27 March 2026
EMA/503781/2024
Committee for Human Medicine Products (CHMP)

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

Draft agreed by Methodology Working Party (MWP)	September 2024
Adopted by CHMP for release for consultation	November 2024
Start of public consultation	29 November 2024
End of consultation (deadline for comments)	28 February 2025
Agreed by NDSG	12 December 2025
Agreed by MWP	18 December 2025
Adopted by CHMP	16 March 2026

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

Table of contents

Executive summary

This document describes the Real-World Data (RWD) specific recommendations as derived from the European Medicines Regulatory Network (EMRN) 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 Data Quality (DQ) in the regulatory context. It is therefore recommended to use the EMRN DQF as a companion document when reading this document. Unless otherwise specified, the definitions provided in the EMRN DQF also apply to this document.

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, as well as data reliability assessment 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 supporting both the regulatory submission and assessment process and ultimately facilitate the use of RWE for regulatory purposes.

The topics addressed in this document are organised into six sections:

1. Introduction to RW-DQF, including its purpose, scope and structure (Section 1)
2. Introduction to RWD (Section 2)
3. Key DQ considerations on RWD quality (Section 3)
4. Practical recommendations for characterising systems and processes that underpin data (Section 4)
5. A structured set of metrics for assessing DQ dimensions (Section 5)
6. A guideline on how to assess DQ in relation to a research question, including an illustrative example (Section 6).

The RW-DQF also includes three practical elements designed to assist the reader in its implementation (Figure 1).

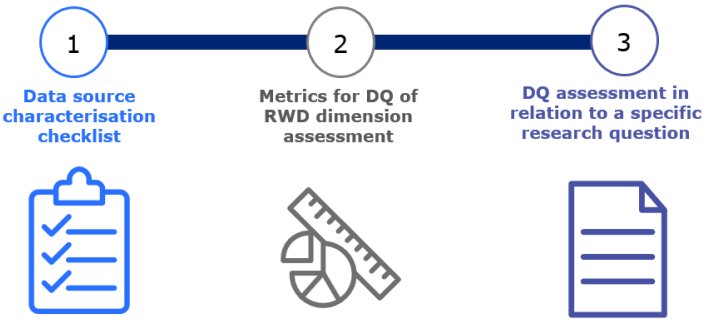

Figure 1 - Representation of the key points of the RW-DQF

1. Application of the EMRN DQF to RWD

1.1. Purpose of this document

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 builds upon existing regulatory guidance, such as EMA's recommendations on RWE and ICH E6 (2) DQ principles, by offering specific methodological approaches tailored to secondary data use in regulatory decision-making.

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 with consultations throughout the drafting period. These initiatives aim to address the wider use of health data, whereas the RW-DQF specifically focuses on the challenges faced when using RWD within the medicine regulation assessment.

The RW-DQF serves as a prescriptive 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 to assess the suitability of a dataset through a fitness-for-use assessment in relation to a specific research question. In addition, example checklists that readers may choose to use as-is or adapt to suit their own case are also provided.

EMA strongly encourages the application of the RW-DQF as a best practice framework for guiding the assessment of RWD quality in regulatory contexts. While the RW-DQF provides structured methodologies and tools to support robust and transparent evaluation of RWD, it does not prescribe fixed data quality thresholds or acceptance criteria, nor prioritises specific metrics. Instead, it promotes a contextual approach that enables assessors to determine the appropriateness of a dataset considering the context of their research purposes. This flexibility ensures that the framework can be applied across diverse data sources and use cases, while maintaining alignment with evolving regulatory expectations.

The RW-DQF is intended for stakeholders and actors involved in regulatory processes, such as members of the EMRN, the Data Analysis and Real-World Interrogation Network (DARWIN EU®), pharmaceutical industry, contract research organisations, and other data holders. Additionally, the document addresses other interested stakeholders involved in various health data use cases, such as academia, patient advocacy groups, and health technology assessment (HTA) bodies.

1.2. Scope of the RW-DQF

This document applies to RWD recorded within routine clinical practice 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 (3) and the ICH M14 Guideline (4), including datasets available through the Data Analysis and Real-World Interrogation Network (DARWIN EU) and the EHDS. The following are examples of clinical practice data which could be considered in scope: administrative or claims data, Electronic Medical/Health Records (EMRs/EHRs), pharmacy/prescription data, patient registries, etc. Existing guidance specific to particular data types (e.g., guidance on registry-based studies (5) for the use of patient registry data) still applies and the RW-DQF should be read in conjunction with any other relevant guidance documents.

RW-DQF is intended to serve as a structured guide on the DQ elements to be assessed and documented for regulatory purposes. It is not designed to function as a detailed SOP or methodology manual. Methodological guidance is already covered in other existing guidance documents — such as the ENCePP guides — which are referenced throughout this document, where relevant.

For the current version, the following should be considered out of scope:

- DQ concerning RWD arising from repurposing of previously published analyses, e.g., meta-analyses
- Quality of certain specialised data, e.g., direct-from-patient (direct-from-user) data e.g., PROs, patient engagement data, patient preferences, mobile health data, social media data), in-vitro diagnostics, herbal products, – as these may have peculiarities in terms of measuring DQ. Therefore, while such data may fall within the broader scope of RWD/RWE, they are not the primary focus of the current RW-DQF chapter.
- Comparative assessment of DQ based on data origin or collection approach (e.g., primary vs. secondary, prospective vs. retrospective) is not within the scope of this chapter. The RW-DQF focuses on evaluating DQ itself, independent of how or why the data were collected.

While the value of different data sources is recognised, the RW-DQF is intended to be neutral and applicable across all types of RWD sources. Therefore, this chapter does not focus on specific types of data source (being generally applicable), and the reader should consult different guidance for this purpose. For example, guidance on the use of registries in regulatory contexts is provided in the EMA Guideline on Registry-Based Studies (5).

1.3. Structure of the RW-DQF

The document comprises six sections. The practical guidance is organised into three core components, presented in Sections 4–6::

- Guidelines to characterise systems and processes (and their impact) (Section 4)
- Metrics to appraise different aspects of DQ within a given dataset (Section 5)
- Guidelines to assess the suitability of a dataset for answering a specific research question (Section 6).

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. While these foundational principles remain consistent, this document adapts them to the unique challenges of RWD, such as heterogeneous data sources, retrospective data use, and varying levels of quality control at the point of data entry. The general principles 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 document, 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 technical database specification documents (6). In addition, the impact of systems and processes is usually considered with respect only to data collection (7) and then approximated (8). This document considers the impact of systems and processes in some detail, as well as the impact along the whole evidence

generation process. Section 3 introduces a checklist that specifies what information should be provided regarding foundational determinants, as to make decision-making more efficient.

- **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 document, intrinsic determinants are considered as **Metrics** that can be used to characterise and quantify DQ, and the document provides guidelines on how to use such metrics. Section 4 introduces a framework and some sample metrics to help build metrics-sets to quantify DQ.
- **Question-specific determinants** are defined as "aspects of DQ that cannot be assessed independently of a research question". This document offers guidance on how to assess the **suitability of a dataset for a specific research question**. Section 5 proposes a 2-step method to assess the suitability of a dataset for a specific research question and analysis method.

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 (9-11). In general, there is a lack of consensus on terminology, among RWD practitioners (researchers, data analysts, data custodians, etc.) 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 relevant external benchmarks (9), but could be also 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 also reported in the glossary of this document (Section 10) for convenience.

Table 1 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 - Impact of processes, metrics and user questions on different dimensions for DQ[1]

	Are data correct? ***Reliability***	**Are data sufficient?** ***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 (detect reliability issues)	Can measure extensiveness	Can measure some aspects of coherence	Can assess timeliness only in limited cases[2]	
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

[1] In the terminology of the EMRN DQF, this table highlights at a high level to what degree the three determinants of DQ (foundational, intrinsic or question specific) affect different DQ dimensions.
[2] Metrics can assess timeliness only if relevant information (e.g.: timestamp) is available in the data or in its available metadata.

2. Background: RWD and Data Quality

2.1. RWD

In the context of RWE studies, RWD are data that describe patient characteristics (including the disease, the treatment, interactions with the healthcare system, as well as social and environmental factors influencing health status) in clinical practice (6, 12, 13).

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). RWD may be collected prospectively—planned and initiated moving forward—or retrospectively, using data already recorded in various sources. Importantly, DQ assessment is expected to be equally relevant and necessary regardless of the data's origin or collection approach—whether primary or secondary, prospective or retrospective.

The secondary use of data, driven by specific research objectives, can involve a single or multiple RWD sources (i.e., multi-database studies) which may require additional standardisation efforts to address variations in terminology, data formats, and regulatory constraints.

Through the analysis of RWD, RWE is generated to answer research questions. RWE can aid medicines regulatory use cases across several phases of the medicinal product's lifecycle (6) such as:

- support planning and validity of studies (e.g., comparison of a study population with patients from the real-world setting to ensure representativeness of the clinical study, generalisability of trial evidence to broader patient populations, inform patient recruitment),

- understand the clinical context (e.g., disease epidemiology, disease prevalence/incidence, description of drug utilisation patterns including switching and off-label use, medication adherence),

- investigate associations and impact (e.g., medicines post-marketing surveillance, impact of policies and regulatory decisions, patient-important outcomes not investigated in trials, assessment of effectiveness of risk minimisation measures, assessment of treatment effectiveness, comparative benefits and harms, in the absence of active-controlled trials) (15)., assessment of outcomes in populations underrepresented in clinical trials (e.g. pregnant women) or outcome assessment by analysis of longitudinal data.

Ensuring the quality of RWD is therefore essential to generate reliable evidence that can confidently inform regulatory decisions across the lifecycle of medicinal products.

Note that while RWD analyses may also be used for other purposes beyond regulatory, discussing these is not the scope of this document.

2.2. Distinctive traits of RWD

RWD have several distinctive traits, that influence their usability and quality. These include:

- source and scope diversity: RWD are derived from multiple sources such as electronic healthcare databases, registries, and claims data; all impacted by the organisation of the healthcare systems.

- structural and semantic heterogeneity: Variability in data formats, terminologies, and coding standards.

- data governance considerations: differences in collection methodologies, recording procedures, data access, and regulatory compliance requirements.

Given this heterogeneity, reproducibility challenges may arise—similar data collected across different datasets can yield varying results. Establishing common data standards and harmonisation practices can help mitigate these issues.

In cases of secondary use, the data may not be suitable for the intended research, as their primary purpose is typically clinical (i.e. medical record data), and different from the purpose of the study. Methods such as data harmonisation, statistical validation, and interoperability frameworks can help mitigate these limitations and improve data suitability for research applications. However, if the data elements collected do not correspond to those required to answer a specific research question, such as population characteristics, exposure, outcomes and covariates, the data source is not fit.

A distinctive trait of RWD is that data access falls under two main models (direct and indirect data access) with different implications on how to manage and describe data quality. Under the indirect data access, data holders may restrict access to patient-level data to third parties. Thus, third parties rely on data holder reported information, documentation, and aggregated results which can limit the evaluation of data quality.

The metadata of RWD for RWE generation is typically included in data catalogues for dissemination, such as the HMA-EMA Catalogue of Real-World Data Sources [3]. These metadata provide important context for fitness-for-use without referring to a specific research question.

2.3. Quality control of RWD

The main 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, preparing for subsequent secondary use of data. The capture/collection of RWD for primary use should be undertaken in accordance with the principles of a Quality Management System (QMS) as explained in more detail in the EMRN DQF, including dimensions and sub-dimensions (1). Furthermore, depending on its purpose and proportionate to its impact on data interpretation and/or regulatory decision-making, sponsor oversight and patient-level data may be needed for quality control and assurance and regulatory inspections (e.g. external control arms).

When RWD are leveraged for RWE generation - whether as a primary or secondary use - they can sometimes not be produced through a standardised regulatory quality-controlled process. However, certain data governance frameworks and validation mechanisms may still apply, depending on the data source and intended use. Current QMS such as the ISO 9000 family, Good Clinical Practices, Good Laboratory Practices or Good Manufacturing Practice is appropriate for assuring quality related to the primary collection of data and are not applicable for the secondary use of data.

2.4. 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, RWD quality is a distributed responsibility (see Table 2). The responsibility to ensure that DQ is properly characterised is divided among all actors in the RWD lifecycle. 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.

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

Table 2 – High-level responsibility matrix for RWD quality, divided by stakeholder group[4]

Data holder	Data user	Regulator
• Document data provenance and implement QMS (systems and processes) • Provide DQ metrics about the data • Publish metadata (variables, coverage)	• Assess fitness-for-use • Document DQ metrics • Validate and interpret the data	• Provide guidance for DQ • Check data validation processes • Check data interpretation

The RWD lifecycle typically involves multiple steps, as illustrated in Figure 2, ranging from defining data requirements and preparation steps for data consumption. Defining the research question and designing the study are essential steps that must be completed before determining the data requirements. It also includes procurement, testing and acceptance, and delivery for consumption. Additional activities such as transformation, cleaning, aggregation, and metadata creation may occur at various stages, within a single organisation or across multiple organisations. This complexity underscores the need for clear governance and robust quality controls to ensure data integrity and consistency throughout the entire lifecycle.

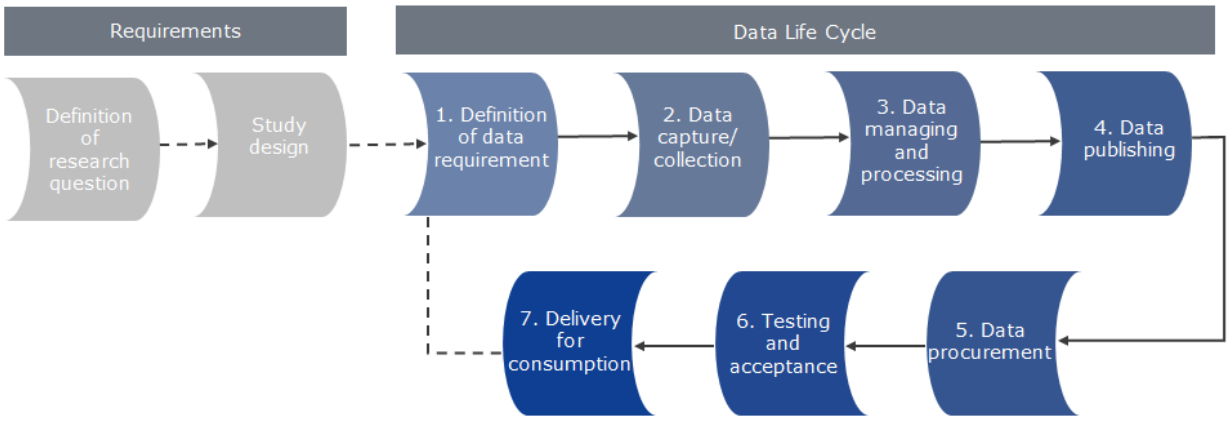

Figure 2 - Data life cycle for secondary use of RWD

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 (3). These responsibilities can be divided into two groups: data holders (e.g., hospitals, registries) and data users (e.g., sponsors, researchers, end-users). Data holders should indicate the purpose for which their data source can be used and document data provenance, which helps data users to select the best data source to answer their research question. It also stimulates better collaboration between data holders and data users as it increases transparency and accountability in DQ management. Without role clarity, it is difficult for

[4] This table provides an example of high-level responsibility matrix for RWD quality by stakeholder group.

regulators to verify if gaps (e.g., missing confounders) stems from data source limitations or user misapplication.

3. Data Quality considerations for the secondary use of RWD

Secondary use of RWD has a significant impact on DQ for the following reasons:

Decoupling of data collection and purpose: (Electronic) health data for secondary use are, by definition, the processing of health data for other purposes rather than primary use such as national statistics, education/teaching, scientific research etc. The secondary purpose may consist of different individual research questions usually not known at the time of data collection. Fitness-for-use depends on a defined research question and therefore cannot be controlled or imposed to a source at the time of data collection. 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. Approaches such as comprehensive metadata documentation, data enrichment techniques, and statistical imputation can help improve secondary data usability.

Anonymisation and pseudonymisation: Sensitive patient-level healthcare data will often need to be anonymised or pseudonymised to facilitate data protection, publication, and re-use by different end users. This process may involve masking, replacing or removing information. Consequently, quality and operational aspects of quality control may be affected given that people with access to identifiable information are usually tightly restricted and may not include staff involved in quality determination. Techniques such as differential privacy and secure multi-party computation can help balance privacy protections with data usability.

In the case of the regulatory use of RWD with significant influence on regulatory decisions (e.g. external control arm of a small pivotal clinical trial for the treatment of a rare disease), quality assurance, quality control and regulatory inspection may become necessary. Since people with access to identifiable information are usually tightly restricted and may not include staff involved in quality determination. Therefore, it may become necessary to obtain informed consent retrospectively from the respective patients for direct access to their health data at the patient level by both sponsor and regulatory authorities.

Data drift: Data distributions may change over time due to evolving clinical practices, coding modifications, or shifts in population characteristics. Consequently, predictive models may lose accuracy because they do not know these new patterns. This change in data patterns is known as data drift (16). This can affect the validity of predictive models and regulatory decisions. To ensure data reliability over time, mechanisms for detecting and mitigating data drift should be implemented. This includes continuous monitoring of changes in data distributions, adjustments for evolving clinical practices and coding modifications, and validation of predictive models to maintain their regulatory relevance.

Data linkage: RWD captured from a single source may not always 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), are often accompanied by data cleaning steps such as *de-duplication of records.* While essential for data linkage, these processes may introduce quality issues. For example, probability-based matching of patients based on non-identifiable or incomplete information can inadvertently combine records from two different patients, resulting in inaccurate information. Linking only a subset of the data may introduce loss of precision. De-duplication may have the effect of introducing incoherence, since both duplicate records are valid in each of the sources. De-duplication may impact accuracy when similar records that represent a different event are inadvertently combined.

On the other hand, when identifiers are available, techniques such as deterministic linkage are typically preferred; in cases where such identifiers are not available, probabilistic linkage methods may be used, though they can vary in reliability depending on the data quality and structure. (7). A methodological framework should be established for integrating RWD from diverse sources, including administrative databases and disease registries. Techniques such as record linkage, standardisation protocols, and best practices for resolving discrepancies should be clearly defined. Where feasible, mechanisms for patient validation of their data (e.g., patient portals for reviewing EHR entries) should be implemented to enhance reliability and trust. Linkage should be carefully evaluated.

Data pooling/merging: While data linkage aims to fill gaps in patient-level information by connecting complementary data sources (e.g., linking EMR and claims data), data pooling or merging focuses on expanding the breadth and scale of available data. This approach involves combining datasets from multiple sources that may cover different populations, geographies, healthcare settings, or disease areas. Pooling enhances statistical power, supports subgroup analyses, and enables broader generalisability of findings. However, it also introduces challenges such as heterogeneity in data models, coding systems, and variable definitions. Harmonisation strategies—such as common data models, standardised terminologies, and robust data transformation protocols—are essential to ensure consistency and comparability across pooled datasets. Unlike linkage, which seeks to enrich individual patient records, pooling aims to increase the volume and diversity of data to support more comprehensive and representative real-world evidence generation.

Common data models (CDMs): 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 allowing implementation of consistent processes for data cleaning, profiling, reporting, analysis applied to different sources (17-19). The application of a CDM occurs after the initial data collection phase. The process to transform an original source to a CDM, called ETL (extract, transform, load) can improve coherence, facilitate standardisation, however, may also introduce data loss or other alterations. For instance, reliability can be affected, both because any transformation increases the risks of error (accuracy) and secondly because often the CDM will define some level of precision that is lower than 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, if the transformation process introduces delays. Implementing traceability mechanisms and validation checks throughout the ETL process can mitigate unintended transformations and ensure data fidelity.

3.1. Impact on Reliability

In a secondary use of data scenario, there is no possibility to control most DQ factors of reliability (e.g., accuracy and precision) at the source (point of data recording), and even the ability to assess reliability is limited. Therefore, the primary focus of DQF implementation is the identification of reliability gaps that could lead to record removal or amendment with approximated values (only in some limited cases it can lead to the correction of the data capture processes) and to maximise the amount of information supporting a reliability assessment.

For RWD datasets that are in the category of "Big Data" as defined by the HMA/EMA joint Big Data Steering Group (20), error detection cannot be practically achieved through manual checking of all data records. Therefore, automation[5] plays a key role in error detection and can sometimes help identify inconsistencies or outliers in the data. As mentioned above, the use of CDM[6] and standardised

[5] By automation we consider the automatic execution of computation checks as part of data processing. As an example, the execution of error detection scripts looking at statistical trends in the data at the time data is ingested into platforms.
[6] The harmonisation to a common data model itself has an impact on precision and potentially accuracy.

analytics can reduce variability in data representation and support consistent analysis (21). However, coding errors or inconsistencies introduced during the data transformation process may persist in the converted dataset – especially if they go undetected and if conversion back to the source data is not possible. Plausibility metrics EMRN DQF (1)[7] also play a key role in assessing the reliability of RWD, as they may provide an alternative to validation against primary institution source records. They cannot fully replace validation, as they can only detect some error classes, but they can still be applied to data for which controls at source are unknown.

Data traceability is an important aspect of reliability and might pose challenges in the context of secondary use of data, given that the data capture system is not developed specifically for the research question.

If data control is deemed necessary due to the importance of the use of the data for regulatory decisions, it can be enabled by means of a subsequent data protection-related informed consent. This can already be communicated in scientific advice procedures for e.g. studies to be used for marketing authorisation applications. Due to the effort involved and the possibility that patients may no longer be contactable or may not wish to give their consent, the proposed risk-based approach is balanced and justifiable.

3.2. Impact on Extensiveness and Representativeness

In a secondary use of RWD scenario, it is possible to assess dataset completeness, but linking this to the data recording process can be challenging, as RWD relies on the availability of recorded data. It is often unknown whether the absence of such data indicates that the outcome did not occur, or it simply was not recorded[8] as for example when an ordered test, was never performed or a prescribed medication was not filled and often requires the assumption that non-present outcomes did not occur. Such an assumption may be more appropriate for major clinical outcomes that have a high likelihood of consistently being recorded versus activities such as routine clinical testing (e.g. cardiovascular event versus routine A1C testing for diabetes diagnostics).

These presumptions, which should be explicit, have implications for the analytical methods of generating RWE and are fundamental for the assessment of DQ (22). These presumptions should also be carefully reflected in the study protocol and report using accurate phrasing (e.g. specifying that a study is investigating the "number of people diagnosed with diabetes" as opposed to the "number of people who have diabetes").

Representativeness is a key aspect of RWD quality. Data collected through routine clinical practice often offers greater real-world representativeness than data generated through primary collection in clinical trials, which typically involve more selective inclusion criteria. The inherent diversity of RWD can enhance the generalisability of insights derived from its use. However, a potential challenge in secondary use of RWD is that the data may not fully reflect the target population, which can affect the generalisability of insights derived from it. For instance, data from a primary care database may not capture the full spectrum of a disease across the population. It is important to assess how the characteristics of the dataset—such as coverage, demographic distribution, and geographic scope—may introduce bias. This issue is not exclusive to secondary use; representativeness can also be a limitation in primary data collection. As part of data quality evaluation, considerations should include the extent to which the dataset reflects the intended population and how potential biases (e.g., demographic imbalances, missing data, regional disparities) are identified and if possible, addressed.

[7] The EMRN DQF provide examples of plausibility metrics, categorised for by their temporal or atemporal aspects. Additional classifications of plausibility metrics are possible (e.g.: based on biological plausibility or based on logical consistency).
[8] This is unlike clinical trial data, where the absence of events is explicit.

Further clarification on the impact of extensiveness (specifically missing data) on data bias and confounding is provided in the literature (7).

Temporal completeness may be affected by change in recording practices, standards and coding.

3.3. Impact on Coherence

RWD are often recorded from different healthcare actors and is varied both due to different data representation, e.g., format, structure, content, and due to differing processes for data collection or data recording and DQ control. This is even more relevant when RWD data is used across geographies and healthcare systems (example being using RWD for an external comparator).

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[9] or prior to consumption[10]. 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.

3.4. Impact on Timeliness

Timeliness reflects whether RWD are available with acceptable currency (i.e. how up-to-date are the data), which is an important consideration for relevance. Different RWD sources are updated and made available to researchers at different time points and timeframes. Furthermore, the data may not be updated in real-time but still be relevant. For example, historic data may lack currency but still be timely for retrospective RWD studies. These considerations on timeliness need to be accounted for in fitness-for-use evaluation.

4. Guidelines for the characterisation of systems and processes underpinning RWD

The quality of RWD 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 feasible if

[9] As for the EMRN DQF life cycle, this corresponds to the time a data provider makes available data for consumption (intended as a responsibility of the data provider, that information provided be coherent within itself).
[10] Prior to consumption refers to the phase "Data procurement" and "Testing and acceptance", intended as a responsibility of the consumer to assess that data is coherent before consumption. This needs to happen latest at the "test and acceptance" phase, after datasets from different sources may be aggregated. It may happen at the "data procurement" phase as well, for instance if a single dataset is not coherent with a larger pool of data and discarded.

there is sufficient 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).

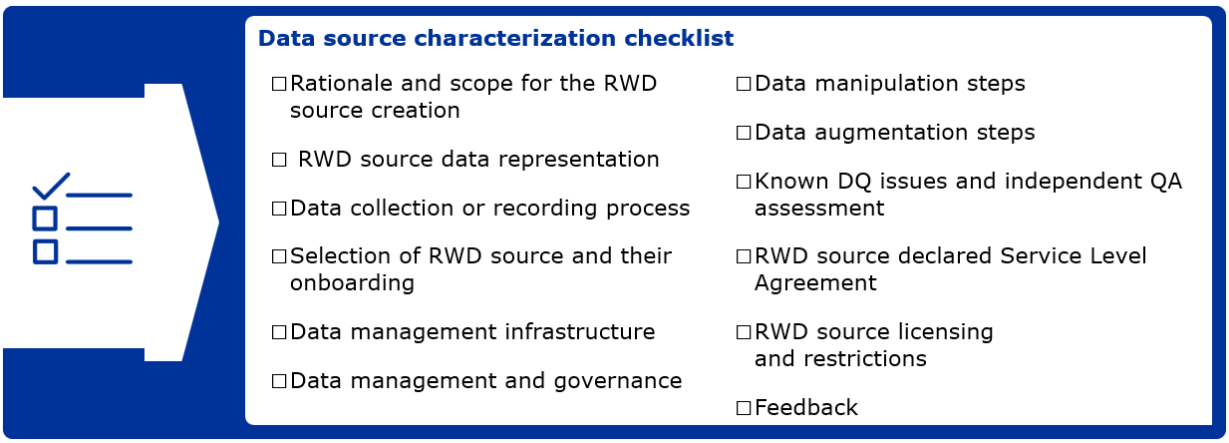

Figure 3 - Overview of system and process characterisation checklist

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 prone to accidental or human errors. Therefore, the maturity model presented here focuses on what information should be provided. Information on how it could be standardised and automated is also described.

4.1. Systems and process characterisation checklist and maturity model

This section suggests how to document and 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, that can be used as checklist to verify if foundational information necessary to assess DQ is provided. Data analyses may sometimes be needed to understand RWD quality. If analyses cannot be performed at the stage of the evaluation, it is important to acknowledge uncertainties and their impact on the evidence submitted (e.g., in the feasibility assessment and/or study protocol).

In line with the EMRN DQF, there are three levels of maturity models (1)[11]:

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) linked to data quality (e.g.: level of duplication). Established standards may not be comprehensive of all data quality considerations described in this document.

[11] The maturity model in the RWD Chapter differs from the EMRN DQF in that "feedback" is not considered as a maturity level, but as an element of systems and processes.

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 (5) or the REQuEST tool (23), or more general frameworks and standards e.g.,: (24). SOPs and KPIs, when reported, follow established standards and guidelines.

Level 3: Implemented and automated[12]. The information should be derived in a way that guarantees higher DQ by design. This means DQ assessment 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. Unstructured data are also collected by automated health data systems and will also require data quality assessment.

Maturity levels are not intended as a classification system for data, but as a guidance for information providers (all actors involved in the data lifecycle) on how to annotate their data, from minimum and accessible[13] requirements, to more mature documentation that can facilitate and accelerate the regulatory process.

For these reasons, in structuring Table 3, we have chosen to indicate what information should be "Documented" in terms of minimal requirements: such information can be provided in any format, including text or web links. The model should account for incremental adoption, enabling data providers to improve their processes over time rather than requiring full compliance from the outset.

What information is considered relevant and is required to be provided depends on the workflows in the data lifecycle (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 considered[14].

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. Therefore, the table proposes to refer to "upstream" quality assessments for all RWD sources under the responsibility of RWD holders. To allow an adequate DQ assessment, data users or RWD end users should be responsible for providing the available evidence on DQ when suitable or required, as part of their submission for regulatory approval.

Finally, the table below proposes to capture information on what feedback mechanisms are provided. The concept of feedback mechanisms is derived from the EMRN DQF overall maturity model, that suggests that mechanisms should be in place to correct DQ issues identified while using processing or using data. Given the complexity and heterogeneity of RWD capture, it may be unfeasible to detail such feedback loops for all aspects of DQ. However, a minimum standard should be established, such as requiring data submitters to provide a description of what feedback mechanisms are made available, should some DQ issues be detected, as well as information on how feedback was

[12] In the EMRN DQF we distinguish four level of maturity: documented, standardised, implemented and automated, where "implemented" requires that DQ information is generated as part of implemented processes, and "automated" requires that such implementation generate "machine readable artifacts". Given the prevalence of automation in RWD, and the role of CDM in standardisation, we considered the distinction between these two levels minimal. When a standard is adopted and implemented, this would likely be based on a CDM ensuring machine readability and automation. Therefore, we considered one single maturity level here.

[13] By requiring documentation in any form, we intend to be inclusive respect to different healthcare realities, while guarantees DQ is suitable for regulatory purposes.

[14] 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.

implemented and processed, e.g.: a structured DQ issue log or periodic validation reports, ensuring a baseline level of feedback on RWD quality.

Table 3: Detailed description of the 12 areas of interest of the system and process characterisation checklist

Reporting Item	**Rationale**		
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 a RWD source.	Minimum requirements	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
		Further suggestions for reporting	1) Provide information using standardised / widely used templates, to make the relevant information easy to digest and interpret[15] 2) Provide information as Metadata, in a standard format and with clear definitions[16], to allow Metadata to be automatically processed, and their quality (e.g.: completeness) be adequately verified.
II. The RWD source representation	Descriptive of the intended coherence of a dataset and its metadata.	Minimum requirements	A) Description of the raw data B) Description of the data model used (including detailed description of non-standard data model) C) Data dictionary and ontologies (vocabularies) in use, with details of the specific versions used. When non-standard dictionaries are used, provide a rationale, and make the full dictionary available D) Information whether patient-level data or aggregate data are captured in the data set
		Further suggestions for reporting	When relevant, dictionaries are provided using standard formats that facilitate the mappings across different vocabularies and across languages.
		Implementation recommendations	*When possible, make use of standard models such as FHIR, OMOP, I2B2, a subset or eventually an extension of such, as well as standard vocabularies (such as ICD-9 to ICD-10 diagnosis).*
III. 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).	Minimum requirements	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.

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

	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).		D) Characteristics of key data elements captured, e.g.: • core, optional elements • planned size • planned coverage over time
		Further suggestions for reporting	• Report information (e.g.: key data elements) in a structured way and making use of standard vocabularies. • Make SOPs available[17] • SOPs are based on common shared standards. • Documentation of patient involvement in data design/use (e.g., advisory boards) • SOPs for integrating patient feedback into data curation • Automated patient surveys linked to data updates.
		Implementation recommendations	*Ideally data collection follows widely shared standards (to facilitate interpretation) that specify key-performance indicators (KPIs) so that adherence to these KPIs can be monitored and reported*
IV. 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 relevance (as well as other dimensions if relative constraints are formulated in inclusion/exclusion (I/E) criteria)	Minimum requirements	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 version of the dataset (or equivalent information that can characterise its state) should be reported.
		Further suggestions for reporting	The DQF assessment includes this checklist for each RWD source, so that a full "chain of evidence" of an RWD resource can be provided in a homogenous way.
V. The data management infrastructure	Essential for reliability regarding data alterations resulting from	Minimum requirements	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[18] testing and QA processes in place.

[17] If SOPs cannot be made available, provide justification
[18] We assume the HW testing is not an issue as necessarily performed a-priori.

	system accidents, software errors or malicious intervention.		C) Measures to prevent accidental physical data alterations (e.g.: backups, redundant systems, checksums).
		Implementation recommendations	*The hardware or software implementation complies with recognised quality standards that can be reported.*
VI. Data management and governance	Data reliability is impacted by data management and governance aspects. These elements also have a broad impact on Metadata quality[19].	Minimum requirements	A) A description of the overall data management principles adopted (e.g.: ALCOA+, FAIR) 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
		Further suggestions for reporting	The representation of metadata follows FAIR standards (25).

		Implementation recommendations	• SOPs and data management processes adhere to standards that can be referred to: e.g., GCP, ENCePP, ISO 25012, ISO 25011, ISO 8000-6x, ISO 25024:2015.
			• Data management and governance is implemented in the data platforms 'Digital Quality Measures' (DQMs) so that reports of performance and deviations are automated. Submitted metadata are generated "by design"
			• Best practices for data representation with a direct impact on DQ are adopted, e.g.:
			  ○ explicit negation is used to disambiguate missing or negative values
			  ○ absolute values are reported when they are a more robust alternative to changing values (e.g.: D.O.B instead of age as a person characteristic)
			  ○ Time based information is associated to related events (e.g.: distance of a relapse from the last visit).

VII. Data manipulation steps[20]	Impacts reliability both in terms of accuracy (possible errors) and precision (i.e., the degree of	Minimum requirements	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

[19] Systems and processes impact all DQ dimensions for metadata, as they are its source.
[20] By "data manipulation" we consider transformations that, in the absence of error, don't affect data reliability: e.g.: unit of measures conversion.

	approximation by which data represents reality). Essential to ensure traceability of information. Also impacts coherence and potentially timeliness.		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). • 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) If AI/ML has been used to transform data, steps that have been taken to guarantee the correctness of outputs. E) Lineage information • Provide justification for the level of data manipulation. • Provide lineage information to specified level sought.
		Further suggestions for reporting	• KPIs for data cleaning (e.g., data duplications, mislabelling, etc.) are provided. • Transaction logs are available including deviations and actions that required manual intervention • Data mapping tables and algorithms are described with a standard characterisation of their performance. • Lists of standard test batteries used to detect loss of accuracy or precision are provided • Provide data transformation scripts, with documented logic, via version control systems (e.g.: git), with full auditing (actual data transformation code is accessible and verifiable). • All lineage information is provided as metadata associated to the dataset
		Implementation recommendations	• *Quality checks, KPIs reported, and lineage information are automatically generated by the data platform (e.g.: unit testing)* • *Tests performed follow some standard or shared set of tests, that can be re-used across RWD sources.*
VIII. Data augmentation steps[21]	Data augmentation steps impact accuracy.	Minimum requirements	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.

[21] We consider here data transformations that produce new information subject to reliability issues: e.g.: imputation of missing values, augmentation, extraction of codes via natural language processing. These transformations may restrict regulatory use of data.

			• Algorithm such as name, source description and justification for use. B) If AI/ML has been used to transform data, steps that have been taken to guarantee the correctness of outputs.
		Further suggestions for reporting	• 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, or data dictionary). • When AI/ML have been used, references to source information is provided.
IX. Known quality issues and independent[22] QA assessment of the RWD source	Explicit description of known DQ issues, as well as external validation performed (all dimensions affected)	Minimum requirements	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
X. The RWD source declared Service Level Agreements (SLAs)	Descriptive of guaranteed timeliness and possible variations of extensiveness/reliability provided.	Minimum requirement	A) Data resource declared SLAs • 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 for data enhancement of variables already collected, or for the collection of data on variables not captured thus far.
		Further suggestions for reporting	• Provide details of established data processes followed by the SLA provider • SLA compliance is automatically assessed and reported
XI. The RWD source licensing and restrictions	Descriptive of aspects that can limit extensiveness 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.

[22] Independent: distinct assessment, from another entity, that can be used as additional evidence to support data quality (not intended as a request for audit)

		Implementation recommendations	*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	Minimum requirements	A) Provide a contact for QA and follow-up on DQ issues detected.
		Implementation recommendations	• *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).

4.2. General considerations on the characterisation of systems and processes

Since the recording or processing of data may have an impact on DQ, every stakeholder involved in any such process should therefore ensure that their actions adhere to the checklist above. To enhance compliance, a structured validation mechanism should be established, including self-assessment procedures, external audits, and periodic regulatory reviews to verify adherence to data quality requirements.

At a time of submission, the above checklist should be provided as to cover all the lineage of the presented data. When some information is deemed confidential (e.g. SOPs) such information should nevertheless be present and made available if required for audit or inspections.

Generally, an end user preparing a dataset to support regulatory activities would provide the above checklist for any eventual processing they did on the datasets, plus a checklist for each data source used, as 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.

5. Data quality metrics for RWD

Metrics are considered in the DQF as "intrinsic determinants" and defined as all DQ aspects that could be assessed, 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. We first introduce an overall framework that groups DQ metrics in terms of the dimension they assess, and in terms of what information they can leverage (e.g.: only the dataset, additional information available as metadata, general know-how that can be used to assess plausibility, or reference data or standards to compare to). In practice, not all such approaches are possible, and the intent of this framework is to suggest different ways to produce metrics, to 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 (9) 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 the characterisation of RWD, along with RWD-specific examples to illustrate the potential outputs of how these metrics can be applied.

5.1. 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. Figure 4 illustrates the framework as a table of example metrics, organised by the type of checks applied and the DQ dimension they assess.

The figure is intended to serve as a flexible guide to identify possible metrics ("*what should be measured and how it could be done*"), or to assess existing test sets. Note that not all metric groups apply to every dimension. The figure omits "relevance" as a dimension as this applies to all dimensions. While some metrics could be devisable (e.g.: % of protocol endpoints covered by the data source), metrics would need to be defined with respect to classes of problems that go beyond this version of this document.

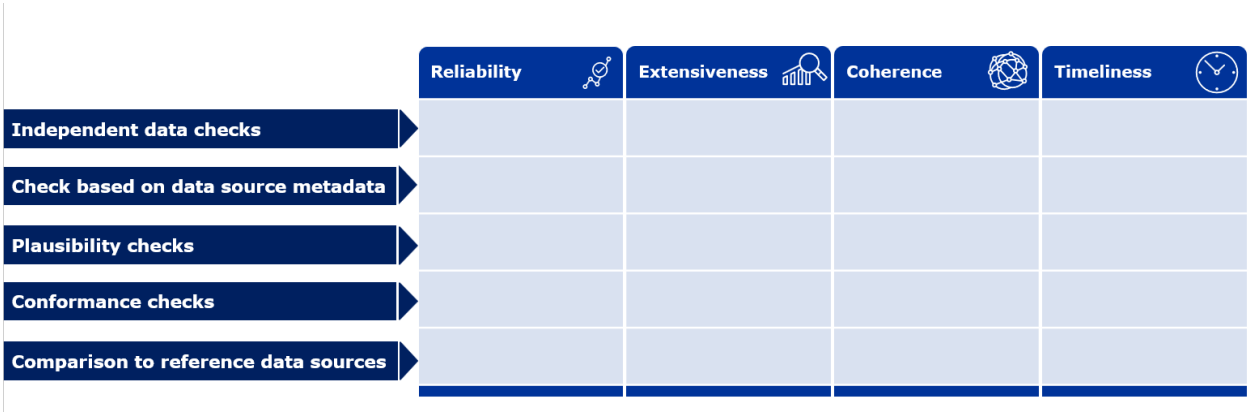

Figure 4 - Illustrative example of a customisable framework for metrics identification

Readers are encouraged to adapt and expand the framework based on the specific context and relevance of each metric to their quality dimensions. DQ dimensions that can be assessed are reported on the Y axis. On the X axis, different classes of metrics are reported, categorised according to the type of additional information (respect to the dataset) they can be based on. The classes are outlined below:

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.

Check based on data source metadata. This class is based on the deep knowledge of a specific dataset, such as what is provided in metadata or supporting documentation. In some cases, it may be feasible 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 metadata detailing the level of confidence behind a data point (e.g.: when some value is mapped to a standard code algorithmically). Metrics considering such annotations could then be used to assess the reliability[23] of a dataset. In principle such metrics could measure, by direct verification, the effect of a full data process.

In the spirit of this framework, this class is introduced to consider how additional information in metadata could be used to develop metrics. This may not be often practically feasible, due to the lack of standards. Furthermore, as discussed in the EMRN DQF, the distinction between Metadata and Data is also not well defined, and largely vanished when self-describing formats are in use. Nevertheless, as CDMs and Metadata standards are evolving, it is useful to consider this potential class of metrics.

Note that although they are related, metadata and conformance are distinct concepts. Metadata are data that describe the data. These provide information to understand, manage, and use the data

[23] 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")

effectively. Conformance checks, on the other hand, represent whether and to what extent the data adhere to specific standards: e.g., data structures, formats.

Other metrics can instead embed knowledge about general characteristics of 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 data: 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 temporal patterns that are impossible (e.g.: drug prescription occurring after death).

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.

Comparison to reference data sources. Comparison against reference data sources can support the assessment of extensiveness and reliability. Reference datasets should be selected based on predefined criteria and ideally demonstrate high quality across all dimensions. These comparisons can be performed at the individual level (e.g., comparing specific data elements across sources) or at the aggregate level (e.g., comparing overall distributions or trends), depending on information available for the reference data source. Metrics derived from reference RWD sources with established quality assurance – such as the proportion of missing data – can help identify potential bias in data collection or recording. In some rare cases, accuracy or validity can be assessed by comparing the same data type for the same patient across real-world and clinical trial settings, using the latter as a reference standard. Alternatively, comparisons may be made against clinical knowledge or expected trends, which serve as a different type of benchmark. RWD may also be compared to other recognised RWD sources that have been extensively validated and recognised in published literature.

These metrics do not include results of validation of accuracy against original data, as that is expected to be covered in foundational documentation. 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 chronic disease is unlikely to grow drastically 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 of checking the data against clinical expectations addresses a concept also called clinical validity.

In the following section, some examples of metrics are provided in relation to this framework. For scenarios where RWD is expected to evolve over time, the metrics framework should be reviewed and updated regularly to reflect changes in data availability, structure, and relevance. The frequency of data quality checks cannot be universally prescribed, as it depends on several context-specific factors. These include the nature and intended use of the data, the data source's update cycles and governance practices, and the level of risk associated with data quality issues in the given context. As such, frequency should be determined on a case-by-case basis to ensure relevance and proportionality to the data's purpose and environment.

Tables 4-7 show examples of DQ metrics, organised following the framework illustrated in Figure 4. The tables refer first to the EMRN DQF and more broadly to commonly used DQ checks (26). This document 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.

5.2. Metrics for DQ assessments

Reliability dimension

These metrics are intended to measure the degree to which data correspond to what they intend to represent. Where applicable, reference thresholds should be provided to define acceptable data quality levels, ensuring standardisation across different regulatory and research applications.

Table 4 - Overview of reliability metrics by sub-dimensions with examples

Sub-dimension	Metric group	Metrics	Example
Accuracy	Plausibility checks[24] (temporal and atemporal)[25]	• Number and percent of records where data values don't agree with standards / knowledge / expectations / 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	• 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 on the same day • 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 that are based on imputation or derivation • Number and percent of records where data values don't agree with standards / knowledge / expectations / feasible ranges	• For X% of variables, the number of missing is listed and if the variable contains imputed or derived data • 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

[24] 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.
[25] While we don't distinguish between temporal and atemporal aspects of plausibility checks here, this is an important aspect and we refer to the EMRN DQF (6.1.3) for some examples highlighting this distinction.

		• 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	nurses within the same facility using the same equipment on the same day • X% of records of pregnancy were attributed to males
	Comparison to other data sources	• Comparison with databases of other populations to check if rates/proportion/counts of records with a specific diagnose, or medicine are in a similar range.	• The prevalence/incidence of disease X is in line with other populations e.g. disease or age distribution compared with national statistics form the country/region.
Precision	Independent data checks	• The number of decimal points used in data values, and their distribution • Number and percent of records where corresponding variables yield identical results across other databases • Number and percentage of record mapped to a specific vocabulary • Number and percentage of records with exact date or date of procedure/treatment as opposed to approximation • Number and percentage of exact birth dates as opposed to approximation • Age distribution in comparison with databases of other populations	• "Height" in meters recorded with two decimal digits, but the last being always 0. • X% of records mapped to oncology OMOP CDM • X% of records with exact data of procedure/treatment e.g. surgery date • X% of records with exact date of birth • The age distribution is in line with other national statistics data form the same country/region.
Traceability	Checks on data source metadata	• Number and percent of variables for which traceability information is available in metadata. • Number and percentage of a CDM vocabulary is traceable to the source data	• 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) • X% of data mapped to CDM vocabulary is traceable to the source data

Extensiveness dimension

These metrics are intended to measure the extent to which data are available.

Table 5 - 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 not missing for a given variable. Key variables are demographics, conditions, treatment	• X% of records have non-missing values for Date of Birth
		• Percentage of patients who have a certain number of measurements for a given variable	• X% of patients have 2 or more blood pressure measurements
	Comparison to other data sources	• Relative percentage of records for which a variable is missing with respect to expected percentage of missing data from a national data source	• X% of patients have date of diagnosis missing for a diabetes database
Coverage	Comparison with external benchmark	• 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

When measuring completeness and coverage, it might prove useful to describe the metrics in a stratified manner, particularly for subpopulations with high clinical relevance (e.g., paediatric patients, rare disease cohorts) or in cases where historical data quality trends suggest potential gaps.

Coherence dimension

These metrics are intended to measure the degree to which data within a dataset are consistent in their representation and meaning.

Table 66 - 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 that conform to formatting requirements	• X% of records have Sex as only one character (e.g., 0 or 1)
		• For relevant variables, % of records where data values conform to allowable values or ranges and follow logical sequences	• X% of records have sex with one of the 3 allowable values "M", "F". or "U". • X% of records do not have drug exposure before birth
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 recorded, 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 that conform to formatting constraints	• 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

This metric is intended to provide an indication of how much data are up to date.

Table 77 - 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[26]	• New data is entered into the database every three months • Prescription data entered daily • Hospitalisation data entered monthly
	Independent data checks	• Data lag (time between last data point collected – also referred to as data cut timepoint – and date when data is made available)	• Data is collected at the end of each month and made available quarterly

Acceptable thresholds for data lag time between collection and availability for analysis should be set in relation to specific use cases. Alerts should be implemented to notify stakeholders of significant delays that could impact decision-making.

It should be noted that updates may not always be complete, and measures should distinguish the lag time to obtain key information with a certain level of confidence.

These metrics have been provided at a general level where one could 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). 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.

[26] As we are considering metrics for intrinsic aspects of a dataset, "updates" is here intended as updated of the final dataset available for research, as opposed to updates to the source database.

5.3. Considerations for the implementation of RWD DQ metrics

5.3.1. Different roles of metrics

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

In regulatory workflows, these roles can function in a sequential manner—starting with a detailed assessment, DQ assurance, followed by reporting and/or in an iterative cycle where feedback from DQ reporting informs subsequent quality assurance improvements.

5.3.1.1. DQ assessment

DQ metrics can be used to reflect the quality of specific data elements, with the view of assessing, among other criteria, 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. Such metrics should be assessed on the population of interest, rather than an overall generic dataset. In some cases, contractual or technical barriers might delay access to the raw data needed to compute such metrics if the information is not readily available.

5.3.1.2. DQ 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 different 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.

5.3.1.3. DQ 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.

5.3.2. 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). The feasibility of implementing these metrics may vary depending on institutional data governance structures, access to standardised reporting tools, and national regulatory frameworks.

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. It is expected that the adoption of higher maturity level will facilitate and eventually make the regulatory decision process more agile.

5.4. Providing supporting information for RWD in regulatory submissions

When using RWD in regulatory submissions, detailed characteristics of the data source that enable DQ assessment should be provided, considering adequate granularity for regulatory assessment. The complete range of dimensions, metrics and other data quality relevant information is described in previous sections. Here, a succinct summary of the minimum necessary elements is provided:

- **Data collection and/or generation context**
    - Information on the data collection or recording process
    - Selection mechanisms (e.g., inclusion of specific patient groups or clinical data types).
- **Quality management practices** – describe standard practices routinely applied to the data, as well as the processes and methods behind the generation of data at the source.
    - Level of automation and use of computerised systems.
    - Data extraction, cleaning, or transformation steps.
    - Data quality checks for logical inconsistencies, erroneous, missing or out-of-range values.
    - Remedial actions taken at the RWD source level.
- **Data enhancement measures – *if applicable*:**
    - Measures taken to improve data completeness (e.g. data collection prerequisites for reimbursement).
    - Linkage information, including data elements used, linking and linkage methodology.
- **Data availability** - Indicate 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-use assessments (27).

To allow for an adequate DQ assessment, it is important that the data holder maintains the information published in the Catalogue of RWD Sources up to date, with the last update occurring within the past 12 months.

Next to this, it is helpful to use appropriate reporting templates depending on study design such as the EQUATOR network (28).

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

6.1. Understanding relevance

A major difference between the EMRN DQF and other proposed DQFs (9-11) for RWD is the definition and the role of relevance. One fundamental principle of the EMRN DQF is that data quality aspects (organised in dimensions) can be measured independently from a research question, while the assessment of the adequacy of such quality can only be defined in relation to a question, or a class of similar questions.

This "suitability" corresponds to a generic notion of relevance found in other frameworks but is captured in terms of "question specific determinants" in the EMRN DQF framework. In other frameworks, relevance is often considered as its own dimension and includes completeness and reliability as sub-dimensions (9). In the EMRNN DQF question specific determinants extend to all dimensions (e.g.: including timeliness and coherence): all aspects of DQ contribute to the suitability of a dataset to support a specific question and method.

What EMRN DQF defines as "relevance" is a dimension that partition DQ aspects, specifically by capturing aspects not already covered by the other dimensions of reliability (is data correct), extensiveness (is data enough), coherence (is data analysable as a whole), timeliness (is data at the right time). This dimension captures the type of data (that may or not be relevant to a specific question)[27].

For example, in a study assessing medication adherence using pharmacy claims data, the relevance dimension would help determine whether refill data are sufficient proxies for patient adherence, rather than evaluating overall data completeness.

In this framework, 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[28].

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 fitness-for-use and justification of datasets used to answer 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.

Fitness-for-use combines all required DQ characteristics as relevance with reliability, coherence, extensiveness and timeliness in an overall evaluation. The emphasis of DQ is ensuring that the data are fit-for-purpose for reliable assessments of decision making to supporting health research and population health.

It is difficult to pre-specify thresholds or minimum criteria for the fitness-for-use assessments as it depends on the study-specific and on disease-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. An example is in rare diseases, where it might be

[27] This is compatible with the ICH definition of relevance, where the term "relevance" includes the availability of key data elements (e.g., exposure, outcomes, covariates) to answer the specific trial question with the specific method

[28] 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.

challenging to identify enough cases via secondary use of data. Another example is in communicable diseases / pandemics, where data completeness and consistency may be compromised due to rapid data collection under emergency conditions, evolving case definitions, and changes in testing or reporting practices.

At a more detailed level, data are considered relevant to a specific question if they capture key elements necessary to address that question (e.g., diagnosis, exposure, outcome, confounders, and covariates and if the number of patients and follow-up times are deemed sufficient to answer the research question (8). 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 as well as a good knowledge of the health care system.

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

Before defining the research question, stakeholders should engage patient representatives when designing research questions, to ensure alignment with patient priorities and lived experiences (29). 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). Prior to proceeding with "Step 1" described below, the minimal data requirements should be established based on the research question to determine which study design is most appropriate. This is necessary to determine whether primary or secondary use of RWD is indeed feasible to address the research question. The result of the fitness-for-purpose assessment could be that the research question of interest cannot be answered with a specific dataset and not all research questions can be refined to fit a dataset (e.g. prevalence of a genetic mutation cannot be ascertained in a database without genetic information).

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[29], but it can be a criterion when multiple RWD sources are available. Techniques such as data harmonisation, mapping to standard ontologies (e.g., SNOMED CT, ICD-10), and algorithmic corrections can enhance coherence where inconsistencies exist.

In terms of coherence, when data originate from multiple geographies or healthcare systems, special attention should be paid to the coherence of all datasets, in terms of the underlying systems and processes and biases and that they may be introduced.

This preparatory step is typically done by inspecting documentation and metadata (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.

[29] There is a potential loss of precision when data are harmonised to a common standard. Therefore, if data do not come in a coherent representation for the intended analysis process, attention must be paid to the reduction in precision when assessing specific variables (later in this document).

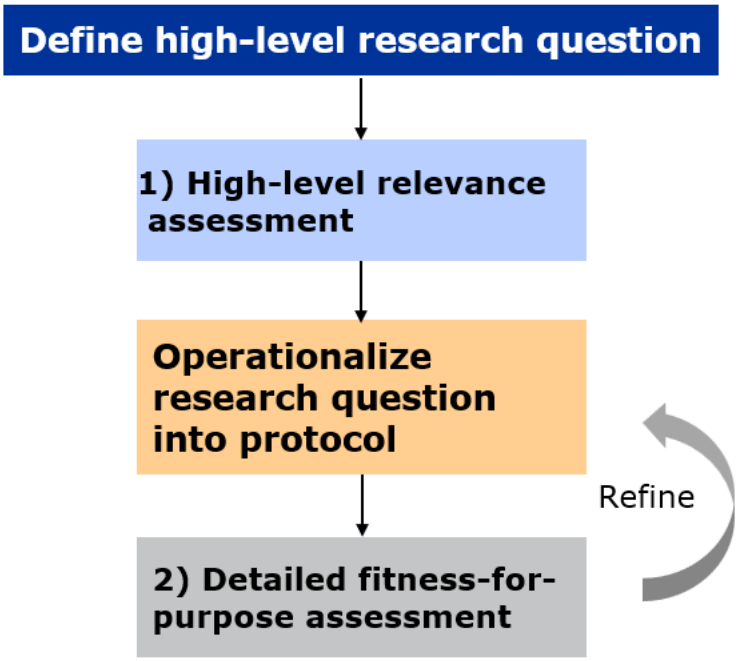

Figure 5 5- General principles for assessment of DQ in relation to a research question

Step 1 – 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), as well as previous studies conducted using these data sources can help in framing the research question more explicitly to guide the choice of a specific data source. This process can be further facilitated by consulting the data sources in relevant online repositories, such as the HMA-EMA Catalogue of Real-World Data Sources [30].

Step 1 is foundational in the feasibility assessment of a data source before direct access to the data is possible, to enable informed decisions about whether a data source is likely to meet the intended analytical or operational needs. This early evaluation helps ensure that resources are directed toward data sources with a higher likelihood of supporting robust and credible insights.

Note that one plausible outcome of the high-level assessments described in Steps 1 and 2 could be that there are at present no suitable RWD sources to answer the research question and new data sources or study designs should be considered.

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

Step 2 – 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 (i.e. the Estimand) and the relevant design elements such as study population, study period, study design (e.g., case-control, cohort), treatment (exposure) group, comparator group, outcome(s), ideal length of follow-up, study population size and representativeness, relevant healthcare setting, availability and completeness of needed variables.
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., 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, prespecified 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 RWD source can be evaluated to determine whether 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 while an initial fitness-for-use assessment may be informed by metrics and metadata reported for an overall RWD source, a thorough evaluation should be conducted 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 is one of the key elements that would require re-assessment 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, the fitness for use assessment being an 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.

6.3. Framework for detailed fitness-for-use assessment

This framework is inspired by The Structured Process to Identify Fit-For-Purpose Data (SPFID) (8). 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. Instead, it focuses on assessing whether a specific data source meets predefined regulatory criteria, thereby ensuring a structured but

flexible evaluation tailored for regulatory decision-making. However, comparative ranking approaches may still be used for study feasibility assessments or exploratory research to ensure the most suitable data source is selected.

Table 8 provides a template to be filled in during the Step 2 of the fit-for-use assessment. 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. In cases where quality metrics provide conflicting signals, a structured decision framework—such as weighted scoring or hierarchical prioritisation of critical elements—should be applied to ensure a balanced assessment.

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

An example is provided in Table 9 for a Chronic Lymphocytic Leukaemia (CLL) 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. While this example focuses on CLL, the structured approach can be adapted for fitness-for-use assessments across different disease areas for methodological consistency.

Table 8 - Fitness-for-use assessment to be filled in during the suitability assessment of a RWD source (non-exhaustive[31])

Design elements *(to be pre-specified)*	**Operational definition / threshold** *(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 sample population vs. 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
	Active treatment cohort size	Minimum number	Low/Medium/High	N/A (to be assessed on a research question basis)	N/A	N/A

[31] Some items may be applicable to multiple domains (e.g., cohort size applicable to study population, but also to exposure / comparator / endpoint.)

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	Time between last data point collected – also referred to as data cut timepoint – and date when data is made available	Maximum lag time	Low/Medium/High	N/A	Timeliness metrics (overall or for specific variables if relevant)	N/A

Note that general guidance on thresholds cannot be uniformly applied, as appropriate values will vary depending on therapeutic area, unmet needs, research objectives, and timing. These factors influence what is considered acceptable or meaningful in a given context, making threshold determination inherently case-specific.

Table 9 - Illustrative fitness-for-use assessment for a Chronic Lymphocytic Leukaemia study *(non-exhaustive)*

Design elements *(pre-specified)*	**Operational definition / threshold** *(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)	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 treated study participants	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 for 20% of patients	N/A	N/A
	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	N/A	Documentation detailing re-assessment of response by adjudication

				40% with 3 assessments or more		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 females 100% of patients have sex captured as one ASCII character (0 or 1)	Documentation detailing method for sex capture

	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)	Duration of monitoring after index date	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	Time between last data point collected – also referred to as data cut timepoint – and date when data is made available	2 years maximum	Medium	N/A	There is on average a 3-year lag time to perform data curation and linkage	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.

7. 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. The guidelines provided in this document offer actionable and focused recommendations for assessing DQ of RWD, aiming to enhance the usefulness of RWE for regulatory purposes.

Looking ahead, future iterations of the RW-DQF may expand to address additional data types and emerging technologies that are increasingly relevant to regulatory decision-making. These could include direct-from-patient (or direct-from-user) data such as PROs, patient engagement and preference data, or non-prescription medicinal products. The evolving landscape of RWD will likely be also shaped by innovations in artificial intelligence, wearable technologies, mobile health inputs, and social media-derived insights. While all these areas are not within the scope of the current version, they represent important frontiers for future guidance, which will be addressed in the future, in alignment with regulatory needs and scientific advancements.

8. 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. ICH E6 (R3) Guideline for good clinical practice (GCP), EMA/CHMP/ICH/135/1995. 2025.
3. The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). Guide on Methodological Standards in Pharmacoepidemiology (Revision 11). EMA/95098/2010. Available at https://encepp.europa.eu/document/download/f6e403a6-8033-4c22-a5ff-195ba3666299_en?filename=01.ENCePPMethodsGuideRev.11.pdf.
4. ICH M14 Guideline: General Principles on Planning, Designing, Analysing, and Reporting of Non-interventional Studies That Utilise Real World Data for Safety Assessment of Medicine. . 2025.
5. 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.
6. Reflection paper on use of real-world data in non-interventional studies to generate real-world evidence for regulatory purposes. Available from: https://www.ema.europa.eu/en/documents/other/reflection-paper-use-real-world-data-non-interventional-studies-generate-real-world-evidence-regulatory-purposes_en.pdf. 2025.
7. Rudrapatna VA, Butte AJ. Opportunities and challenges in using real-world data for health care. J Clin Invest. 2020;130(2):565-74.
8. Gatto NM, Campbell UB, Rubinstein E, Jaksa A, Mattox P, Mo J, et al. The Structured Process to Identify Fit-For-Purpose Data: A Data Feasibility Assessment Framework. Clin Pharmacol Ther. 2022;111(1):122-34.
9. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, 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):1244.
10. Schmidt CO, Struckmann S, Enzenbach C, Reineke A, Stausberg J, Damerow S, 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):63.
11. 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.
12. Cave A, Kurz X, Arlett P. Real-World Data for Regulatory Decision Making: Challenges and Possible Solutions for Europe. Clin Pharmacol Ther. 2019;106(1):36-9.
13. 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.
14. Compassionate use. European Medicines Agency 2024 [Available from: https://www.ema.europa.eu/en/human-regulatory-overview/research-development/compassionate-use.
15. 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.
16. Lu; J, Liu; A, Dong; F, Gu; F, Gama; J, Zhang; G. Learning under Concept Drift: A Review. IEEE Transactions on Knowledge and Data Engineering. 2019;31(12):2346-63.
17. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. J Am Med Inform Assoc. 2021;28(10):2251-7.
18. Blacketer; C, Voss; E, DeFalco; F, Hughes; N, Schuemie; M, Moinat; M, et al. Using the Data Quality Dashboard to Improve the EHDEN Network. Appl Sci 2021.11(24).
19. Callahan TJ, Bauck AE, Bertoch D, Brown J, Khare R, Ryan PB, et al. A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks. EGEMS (Wash DC). 2017;5(1):8.
20. 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.
21. Kent S, Burn E, Dawoud D, Jonsson P, Ostby JT, Hughes N, et al. Common Problems, Common Data Model Solutions: Evidence Generation for Health Technology Assessment. Pharmacoeconomics. 2021;39(3):275-85.

22. Rivera DR, Henk HJ, Garrett-Mayer E, Christian JB, Belli AJ, Bruinooge SS, 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):283-92.
23. REQueST Tool and its vision paper [Internet]. EUnetHTA. 2019. Available from: https://catalogues.ema.europa.eu/system/files/2025-02/05.01.0301%20Feasibility%20Documentation%20%20-%20Registry%20Evaluation%20and%20Quality%20Standards%20Tool%20%28REQueST%29%20_%2010-Sep-2023_Redacted.pdf.
24. Documentation Getting started with OHDSI (observational health data sciences and informatics) https://www.ohdsi.org/web/wiki/doku.php?id=documentation:getting_started.
25. FAIR principles [Available from: https://www.go-fair.org/fair-principles/.
26. Observational Health Data Sciences, Informatics. Chapter 15 Data Quality. 2021. Available from: https://ohdsi.github.io/TheBookOfOhdsi/DataQuality.html.
27. Good Practice Guide for the use of the HMA-EMA Catalogues of real-world data sources and studies. 2022. Available from: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/good-practice-guide-use-hma-ema-catalogues-real-world-data-sources-studies_en.pdf.
28. EQUATOR network - Enhancing the Quality and Transparantcy of Health Research, Accessed Septemer 2025 [Available from: https://www.equator-network.org/reporting-guidelines/.
29. Arlett P, Umuhire D, Verpillat P, Foggi P, Wandel Liminga U, Sepodes B, et al. Clinical Evidence 2030. Clin Pharmacol Ther. 2025;117(4):884-6.
30. Guideline on good pharmacovigilance practices (GVP) Annex I - Definitions (Rev 5), MA/876333/2011 Rev 5*. 2024.

9. Definitions

Abbreviation	Meaning
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
ICH	International Council for harmonisation of technical requirements for pharmaceuticals for human use
I/E criteria	Inclusion/Exclusion criteria
KPI	Key Performance Indicator
MAH	Marketing Authorisation Holder
QA	Quality Assurance
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

10. 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:

Term	Definition
Coherence	Coherence (also referred to as Consistency) is defined as the dimension that expresses how different parts of an overall dataset (eventually composed of data from multiple sources) are consistent in their representation and meaning.
Data interoperability	The ability of different systems to exchange and interpret data consistently
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 end users 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 (e.g., for rare diseases) or whole population coverage (e.g., for pandemic response planning).
Data provenance	The history and lineage of a dataset, detailing the origins, transformations, and processes that the data has undergone from its initial collection to its current state
Data user/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. RWD end users include also regulatory agencies, payers, HTAs, healthcare providers, and patient advocacy groups, who use RWD for assessing treatment effectiveness, safety, and health outcomes across populations.
Extensiveness	Extensiveness is defined as the dimension capturing the amount of data available. When considering the "fit for purpose" definition of quality, extensiveness covers how sufficient the data are for their intended use. The sufficiency of data can vary depending on the specific research question being addressed, as different questions may require different amounts of data to be adequately answered.
Fit for purpose / fit for use	Possessing all required data quality characteristic needed to address a specific goal. The emphasis of data quality is ensuring that the data are fit for purpose for reliable assessments of whether the data are fit for the purpose of decision making to supporting health research and population health. Sometimes also referred to as "suitable".
Foundational determinants	A characterisation of the systems and processes underpinning data generation and manipulation that have an impact on DQ.

Term	Definition
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	A maturity model is a framework for assessing processes, technology and structure of an organisation or function. It provides a structured approach to evaluating how well an organisation or a function manages its data quality processes, policies, and practices. The model defines key characteristics at each level to guide measure continuous improvement in data quality over time.
Metadata	Metadata are defined as "data about data" providing context about their purpose and generation. It's a set of data that describes and gives information on other data providing context about their purpose, location, key-variables, generation, format, and ownership of a dataset. Metadata are often published in data catalogues, which have the purpose of allowing data to be discoverable and checked for fitness for purpose, without revealing the data themselves.
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. Similarly, if a patient is recorded as receiving a contraindicated medication in multiple visits without intervention, this may indicate a data entry or processing error requiring further validation.
Primary data collection	Collection of data directly from patients, caregivers, healthcare professionals or other persons involved in patient care (GVP annex I) (30).
Question-specific determinants	Aspects of DQ that cannot be assessed independently of a research question.
Relevance (EMA DQF definition)	Relevance is defined as the extent to which a dataset presents the data elements useful to answer a given research question. This definition is narrower, as it identifies the DQ dimension, assessing the usefulness of a dataset to a research question, that is not covered by other dimensions (e.g.: extensiveness, reliability,)
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 population it is meant to represent.

Term	Definition
RW-DQF	Application of the European Medicines Regulatory Network (EMRN) Data Quality Framework (DQF) for EU Medicines regulation to real-world (RW) data.
RWD holder	People owning and or holding the RWD
RWD practitioners	People involved in the RWD collection or recording process such as researchers, data analysts and data custodians.
Sample population	The group of individuals whose data are captured in the data source and used for analysis.
Secondary use of data	Use of existing data for a different purpose than the one for which it was originally collected (GVP, annex 1) (30).
Target population	The broader group of individuals that the study aims to draw conclusions about. This is defined by the research question and may include people beyond those captured in the data source (e.g., all patients with a condition in a country or region)
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.
Validation	For the scope of this document, validation refers to checking whether the data correspond to the source.