, values, time components and underlying calculations, the presence of unique identification numbers for each person and the documentation of standardised processes leading to entry and exit of person; this aspect of quality is a characteristic of the data source independent from its use for a specific study.

- quality in relation to the *relevance* of the data source to provide adequate and valid evidence informing a specific research question following the application of appropriate epidemiological and statistical techniques; this aspect requires adequate information on the format and content of the data source, such as the presence of the data needed for the study, the numbers of individuals included, population characteristics, coding terminologies, the availability and completeness of data elements and the time span of the data; this aspect of quality is partly dependent on the research question as some data characteristics (such as some data elements or age range of the population) may be required for some studies and not for others.

Several data quality frameworks have been proposed to help understand the strengths and limitations of a data source to answer a research question and the impact they may have on the suitability of data sources for a specific study[6,7,8]. These data quality frameworks differ as to the specific dimensions included (with varying levels of details and names used to describe these dimensions) and the methods used to assess them, and some frameworks address both the data reliability and relevance or only one of these. In Europe, the Towards European Health Data Space (TEHDAS) project has set out and defined six dimensions deemed the most important ones at data source level: reliability, relevance, timeliness, coherence, coverage and completeness.[4]

4.2. Assessing suitability of data sources with the catalogue

Reliability

The metadata catalogue provides information allowing an initial evaluation of the suitability of data sources. Information on the following aspects of *reliability* is provided:

- Data management, including the possibility of data validation (elements C2.7, C2.9, C8.5 and C8.5.1), the mapping to a CDM (D1.2.1.1, D1.2, D1.2.1, D1.4 and D1.7)
- The data source ETL process and status (B7.1 to B7.5)
- Any qualification received (C3.1, C3.1.1)
- Governance details as regards data capture and management, data quality checks and validation of results (C2.3)
- The process of collecting and recording the data (C4.3), linkage information (B5.2, B.5.2.1, B5.3, B4.1)

[4] ENCePP Guide on Methodological Standards in Pharmacoepidemiology, 10th Rev. (2022). Chapter 12.1 General principles of quality management
[5] Wang S., Schneeweiss S. Assessing and Interpreting Real-World Evidence Studies: Introductory Points for New Reviewers. Clin Pharmacol Ther. 2022;111(1):145-149.
[6] ENCePP Guide on Methodological Standards in Pharmacoepidemiology, 10th Rev. (2022). Chapter 12.2. Data Quality Frameworks.
[7] TEHDAS. European Health Data Space Data Quality Framework (2022).
[8] HMA/EMA. Data Quality Framework for EU medicines regulation (2022).

- All vocabularies used in the data source
- A link to the publications describing the data sources (e.g. validation, data elements, representativity).

Access to raw data and computational resources would be required for a more in-depth assessment of reliability, for example a verification of the records and values, data validation against reference or plausible values and other computations. Such assessment should be performed by the data holders and periodically updated. The data holders should make the methods and the results of the assessment publicly available for consultation to support the assessment and replication of studies.

Relevance

The metadata catalogue is also suitable for an initial evaluation of the *relevance* of the data sources to generate valid evidence informing a specific research question based on the study design, e.g. to implement step 3 of the Structured Process to Identify Fit-for-Purpose Data (SIFPD)[9] or the Population, Intervention, Comparison, Outcome and Time horizon (PICOT) format.[10] The catalogue also provides the data elements to be included in the table of data sources recommended by the HARmonized Protocol template to Enhance Reproducibility (HARPER).[11] The assessment of relevance is supported by the availability of the following variables:

- Setting: county(-ies) (C1.5), region(s) (C1.5.1), type of data source (C5.1 and C5.1.1), care setting (C1.14).
- Population: total and active population size (C7.1), percentage of the population covered by the data source in the catchment areas (C1.11.2) and 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), 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), biomarker data (C6.26).
- Outcomes: availability of data on hospital admission or discharge (C6.10), ICU admission (C6.10.1), death 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).

Links to the EU PAS Register also allow to identify studies that have been performed with the same data source, allowing an evaluation of the analyses that can be performed.

[9] Gatto, N. M., Campbell, U. B., Rubinstein, E., Jaksa, A., Mattox, P., Mo, J., & Reynolds, R. F. (2022). The Structured Process to Identify Fit-For-Purpose Data: A Data Feasibility Assessment Framework. Clin Pharmacol Ther. 2022;111(1), 122–134. https://doi.org/10.1002/cpt.2466

[10] Brown, P., Brunnhuber, K., Chalkidou, K., Chalmers, I., Clarke, M., Fenton, M., Forbes, C., Glanville, J., Hicks, N. J., Moody, J., Twaddle, S., Timimi, H., & Young, P. How to formulate research recommendations. BMJ. 2006;333(7572), 804–806. https://doi.org/10.1136/bmj.38987.492014.94

[11] Wang S, Pottegard A, Crown W et al. HARmonized Protocol Template to Enhance Reproducibility (HARPER) of Hypothesis Evaluating Real-World Evidence Studies on Treatment Effects: A Good Practices Report of a Joint ISPE/ISPOR Task Force. Pharmacol Drug Saf. 2022;

In order to provide adequate evidence, appropriate epidemiological and statistical methods must be applied to the study design and the analysis and interpretation of data generated from a real-world data source. These methods are not addressed by the metadata catalogue but are described in other guidance, e.g. the ENCePP Guide on Methodological Standards in Pharmacoepidemiology, 10th Rev. (2022).

4.3. Use cases

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 catalogue to identify relevant data sources fulfilling the specifications of the research question or, if there is a prior interest in using a specific data source, to access the record for this data source and consult the available information. The search may initially use the data elements useful to assess pre-defined PICOT criteria (see section 4.1) in order to identify possibly suitable data sources.

2. In a second step, the investigator accesses the record of each potential data source and screens more detailed 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 data sources (C2.3) to determine whether they are accessible, as well as the conditions related to this 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 3.1). 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 link providing access to the EU PAS Register of studies that have been performed with the same data source and addressed research questions similar (as to the topic or study design) to the current one. After selecting studies with a similar topic or design as for the planned study, 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 of the current 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;

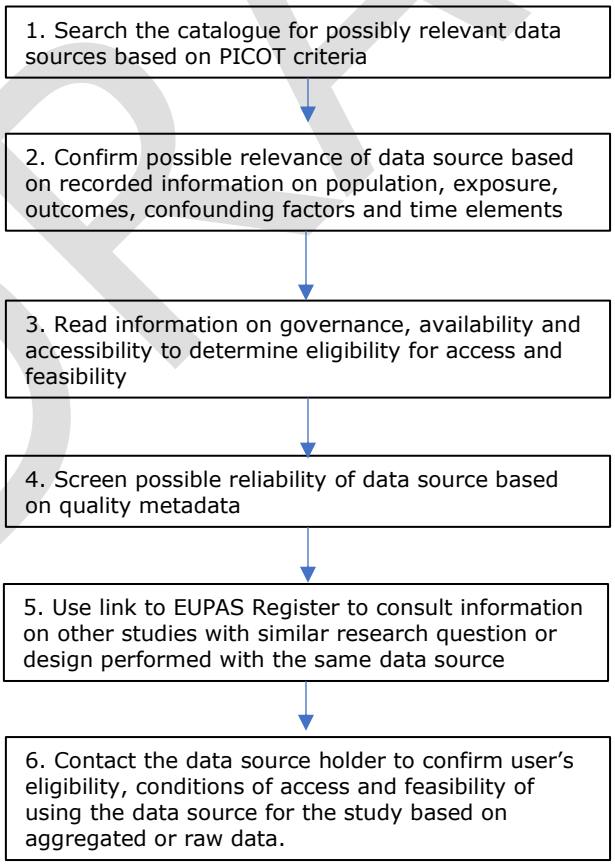
- check in the study protocol or study report (if available) the algorithms that have been used to identify diseases or outcomes of interest and their severity (for example persons with a rare disease if applicable) and which prompts, and contents were used in such algorithm(s);
- learn about the strengths and weaknesses of the data source encountered in the study conduct; in case a limitation is acknowledged 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-review journals and comments made on study limitations.

If there are remaining uncertainties as regards the reliability and relevance of the data source for the specific study, the investigators of similar studies in terms of topic or study design can be contacted to gather additional information.

If past studies using the same data source cannot be found, it may be preferable to investigate the information available for another relevant data source.

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.

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



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 user may verify if the data source has been registered in the catalogue.[12] Depending on the information that is already available in the protocol, that is missing or that needs verification, the user accesses different sections of the catalogue. In order to verify the representativeness of the study population described in the report, the user may verify 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 and covariates collected in the data source and identify those that have not been proposed to be extracted but could be useful to include for 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 EU PAS register 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

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 in the study to interpret its 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 contains results and generally quantitative information on the characteristics of the study population originating from the data source. The assessor may therefore identify, and investigate if needed, differences between the information provided in the study report and in the metadata catalogue.

Some verification may be applied to the description of the study population, the sample size originating from the data source included in the report, the nature and categories of variables included in the analysis and the coding system provided. 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 he needs to describe the data source(s) proposed to be used or used in the study. The information on the data source he finds in

[12] Except of specific circumstances, there is no legal obligation to register a data source into the metadata catalogue. It is however expected that data source holders will register their data source, and update the record, whenever it will be used for public health or regulatory purpose, as absence of public information on the data source 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 record updated, in the metadata catalogue as part of the agreement.

other publications or other documentation is heterogeneous, and a comparison between the characteristics of several databases used in to be used or used in the study is difficult to perform.

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

If a data source is not registered in the metadata catalogue, this registration can be made simultaneously to the writing of the protocol or report. If access to the data source has been obtained through a contractual agreement, this agreement could provide for the registration of the data source, or updating of its record, before the study commences.

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 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 metadata catalogue provides:

- a harmonised description of the characteristics of each data source that allow to compare differences, e.g., in age groups 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 the study is implemented in a CDM, the analyst may find in the catalogue the specifications of the ETL procedure from the data source to the CDM. Irrespective of whether the data holder has converted to the CDM the entire data source, or only an extraction thereof, this information supports the programmer in developing the study script. Using the link to the EU PAS Register, the analyst can also access detailed information on the studies performed with the same data source and registered in the EU PAS Register, and select the studies that investigated the same topic and/or 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.

At the end of the analysis, the analyst should also record the script of the analysis in a public repository and provide the link in the EU PAS Register, thus enabling transparency and quality control and facilitating reproducibility.

User guides

5. Description of the metadata list and definitions

Following several prioritisation exercises and consultations with stakeholders, the below metadata elements, which have been published by HMA/EMA in the List of metadata for Real World Data catalogues[13], have been selected for a first iteration of this process. The data elements aim to describe the data sources, with a view of facilitating the choice of data source for the specific use cases listed in chapter 4.

5.1. Metadata characterising the 'data source'

A data source is described by the data holder that sustains the collection of records in the data source, the underlying population that can potentially contribute records to a data source, and the prompt that leads to creation of a record in the data source.

5.1.1. Data source – Administrative details

5.1.1.1. Name of the data source (C1.2)

The name of the data source, as used in European projects, must be provided. If the database is widely known by several names, these can be provided in this field, separated by a '/' sign. Where the name of the data source is in a local language, the English translation should also be provided, using parentheses.

5.1.1.2. Data source acronym (C1.3)

Where the data source is generally known under a specific acronym, this should be provided.

5.1.1.3. Data holder (C4.1)

The data holder must be provided, selecting one of the existing entries from the 'institutions' available look-up. For the purpose of this catalogue, a data holder is defined as an organisation that sustains the collection of records in a data source.

5.1.1.4. Data source contact name (M1.3)

A contact name should be provided for queries related to the data source. The contact details would be visible in the publicly available catalogue.

5.1.1.5. Data source contact email (M1.6)

An e-mail contact should be made available for queries related to the data source. This information will be visible in the publicly available catalogue.

5.1.1.6. Data source countries (C1.5)

The country where the data originate should be selected from the list of country codes (ISO 3166-1).
Where needed, multiple countries can be selected.

[13] HMA/EMA. List of metadata for Real World Data catalogues (2022).

5.1.1.7. Data source language(s) (C6.2)

The data source language should be specified using the appropriate ISO 639 code.

5.1.1.8. Data source regions (C1.5.1)

The geographical regions that the data source covers should be provided using regions codes (ISO 3166-2). Multiple regions can be selected where required.

5.1.1.9. Date when the data source was first established (C4.5)

The date when the data source was first set-up. This date can be different from the 'first collection date' (C1.12).

5.1.1.10. First collection date (C1.12)

The date when data started to be collected or extracted.
It is expected that this information is populated only once, when the data source is first described (with the exception of error corrections from the initial submission).

5.1.1.11. Last collection date (C1.13)

Where applicable, the date when the data collection ended. This information should only be provided for data sources where the data collection has stopped permanently.

5.1.1.12. Data source website (C11.1)

Where such an information is available, a link to the dedicated webpage describing the data source should be provided. The information listed would capture information such as data content, release notes etc.

5.1.1.13. Data source publications (C11.2)

A list of peer-reviewed papers or documents describing the data source (validation, data elements, representativity) or its use for pharmacoepidemiologic research

5.1.1.14. Data source qualification (C3.1, C3.1.1)

If the data source has successfully undergone a formal qualification process (e.g., from the EMA, or ISO or other certifications), this should be described.

5.1.1.15. Main financial support (C4.6)

The source of finance for the data source in the last three years should be specified using the below categories:
- Funding by own institution
- National, regional, or municipal public funding
- European public funding
- Funding from industry or contract research organisation
- Funding from public-private partnership
- Funds from patients organisations, charity or foundation

5.1.1.16. Data source type (C5.1, C5.1.1)

Data source may fit in one of more of the following categories:

Administrative
- population registry
- death registry

- registration with healthcare system
- exemptions from co-payment
- diagnostic tests or procedures reimbursement
- administrative healthcare claims

Primary care
- primary care medical records
- pharmacy dispensation records

Secondary care
- hospital discharge records
- hospital inpatient records
- hospital outpatient visit records
- emergency care discharge records
- specialist ambulatory care records

Registries
- birth registry
- induced terminations registry
- congenital anomaly registry
- cancer registry
- disease registry
- vaccination registry
- drug registry

Other
- biobank
- spontaneous reporting of adverse drug reactions

If none of the listed categories apply to the data source, it's type should be described in the available free text field (C5.1.1).

5.1.1.17. Care setting for data source (C1.14)

Where the data source describes a care setting, this can be further characterised as:

- primary care – GP, community pharmacist level
- primary care – specialist level (e.g. paediatricians)
- secondary care – specialist level (ambulatory)
- hospital inpatient care
- hospital outpatient care

5.1.2. Data source – Data elements collected

5.1.2.1. Data source characteristics

To characterise the content of the data source the specific data elements should be selected as applicable.

Value (yes/no)	Description
Specific diseases (C1.10)	Data source collects information with a focus on specific diseases. This might be a patient registry or other similar initiatives. Where this is applicable
Hospital admission discharge (C6.10)	Information on hospital admission and/or hospital discharge is available in the data source.
ICU admission (C6.10.1)	Information on intensive care admission available.

Value (yes/no)	Description
Cause of death (C6.11)	The cause of death is captured, either as structured or unstructured information
Rare diseases (C6.12)	The data source captures rare diseases, where the prevalence of the condition in the EU is less than 5 in 10,000
Prescriptions and/or dispensing (C6.13)	The data source contains information on prescriptions or dispensing of medicines
ATMP (C6.16)	A medicinal product for human use that is either a gene therapy medicinal product, a somatic cell therapy product or a tissue engineered products as defined in Regulation (EC) No 1394/2007 [Reg (EC) No 1394/2007 Art 1(1)].
Contraception (C6.17)	Any information on use of any type of contraception (oral, injectable, devices etc.)
Indication for use (C6.18)	Therapeutic indication for the use of medicinal product
Administration of vaccines (C6.19)	Information on any vaccines administered
Administration of other injectables (C6.19.1)	Information on medicinal products administered via an injectable route (e.g.: solutions for perfusion, solutions for injection)
Medical devices (C6.20)	Where data source captures information on medicinal devices (e.g.: pens, syringes, inhalers)
Procedures (C6.21)	Medical procedures (e.g. surgical interventions, tests)
Clinical measurements (C6.23)	Information on clinical measurements (e.g.: BMI, blood pressure, height)
Healthcare provider (C6.24.1)	Data on individual health professionals or a health facility organization licensed to provide health care diagnosis and treatment services including medication, surgery and medical devices
Genetic data (C6.25)	Data related to genotyping, genome sequencing
Biomarker data (C6.26)	The term "biomarker" refers to a broad subcategory of medical signs – that is, objective indications of medical state observed from outside the patient – which can be measured accurately and reproducibly. For example, haematological assays, infectious disease markers or metabolomic biomarkers.
Patient-generated data (C6.27)	Health-related data created, recorded, or gathered by or from patients (or family members or other caregivers) to help address a health concern
Units of healthcare utilisation (C6.29)	Quantification of the use of services for the purpose of preventing or curing health problems (e.g.: number of visits to GP per year, number of hospital days)
Unique identifiers for persons (C6.4)	Where applicable, if patients are uniquely identified
Diagnostic codes (C6.9)	If diagnostic codes are captured; further information will be captured in section 5.1.5.11

Value (yes/no)	Description
Pregnancy and neonates (C1.9)	Where data on pregnant women and neonates (under 28 days of age), infant, and child development

515

5.1.2.2. Disease information collected (C1.10.1)

The disease or diseases for which information is collected should be specified in this field, using MedDRA terminology.

5.1.2.3. Population age groups (C1.8)

The information on the following age groups are being captured separately:

- newborn infants (0 to 27 days),
- infants and toddlers (28 days to 23 months),
- children (2 to 11 years),
- adolescents (12 to 17 years),
- adults (18 to 45 years),
- adults (46 to 64 years),
- adults (65 to 74 years),
- adults (75 to 84 years),
- adults (85 years and over)

5.1.2.4. Family linkage (C6.6, C6.6.1)

Where family linkage is made available in the data source this should be characterised using one or more of the following values: household (where the information on individuals sharing a household can be identified), mother-child, father-child, sibling.

If family linkage is not available, it should be specified if familial linkage can be created on an ad-hoc basis (C6.6.1).

5.1.2.5. Sociodemographic information collected (C6.7)

Where one or more of the following specific sociodemographic information are captured by the data source, these should be selected:

- age
- gender
- ethnicity
- country of origin
- indicator of socioeconomic status
- marital status
- education level
- type of residency
- living in rural area
- health area
- deprivation index

5.1.2.6. Lifestyle factors (C6.8)

Where the data source captures this information, one or more of the following lifestyle factors can be selected:

- tobacco use
- alcohol use
- amount of physical exercise
- diet

5.1.2.7. Population covered by the data source (C1.11.2)

The percentage of the population covered by the data source in the catchment area should be specified.

5.1.2.8. Population not covered by the data source (C1.11.1)

The description of the population covered by the data source in the catchment area whose data are not collected, where applicable (e.g.: people who are registered only for private care).

5.1.3. Data source - Quantitative descriptors

This section aims to collect a limited amount of data elements that look at the quantitative details of the data source. In future iterations of the data source catalogue this section can be further expanded as found useful.

5.1.3.1. Population size (C7.1)

The total number of unique individuals with records captured in the data source.

5.1.3.2. Population size by age (C7.3)

Where this information can be extracted, the number of unique individuals split by age groups should be captured.

5.1.3.3. Active population size (C7.1.1)

An active population for administrative healthcare data refers to the collection of patients for which there is an active record in the practice, i.e. the record was created and not closed (because patient moved or died).

5.1.3.4. Active population size by age (C7.3.1)

Where this information can be extracted, the number of unique active individuals split by age groups should be captured.

An active population for administrative healthcare data refers to the collection of patients for which there is an active record in the practice, i.e. the record was created and not closed (because patient moved or died).

5.1.3.5. Median time (B6.3)

The median time, in years, between first and last available records for unique individuals captured in the data source.

5.1.3.6. Median time active (B6.3.1)

The median time, in years, between first and last available records for unique ***active*** individuals (alive and currently registered) captured in the data source.

An active population for administrative healthcare data refers to the collection of patients for which there is an active record in the practice, i.e. the record was created and not closed (because patient moved or died).

5.1.4. Data source – Data flows and management

5.1.4.1. Governance details (C2.3)

Description of the documents or links to webpages that describe the overall governance, processes and procedures for data capture and management, data access, data quality check and validation results, utilisation for research purposes.

5.1.4.2. Follow-up (C2.13, C2.13.1, C2.7)

If further follow-up would be needed, the availability of below access options should be specified:
Accessing biospecimens: if this is possible (C.2.13) then also the biospecimen access conditions should be described (or a reference source can be added) (C2.13.1)
Contacting patients or practitioners (C2.7)

5.1.4.3. The process of collection and recording (C4.3)

The process or manner in which recording of data in the data source occurs should be described; this could include the tools used, such as surveys, or a description of the system that the data holder uses to gather data and store it the data source.

5.1.4.4. Record creation (C5.2)

The event triggering the creation of a record in the data source should be described (e.g.: hospital discharge, specialist encounter, medicinal product dispensing).
This refers in general to the creation of a record in the data source (and not to the registration of a person, see below).

5.1.4.5. Registration of a person (C1.6, C1.6.1)

The event triggering registration of a person in the data source should be selected from the following available values:
- Birth
- Immigration
- Residency obtained
- Start of insurance coverage
- Disease diagnosis
- Start of treatment
- Practice registration

Where none of the above values apply, the triggering event for a person to be registered in the data source should be described separately (C1.6.1).

5.1.4.6. De-registration of a person (C1.7, C1.7.1)

The event triggering de-registration of a person in the data source: The event triggering de-registration of a person in the data source should be selected from the following available values:
- Death
- Emigration
- End of insurance coverage
- Practice deregistration
- Loss to follow up
- End of treatment

Where none of the above values apply, the triggering event for a person to be de-registered in the data source should be described separately (C1.7.1).

5.1.4.7. Linkage (B5.2, B5.2.1, B5.3, B4.1)

Where the data source is created by the linkage of other data sources, the elements of the linkage should be briefly captured as follows:

- The linkage strategy (B5.2): whether the linkage is deterministic, probabilistic or a combination of the two.
- The linkage variable used (B5.2.1) (e.g.: patient ID, date of birth etc.)
- The completeness of the linkage (B5.3), described as a percentage along with the reference used
- Names of the linked data sources (B4.1). Where these data sources are available in the data source catalogue, these should be cross-referenced.

5.1.4.8. Data management specifications (C2.7, C8.5, C8.5.1, C2.9):

The following information related to data management specifications should be selected, as applicable to the data source:
- Whether or not the data source allows data validation (e.g.: access to original medical charts)
- If the records are preserved indefinitely (C8.5)
- Where the records are not indefinitely preserved, the number of years for which the records are kept should be specified (C8.5.1)
- Whether approval is needed for publishing results of a study using its data (C2.9)

5.1.4.9. Informed consent for use of data for research (C2.5, C2.5.1)

The need for informed consent in the context of research should be captured here. The type of informed consent could be categorised as:
- Not required
- Required for general use of the data source
- Required for all studies run on the data source
- Required for intervention studies only
- Waiver

Where the informed consent does not fit in the above categories, the value 'Other' can be used and further details should be provided (C2.5.1).

5.1.4.10. Data source refresh (C8.2)

Where the data source is refreshed on fixed dates around the year, this should be provided by selecting the month as applicable (e.g.: every June). The field can be repeated where the refresh happens more often than once a year (e.g.: every May and November).

5.1.4.11. Data source last refresh (C8.3)

Where the data source is refreshed at particular times throughout the year, the date when the last refresh of the data source occurred should be provided.

5.1.4.12. CDM (Common Data Model) specifications (D1.2.1.1, D1.2, D1.2.1, D1.4, D1.7)

The following data elements should be captured for data sources being transformed using a Common Data Model (CDM), (D1.2.1.1) as follows:
- The CDM name should be selected from the existing predefined list as follows: OMOP, ConcepTION, Nordic, Sentinel, PCORnet, VSD, i2b2, CDISC SDTM, PEDSnet (D1.2).

Where the common data model used is not listed in the values offered, further details should be provided (D1.2.1)
- The CDM website reference should be provided where available (D1.4)
- The CDM release frequency, in number of months, should be provided (D1.7)

5.1.4.13. Data source ETL to a CDM (B7.1, B7.5, B7.3, B7.4)

Where applicable, further information on the data transformation (ETL) to a common data model (CDM) should be provided as follows:
- The status of the transformation (ETL) of the data source should be described as either: planned, in progress or completed.
- The frequency in months of the ETL frequency
- The version(s) of CDM(s) to which the data source has been ETL-d
- Data source ETL specifications: documents describing the mapping of the data source to the CDM (including codes and scripts to transform original data to CDM)

5.1.5. Data source – Vocabularies and standardised dictionaries

5.1.5.1. Medicinal product information available

The type of information captured with regards to the medicinal product should be selected from the values described in the table below.

Vocabulary	Description
Brand name	Specific name or trademark under which a medicine is sold
Batch number	The designation printed on the medicine label that allows the history of its production to be traced
Formulation	Pharmaceutical form of the medicinal product (e.g.: tablets, capsules etc.)
Strength	The amount of active ingredient contained in the medicinal product.
Package size	Number of individual formulations contained in a package (e.g.: 30 tablets per package)
Dose	The medicinal product dose prescribed or administered to the patient
Dosage regime	The schedule of doses of a medicinal product per unit of time (e.g.: every 6 hours)
Route of administration	The manner in which a medicinal product enters the body (e.g.: oral, intravenous)

5.1.5.2. Medicinal product vocabulary used (C6.15.1)

Vocabulary	Description
Art 57	Authorised medicines information in EU and EEA. Further reference here.
IFA GmbH	Informationsstelle für Arzneispezialitäten. Further reference here.

Vocabulary	Description
EDQM	European Directorate for the Quality of Medicines. Further reference here.
SPN	Standard Product Nomenclature. Further reference here.
MTHSPL	FDA Structured Product labelling. Further reference here.

Where the medicinal product information is not coded (i.e.: provided as free text) this should be marked accordingly.

If other dictionaries than the listed ones are used, the value 'Other' should be used.

5.1.5.3. Cause of death vocabulary

Vocabulary	Description
ICPC	International Classification of Primary Care. Further reference here.
ICD9	International Classification of Diseases, 9th revision. External reference here.
ICD10	International Classification of Diseases, 10th revision. External reference here.
ICD1	International Classification of Diseases, 1st version. External reference here.
Read	External reference here.
SNOMED	Systematized Nomenclature of Medicine. Further reference here.
SNOMED CT	Systemized Nomenclature of Medicine – Clinical Terms. Further reference here.
MedDRA	Medical Dictionary for Regulatory Activities. Further reference here.
OPCS	Classification of Interventions and Procedures. Further reference here.

Where the cause of death is not coded (i.e.: provided as free text) this should be marked accordingly.

If other dictionaries than the listed ones are used, the value 'Other' should be used.

5.1.5.4. Quality of life measurements (C6.28, C6.28.1)

Vocabulary	Description
AQoL-8D	Assessment of Quality of Life 8-Dimension. Further reference here.
QOLS	Quality of Life Scale. Further reference here
MQOL	McGill Quality of Life Questionnaire. Further reference here.
MQOL-E	The McGill Quality of Life Questionnaire – Expanded. Further reference here.
HRQOL	Health-related quality of life. Further reference here.
WHOQOL	World Health Organization External Measuring Quality of Life. Further reference here.
EQ5D	Standardised measure of health-related quality of life developed by the EuroQol Group. EQ-5D assesses health status in terms of five dimensions of health. Further reference here.
15D	The 15D is a generic 15-dimensional self-administered instrument for measuring HRQoL (Health-related quality of life). Further reference here.
SF-36	The Short Form (36) Health Survey is a 36-item, patient-reported survey of patient health. Further reference here.
SF-6D	An abbreviated variant of SF-36 commonly used in health economics as a variable in the quality-adjusted life year calculation to determine the cost-effectiveness of a health treatment. External reference here.
HUI	Health Utilities Index. Further reference here.

Where the quality of life is captured but not coded (i.e.: provided as free text) this should be marked accordingly.

If other dictionaries than the listed ones are used, the value 'Other' should be used. In this case, the name of the 'quality of life' scale used should be provided in the free text field accordingly (C6.28.1).

5.1.5.5. Prescription vocabulary (C6.13.1)

Vocabulary	Description
ATC	Anatomical Therapeutic Chemical code. Further reference here.
RxNorm	A normalized naming system for generic and branded drugs. Further reference here.
EphMRA	Anatomical Classification of Pharmaceutical Products maintained by EphMRA. Further reference here.

Vocabulary	Description
DrugBank	DrugBank Online is a comprehensive, free-to-access, online database containing information on drugs and drug targets. Further reference here.

Where the prescription is captured but not coded (i.e.: provided as free text) this should be marked accordingly.

If other dictionaries than the listed ones are used, the value 'Other' should be used.

5.1.5.6. Dispensing vocabulary (C6.14.1)

The dictionary used to code the dispensing information captured in the data source should be selected.

For further details on the available values, see section 5.1.5.5.

Where the dispensing information is captured but not coded (i.e.: provided as free text) this should be marked accordingly.

If other dictionaries than the listed ones are used, the value 'Other' should be used

5.1.5.7. Indication vocabulary (C6.18.1, C6.18.2)

The dictionary used to code the therapeutic indication captured in the data source should be selected.

For further details on the available values, see section 5.1.5.3.

Where the therapeutic indication is captured but not coded (i.e.: provided as free text) this should be marked accordingly.

If other dictionaries than the listed ones are used, the value 'Other' should be used. In this case, the name of the 'quality of life' scale used should be provided in the free text field accordingly (C6.18.2)

5.1.5.8. Procedures vocabulary (C6.22)

The dictionary used to code the procedures captured in the data source should be selected.

For further details on the available values, see section 5.1.5.3.

Where the procedure is captured but not coded (i.e.: provided as free text) this should be marked accordingly.

5.1.5.9. Genetic data vocabulary (C6.25.1)

Vocabulary	Description
OGG	A biological ontology in the area of genes and genomes. Further reference here.
GO	Gene Ontology. Further reference here.
EGO	Eukaryotic Gene Orthologues. Further reference here.
SOPHARM	Suggested Ontology for Pharmacogenomics. Integrates OBO ontologies and formalizes specific gene variants. Further reference here.

Vocabulary	Description
PHARE	PHArmacogenomic RElationships Ontology. Further reference here.

Where the genetic data is captured but not coded (i.e.: provided as free text) this should be marked accordingly.

If other dictionaries than the listed ones are used, the value 'Other' should be used

5.1.5.10. Biomarker data vocabulary (C6.26.1)

Vocabulary	Description
SMASH	Semantic Mining of Activity, Social, and Health data. Further reference here.
FOBI	Food-Biomarker Ontology. Further reference here.

Where the biomarker data is captured but not coded (i.e.: provided as free text) this should be marked accordingly.

If other dictionaries than the listed ones are used, the value 'Other' should be used.

5.1.5.11. Diagnosis/ medical event vocabulary (C6.9.1)

The dictionary used to code the diagnosis, or any other medical event captured in the data source should be selected.

For further details on the available values, see section 5.1.5.4.

Where the diagnosis or medical event is captured but not coded (i.e.: provided as free text) this should be marked accordingly.

If other dictionaries than the listed ones are used, the value 'Other' should be used.

6. Registering a data source in the Data source catalogue

A Data holder would be able to request to register on a voluntary basis a data source in the Data source catalogue via a dedicated webform to be made available in the second half of 2023. An e-mail address supporting this process is available: metadata@ema.europa.eu.

Additionally, EMA is proactively contacting data holders requesting the addition of the metadata information in the catalogue, looking to current data sources registered in ENCePP Resources Database.

7. Maintenance of information in the Data source catalogue

It is important that the metadata information is kept up-to-date; this refresh of information is expected to be run on a yearly basis or more often for particular data sources if found necessary.

The data holder will be provided with the technical means to update the information provided directly, via a dedicated webform. This will be made available in the second half of 2023.

References

Brown, P., Brunnhuber, K., Chalkidou, K., Chalmers, I., Clarke, M., Fenton, M., Forbes, C., Glanville, J., Hicks, N. J., Moody, J., Twaddle, S., Timimi, H., & Young, P. How to formulate research recommendations. BMJ. 2006;333(7572), 804–806. https://doi.org/10.1136/bmj.38987.492014.94

ENCePP Guide on Methodological Standards in Pharmacoepidemiology, 10th Rev. (2022). Chapter 12.1 General principles of quality management

ENCePP Guide on Methodological Standards in Pharmacoepidemiology, 10th Rev. (2022). Chapter 12.2. Data Quality Frameworks.

FAIR Principles. https://www.go-fair.org/fair-principles/

HMA/EMA. List of metadata for Real World Data catalogues (2022).

Gatto, N. M., Campbell, U. B., Rubinstein, E., Jaksa, A., Mattox, P., Mo, J., & Reynolds, R. F. (2022). The Structured Process to Identify Fit-For-Purpose Data: A Data Feasibility Assessment Framework. Clin Pharmacol Ther. 2022;111(1), 122–134. https://doi.org/10.1002/cpt.2466

MINERVA: Strengthening Use of Real-World Data in Medicines Development: Metadata for Data Discoverability and Study Replicability (2022). EUPAS39322

TEHDAS. European Health Data Space Data Quality Framework (2022).

Wang S, Pottegard A, Crown W et al. HARmonized Protocol Template to Enhance Reproducibility (HARPER) of Hypothesis Evaluating Real-World Evidence Studies on Treatment Effects: A Good Practices Report of a Joint ISPE/ISPOR Task Force. Pharmacol Drug Saf. 2022;

Wang S., Schneeweiss S. Assessing and Interpreting Real-World Evidence Studies: Introductory Points for New Reviewers. Clin Pharmacol Ther. 2022;111(1):145-149.