EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

4 November 2024
EMA/503781/2024
Committee for Medicinal Products for Human Use (CHMP)

Data Quality Framework for EU medicines regulation: application to Real-World Data

Draft

Draft agreed by Methodology Working Party (MWP)	September 2024
Adopted by CHMP for release for consultation	4 November 2024
Start of public consultation	29 November 2024
End of consultation (deadline for comments)	31 January 2025

Comments should be provided using this EUSurvey form. For any technical issues, please contact the EUSurvey Support.

Keywords	***Data quality, framework, real-world data, real-world evidence, use of data, primary, metadata, reliability, extensiveness, coherence, timeliness, relevance, maturity models, validation***

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

Data Quality Framework for EU medicines regulation: application to Real-World Data

Table of contents

Executive summary 3

1. Background: Real-World Data and Data Quality 4

1.1. Definition of Real-World Data 4

1.2. Distinctive traits of RWD 4

1.3. RWD use-based quality control 5

1.4. Impact of secondary use of RWD on data quality 5

1.4.1. Impact on Reliability 6

1.4.2. Impact on Extensiveness and Representativeness 6

1.4.3. Impact on Coherence 7

1.5. Responsibility for DQ in RWD 7

2. Application of the EMRN DQF to RWD 7

2.1. Purpose of this document 7

2.2. Scope of the RW-DQF 8

2.3. Structure of the RW-DQF 8

2.3.1. Understanding relevance 9

3. Guidelines for the characterisation of systems and processes underpinning data 10

3.1. Systems and process characterisation checklist and maturity model 10

3.2. General considerations on the characterisation of systems and processes 17

4. Data quality metrics for RWD 18

4.1. Framework for the categorisation and identification of metrics 18

4.2. Metrics for DQ assessments 20

4.3. Considerations for the implementation of RWD DQ metrics 23

4.3.1. Different roles of metrics 23

4.3.2. Additional considerations on level of application and maturity for metric assessments 24

5. Guidelines to assess quality in relation to a specific research question 24

5.1. General principles for assessment of data quality in relation to a research question 24

5.2. Framework for detailed fitness-for-use assessment 27

5.3. Illustrative example for detailed fitness-for-use assessment 28

5.4. Toward a generalisation of question-specific aspects 31

5.5. Providing supporting information for RWD in regulatory submissions 32

6. Concluding remarks 32

7. References 32

Definitions 34

Glossary 34

Executive summary

This document describes the Real-World Data (RWD) specific recommendations as derived from the Data Quality Framework (DQF) for EU Medicines regulation endorsed by the Committee for Medicinal Products for Human Use (CHMP) [1] (hereafter referred to as EMRN DQF). The EMRN DQF sets out the principles, concepts, and definitions as intended to be applied widely across datasets used in medicine regulatory use cases. It also provides examples and in-depth clarifications on the developed framework elements for characterising, assessing, and assuring DQ in the regulatory context. It is therefore recommended to use the EMRN DQF as a companion document when reading this chapter.

The application of the EMRN DQF to RWD (hereafter referred to as RW-DQF) sets out the specificities of RWD and enable regulators to evaluate the quality of data underpinning Real-World Evidence (RWE) as used in the regulatory assessment. It also provides guidance on the relevance assessment of such data to a research question based on DQ metrics and evidence of systems and processes underpinning data. These parts provide actionable and focused recommendations for assessing DQ of RWD, with the goal of improving the usefulness of RWE for regulatory purposes. The RW-DQF is intended for the use of stakeholders involved in regulatory processes, primarily aimed at members of the European Medicines Regulatory Network (EMRN), but also other actors involved in this process, such as the Data Analysis and Real-World Interrogation Network (DARWIN EU®), pharmaceutical industry, academia, contract research organisations, and data holders.

With a view of maintaining the consistency with parallel activities ongoing in the context of European Health Data Space (EHDS), the document was developed in close collaboration with the Towards European Health Data Space (TEHDAS) and QUANTUM (The Health Data Quality label) projects. These initiatives aim to address the wider use of health data, whereas the RW-DQF specifically focuses on the challenges faced when using this data within the medicine regulation assessment.

The topics addressed in this document are: an introduction to RWD key considerations on quality (Chapters 3 & 4), practical recommendations on characterisation of the systems and processes that underpin data (Chapter 5), a set of metrics to assess data quality (DQ) dimensions (Chapter 6), and a guideline on how to assess DQ in relation to a research question via the use of a framework and an illustrative example (Chapter 7) (Figure 1).

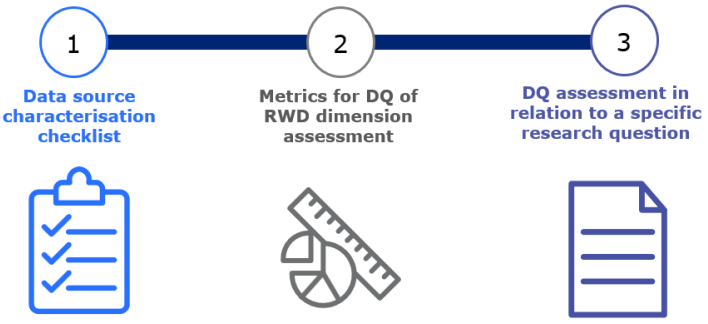

Figure 1- Representation of the key points of the DQF for EU medicines regulation: application to RWD.

1. Background: Real-World Data and Data Quality

1.1. Definition of Real-World Data

In the context of RWE studies, Real-World Data (RWD) are data that describe patient characteristics (including treatment utilisation and outcomes) in routine clinical practice [2, 3, 4]. In broader terms, RWD represent data captured in routine care which are not collected in a clinical trial and are relevant to the subject (e.g., age, sex, ethnicity etc.), the disease, the treatment, interactions with the healthcare system, as well as social and environmental factors influencing health status. RWD may originate from primary data collection (primary use of data), i.e., data collected specifically for the study in question, or secondary use of data initially recorded in the context of different primary purposes (such as the clinical management of patients or for administrative reasons). The secondary use of data, driven by specific research objectives, can involve a single or multiple RWD sources (i.e., multi-database studies). For example, the assessment of treatment pathway disparities in a given region and for a given indication could entail combining and analysing RWD from multiple sources.

Through the analysis of RWD, Real-World Evidence (RWE) is generated to answer research questions. The main areas where RWD analyses can aid medicines regulatory use cases across a medicinal product's lifecycle are to [2]: 1) support planning and validity of applicant studies (e.g., comparison of a study population with patients from the real-world setting to ensure representativeness of the clinical study, patient recruitment), 2) understand the clinical context (e.g., disease epidemiology, disease prevalence/incidence, description of drug utilisation patterns including switching and off-label use), 3) investigate associations and impact (e.g., medicines post-marketing surveillance, assessment of effectiveness of risk minimisation measures) [5].

1.2. Distinctive traits of RWD

RWD have several distinctive traits, including the variety of sources capturing them, their structure, format, variables collected, terminologies used, processes related to data collection or data recording. RWD can be leveraged for their primary intended purpose or for secondary use [6, 7]. Types of RWD sources include, for example, electronic healthcare databases (e.g., from primary care, specialist care, and hospital care settings, claims databases), longitudinal drug prescription, dispensing or other drug utilisation data or patient registries. The latter is defined as a system which records uniform data (clinical and other) to identify specified outcomes for a population defined by a particular disease, condition or exposure [13]. Data captured in a registry can involve data collection with the primary purpose of generating RWE or secondary use of data. In addition, RWD can be collected via compassionate use programmes on products in development for patients who cannot enter clinical trials, as they can facilitate the understanding of the best conditions of medication use and which patients can be benefited the most [22].

The RWD lifecycle may involve several manipulations with or without changing hands between organisations, e.g., aggregation, transformation, cleaning, metadata creation, publishing (see Figure *1*). In cases of secondary use of RWD, the data may not be entirely fit for the study at hand, as their primary purpose might differ.

The metadata of RWD for RWE generation is typically included in data catalogues for dissemination. These "research-ready" data are usually optimised for quality and usability as much as possible without referring to a specific research question.

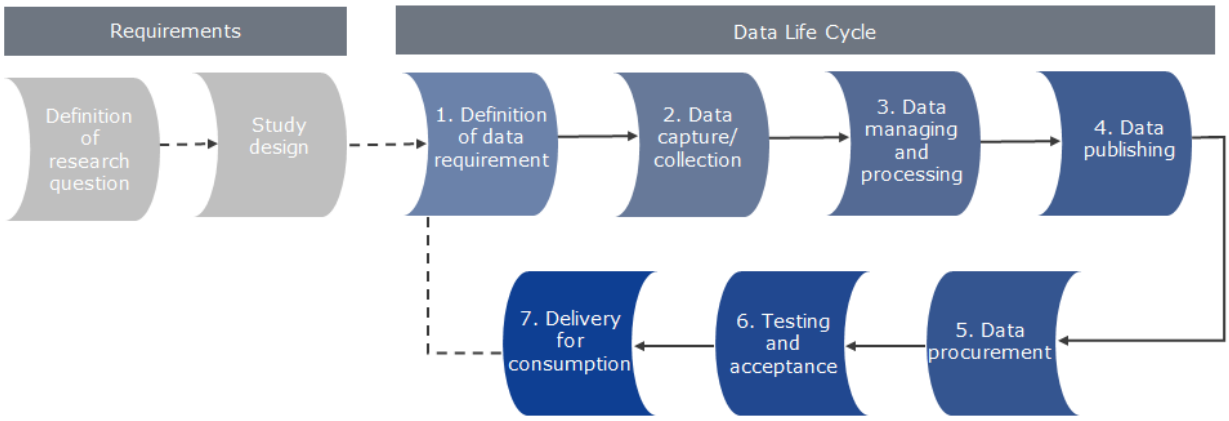

Figure 1 Data life cycle for secondary use of RWD. The definition of research question and the study design are required steps preceding the definition of data requirement (Step 1).

1.3. RWD use-based quality control

The primary purpose of recording RWD is the provision of health services to assess, maintain or restore the state of health of the person that the data belong to, and in some cases, RWE generation. The quality of RWD capture should be under the control of a Quality Management System (QMS) as explained in more detail in the EMRN Data Quality Framework (DQF) [1].

When RWD are leveraged for RWE (either when that is their primary or secondary use) and for other secondary purposes, they must be considered as uncontrolled, i.e., not produced through a defined process with feedback loops to detect and correct errors and prevent their future occurrence. A QMS such as the ISO 9000 family, Good Clinical Practices, Good Laboratory Practices or Good Manufacturing Practice can only be expected to assure quality with respect to the primary use of RWD.

1.4. Impact of secondary use of RWD on data quality

Secondary use of RWD has a significant impact on data quality (DQ) for the reasons outlined here:

Decoupling of data collection and purpose: Data for secondary use are, by definition, originally recorded for other purposes than generation of RWE, and the secondary purpose consists of many different individual research questions not always known at the time of capturing and publishing research-ready data. Fitness-for-use depends on a defined research question, and therefore cannot be controlled or imposed to a source at the time of data recording. In other words, DQ can be measured at source, but cannot be fully assessed for adequacy at the time of data collection or data recording.

Anonymisation and pseudonymisation: Sensitive patient-level healthcare data need to be anonymised or pseudonymised to facilitate data protection, publication, and re-use by different researchers. This process sometimes involves masking, replacing or removing information. Consequently, quality and operational aspects of quality control may be negatively affected given that people with access to identifiable information are usually tightly restricted and may not include staff involved in quality determination.

Data linkage: RWD captured from a single source may not systematically provide a comprehensive view over the whole lifetime of the patient. Data linkage can be used to address this problem by combining RWD from several individual data sources. Linkage methodologies such as *person matching* (after pseudonymisation) and *de-duplication of records* are often used but may create quality issues. For example, probability-based matching of patients based on non-identifiable or incomplete

information can combine two different patient records, resulting in inaccurate information. De-duplication may have the effect of introducing incoherence, since both duplicate records are valid in each of the sources. On the other hand, identifier-based techniques such as deterministic and probabilistic linkage are often adequate [7].

Additionally, RWD originating from multiple RWD sources are often highly disparate, with structure and terminology that are not standardised. Given this variety, the use of Common Data Model (CDMs) plays an important role in RWD, facilitating the systematic implementation of some aspects of quality control and the development of consistent methodologies (e.g.: data cleaning, profiling, reporting, analysis) applied to different sources. The process to transform an original source to a CDM, called ETL (extract, transform, load) can improve coherence, but at the same time have a negative impact on other dimensions of DQ. For instance, reliability can be affected, both because any transformation increases the risks of error (accuracy) and because the CDM will define some level of precision that may be lower to that of the original source. Extensiveness can also be affected if for instance, some data are removed as non-conformant to the target model, as well as timeliness, which is also affected as the transformation process introduces delays.

1.4.1. Impact on Reliability

In a secondary use of data scenario, there is limited possibility to control most DQ factors of reliability (e.g., accuracy and precision) at the source (point of data recording). Therefore, the primary focus of DQF implementation is error detection that could lead to record removal or amendment with approximated values, while only in some limited cases it can lead to the correction of the data capture processes.

For RWD datasets that are in the category of "Big Data" as defined by the HMA/EMA joint Big Data Steering Group [9], error detection cannot be practically achieved through manual checking of all data records. Therefore, automation plays a key role in error detection and can sometimes help identify inconsistencies or outliers in the data. The use of common data models (CDM)[1] and standardised analytics can prevent coding errors and differences in data curation [10]. However, if these errors are missed during the conversion process, they will remain in the converted data and may be hidden if conversion back to the source data is not possible. Plausibility metrics also play a key role in assessing the reliability of RWD, as it is possible to provide an alternative to validation against primary institution source records.

1.4.2. Impact on Extensiveness and Representativeness

In a secondary use of RWD scenario, it is possible to measure the amount of information in a dataset (e.g.: measuring completeness), but it can be challenging to characterise how this relates to the data recording process. For instance, DQ control on secondary use of data often cannot adequately detect missing information, especially when this relates to an outcome or event that is not necessarily expected to be present[2]. This is made even more challenging as missing data can be the result of flawed data transfer rather than incomplete RWD capture.

In addition, RWD tends to be interpreted following a closed-world assumption that means that an outcome or event is assumed as non-present because it has not been recorded[3] (an example is a patient with cardiovascular disease and type II diabetes, who is considered however non-diabetic in their medical records because they were never tested for the disease).

[1] The harmonisation to a common data model itself has an impact on precision and potentially accuracy.
[2] In some cases (e.g.: age), where the information is known to necessarily exist, the lack of such information can be clearly detected as missing information.
[3] This is unlike in clinical trial data, where absence of events is explicit, and there is no concept of missing values.

These presumptions, which are rarely made explicit, have implications for the analytical methods of generating RWE and are fundamental for the assessment of DQ [11].

When it comes to secondary use of RWD, lack of representativeness of the target population for the study objective may lead to biased outcomes in some types of studies. For example, a study on disease prevalence using RWD from a primary care database may produce skewed results if individuals in the data source are not representative of the entire population.

1.4.3. Impact on Coherence

RWD are often recorded from different healthcare actors and is varied both due to different data representation, i.e., format, structure, content, etc., and due to differing processes for data collection or data recording and DQ control. Coherence, which refers to the homogeneity/uniformity and consistency of data within a single source or across multiple RWD sources, is a critical aspect that needs to be assessed at the time of data publication or consumption. Additionally, coherence is essential to be re-assessed whenever new RWD sources or elements are introduced, especially in the case of data linkage, to ensure data integrity.

Though assessment of several aspects of coherence can be facilitated by measuring conformance vs a specific (common) target model (e.g.: format coherence or structural coherence), some aspects are more subtle:

- Semantic coherence may vary as diverse sources adopt different approaches to map between terminologies. For instance, the term "anuria" can describe a condition of total cessation of urine production in one source, while the same term in another source can be used to specifically note instances where the measurement of urine output is below a specific threshold. The mapping strategy of each source to a target model, coupled with the limitations of terminologies to fully capture the semantic meaning of a mapped term, can lead to coherence issues across diverse sources.
- Temporal coherence can be an issue for long-term datasets, as medical practices (and therefore the meaning of data) may change along the data recording timeline.

1.5. Responsibility for DQ in RWD

Given the variety and complexity of the processes related to RWD recording and utilisation, the variety of actors and data processing involved, DQ in RWD is a distributed responsibility. The responsibility to ensure that DQ is properly characterised is divided among all actors in the RWD life cycle. This includes the measures by which any processes involved in the various steps of the data life cycle (e.g., data capture, aggregation, processing) can impact DQ. To allow an efficient DQ assessment, each party is responsible for making evidence on DQ available when suitable or required, as well as to maintain DQ within declared or acceptable standards, while documenting the processes followed and the data tools used. More detailed information regarding DQ responsibility in RWD can be found under subsection 5.1. (Systems and process characterisation checklist and maturity model).

2. Application of the EMRN DQF to RWD

2.1. Purpose of this document

This document, hereafter referred to as the RW-DQF, extends the EMRN DQF to provide more actionable and focused recommendations for assessing the DQ of RWD with the goal of improving the quality of RWD for regulatory use. It also serves as a guide for enhancing the assessment and

documentation of RWD quality by providing guidelines for characterising the systems and processes underpinning data and their impact, key metrics to evaluate different aspects of DQ within a dataset, and guidelines for using these metrics to assess the suitability of a dataset through a fitness-for-use assessment in relation to a specific research question.

2.2. Scope of the RW-DQF

This document focuses on the subset of RWD recorded within routine clinical practice, i.e., administrative or claims data, EHRs, pharmacy/prescription data, patient registries etc., when used in the context of a specific research question and in line with the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) guide on Methodological Standards in Pharmacoepidemiology [12]. Existing guidance specific to particular data types (e.g., guidance on registry-based studies [13] for the use of patient registry data) still applies and the RW-DQF should be read in conjunction with any other relevant guidance documents.

The following should be considered out of scope:

- DQ concerning RWD arising from repurposing of previously published analyses, e.g., meta-analyses etc.
- DQ of direct-from-patient data, i.e., PROs, patient engagement data, patient preferences, mobile health data, social media data, etc., as these may have peculiarities in terms of measuring DQ and would be subject to further guidance.

2.3. Structure of the RW-DQF

The RWD chapter is composed of three parts, presented in the three following sections:

- Guidelines to characterise systems and processes (and their impact)
- Metrics to appraise different aspects of DQ within a given dataset
- Guidelines to assess the suitability of a dataset for answering a specific research question.

The RW-DQF inherits the concepts and design of the EMRN DQF. This includes the categorisation of quality aspects into three types of determinants (foundational, intrinsic and question-specific), the maturity model and the definitions of the DQ dimensions. Such concepts are presented in this document using a varied terminology, that is more commonly understood among RWD stakeholders. They are also further specialised and altered to the RWD context.

In particular:

- **Foundational determinants** are defined as "everything that impacts DQ, but it is not related directly to the dataset and does not depend on any specific research question". In this chapter, foundational DQ aspects are referred to as the characterisation of the **systems and processes** that have an impact on DQ. In most establishments where RWD are captured, processed and/or consumed, information on foundational determinants is often limited to onboarding documents [2]. In addition, the impact of systems and processes is usually considered with respect only to data accrual [7] and then approximated [14]. This chapter considers the impact of systems and processes in some detail, as well as the impact along the whole evidence generation process.
- **Intrinsic determinants** are defined as "DQ aspects that can be observed only on the basis of a given dataset, without requirement for information about how the data were captured, or about its primary/intended use". In this chapter, intrinsic determinants are considered as **Metrics** that can be used to characterise DQ, and the chapter provides guidelines on how to use such metrics.

- **Question-specific determinants** are defined as "aspects of DQ that cannot be assessed independently of a research question". In this chapter, they are considered in terms of how to assess the **suitability of a dataset for a specific research question**.

As for dimensions and sub-dimensions, the terminology introduced in the EMRN DQF sometimes differs from what is found in other DQFs focused on RWD [15-17]. In general, there is a lack of consensus on terminology, among RWD practitioners (researchers, data analysts, data custodians) and more broadly among people involved in the RWD recording process. For instance, the term "validation" refers to checking whether the data correspond to the source [15], but is commonly understood as "checking that the data conform to a schema" by database administrators.

This RW-DQF will use the terminology introduced in the EMRN DQF [1]. These are reported in the glossary of this document (Chapter 9) for convenience.

2.3.1. Understanding relevance

In the EMRN DQF, the notion of relevance (i.e., the extent to which a dataset presents the data elements useful to answer a given research question) is considered as something cross-cutting through all DQ dimensions and applying to each of them. All aspects of DQ can be measured by metrics independently of a research question, while no thresholds or acceptance criteria can be established independently from it[4].

Error! Reference source not found. summarises the interplay between systems and processes characterisation, metrics, and suitability to a research question across five dimensions for DQ. It can be seen in this table that "relevance" is the only dimension purely determined by the research question.

Table 1- Interplay between definitions and dimensions for DQ

	Are data correct? ***Reliability***	**Are data enough?** ***Extensiveness***	**Are data homogenous?** ***Coherence***	**Are data timely?** ***Timeliness***	**Is this the right type of data?** ***Relevance***
Systems and processes	Determine reliability	Determine extensiveness	Enable coherence	Determines timeliness	
Metrics	Can assess reliability	Can measure extensiveness	Can measure and improve coherence		
Suitability to a research question	Defines "acceptable" reliability	Defines if data are sufficient	Defines if the level of coherence is adequate	Defines acceptable timeliness	Defines if the content of the data is what is needed

A major difference between the EMRN DQF and other proposed DQFs [15-17] for RWD is the role of relevance. In other frameworks, relevance is often considered as its own dimension, and includes completeness and reliability as sub-dimensions [15]. The notion of "Relevance DQ dimension" introduced in the EMRN DQF is much more restricted, and only meant to capture some DQ aspects not covered by other dimensions (corresponding to the question: is the type of data fit-for-use?).

[4] As discussed in the EMRN DQF, there may be "general questions" that a dataset may be expected to be used for, and from which some quality threshold could be derived. However, establishing such threshold is easily discretional without a clear definition of such target uses. Even in this case, an "unqualifying" dataset may still be useful in a different use case, e.g.: if data are very scarce and critical.

In the context of RWD, reference to the Relevance DQ Dimension is omitted in the discussion around metrics but is considered in the context of assessing the suitability of datasets used to answer a specific research question.

3. Guidelines for the characterisation of systems and processes underpinning data

The quality of data cannot be assured unless the systems and processes responsible for their collection or recording and transformation are reliable and offer the necessary guarantees. If RWD are considered for regulatory use, no matter the content of an RWD source, its use would be unfeasible unless there is some reasonable evidence that the information provided is true and not accidentally or intentionally altered. This section provides guidelines on how to characterise systems and processes, so that their effect on DQ can be assessed.

This section builds on the "Foundational Determinants" definition of the EMRN DQF [1], and the related maturity model. In the RWD context, the maturity model is adapted and takes the form of a practical checklist (Figure 3).

At its core, maturity depends on the ability to produce documented evidence of good DQ practices and quality-related actions. Maturity advances if this documented evidence is standardised and systematic by nature, to allow it to be coherently interpreted across different RWD sources.

Maturity depends on the level of automation, as automated DQ processes can be both more extensive and less subject to accidental or human errors. Therefore, the maturity model presented here focuses on what information should be provided, how it could be standardised, and how it could be automated.

3.1. Systems and process characterisation checklist and maturity model

This section suggests how to characterise aspects of systems and processes that have an impact on DQ. These aspects are grouped by areas that reflect different steps found in the data life cycle (see Figure 2) and are presented in Table 2 that can be used as checklist to verify if foundational information necessary to assess DQ is provided.

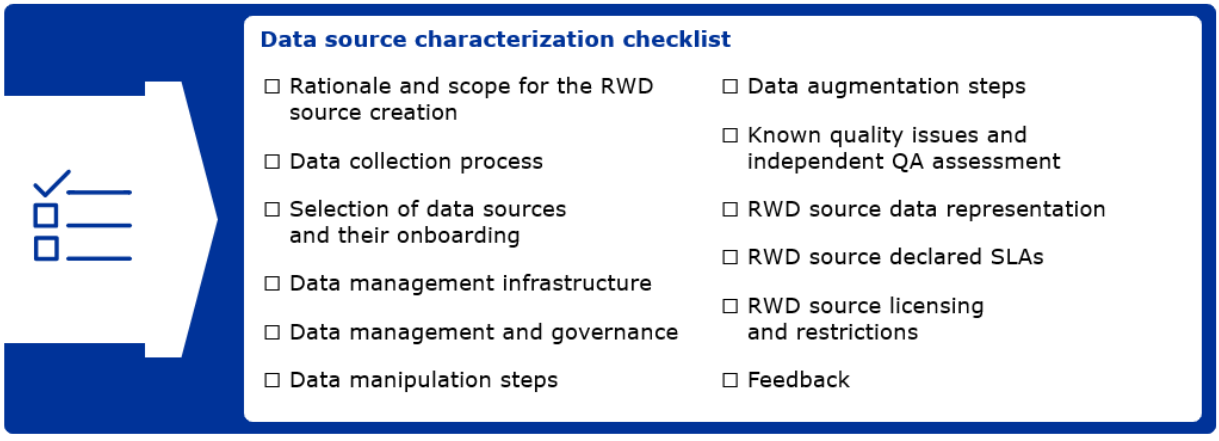

Figure 2 - RWD source characterisation checklist overview. SLA – service level agreement.

For each area of interest, the table lists what information should be provided together with how this information should generally be generated, following the maturity model derived from the EMRN DQF [1][5]:

Level 1: Documented. Some information should be provided at least as simple documentation (e.g.: short text) and/or supporting links. More extensive documentation can include standard operating procedures (SOPs), as well as key performance indicators (KPIs).

Level 2: Formalised. The information should be provided in a way that follows established or emergent standards (e.g.: following recommendations such as the ones in the EMA Guideline on registry-based studies [13] or the REQuEST tool [18], or more general frameworks and standards e.g.,: [19]. SOPs and KPIs, when reported, follow established standards and guidelines.

Level 3: Automated. The information should be derived in a way that guarantees higher DQ by design. This means the data are generated by systems and platforms by a computed process rather than being entered ad-hoc or a-posteriori. For example, in the case of data lineage, provenance information is generated by an ETL engine or derived from some executable electronic specification instead of being provided independently from the system by manual documentation.

These three maturity levels are reflected in table 2, where the "Documented" column lists what information should be provided. The "Formalised" column provides suggestion of how data can be presented in a standardised manner. The "Automated" column provides suggestions of ways to improve and verify data captured through automated processes.

Not all required information is relevant for all workflows (e.g.: a registry implies different processes than a source aggregating and repurposing claims or prescription data). For pragmatic reasons, alongside information describing systems and processes in place, descriptions of the intended characteristics of RWD are also included in Table 2[6].

Achieving higher levels of DQ is a distributed responsibility across the data lifecycle from data capture to data use. Essential DQ checks cannot be fully performed solely by a single stakeholder as for instance the responsible person for a study submission. In addition, data characterisation such as data pertaining to the data's fitness-for-use, cannot be anticipated by the stakeholder capturing the data or from the data holder alone, however, the data holder has usually knowledge whether a research question can be answered. Therefore, Table 2 proposes to refer to "upstream" quality assessments for all RWD sources under the responsibility of RWD holders. To allow an adequate DQ assessment, the RWD submitters or the RWD end users, i.e., the stakeholders that get access to RWD and use the RWD for secondary purpose, should be responsible for providing the available evidence on DQ when suitable or required.

Finally, this table proposes to capture information on what feedback mechanisms are provided. The concept of feedback mechanisms is derived from the EMRN DQF overall maturity model. Given the complexity and heterogeneity of RWD capture, detailed feedback loops on all aspects of DQ may be unfeasible. Therefore, feedback is here proposed as an explicit aspect to report for the overall RWD source, rather than as an extra maturity level.

[5] The maturity model in the RWD Chapter differ from the EMRN DQF in that "feedback" is not considered as a maturity level, but as an element of systems and processes.
[6] Sections IX, X, XI describe RWD source characteristics pragmatically useful in a "data checklist": even if not strictly affecting DQ, they are relevant for actions supporting DQ assessment.

374 **Table 2: RWD Source characterisation checklist with 12 areas of interest**

❖		Level 1: Documented	Level 2: Formalised	Level 3: Automated
Item	Rationale	Documentation to be provided: free text and/or online link(s)	Suggestions for standardised documentation to enhance interpretability	Suggestions for robust documentation generation "by design" and machine readable
I. Rationale and scope for the RWD source creation	Relevant for all DQ dimensions as it provides a general understanding of the strengths and limitations of an RWD source.	A) The primary purpose(s) for which data are collected B) The justification or criteria used for the selection of the data being collected (or integrated) C) Publications describing this RWD source	Provide information using standardised / widely used templates, to make the relevant information easy to digest and interpret[7]	Provide information as **Metadata**, in a standard format and with clear definitions[8], to allow Metadata to be automatically processed and their quality (e.g.: completeness) be adequately verified.
II. The data collection or recording process	Essential to understand coverage and to assess reliability (that can be affected by errors or biases in the collection process). Also, essential to evaluate SOP for data collection or recording practices that may impact coherence (e.g., where "curation at source" is involved and provide hard constraints for timeliness).	A) Description of the data provider, including: • its nature (patients self-reported, carers or third parties, healthcare professionals with specified speciality) • its geographical and organisational setting. B) Description of data collection or recording SOPs, including the rationale for the SOP design. C) Information of how SOPs are implemented, and their execution monitored. D) Characteristics of key data elements captured, e.g.: • core, optional elements • planned size • planned coverage over time	In addition to **I**: make use of standard vocabularies where available. SOPs are made available and are based on common shared standards.	In addition to **I**: SOPs specify KPIs so that adherence to these KPIs can be monitored and reported.

[7] An example of a such a template can be found at [REQuesT]. We encourage the development of such shared template, when not available.
[8] In essence, information provided should follow FAIR standards, when feasible.

III. The selection of RWD sources and their onboarding (Applies to RWD sources that integrate or repurpose other RWD sources)	When data are provided by a data aggregator, ensure that all the available evidence related to systems and processes potentially affecting DQ can be followed. Provide information of impact on both reliability and evidence (as well as other dimensions if relative constraints are formulated in inclusion/exclusion (I/E) criteria)	A) The data providers' selection processes and criteria, e.g.: • Inclusion and exclusion criteria for the acceptance of a RWD source B) A comprehensive DQ assessment of RWD sources being consumed (as a reference, or as evidence of the frameworks being followed)	The DQF assessment includes this checklist for each RWD source. Reference to datasets with unambiguous identifiers that can distinguish between datasets versions, when relevant.	(Aspirational) The DQ assessment provided can be processed so that upstream DQ can be incorporated, to reconstruct a full "chain of evidence" of an RWD resource.
IV. The data management infrastructure	Essential for reliability regarding data alterations resulting from system accidents, software errors or malicious intervention.	A) The list of systems used to manage the RWD source, from data collection or recording to processing to making it available (version, features used). B) The software[9] testing and QA processes in place. C) Measures to prevent accidental physical data alterations (e.g.: backups, redundant systems, checksums).	The hardware or software implementation complies with recognised quality standards that can be reported.	
V. Data management and governance	Data management and governance impact reliability, as well as all quality	A) A description of the overall data management principles adopted (e.g.: ALCOA+, FAIR)	SOPs and data management processes adhere to standards that can be referred to: e.g., GCP, ENCePP, ISO 25012, ISO 25101, ISO 8000-	Data management and governance is implemented in the data platforms 'Digital Quality Measures' (DQMs) so that reports of performance and

[9] We assume the HW testing is not an issue as necessarily performed a-priori.

	dimensions for metadata.	B) A description of data management processes in place: * SOPs in place * Responsibilities and roles * DQ controls * KPIs C) Measures to prevent unauthorised data alterations (e.g.: cybersecurity approach) D) Monitoring, auditing, and quality improvement procedures in place. E) Metadata management practices and SOPs	6x, ISO 25024:2015. The representation of metadata follows FAIR standards. The use of best practices with a direct impact on DQ is explicitly reported (e.g.: variables for which explicit negation is used, variables for which absolute values are reported).	deviations are automated. • Submitted metadata are generated "by design"
VI. Data manipulation steps[10]	Impacts reliability both in terms of accuracy (possible errors) and precision (i.e., the degree of approximation by which data represents reality). Essential to ensure traceability of information. Also impacts coherence and potentially timeliness.	A) A description of data onboarding procedures, e.g.: * Frequency and modality of updates * "acceptance tests" performed on RWD sources. e.g.: sources are monitored over time for sudden variation of content, as a proxy to detect process errors B) A description of data manipulation steps, including: • Data transformations performed (e.g.: unit of measure conversions, formatting, pivoting, deriving new values, such as BMI from weight and height).	Tests performed follow some standard or shared set of tests, that can be re-used across RWD sources. Key performance indicators (KPIs) for data cleaning (e.g., data duplications, mislabelling, etc.) are provided. Data mapping tables and algorithms are described with a standard characterisation of their performance. Lists of standard test batteries	Information about data onboarding is directly provided by the platform, e.g.: * Transaction logs are available including deviations and actions that required manual intervention Actual data transformation code is accessible and verifiable. Quality checks and KPIs reported are automatically generated by the data platform (e.g.: unit testing) Lineage information is

[10] By "data manipulation" we consider transformations that, in the absence of error, don't affect data reliability: e.g.: unit of measures conversion.

		• Data cleaning steps (e.g.: duplicate detection) • Data mapping steps (e.g.: terminology mapping). • Include information about loss of precision expected (e.g., loss of time detail if time of data capture is rounded up to nearest minute; or loss of precision resulting from terminology mapping). C) A description of testing procedures • SOP for testing (e.g.: test of pipelines vs test of executions) D) Lineage information • Provide justification for the level of data manipulation. • Provide lineage information to specified level sought.	used to detect loss of accuracy or precision are provided. All lineage information is provided as metadata associated to the dataset	automatically generated by the processing platform
VII. Data augmentation steps[11]	Data augmentation steps impact accuracy.	A) Information on data augmentation steps (e.g.: imputation or linkage) • Justification, methods (algorithms), assumptions, excepted error rate • Detail on where such methods are applied. • algorithm such as name, source description and justification for use.	• Algorithms are published, shared and their performance documented. Reference to algorithms is to a specific version. • Information on which values result from imputation is provided as part of the dataset (e.g.: in metadata,	

[11] We consider here data transformations that produce new information subject to reliability issues: e.g.: imputation of missing values, or extraction of codes via natural language processing.

			or data dictionary).	
VIII. Known quality issues and independent QA assessment of the RWD source	Explicit description of known DQ issues, as well as external validation performed (all dimensions affected)	A) Self-reported known DQ issues with an explanation of factors leading to issues (e.g., poor overall completeness in Q3 2020 due to COVID-19) • Include a description of known approximations of loss of precisions in mappings. B) Known independent data validations: * Validation studies • Publications resulting from this RWD source		
IX. The RWD source representation	Descriptive of the intended coherence of a dataset and its metadata.	A) Description of the data model used Description of non-standard data model B) Data dictionary and ontologies (vocabularies) in use.	The description refers to a model such as FHIR, OMOP, I2B2, a subset or eventually an extension of such. Information on data dictionaries is: • Based on standard vocabularies (such as ICD-9 to ICD-10 diagnosis) • Refers to a specific version of vocabularies used. • When non-standard dictionaries are used, a rationale is provided, and full dictionary is made available	Data dictionaries are provided using standard formats that facilitate the mappings across different vocabularies and across languages.
X. The RWD source declared Service Level	Descriptive of guaranteed timeliness and possible variations of	A) Data resource declared SLAs	Provide details of established data processes	SLA compliance is automatically assessed and reported

Agreements (SLAs)	extensivenes s/reliability provided.	• Guaranteed frequency of updates • Guaranteed incident response time (e.g.: corrections in case of errors) B) Processes and resources accompanying data (e.g.: documentation, training material, help desk). C) Extended capabilities related to DQ: e.g.: possibility to collect additional data if needed	followed by the SLA provider.	
XI. The RWD source licensing and restrictions	Descriptive of aspects that can limit extensivenes s and coherence in downstream data aggregations.	A) Details on conditions and processes under which data are made available, such as: • Features of data use agreements that may limit data use or access (consent, limitations of use). • Licensing constraints B) Dataset retention and accessibility policies.	Policies and licensing reported are standardised and applied to a broad range of RWD sources	
XII. Feedback	Descriptive of feedback mechanisms in place to improve all aspects of DQ	A) Provide a contact for QA and follow-up on DQ issues detected.	The contact provided allows tracking of issues and follow-up after standard service support patterns	The feedback mechanism provided includes notification of automatically detected DQ issues.

When filling the above form, all fields should be used, eventually clarifying when something doesn't apply and why (e.g.: no processing of the dataset was done).

3.2. General considerations on the characterisation of systems and processes

Since the recording or processing of data may have an impact on DQ, every actor involved in any such process should therefore ensure that their actions adhere to the checklist above. Generally, an end user preparing a dataset to support regulatory activities would therefore provide the above checklist for any eventual processing they did on the datasets, plus a checklist for each source of data used, while the RWD source provider is the entity better positioned to provide a checklist covering its specific data assets.

Overall, independent of who takes responsibility for the information provided, how DQ information is aggregated, or whether this information is provided in full, summarised, and accessible on demand

(e.g.: in case of audit), all the available evidence related to systems and processes potentially affecting DQ should be clear at the time of submission.

4. Data quality metrics for RWD

Metrics are the most obvious case of what is introduced in the DQF as "intrinsic determinants". This is defined as all DQ aspects that could be assessed based on a dataset itself, without information on how data have been produced, nor its intended usage.

This section introduces metrics that can be used to measure different aspects of DQ. An overall framework that groups DQ metrics in terms of their requirements and dimensions is introduced in a way that can help assemble and systematise existing quality metrics into balanced sets, as well as identify gaps in existing metric sets.

This section also provides a list of example metrics. The presented list is not meant to be exhaustive: there are many DQ metrics outlined in the literature [15] and many more that could be created based on the individual characteristics of a data type. In the EMRN DQF [1], metrics were presented at an abstract level and were covering a very wide range of scenarios, including examples beyond clinical RWD, mostly with the goal of illustrating each quality dimension. What is presented here, is a sample of concrete metrics that are highly relevant and broadly applicable for characterisation of RWD, together with RWD-specific examples to illustrate a potential output these metrics can be applied.

4.1. Framework for the categorisation and identification of metrics

To categorise and identify metrics, a simple framework can be used to test the completeness of test sets in use, as well as to identify gaps, redundancy, or complementary metrics (See Figure 4). This framework can be visualised as a simple table, with dimensions in the columns and metric groups in the rows. Each dimension shown in Figure 4 consists of multiple sub-dimensions, detailed in Tables 3 to 6. Note that not all metric groups apply to every dimension in this table.

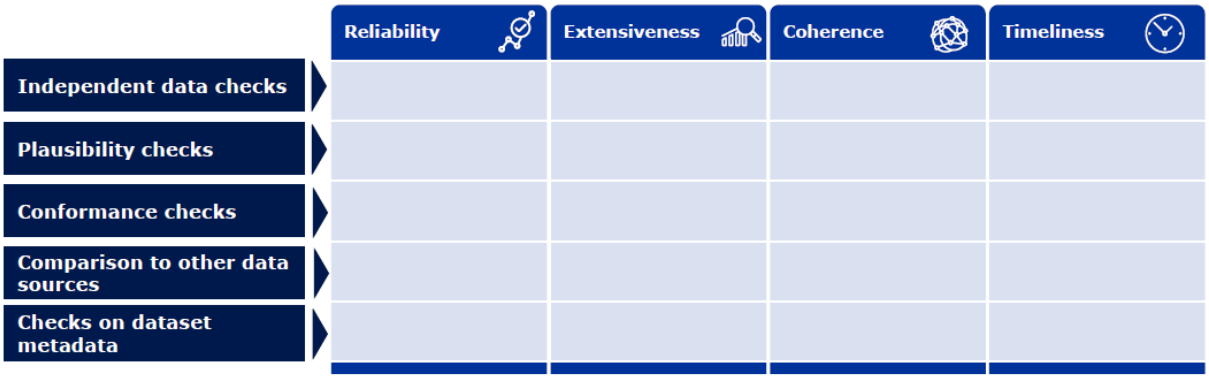

Figure 4 – Proposed 2-dimensional framework for metrics identification.

These dimensions are classes of the DQ features that the metrics are meant to measure. Requirement for metrics assessment is on the other axis. These metric assessment requirements identify what resources are needed and what additional information or know-how a metric embeds. In terms of requirements, this section outlines the following metric groups:

Independent data checks. These are metric groups for which no additional knowledge or information on the content of the dataset is required. Examples may include the number of empty or corrupted fields or the number of potential duplicates. Independent data checks can be designed and applied to a broad range of data.

Other metrics can instead embed knowledge about general know-how on the data being measured. Key examples are plausibility metrics and conformance metrics introduced below.

Plausibility checks. These are metrics that capture DQ aspects based on general knowledge about the world represented in data. For instance, the number of (un)reliable values could be assessed by detecting patterns that are impossible to be present in the Real-World: e.g.: female patients that have observations only occurring in males, measured quantities that exceed a certain magnitude (e.g.: a blood pressure of 1000/500 mmHg), or patterns that are impossible (e.g.: the timing of a causal effect occurring after its effect), etc.

Conformance checks. Metrics assessing conformity to standards dictating data structures, dictionaries, or format, e.g., all values to represent a condition come from a prescribed terminology source.

Checks on dataset metadata. This class is based on the know-how on a specific dataset, such as what is provided in metadata or supporting documentation. In some cases, it is useful to consider metrics that are based on the 'descriptors' that come with a dataset (e.g., metadata) that reflect the processes or the standards behind a dataset. For instance, a dataset could be provided with additional dataset detailing what values are recorded, and what is imputed. Metrics summarising the percentage of imputed data could then be used to assess the reliability[12] of a dataset. It is also useful to verify how data values match expectations with respect to metadata constraints. In principle such metrics could measure, by direct verification, the effect of a full data process.

Comparison to other data sources. Metrics resulting from the comparison against reference RWD sources can support extensiveness and reliability assessments, particularly against broadly recognised RWD sources with demonstrated quality assurance. For example, it is useful to compare the proportion of missing data to a reference dataset to gain an understanding of the possibility of bias in data collection or recording. This kind of metrics can be used to determine the true accuracy or validity of data only in rare cases, for instance when the same type of data has been collected for the same patient in real world and in a randomised clinical trial (RCT) and the latter can be leveraged as gold standard to assess the validity of the RWD.

These metrics don't include results of validation of accuracy against original data, as that is expected to be covered in foundational documentation (see section 3.1. . Certain data can also be valid when observed individually, but the collective trend of all data of a kind should follow expected distributions or trends, based on clinical expectations. For example, the prevalence of a disease is unlikely to grow drastically (e.g., from 2% to 80%) in a population from one year to another. In that case, metrics are difficult to determine. Instead, a visual representation of data may be needed to detect abnormal trends and data with low plausibility. This process is also called clinical validity.

In the following section, some examples of metrics are provided in relation to this framework. Tables 3-6 refer first to the EMRN DQF and more broadly to commonly used DQ checks [20] . This chapter does not refer to the verification (within data) versus validation distinction (compared to other RWD sources), as this is made more detailed and operational by the above implementation categories.

[12] We note that the term "reliability" is used here with the definition presented in the EMRN DQF ("that data correspond to reality"), that differs from its interpretation in statistics ("consistency of repeated measures")

4.2. Metrics for DQ assessments

Reliability dimension

These metrics are meant to measure the degree to which data correspond to what they intend to represent.

Table 3 - Overview of reliability metrics by sub-dimensions with examples

Sub-dimension	Metric group	Metrics	Example
Accuracy	**Plausibility checks[13]**	• Number and percent of records where data values don't agree with standards or knowledge or feasible ranges • Number and percent of records where values of repeated measurement of the same fact don't show expected variability • Number and percent of records with logical inconsistencies • Number and percent of records where observed or derived values don't conform to expected temporal properties • Number and percent of records where duplicate records aren't flagged. • Number and percent of records where data values don't agree with common expectations (e.g., human with 4 arms)	• For X% of records/rows, systolic blood pressure values are higher than 250 mmHg • X% of records showed >2kg difference when weight was measured by separate nurses within the same facility using the same equipment • X% of records of pregnancy were attributed to males • For X% of records, discharge date happens before admission date
	Checks on dataset metadata	• Number and percent of variables/datasets that are based on imputation or derivation	• End of treatment date is derived for X% of patients from treatment start date and treatment cycle length
	Comparison to other data sources	• Number and percent of records where corresponding variables yield identical results across independent or dependent databases	• X% of EHR records had a date of birth value that matched the value in a birth

[13] Accuracy metrics based on general knowledge are typically plausibility metrics, where a dataset is assessed regarding its likelihood to be correct, based on common expectations regarding data distribution or general constraints between different values.

			registry for the same patient
Precision	**Independent data checks**	• The number of decimal points used in data values, and their distribution	• "Height" in meters recorded with two decimal digits, but the last being always 0.
Traceability	**Checks on dataset metadata**	• Number and percent of datasets/variables for which traceability information is available in metadata.	• Metadata regarding traceability are available for only 2 out of the 3 datasets feeding into an overall disease registry (treatment data and death records contain traceability-related metadata, but not medical history data)

463 ***Extensiveness dimension***

464 **Table 4 - Overview of extensiveness metrics by sub-dimensions with examples**

Sub-dimension	Metric group	Metrics	Example
Completeness	**Independent data checks**	• Percentage of records for which data are missing for a given variable	• X% of patients have a value recorded for their Date of birth
		• Percentage of patients who have a certain number of measurements for a given variable	• X% of patients have 2 or more ECOG assessments
	Comparison to other data sources	• Relative percentage of records for which a variable is missing with respect to a trusted source of knowledge	• X% of patients have date of diagnosis missing for a diabetes database, compared to a National and institutionally validated diabetes registry

Coverage	**Comparison to other data sources**	• Percentage of a target population present in a database	• X% of diabetic patients covered in a regional diabetes registry as compared to the National Patient Registry

465 ***Coherence dimension***

466 **Table 5 - Overview of coherence metrics by sub-dimensions with examples**

Sub-dimension	Metric group	Metrics	Example
Format coherence (conformance)	**Conformance checks**	• For relevant variables, % of records where that conform to formatting constraints	• X% of records have Sex as only one ASCII character (0 or 1)
		• For relevant variables, % of records where data values conform to allowable values or ranges	• X% of records have sex with one of the 3 allowable values "M", "F". or "U".
Relational coherence (conformance)	**Independent data checks**	• Data values conform to relational constraints based on external standards	• X% of records having the Provider ID in the Drug exposure data correspond to the record in the Provider table
	Conformance checks	• For computed values, % of records where computed values conform to programming specifications	• For X% of patients, database calculated, and hand calculated BMI (body mass index) values are identical at a 0.2 margin of error
Semantic coherence (conformance)	**Conformance checks**	• For relevant variables which employ code lists according to external standards, % of patient records which use a given code list	• X% of diagnoses are coded with ICD-10 (as required by CDM)
Uniqueness	**Independent data checks**	• Number of records flagged as potential duplicates	• Out of X records, 2 are flagged as potential duplicates: William Smith's DOB and ID matches with Bill Smith's DOB and ID.

Timeliness dimension

Table 6 - Overview of timeliness metrics by sub-dimensions with examples

Sub-dimensions	Metric group	Metrics	Example
Currency (Is your data acceptably up to date?)	**Independent data checks**	• Average time of updates in a database	• Timestamps between 2 consecutive forms indicate patient records are updated on average every 3 months

These metrics have been provided at a general level where one would apply the metric to all records, however, there can be some hypothesis-driven stratification to look at the data with more granularity (in context of a particular question/context, see below section 7). E.g., for completeness and coverage, one may want to look at the metrics in a stratified way, where there may be a sub-population of particular interest/criticality or where there is an expectation for lower quality.

4.3. Considerations for the implementation of RWD DQ metrics

4.3.1. Different roles of metrics

This section distinguishes different primary roles of DQ metrics: such roles correspond to different optimal sets of metrics.

4.3.1.1. Quality assurance

When metrics are used for DQ assurance, the intention is to identify issues with the aim of correcting these issues when possible. Such metrics are naturally automated and tend to be as extensive as possible. Test sets comprising hundreds of metrics are possible: anomalies and unexpected values detected can then be screened and lead to follow-up actions, including inspection of sources of data pipelines to identify errors. Often such issues can be prioritised with respect to frequency and severity, hence there are little downsides on test set being extensive, especially when automation is in place.

4.3.1.2. Quality reporting

When metrics are used for DQ reporting, they are meant to provide some high-level assessment of quality that can be used for an assessment of DQ. In this case, metrics should be more high-level and limited in number, and such that some relative assessment of DQ among datasets is possible (e.g.: typical metrics would be average completeness, or average conformance). The value of such high-level characterisation is limited but useful when datasets are presented in a catalogue for a first characterisation of DQ.

4.3.1.3. DQ assessment

More precise than reporting, DQ metrics can be used to reflect the quality of specific data elements, with the view of assessing whether a dataset is or is not suitable to answer a specific research question, e.g.: whether the precision of age is suitable for a paediatric study. In this case metrics need to be presented at a level of detail that makes them unsuitable for high-level reporting. Furthermore, such metrics should be assessed on the population of interest, rather than an overall generic dataset.

4.3.2. Additional considerations on level of application and maturity for metric assessments

As per the above description of the different roles of metrics, metrics may be used at different points in the chain of RWD creation and aggregation. For example, for a registry collecting data from multiple hospitals, metrics can be generated at different levels: within the individual hospitals, or at the whole registry level, varying in relevance of metrics (e.g., coherence between sites).

As described in the EMRN DQF, metrics may be assessed and reported on with varying levels of maturity. For example, at the lowest level, metrics may have to be estimated and self-reported by the data owner with approximate knowledge of general data trends ('qualitative assessment') and may be generated "ad-hoc". While at higher levels of maturity, 'quantitative assessments' (based directly on the data) should be fully automated. Fully automated checks should take place in capture systems during data collection or recording and then throughout the generation system.

5. Guidelines to assess quality in relation to a specific research question

This framework distinguishes the measurement of DQ (in metrics, but also via supporting evidence) from the evaluation of its fitness-for-use, as this relates, by definition, to the relevance of the dataset for a specific question addressed.

It is in fact difficult to pre-specify thresholds or minimum criteria for the fitness-for-use assessment: generally, a lot depends on the type of study, and on disease-dependent and analysis-dependent factors. In addition, there may be some other considerations, such as lack of other RWD sources in a therapeutic area or disease frequency that have an impact on setting acceptability thresholds (e.g. for rare diseases, it might be challenging to identify enough cases via secondary use of data).

At a more detailed level, data relevance to a specific question is demonstrated if the data capture key data elements of the research to address such question (e.g., diagnosis, exposure, outcome, and covariates) in a reliable, coherent and timely way, or if the number of patients and follow-up time are sufficient to demonstrate the impact of the intervention/determinant under investigation [14]. To assess the relevance of a RWD source, an in-depth and systematic evaluation of the data source in relation to its design elements is required.

In the next section, guidelines are provided to assess the relevance of data to a research question based on DQ metrics and the evidence of systems and processes provided. Note that relevance is not limited to accepting thresholds or evaluating the "usability" of a study. The quality characterisation of a source is also useful input to define applicable methods as well as additional RWD sources that may be required to answer the research question.

5.1. General principles for assessment of data quality in relation to a research question

Once a research question has been defined, a set of steps that guide the assessment of the suitability of a RWD source with respect to quality can be considered (See Figure 5).

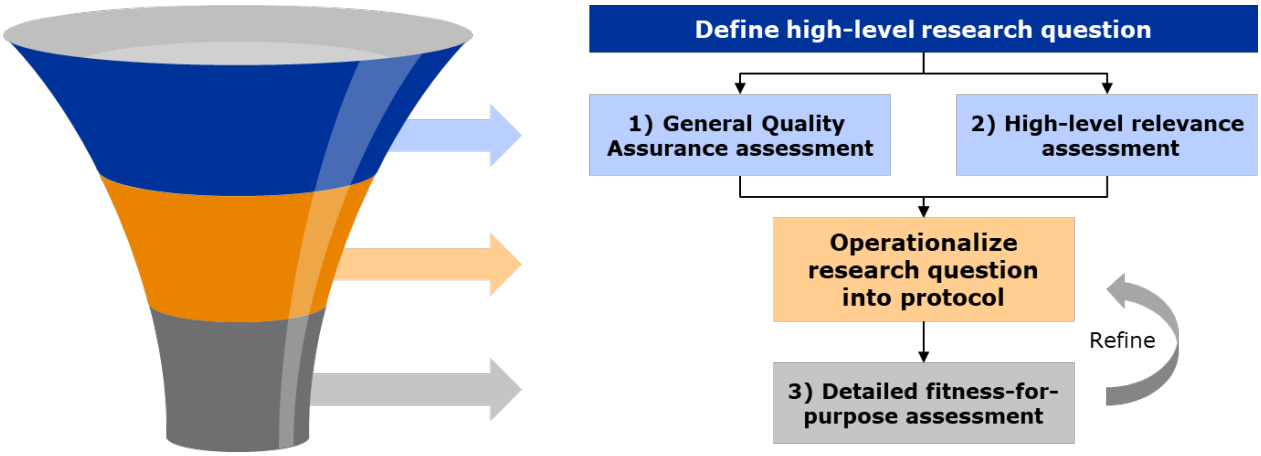

Figure 5 - General principles for assessment of DQ in relation to a research question

Step 1 – General Quality Assurance assessment

To determine if a RWD source may be relevant for regulatory purposes, the first step is to ensure that the available information and documentation meet the overall quality requirements in terms of reliability and, if applicable, timeliness.

Since some research questions (e.g., related to pharmacovigilance) are time-sensitive, the overall characterisation of a source with respect to timeliness (e.g., overall time lag) can be a key criterion for its acceptability.

It is also possible to assess how much a dataset is represented in a coherent way that facilitates analysis. Documentation on terminologies and standards used, can help in assessing the fitness of a dataset to a specific analysis goal and process. Coherence assessments do not generally result in yes/no decisions on the suitability of a dataset, as it can be usually improved with extra efforts[14], but it can be a criterion when multiple RWD sources are available.

This first step of the assessment is typically done by inspecting documentation (e.g.: the systems and processes checklist), for instance to assess the reliability of data one would look at description on the presence of QA processes, documentation of any data curation, data transformation/enrichment steps, etc.

Step 2 – High-level relevance assessment

The next step is to assess the "relevance" of a dataset to a question. In the EMRN DQF, the definition of "relevance" is narrowed to data having the right kind of variables for the question at hand. To assess the "fitness-for-use" aspect of a dataset to a specific question, an essential step is indeed to identify if the content of the information fits the requirements posed by the question. Whether a dataset presents the right kind of variables can be assessed at high-level based on the overall data documentation (e.g.: how data are collected, the purpose, the data dictionary).

A preliminary assessment of relevance can be conducted by inspecting metadata, without directly accessing the data or relying on detailed metrics. Increasing knowledge of available data sources (e.g., EHR, administrative claims) can help in framing the research question more explicitly to guide the

[14] There is a potential loss of precision when data are harmonised to a common standard. Therefore, if data are not coming in the coherent representation for an intended analysis process, some attention must be put on precision degradation, when assessing specific variables (later in this document).

choice of a specific data source. This process can be further facilitated by registering data sources in relevant online repositories, such as the HMA-EMA Catalogue of Real-World Data Sources [15].

Step 3 – Detailed fitness-for-use assessment

Once the DQ of a RWD source is determined acceptable at an overall level, a specific RWD source inspection is required. To do so, one must first:

1. Articulate the research question and the relevant design elements such as study population, study sampling (e.g., case-control, cohort), treatment/exposure group, comparator group, primary and secondary outcome(s), length of follow-up, data lag time, confounders.

2. Operationalise data elements into variables depending on the specificities of the research question to get a better understanding of the disease area (e.g., rate of evolution of standard-of-care, i.e., how frequently the standard-of-care changes for a given indication, time-to-disease progression).

    - Where possible, pre-specify the importance of the quality of data elements in the protocol – this assessment should be done in anticipation of the analysis methods (e.g., sample size calculations, use of time-to-event endpoints, sensitivity analyses, statistical adjustment for measurement error), which will impact what is considered acceptable for missing data or errors. While not part of the quality assessment itself, anticipating methods is important to provide context for performing a quality assessment.

After this phase, the qualification of the RWD source can be performed by assessing if the data quality of the variables of interest is adequate for the intended analysis. This entails assessing the extensiveness (e.g.: completeness) of the required design elements as well as the coherence, reliability, and timeliness of those elements.

Note that the fitness-for-use assessment could be done based on metrics and metadata that are reported for an overall RWD source or could be performed on the final (sub)dataset selected for the study (e.g., specific data cut/sub-population/aggregation of RWD sources).

In general, all summary metrics may change when a subset of a population is considered (e.g.: the precision of "age" may change if a subset of a population focusing on paediatric patients is considered). While this is rare for accuracy and timeliness, extensiveness is often affected: for an identified data (sub)set of interest, a fit-for-use assessment also entails seeing if the sample size of planned patient population is enough to guarantee robust evidence, and whether data are representative of the target population when relevant. Coherence in particular needs to be re-assessed each time a new data source is introduced in an analysis.

Generally, the RWD source should be chosen to match the research question, rather than adjusting the research question to fit the RWD source. It is important to note that, in some cases, the metrics and characterisation can lead to changes in the study design to accommodate limitations in the data (iterative process). For example, if a rare disease is insufficiently captured in a RWD source or in the patients of interest included in the RWD source, but a broader concept that is also of interest is well-captured, the study may focus on the broader concept instead. In contrast, in a causal study, if important confounders are not captured in the RWD source, it may be necessary to find an alternative RWD source to conduct the study.

[15] https://catalogues.ema.europa.eu/catalogue-rwd-sources

5.2. Framework for detailed fitness-for-use assessment

This framework is inspired by The Structured Process to Identify Fit-For-Purpose Data (SPFID) [14]. However, it differs from it in that the aim of this RW-DQF is not to exhaustively look for different RWD sources and rank them comparatively for their fitness for purpose. It rather provides a guideline to assess if a data source is suitable for regulatory use.

Table 7 provides a template to be filled in during the Step 3 of the fit-for-purpose assessment.

Table 7 - Fitness-for-use assessment to be filled in during the suitability assessment of a RWD source

Design elements *(to be pre-specified)*	**Operational definition** *(to be pre-specified)*	**Data elements for valid capture** *(to be pre-specified)*	**Criticality of the quality of the element, including justification where relevant** *(to be pre-specified)*	**Suggested extensiveness assessment** *(to be filled in during assessment)*	**Suggested assessment of other quality dimensions** *(to be filled in during assessment)*	**Suggested substantiation by documentation** *(to be filled in during assessment)*
Study population	Inclusion criteria *Criterion 1* … *Criterion n* Exclusion criteria *Criterion 1* … *Criterion n*	Data elements required for I/E criteria	Low/Medium/High	Completeness metrics	Reliability metrics (Precision)	As relevant
	Cohort size	Minimum cohort size	Low/Medium/High	N/A (to be assessed on a research question basis)	N/A	N/A
	Representativeness of the target population	Population characteristics for which similarity to those of the studied sample is important	Low/Medium/High	Coverage metrics	N/A	As relevant
Treatment/exposure		Data elements required	Low/Medium/High	Completeness metrics	Reliability metrics	As relevant
	Newly treated population size	Minimum number	Low/Medium/High	N/A (to be assessed on a research question basis)	N/A	N/A
Comparator group (if relevant)		Data elements required	Low/Medium/High	Completeness metrics	Reliability metrics	As relevant
	Size of comparator sample	Minimum number	Low/Medium/High	N/A (to be assessed on a research question basis)	N/A	N/A
Key endpoint	Key endpoint 1 … Key endpoint n	Data elements required	Low/Medium/High	Completeness metrics (overall and over time)	Reliability metrics Coherence metrics Timeliness metrics	As relevant

Confounders (if relevant)	Confounder 1 … Confounder n	Data elements required	Low/Medium/High	Completeness metrics (overall and over time)	Reliability metrics Coherence metrics	As relevant
Follow-up time (if relevant)		Minimum follow-up	Low/Medium/High	Coverage metric on follow-up time	Timeliness metrics (overall or for specific variables if relevant)	As relevant
Lag time		Maximum lag time	Low/Medium/High	N/A	Timeliness metrics (overall or for specific variables if relevant)	N/A

5.3. Illustrative example for detailed fitness-for-use assessment

An example is provided here for a Chronic Lymphocytic Leukaemia External Comparator study based on multi-site EHR. Note that this is purely an illustrative use case to demonstrate how to use the framework for step 3 (See

Table 8).

Table 8 - Detailed fitness-for-use assessment for a Chronic Lymphocytic Leukaemia study

Design elements *(pre-specified)*	**Operational definition** *(pre-specified)*	**Data elements for valid capture** *(pre-specified)*	**Criticality of the quality of the element, including justification where relevant** *(pre-specified)*	**Extensiveness assessment** *(filled in during feasibility assessment)*	**Other quality assessment** *(filled in during feasibility assessment)*	**Documentation** *(filled in during feasibility assessment)*
Study population	Age >18 years at time of enrolment	Date of birth (month)	High	100% of patients have DOB available, or age at registration to the RWD source	100% of patients have DOB captured in the same month and year format (MM/YYYY)	Documentation on the RWD source's target population age range and format of age/DOB capture (if available)
	Confirmed diagnosis of CLL	Physician diagnosis (ICD 10 code or equivalent)	High	100% of patients have a CLL diagnosis	100% of diagnoses have been mapped to ICD-10	Documentation on mapping of different coding systems to ICD-10
		Lab results	Medium	40% of patients have confirmatory lab results	100% of lab results are within a plausible range	Documentation on consistency of lab assessments

						across different sites
	Known 17p deletion status	17p deletion status	Low (possibility of using *probabilistic bias methods using published prevalence of 17p deletion and its association with selected endpoints to derive subgroup estimates)*	70% of patients have known 17p deletion status	Time lag (6 months) between 17p deletion availability and initial diagnosis date	N/A
	Cohort size	5000 patients	Medium	6000 patients after application of I/E criteria of interest	N/A	N/A
	Representativeness vs target emulation population (from RCT)	Average age in acceptable +/- range compared to target population	High *(bias towards worse outcomes if older population)*	Average age is 83 vs 82 in target population	N/A	N/A
Treatment/ exposure	Received a BTKi	Treatment information (BTKi)	High	BTKis are in the list of drugs covered by this database, and 90% of patients have at least 1 record of treatment	95% of cancer treatment information has a date after diagnosis of CLL. 90% of treatment records pass uniqueness checks	N/A
	Number of newly treated	300 patients receiving BTKi after confirmed diagnosis of CLL	High	315 patients	N/A	N/A
Comparator group (if relevant)	Received best supportive care	Absence of anti-cancer treatment	High	Explicit negation of treatment received only	N/A	N/A

				for 20% of patients		
	Number in comparator	300 patients who are on best supportive care	Medium	270 patients	N/A	N/A
Key endpoint	Overall Response Rate	Response per criteria at intervals	High *(consistency of response assessment essential for primary outcome)*	85% of patients with at least one assessment 40% with 3 assessments or more	N/A	Documentation detailing re-assessment of response by adjudication committee for homogeneous assessment
		Treatment regimen and/or cycle start date	Medium	90% of patients with start date available	Variable only available at month level for 50% of records	N/A
	Overall Survival	Date of death	High	20% of patients with known death have date of death	Statistical checks for reliability of linkage to death registry passed for 100% of patients	Linkage process documentation Traceability documentation
		Treatment regimen and/or cycle start date	Medium	90% of patients with start date available	Variable only available at month level for 50% of records	N/A
	Number of participants with AE	AE capture across patients during follow-up period	High	30% of patients with AE data available	N/A	Documentation detailing method for AE capture, and which AEs are to be captured
Confounders (if relevant)	Sex	Sex, reported as male or female	High	100% of patients have sex information available	100% of patients reported with a pregnancy are classified as	Documentation detailing method for sex capture

					females 100% of patients have sex captured as one ASCII character (0 or 1)	
	Cancer stage	Stage I-IV	Medium	80% of patients have at least one stage record 30% have stage at each line of therapy	Distribution of staging observed to be different pre- and post-2017 (due to update in guidelines)	Documentation of internal guidelines for consistent stage assessment
Follow-up time (if relevant)	6 months	6 months follow-up	Medium	80% of patients have >6 months follow up	N/A	Documentation of internal guidelines for typical follow-up time (if available)
Lag time	2 years	2 years maximum	Medium	There is on average a 3-year lag time to perform data curation and linkage	N/A	Documentation of the RWD source about lag time

DOB: date of birth; CLL: Chronic Lymphocytic Leukaemia; N/A: not assessed; I/E: in- and exclusion criteria; BTKi: Bruton tyrosine kinase inhibitor; AE: adverse event.

When assessing the fitness-for-use of the RWD source, this table can provide guidance in making a final decision on the suitability of a dataset for a given study, or prompt changes in the method/protocol when necessary to leverage available data.

5.4. Toward a generalisation of question-specific aspects

The maturity model for "question-specific determinants" in the EMRN DQF suggests the definition of typical use cases that could be the basis for some more standardised approach to DQ acceptance processes (e.g.: more automated domain-based quality requirement and test packages, as well as standardised reporting). As a step in this direction, the design elements to be pre-specified (outlined in Tables 7 and 8) (i.e., study population, treatment/exposure, comparator group, key outcomes, confounders, follow-up time, lag time) are relevant for different study questions to guide the specific DQ assessment. For example, confounders can be considered a relevant design element across various study types (e.g., clinical management and unmet needs, drug utilisation, disease epidemiology, etc), while comparator group(s) may only be relevant for comparative effectiveness studies. These typical design elements, which might also impact regulatory actions, are important but not prescriptive.

5.5. Providing supporting information for RWD in regulatory submissions

Where RWD are used for regulatory submissions, the data source characteristics allowing for DQ assessment should be provided, with the adequate granularity for regulatory assessment. Relevant characteristics include:

- Information on the standard quality management practices routinely applied to the data, as well as the processes and methods behind the generation of data at the source. This includes details on the level of automation and the use of computerised systems and can be relevant to:
    - Data cleaning, extraction, or transformation steps.
    - Data quality checks to detect logical inconsistencies and erroneous, missing or out-of-range values.
    - Any remedial actions taken at the RWD source level. Information on the data collection or recording process and any selection mechanisms involved (e.g., inclusion of specific patients, capturing of specific clinical data).
- Measures taken to improve the completeness of data elements (e.g. data collection prerequisites for reimbursement);
- Any linkage performed on the data, including details on the data elements used for linking and the linkage methodology;
- Information on whether patient-level data or only aggregate data are available.

This information can be made publicly available for transparency by the data holder by registering the RWD source in the HMA-EMA Catalogue of RWD Sources. The Catalogue is a repository of metadata collected from RWD sources and contains information on the systems and processes behind data capture, as well as descriptors of the data. It is intended to capture the extent of variety of existing RWD sources and facilitate fit-for-purpose assessments [21].

To allow for an adequate DQ assessment, it is important that the information published in the Catalogue of RWD Sources remains up to date, with the last update occurring within the past 12 months. This provision may be included in the contractual agreement between the MAH or Applicant and the data holder, as relevant.

6. Concluding remarks

This document is an extension of the generic EMRN DQF for medicines regulation, focusing on RWD specificities. The RW-DQF provides background on how DQ can impact the use of RWD to generate RWE for regulatory assessment. It further provides guidance for the characterisation of systems and processes underpinning data and their impact, key metrics to assess different aspects of DQ within a dataset as well as guidance to use metrics to assess the suitability of a dataset by a fitness-for-use assessment in relation to a specific research question. These parts provide actionable and focused recommendations for assessing DQ of RWD with the goal of improving the usefulness of RWE for regulatory purposes.

7. References

1. Data Quality Framework for EU medicines regulation. 2022. Available from: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en.pdf.

2. Reflection paper on use of real-world data in non-interventional studies to generate real-world evidence. 2024. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-use-real-world-data-non-interventional-studies-generate-real-world-evidence_en.pdf.
3. Cave, A., X. Kurz, and P. Arlett, Real-World Data for Regulatory Decision Making: Challenges and Possible Solutions for Europe. Clin Pharmacol Ther, 2019. **106**(1): p. 36-39.
4. European Commission. Can we use data for another purpose? 2022. Available from: https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr/purpose-data-processing/can-we-use-data-another-purpose_en.
5. RWE in regulatory assessment and decision-making processes. A report on the experience with regulatory-led RWD studies. https://www.ema.europa.eu/en/documents/presentation/presentation-report-experience-regulatory-led-rwd-studies-s-prilla-ema_en.pdf. 27 June 2023.
6. European Health Data Space Data Quality Framework, Deliverable 6.1 of TEHDAS EU 3rd Health Program (GA: 101035467). https://tehdas.eu/app/uploads/2022/05/tehdas-european-health-data-space-data-quality-framework-2022-05-18.pdf.
7. Rudrapatna, V.A. and A.J. Butte, Opportunities and challenges in using real-world data for health care. J Clin Invest, 2020. **130**(2): p. 565-574.
8. Dusetzina, S.B., Tyree, S., Meyer, A.M., Linking Data for Health Services Research: A Framework and Instructional Guide. Rockville (MD). Chapter 4 An Overview of Record Linkage Methods. Agency for Healthcare Research and Quality (US). Available from: https://www.ncbi.nlm.nih.gov/books/NBK253312/, 2014.
9. Big data use for public health: publication of Big Data Steering Group workplan 2022-25. European Medicines Agency. 2022. Available from: https://www.ema.europa.eu/en/news/big-data-use-public-health-publication-big-data-steering-group-workplan-2022-25.
10. Kent, S., et al., Common Problems, Common Data Model Solutions: Evidence Generation for Health Technology Assessment. PharmacoEconomics, 2021. **39**(3): p. 275–85.
11. Rivera, D.R., et al., The Friends of Cancer Research Real-World Data Collaboration Pilot 2.0: Methodological Recommendations from Oncology Case Studies. Clin Pharmacol Ther, 2022. **111**(1): p. 283-292.
12. The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). Guide on Methodological Standards in Pharmacoepidemiology (Revision 11). EMA/95098/2010. Available from: https://encepp.europa.eu/encepp-toolkit/methodological-guide_en.
13. Guideline on registry-based studies - Scientific guideline. European Medicines Agency. 2021. Available from: https://www.ema.europa.eu/en/guideline-registry-based-studies-scientific-guideline.
14. 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.
15. Kahn, M.G., et al., A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC), 2016. **4**(1): p. 1244.
16. Schmidt, C.O., et al., Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med Res Methodol, 2021. **21**(1): p. 63.
17. NESTcc. Data Quality Framework, A report of the Data Quality Subcommittee of the NEST Coordinating Center - An initiative of MDIC. 2020 [February 14th, 2022]. Available from: https://nestcc.org/nestcc-data-quality-framework.
18. REQueST Tool and its vision paper [Internet]. EUnetHTA. 2019. Available from: https://www.eunethta.eu/request-tool-and-its-vision-paper/.
19. Documentation:Getting_started [observational health data sciences and informatics. Available from: https://www.ohdsi.org/web/wiki/doku.php?id=documentation:getting_started.
20. Observational Health Data Sciences, Informatics. Chapter 15 Data Quality. 2021. Available from: https://ohdsi.github.io/TheBookOfOhdsi/DataQuality.html.
21. Good Practice Guide for the use of the Metadata Catalogue of Real-World Data Sources. 2022. Available from: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/good-practice-guide-use-metadata-catalogue-real-world-data-sources_en.pdf.
22. Compassionate use. European Medicines Agency. 2024. Available from: https://www.ema.europa.eu/en/human-regulatory-overview/research-development/compassionate-use.

Definitions

Abbreviation	Definition
CDM	Common data model
DQ	Data Quality
DQF	Data Quality Framework
EHDS	European Health Data Space
EHR	Electronic Health Record
EMRN	European Medicines Regulatory Network
ENCePP	The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance
ETL	Extract, Transform, Load
I/E criteria	Inclusion/Exclusion criteria
KPI	Key Performance Indicator
QMS	Quality Management System
RWE	Real-World Evidence
RCT	Randomised Clinical Trial
RWD	Real-World Data
SLA	Service Level Agreement
SOP	Standard Operating Procedure
TEHDAS	Towards European Health Data Space

Glossary

The detailed definitions and concepts, with accompanying examples are found in the EMRN DQF [1]. However, to facilitate the reading of this document, a glossary addressing frequently used terms is provided below:

Definition	Explanation
Coherence	Coherence (also referred to as Consistency) is defined as the dimension that expresses how different parts of an overall dataset are consistent in their representation and meaning.
Data linkage	Data linkage is the process of bringing information from different data sources together for the same person/identifier or entity to create a new, richer dataset. Data linkage allows researchers to exploit and enhance existing data sources without the time and cost associated with primary data collection. Linked data can be used to supplement studies by creating population-level cohorts with longer follow-up and can answer questions that require large sample sizes

Definition	Explanation
	(e.g., for rare diseases) or whole population coverage (e.g., for pandemic response planning).
Extensiveness	Extensiveness is defined as the dimension capturing the amount of data available.
Foundational determinants	A characterisation of the systems and processes underpinning data generation and manipulation that have an impact on DQ.
Intrinsic determinants	DQ aspects that can be observed only on the basis of a given dataset, without requirement for information about how the data were captured, or about its primary/intended use.
Maturity model	Provide guidance as to how determinants (foundational, intrinsic, and question-specific) can be characterised in successive levels of maturity.
Plausibility metrics	Indicators of plausibility that can be used as proxy to detect errors. When some combination of information is unlikely (or impossible) to happen in the Real-World this reveals accuracy issues. For example, a weight of a person exceeding 300 kg is possible, but the weight of many or all persons in a dataset exceeding that value is implausible and likely revealing some errors in the measurement or the processing of the data.
Primary use of data	Primary use of (electronic) health data is the processing of personal health data for the provision of health services to assess, maintain or restore the state of health of the person it belongs it, including the prescription, dispensation and provision of medicinal products and medical devices, as well as for relevant social security, administrative or reimbursement services.
Question-specific determinants	Aspects of DQ that cannot be assessed independently of a research question.
Relevance	Relevance is defined as the extent to which a dataset presents the data elements useful to answer a given research question.
Reliability	Reliability is defined as the dimension that covers how closely the data reflect what they are directly measuring.
Representativeness	Representativeness is defined as the data having the same characteristics as the whole it is meant to represent.
RWD end users	People getting access to and using RWD for secondary purposes, such as using RWD from multiple RWD sources as external comparators to a clinical trial arm, in submissions to regulators and payers/HTAs, using RWD from multiple RWD sources to assess the RW safety of a treatment across geographies and ethnicities, etc.
RWD holder	People owning and or holding the RWD

Definition	Explanation
RWD practitioners	People involved in the RWD collection or recording process such as researchers, data analysts and data custodians.
RWD submitter	RWD submitters are the RWD end users or stakeholders that get access to RWD data and use if for secondary purpose to answer a research question.
Secondary use of data	Secondary use of (electronic) health data is the processing of health data for other purposes rather than primary use such as national statistics, education/teaching, scientific research etc. The data used may include personal health data initially collected in the context of primary use, but also electronic health data collected for the purpose of secondary use.
Quality Management System	A QMS is a formalised approach adopted by an organisation that documents processes, procedures, and responsibilities for achieving quality policies and objectives (e.g., Good Clinical Practices, Good Laboratory Practices or Good Manufacturing Practice. It achieves these quality objectives through quality planning, quality assurance, quality control and quality improvement. Standards like the ISO 9000 family define QMS across industries, while more specific QMS have been developed for specific industry or products.

739