30 October 2023
Data Analytics and Methods Task Force
EMA/326985/2023

Data Quality Framework for EU medicines regulation

Draft agreed by BDSG for release for consultation	10 October 2022
End of consultation (deadline for comments)	18 November 2022
Agreed by BDSG and MWP	30 June 2023
Adopted by CHMP	30 October 2023

Keywords	Data quality framework, medicines regulation, data quality dimensions, primary and secondary use of data

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

1. Executive summary

This document is the first release of the EU Data Quality Framework (DQF) for medicines regulation and defines high-level principles and procedures that apply across EMA's regulatory mandate. This framework provides general considerations on data quality that are relevant for regulatory decision making, definitions for data quality dimensions and sub-dimensions, as well as their characterisation and related metrics. It provides an analysis of what data quality actions and metrics should be considered in different use cases and introduces a maturity model to guide the evolution of automation to support data-driven regulatory decision making.

This document is intended to be a general resource from which more focused recommendations can be derived for specific regulatory domains with specified metrics and checks. See figure 1 for a summarised representation of the key points of the DQF.

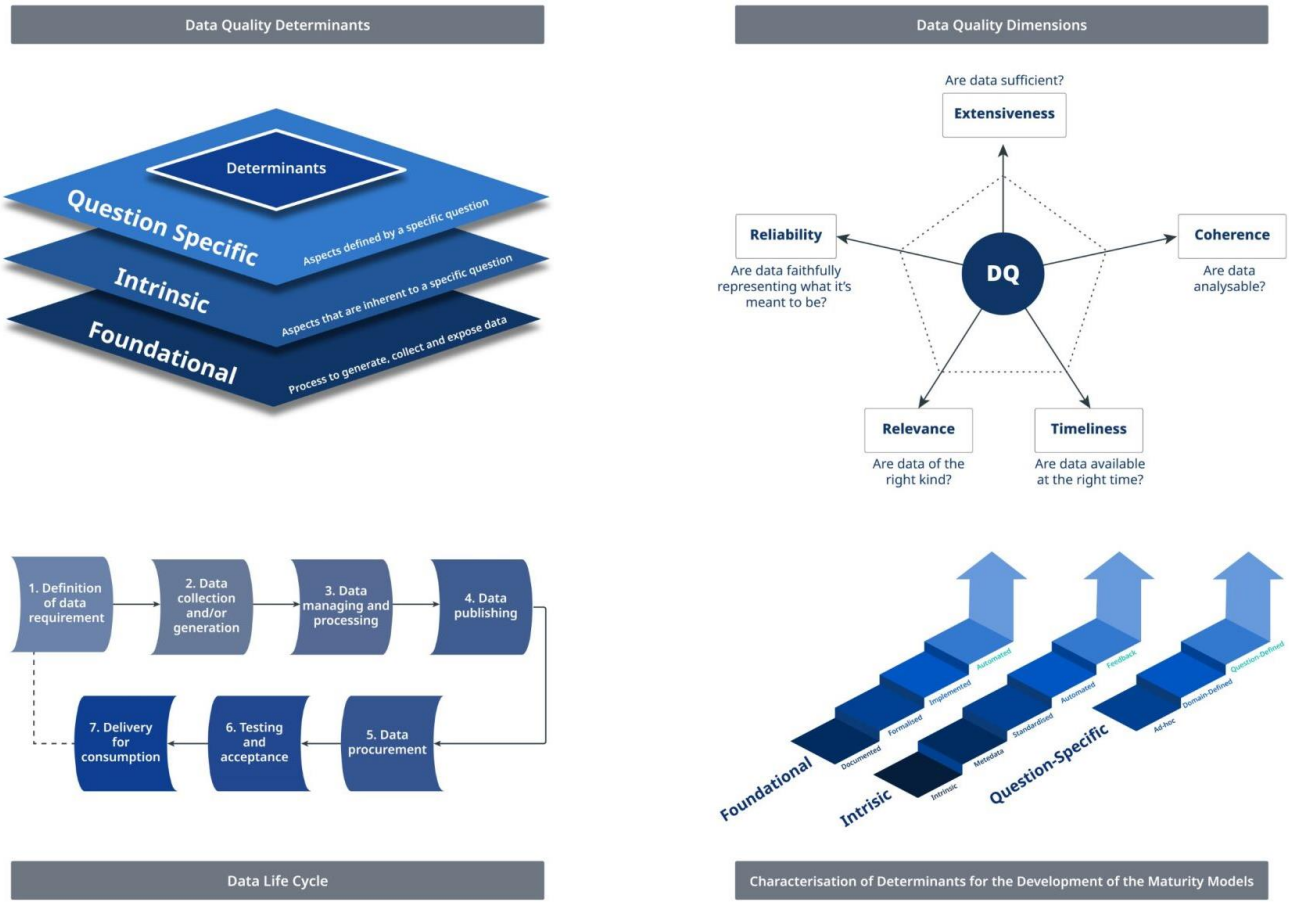

Figure 1 – Representation of the key points of the Data Quality Framework

2. Abbreviations

CDM	Common Data Model
CHMP	Committee for Medicinal Products for Human Use
DQ	Data Quality
DQF	Data Quality Framework
EHR	Electronic Health Record
EHDS	European Health Data Space
EMA	European Medicines Agency
ETL	Extract, Transform and Load
FAIR	Findable, Accessible, Interoperable and Reusable
GxP	Good x Practices, where x stands for laboratory (GLP), clinical (GCP), manufacturing (GMP), distribution or documentation (GDP)
ICSR	Individual Case Safety Reports
ISO	International Organisation for Standardisation
MDM	Master Data Management
QMS	Quality Management System
QSR	Quality System Regulation
RWD	Real-World Data
RWE	Real-World Evidence

3. Background – The need for a Data Quality Framework for medicines regulation

As acknowledged in the recommendations of the HMA-EMA Joint Big Data Task Force and the workplan of the HMA-EMA Joint Big Data Steering Group, establishing an EU framework for data quality (DQ) and representativeness is a critical element for realising the full potential of (big) data and driving regulatory decisions.

In recent years, the EU regulatory assessment process has been exploring a shift from document-based submissions to direct assessments of the data underlying those submissions. To facilitate this potential shift, there is an increased need for standardisation [1], and the need for a framework, which would characterise DQ and would allow the regulator to make reliable assessments of whether the data are fit for the purpose of decision making.

In addition, the progress in digitalisation and information technology creates new opportunities, but also contributes to an increasingly complex landscape for regulatory decision making. While new types of data become available, guidelines or methods to demonstrate whether such data are adequate for

decision making are still scarce. Therefore, a Data Quality Framework (DQF) is needed to guide coherent and consistent quality assessment procedures.

One notable example is healthcare data that are becoming available in increasing quantity to potentially support regulatory decision making for medicines. Information derived from routinely collected Real-World Data (RWD) has for a long time been used to support regulatory decision making on the safety of drugs in the post-authorisation phase. While most traditional pre-approval randomised controlled clinical trials remain the fundamental method of establishing the safety and efficacy of medicines during the pre-authorisation phase, they could potentially benefit from the evidence generated using this data. Insights into the Real-World are also required by downstream stakeholders including Health Technology Assessment bodies, payers and ultimately clinicians and patients. Bridging these gaps, the regulatory network needs to acquire the ability to describe and quantify the degree to which these data are accurate and fit for purpose.

4. Scope of this DQF

This document aims to provide a set of definitions, principles and guidelines that can coherently be applied to any data source for the purpose of characterising, assessing, and assuring DQ for regulatory decision making. This framework is intended to encompass primary and secondary use, as well as metadata and supporting information (e.g., MDM (Master data management), underlying reference Data) applicable to support Committee for Medicinal Products for Human Use (CHMP) decision making. The document is targeted primarily at the EU medicine regulatory network, but the relevance of the content can be of interest to a wider range of stakeholders such as marketing authorisation holders, data source holders, researchers, and patient associations.

As methods, terminologies, metrics, and issues vary across data types and sources, this framework seeks to provide a coherent basis to identify, define, and further develop DQ assessment procedures and recommendations for current and novel data types.

Objectives of this framework are therefore to achieve consistency in DQ related processes, foster the development of horizontal systems[1] for DQ and eventually enable a more adequate and automated use of data for regulatory decision making.

This framework builds on the recommendations of TEHDAS [2] and extends them with a classification of quality dimensions and assessment criteria, as well as guidelines for their application. It builds on the definitions and recommendations that have been proposed in several existing DQ frameworks, including [2-13].

While many examples provided in this framework relate to Real-World Data, the scope of this framework extends to a broad range of regulatory activities and their respective data types, including Real-World Data [14, 15] (including within clinical trials to supplement trial-specific data collection), bioanalytical omics data, animal health data, preclinical data (cell and animal-based laboratory data), spontaneous adverse event reporting data, chemical and manufacturing control data, and more.

[1] A "Horizontal system" provides a specific set of functionalities across a variety of use cases. In this case the intentions to develop systems and approaches to DQ that can be used (and potentially shared) across use cases. "Horizontal system" is defined by contrast to "Vertical system", where all DQ processes and system would be developed ad hoc and targeted to a specific use case.

4.1. Definition of data

In this DQF, data are considered as any information asset that represents measurements or observations and that can be used to support decision making, directly or indirectly through analysis.

4.2. Definition of DQ

In general terms, quality is defined as an attribute of a product or service that defines the degree to which it meets customer and other stakeholder needs within statutory and regulatory requirements or its fitness for intended use[2]. The same principle applies to data and for the purpose of this document, the following definition is adopted:

Data quality is defined as: *"fitness for purpose for users' needs in relation to health research, policy making, and regulation and that the data reflect the reality, which they aim to represent"* [2][2].

Therefore, this DQF restricts its scope to determinants of DQ that are relevant for regulatory decision making.

4.3. Limitations of scope

Following the definition of DQ and the restricted focus on regulatory decision making this framework's scope excludes:

- Analytical methods to derive evidence, i.e., conclusions and insights, from underlying data. This framework focuses on defining guidelines about assessing the level of the quality of the data used for regulatory decisions, not on their actual usage for regulatory decision making and the methods involved. While data quality and methods for evidence generation are effectively a continuum in terms of decision making, when taking the perspective of data collection, dissemination, and re-use, they are distinct.
- Aspects of DQ that do not directly impact regulatory decision making e.g., conciseness or accessibility. For instance, conciseness is a relevant dimension of data quality in that it affects fitness for purpose when transmitting or archiving large datasets (e.g., for genomics data). However, it is not relevant in terms of data being fit for purpose to answer a specific (regulatory) question. Accessibility can also be an important aspect of DQ, but in the context of a regulatory activity, data is by definition accessible to interested parties[3].
- Data transparency, intended as the characteristic of data being used lawfully, traceably and for valid purposes is also excluded from this framework. Issues related to data transparency go beyond data quality assessment in support of decision making[4]. Rather, this framework provides guidelines that can be part of a broader set of recommendations to support Data Transparency. As for accessibility, it should be noted that there may be an indirect impact of transparency to data quality, and this will be addressed, when relevant, in extensions of this Framework.
- Quality of the underlying elements the data refer to e.g., when considering a dataset about the purity of a medicine, this framework will cover the reliability, completeness, and other aspects of

[2] Note that reality is in general not fully observable. In this definition we consider how data reflects aspects of reality that data is designed to capture (e.g.: disease frequency in a population). Context is important to understand how what is observed relates to whole.

- [3] In some cases, aspects of DQ that do not directly relate to decision making, may have an indirect impact on it, e.g., a data source that is broadly accessible will likely be less opaque, more validated, and potentially of higher quality. Such aspects might be considered and possibly quantified in future extensions of this framework.

[4] As an example, the EU GDPR regulation defines transparency as: "The principle of transparency requires that any information addressed to the public or to the data subject be concise, easily accessible and easy to understand". This poses a broader set of requirements than what is strictly related to the use of data for regulatory decision making.

the data, but not aspects of quality (in this case purity or pharmaceutical quality) of the medicine per se.

- Semantic interoperability and standardisation are key DQ aspects for data usability and for the implementation of a DQF, but they do not affect the fitness for purpose of date respect to decision making (e.g.: a "false" dataset would not become less "false" if represented in a standard way). Therefore, the provision of guidelines and recommendations to define and select standards for interoperability shall fall out of the remits of this DQF. It falls within the scope of this document to suggest the application of standards to facilitate DQ assessment.
- Recommendations for the specific design of systems, processes, and responsibilities to guarantee DQ, or specific solutions or products. This framework focuses on the principles that such systems, processes and the resulting data should follow, to enable and optimise the use of such data for regulatory activities.

4.4. Structure of this DQF

The DQF for EU medicines regulation is composed of two parts, reflecting different stages in the specification process.

The first part (general framework) is designed to provide a coherent approach to DQ, encompassing a broad range of data types and extensible to novel use cases[5]. To achieve this, it provides a common ground on different DQ aspects that apply to different data types and scenarios: definitions, DQ dimensions and examples of metrics covering such dimensions. It further identifies general patterns for the applicability of DQ processes and it articulates a set of maturity models designed to drive increased automation of data-driven medicines regulatory decision making.

The second part (framework specialisation) specialises and eventually extends such generic recommendations to cater for specific data types or regulatory questions. This part poses the basis for the derivation of actual implementable guidelines, which will need to evolve as data and technologies change over time.

This document is the first version of the DQF for EU Regulatory Network [15]. It focuses on the general framework and addresses terminology, definitions, and general guiding principles around DQ in the context of medicines regulation. In the upcoming years, the DQF will be updated regularly with further deep dives in regulatory use cases of particular interest. The document will be in line with developments in TEHDAS to further strengthen the EMA data qualification process and the collaboration with the European Health Data Space (EHDS).

A glossary with the main terms and definitions can be found in chapter 9.

5. General considerations underlying the maintenance and assessment of DQ

5.1. DQ determinants for evidence generation

The landscape of data that can be potentially used for regulatory purposes extends to diverse data sources, each generated through different processes and fit for different primary and secondary uses. When considering the overall quality of a dataset at the point of regulatory decision making, it is important to distinguish what contributes to quality, and what can be measured or controlled at what

[5] In the context of this framework, "use-case" is used as a broader synonym of "regulatory question", when referring to a set of related questions and related activities.

stage. In this framework, such elements related to DQ are referred to as "determinants" and classified into three categories (see figure 2):

1. **Foundational determinants** pertain to the processes and systems through which data are generated, collected, processed, and made available. Foundational determinants are what affects the quality of data, but it's not part of the data themselves. As such, they do not depend on, and cannot be completely derived from, the content of a dataset. For data to be trusted for regulatory decision making, the underlying infrastructure and processes that collect, host, transform and move the data must be designed in such a way that the correspondence between data and the real entity it represents is not altered. Examples of foundational determinants are the use of certified software systems to collect and process data, the presence of processes, training, and audit processes to ensure data are properly recorded and documented, the validation and verifiability of data processing steps.
2. **Intrinsic determinants** of data pertain to aspects that are inherent to a given dataset. Intrinsic determinants are what can be derived given a dataset and possibly some external generic knowledge, but without the context in which the data were generated, as well of the context the data will be used in (e.g., a scientific or regulatory question). Examples of intrinsic determinants are coherent or incoherent formatting, the presence of errors (e.g., truncation) or the plausibility of the data.
3. **Question specific determinants** pertain to aspects of DQ that cannot generally be defined independently of a specific question or approach to analysis. Examples of question specific determinants are the acceptability of the completeness of a dataset, or its level of approximation (e.g., date expressed in dates or months) to answer a specific question.

In general, foundational determinants have a direct impact on DQ. When they cannot be controlled, the only option is to control the intrinsic aspects of DQ. The scope of such control is limited in its ability to assure fitness for purpose when a question (or set of typical questions) is not defined.

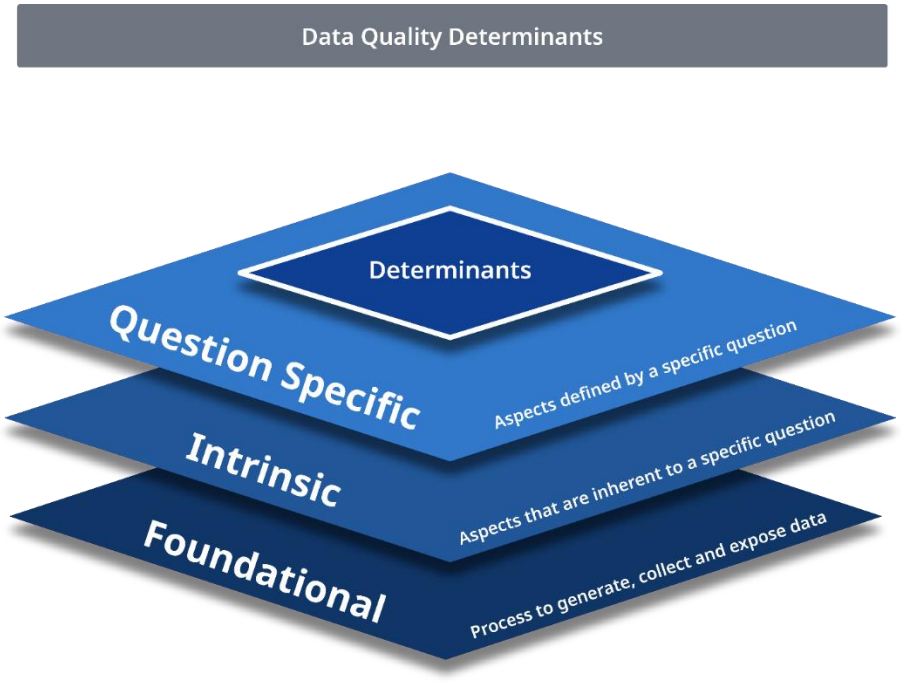

Figure 2 - Determinants of data quality

5.2. DQ along the evidence generation process

Data that are suitable and available for evidence generation go through a process (part of a broader "life cycle"[6]) that is specific to the type of data, the processes and organisations that produce it. For data that is already collected (secondary use), a fit for purpose assessment for the intended use should be done prior to this process. DQ checks occur at various steps of this process and may include iterative feedback loops.

As a reference, a general high-level lifecycle is outlined as follows (see Figure 3):

- **Definition of data requirements**: what data are sought, and what their characteristics should be. For primary data this phase can include elements directly related to evidence generation.
- **Data collection or generation**: gaining data reflecting the observed reality.
- **Data management and processing**: including data transfers, normalisation, and cleansing.
- **Data publishing**: making data available to consumers.
- **Data procurement and aggregation**: sourcing data from one or more consumers.
- **Testing and acceptance**: assessing the suitability of the procured data for intended needs.

Delivery for consumption: using data to support a specific activity, e.g., analysis.

Not all phases here presented are present in all data workflows (e.g., data collected from sensor or social data may be collected on a "what is available" basis, rather than based on specific requirements) and possibly extra phases may apply, and the order may differ.

For the scope of the assessment and management of DQ, it is important to establish what determinants apply at which stage of this process, and what may be the impact. For instance, intrinsic aspects of DQ can be measured and such measures could be used to improve reliability at the stage of data collection and generation, or it could be used to provide an assessment of quality at publication time. Integration with additional data would require re-assessment. Question-specific determinants of DQ need to be assessed each time data are repurposed to answer a question it was not originally collected or designed for.

DQ checks occur at various steps along the evidence generation process and may include iterative feedback loops as indicated by the dashed line in figure 3.

[6] The data life cycle is broader in that it would extend to aspects of data disposal and maintenance beyond usage.

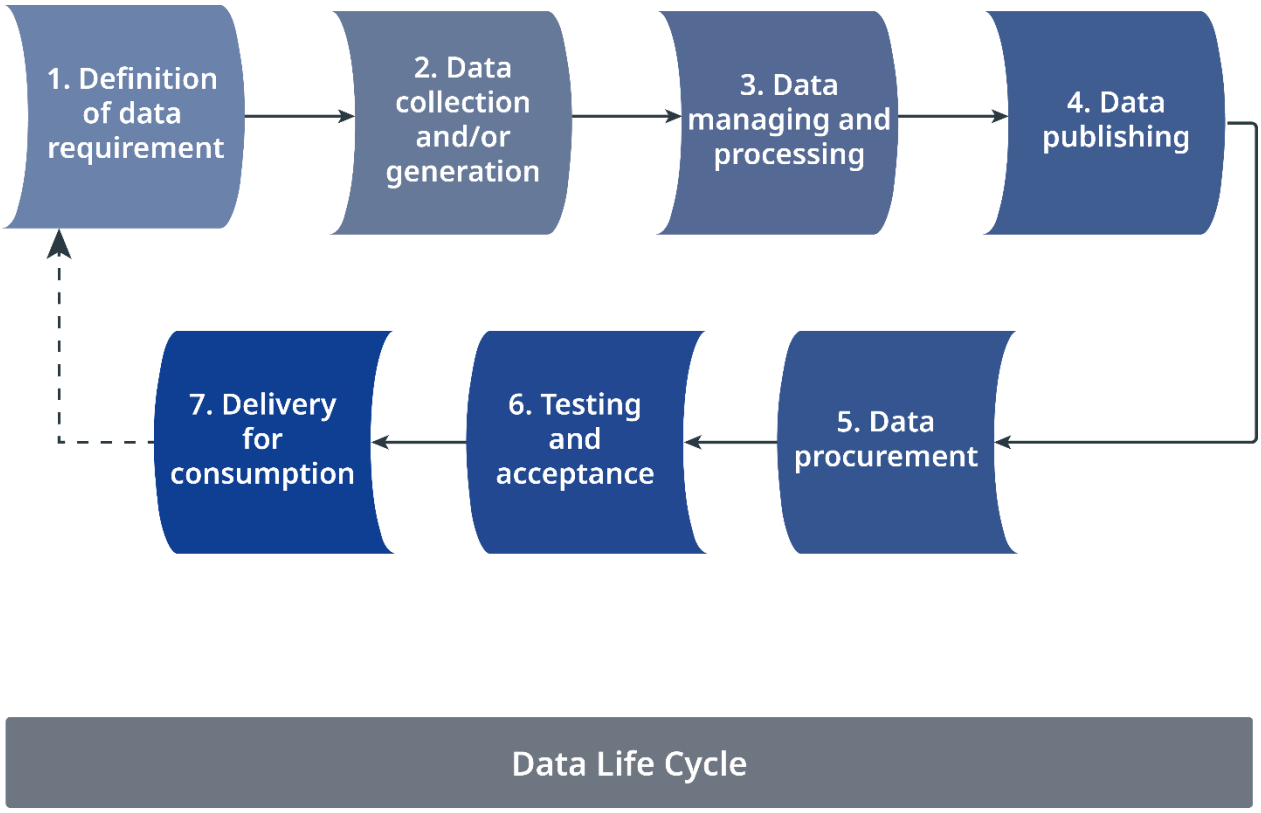

Figure 3 - A typical data processing workflow in the evidence generation process

5.2.1. DQ vs standardisation

From the point of view of regulatory decision making, DQ is distinct from data standardisation: data that are not fit for purpose in terms of answering a regulatory question will not become fit when standardised, and non-standardised data can be still used to answer a regulatory question. DQ also applies to individual and non-standard data sources.

The implementation of systems and processes to assure DQ is largely affected (and in some cases fully determined) by the adoption of standards[7] as well as by data management recommendations (e.g., FAIR data [Findable, Accessible, Interoperable and Reusable]) [16], and the availability of resources such as ontologies and MDM systems that underpin semantic interoperability.

Therefore, recommendations on specific standards and standardisation processes are not included in this Framework, while adoption of standards does drive implementation maturity levels.

5.2.2. Primary vs secondary use of data

Primary data collection is a process of collecting original data (newly collected), directly from the source. It can be gathered from observations, interviews and from biometrics (blood pressure, weight, blood tests, etc.) or surveys (questionnaires). Primary use of data is the use of information for the specific purposes they were collected for, while secondary use of data involves using the data that have initially been gathered for other purposes. See the glossary for an explanation of primary and secondary data.

In the application of guidelines and metrics, an important distinction arises between primary and secondary use of data. When systems are designed to collect and process data for a specified primary

[7] It should be noted that data standardization processes may alter the original information and its semantics. As noted later in this document, "standardization at source" is preferred to "a-posteriori" standardization.

purpose, or when a set of established requirements for secondary use exist, intrinsic and question specific aspects of DQ can already be considered at the time of collection and generation. It is thus possible to design systems and processes that guarantee some quality level required for evidence generation. In fact, much of the analysis decision and process specification can happen at this stage, and downstream analysis focuses on synthesising results and qualifying the level of uncertainty. This is generally not the case for secondary use of data, whether intended or opportunistic, where the quality criteria for usage may not coincide with the ones relevant for the existing purposes of data collection. In these cases, DQ can often only be controlled based on intrinsic determinants.

5.2.3. Publication vs data consumption

Along the data life cycle, data are processed through two different contexts. In one – publication – data are generated or collected, processed (sometimes harmonised), and made available. Examples are the aggregation of multiple sources of data to provide a dataset that is made available for general usage, for instance in a catalogue, or the generation of data by wearables and their "publication" through APIs. In the other context – consumption – data are procured and aggregated to support analysis. One example would be a study where multiple sources are collected, integrated, and harmonised to answer a specific question.

These two contexts may be overlapping (e.g., when data are collected for a specific primary use) or may be very distinct (e.g., when data are collected and published in a catalogue for a range of possible foreseen or unforeseen usages usually for secondary analysis).

It is useful to make this distinction as the purpose and the potential for quality assessment change between these two contexts. Even intrinsic aspects of quality for the same dataset may differ. Detailed specification of quality assessment may be developed separately for the publication or consumption contexts, e.g., for a data catalogue, in terms of acceptable minimal quality for generic usages, or for data procurement, in terms of minimal viability for a specific question.

5.3. Data and metadata

Metadata are traditionally defined as "data about data" providing context about their purpose and generation (e.g., characterisation of sources, data processing steps, lineage, and data elements definitions). When data consist of numeric or unstructured information (e.g., images), metadata are typically provided as an addition to a dataset. In general, the distinction between data and metadata is not well defined: some information appearing as metadata in one context (e.g., instrument provider for a test) can be considered as data in another (e.g., if assessing measurement bias).

For regulatory decision making, metadata should in general be considered similarly as data. More precisely, if some change in metadata would require a revision of derived conclusions, then it should be treated as data from the perspective of DQ [17]. In a DQ context, metadata should not be seen as limited to metrics (including DQ metrics) and summary description of datasets, but should extend to characterisation of sources, processes, and data elements definitions.

All types of metadata are often published in data catalogues[8], which have the purpose of allowing data to be discoverable and checked for fitness for purpose without revealing the data themselves.

[8] Various metadata catalogues are in development such as DARWIN and ENCePP. An example from Statistics Finland can be found here: https://www.aineistokatalogi.fi/catalog.

5.4. Data immutability

Data about some measured or observed aspect of reality may change in time, both reflecting the actual change in the observed entities, and the change of the available information at some given time. For instance, the weight of an individual may change, both due to a change of the individual itself, or because of a more accurate reading superseding a previous measurement.

It is important to distinguish the current data availability (or the current knowledge about reality) from the data as it was available at a given time (a specific record of knowledge about reality). The latest data reflecting what was known at a given time, is "immutable" in that, by definition, cannot change.

For any regulatory purpose, evidence should be based on data intended as the record of knowledge at a given time. In other words, data used to support regulation should be immutable. This doesn't imply that evidence can never be updated: any update should be considered as a distinct (albeit) related dataset. This is a foundational concept implied by most frameworks e.g., ALCOA and FAIR [16].

The consequence of this principle is that data used for decision making should be versioned and unaltered within any given version.

5.5. Data vs information

In its strictest definition, data represent facts or observations that are unprocessed (e.g., as generated by an instrument) while information represents insights originating from such data, once they are understood and processed in their context (e.g., a patient's response to a treatment as opposed to a set of individual readouts).

This Framework focuses on evidence generation that can be provided for decision making and as such it goes beyond a distinction between data and information.

5.6. Frame of reference (validation vs verification)

Some aspects of DQ can be measured in respect to different references, contained within the same dataset, or existing beyond the scope of the dataset either as a generic reference or external gold standard, or as the actual fact in the real world. For instance, the weight of an individual could be verified for quality based on its capture in the data (e.g., as a missing value), based on knowledge of a natural weight range or verified against a recorded weight in a source document.

In some frameworks, the assessment of quality within a dataset is referred to as "verification" while the assessment in respect to a source record or external gold standard is referred to as "validation". We follow these definitions.

Note: this notion of validation should not be confused with validation as a form of coherence checking, see section 6.3.1.

5.7. Granularity of data and DQ

For structured data, DQ can be typically assessed at different levels of granularity:

- The **value level** corresponds to a specific data point (e.g., a weight). This is also referred as **row level** when the focus is on all values relative to the same entity[9].

[9] The term entity is used to denote the subject of a set of values. In an information record, for instance about a sample, each value in the record would be expressing the measurement of a variable that relates to the same subject or entity. The entity is typically identified in a record via an identifier.

- The **variable level** (also referred to as **column level**) covers a data point for a whole set of individuals (e.g., weight as a variable in a clinical study demographics table). Metrics for DQ at the value level are often easily extended to the column level, for instance by converting values to a percentage[10].
- The **dataset level** covers an overall set of related observations[11]. In some contexts, a further distinction can be made, within a dataset, between parts of dataset that are about similar entities. When such distinction is made, such parts are referred to as **table level**, as those parts would normally appear in distinct tables.

The concept of granularity also applies to unstructured data, but the definition of its levels is generally specific to the data type and hence is not addressed in this general Framework.

This DQF will focus on the lowest possible level, i.e., for structured data, the value level. However, some metrics may be defined only at a higher level. For example, the plausibility of a single record of a person with a weight of 300 kg may not trigger a metric violation, but if 80% of the records are above 300 kg, it will.

6. DQ dimensions and metrics

The definition of DQ dimensions and metrics rely on the general definition of dimension, metrics, and measures:

- A **dimension** represents one or more related aspects or features of reality (e.g., for a physical object, its extension, or its durability).
- A **metric** represents a way to assess the value of a specific feature (e.g., absolute length measured in meters under some specified circumstances).
- A **measure** represents a single instance of a metric (e.g., 2 meters). More measures can be combined to derive more general metrics (e.g., average length)[12].

DQ metrics can be defined as indicators that when applied to a data source, can derive an assessment of one of more quality dimensions. A single quality metric can be used as an indicator for more than one dimension as expressed below in the examples for Coherence.

For some metrics, acceptance thresholds (e.g., maximum percentage of missing values) can be defined. In general, and for unintended secondary usages, such thresholds can be defined only depending on the question being asked. However, when data are collected for primary use, or when some well-defined secondary uses are targeted, thresholds may be defined (e.g., minimum/maximum) that apply even at the point of data collection. The quality of data is the sum of several features[13] of data, ranging from their correspondence to reality to their representation. It is useful to categorise such features in dimensions, which is a set of features whose measure reveals independent aspects of DQ. In other words, different dimensions answer different distinct DQ questions.

Several data frameworks propose an organisation of DQ in dimensions that are similar across frameworks, but often inconsistent in the exact definitions. This complicates a coherent assessment of DQ when multiple sources are aggregated.

[10] This is typically the case for binary or categorical data.
[11] In general, a dataset in support of a specific question will be comprised of homogenous data (e.g., a specific measurements) or of disparate types of data, which are related in that they measure (directly or indirectly) the same entities. In this sense different parts of a datasets are "linked" as they will share references to same entities.
[12] Note that measures could be unitless and not necessarily contiguous.
[13] Feature is here intended as a synonym of "aspect" or "characteristic".

This guideline introduces a set of dimensions (see figure 4) that are relevant from a regulatory point of view, complement them with a precise definition, possible metrics, and examples. The intention is to remove ambiguity and provide a useful reference that can help mapping different conceptualisation of quality from a variety of sources to a common denominator that is useful to frame metrics and maturity models for supporting evidence generation.

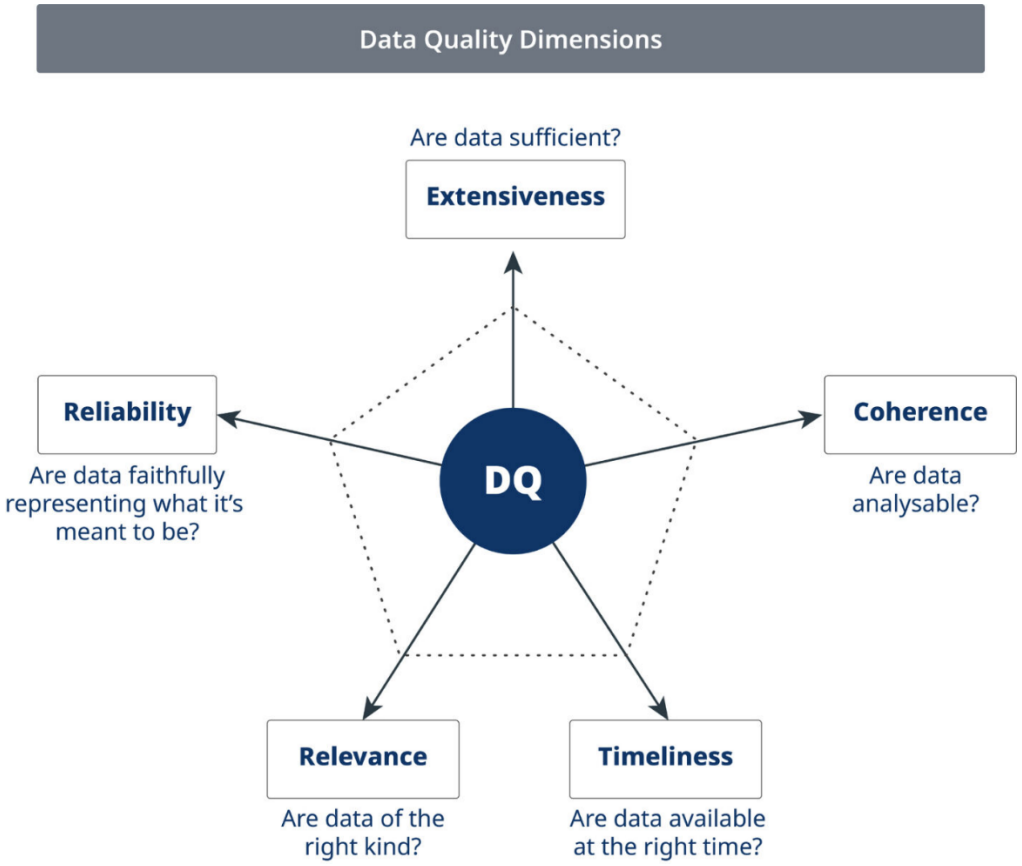

Figure 4 - Dimensions of data quality

6.1. Reliability

Reliability is defined as the dimension that covers how closely the data reflect what they are directly measuring.

The Reliability dimension answers the question: to what degree are data accurate or correctly representing an observed reality? When considering the "fit for purpose" definition of quality, Reliability covers how correct and true the data are.

6.1.1. Reliability sub-dimensions

Given this definition, sub-dimensions can be defined:

- **Accuracy** defined as the amount of discrepancy between data and reality. This definition of accuracy encompasses measures of the amount of wrong information in a dataset (data systematically not reflecting reality) with the formal definition of accuracy in measurements (e.g., the distance between the measurements and the real value). For example, the weight of a person could be incorrect due to a data transcription error, or because a person is measured fully clothed, given a systematic excess weight of 1 to 2 kg.

- **Precision**[14] defined as the degree of approximation by which data represents reality. For instance, the age of a person could be reported in years or months.

6.1.1.1. Other DQ concepts related to Reliability[15]

Strictly related to Reliability is the concept of **Plausibility**, defined as the likelihood of some information being true. Plausibility can be a 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 (unless the foundational determinants indicate otherwise) and likely revealing some errors in the measurement or the processing of the data. Plausibility results from the comparison of a data item to typical or necessary characteristics of the entity it intends to represent and is therefore hard to measure as a pure intrinsic characteristic as it depends on the availability of background knowledge or an external gold standard.

Traceability (also referred to as **data lineage** or **provenance**) refers to data presenting the knowledge of how data came to be, what source it originated from, and what processing it went through before appearing in its current form. Traceability is a feature of data that falls within Reliability in that it connects what is measured with the actual data.

6.1.2. Considerations for Reliability

Reliability fundamentally depends on the systems and process in place for the primary collection of data and its processing and curation in further phases of the evidence generation process both for primary and secondary use cases.

In the absence of errors, accuracy would not decrease along the data aggregation process. Precision may instead decrease when data are harmonised to a Common Data Model (CDM), as this may call for less precise representation than original sources to fit the model[16].

Intrinsic aspects of Reliability are hard to measure in a pure data-oriented framework, however Plausibility measures can provide a way to detect some classes of errors. Reliability is independent from a specific question, though each question, in relation to data, will set a threshold for acceptable Reliability.

[14] This definition of precision encompasses the notion of "reproducibility of values" under repeated measurements, in that it captures how "coarse" is the correspondence between data and the characteristic it intends to measure.
[15] " Other concepts" present relevant aspects of DQ that falls within a dimension or that are in common use, but that don't strictly adhere to definition of the dimension provided.
[16] An example could be a CDM allowing timestamp in seconds where a source may use milliseconds, or a CDM prescribing some terminology, which is less specialised in some areas than the one used in the source.

6.1.3. Examples of reliability metrics

Table 1 – Example of metrics for Reliability

Sub-dimension	Metric group	Abstract metric	Reference	Example
Plausibility (proxy for Accuracy)[17]	Atemporal Plausibility	Data values and distributions agree with internal measurements or local knowledge	Validation	Height and weight are a positive value. Counts of unique subjects by treatment are as expected (respect to an applicable gold standard).
		Data values and distributions for independent measurements of the same fact agree	Verification	Oral and axillary temperatures are similar. Serum glucose measurement is similar to finger stick glucose measurement.
		Logical constraints between values agree with common knowledge	Verification	The patient's sex agrees with sex-specific contexts (pregnancy, prostate cancer).
		Values of repeated measurement of the same fact show expected variability	Verification	Weight values are similar when taken by separate nurses within the same facility using the same equipment.
		Data values and distributions agree with trusted reference standards	Validation	HbA1c values from hospital and national reference lab are statistically match under the same conditions.
		Equivalent values for identical measurements are obtained from two independent databases representing the same observations with equal credibility	Validation	Date of birth value in the EHR is not identical to that in the registry record of the same patient.

[17] Our examples are limited to Accuracy as this is the most common application of Plausibility. In theory, Plausibility could extend to other dimensions of Reliability (e.g., a weight expressed in grams is likely to be imprecise if all values end with three zeros).

Sub-dimension	Metric group	Abstract metric	Reference	Example
		Two or more dependent databases yield similar values for identical variables (e.g., database 1 abstracted from database 2)	Validation	Cancer stage value in the EHR does not corresponds with a NAACCR code in the tumour registry of the same patient.
		Calculated data values agree with common knowledge	Validation	Height and weight of a patient resulting in an implausible BMI value of less than 5 propose an inaccurate height, weight, or both.
	Temporal Plausibility	Observed or derived values conform to expected temporal properties	Verification	Discharge date happens after admission date.
		Sequence of values that represent state transitions conform to expected properties	Verification	Date of primary vaccine administration precedes that of the booster vaccine administration.
		Observed or derived values have similar temporal properties across one or more external comparators (gold standard)	Validation	Length of stay for outpatient procedure conforms to insurance data for similar populations (no more than 1 day).
		Sequences of values that represent state transitions are similar to external comparators (gold standards)	Validation	Immunisation sequences matches that of the EMA recommendations.
		Measures of data value density against a time-oriented denominator are expected based on external knowledge	Validation	Count of immunisation per month shows an expected spike outside of flu season.

In case a gold standard is not available, a metric can be "verified" based on a comparison with a similar metric from another source. This could be equivalent values for identical measurements from two independent databases. This is not necessarily a "gold standard" but may be the best available option. Alternative methods might be explored when no "gold standard" is available (see 8.6 for more details).

6.2. Extensiveness[18]

Extensiveness is defined as the dimension capturing the amount of data available.

The Extensiveness dimension answers the question, "how much data do we have"? When considering the "fit for purpose" definition of quality, Extensiveness covers how sufficient the data are.

6.2.1. Sub-dimensions of Extensiveness

When considering the amount of information available, one can think of expressing this as a percentage relative to the total amount of information that could be available. The distinction between Completeness and Coverage stems from the definition of the scope of totally available information.

- **Completeness** measures the amount of information available with respect to the total information that could be available given the capture process and data format. Data unavailable in the dataset (either due to systematic reasons such as information available in the data source but not included in the data model, or specific entries that are unavailable for a given field) are called "missing". For example, the percentage of non-missing values for a required field (e.g., sex) in a dataset would be a measure for completeness.[19]
- **Coverage** measures the amount of information available with respect to what exists in the Real-World, whether it is inside the capture process and data format or not. Coverage may not be easily measured as the total information may not be definable or accessible. An example of coverage is the percentage of a given population (e.g., a country or a specific demographics) available in a dataset. When considering coverage in its relation to evidence generation methods, it is also referred to as observability [18].

6.2.1.1. Other DQ concepts related to Extensiveness

Two concepts that are often associated to extensiveness are representativeness and missingness. While these concepts describe to a certain extent how much data is available, they are more importantly used to characterise how much data is reflecting reality. **Representativeness** is defined as the data having the same characteristics as the whole it is meant to represent (e.g., whether a set of individuals present in a dataset is representative of a population under study). **Missingness** is meant as the characterisation of what is the impact of incomplete data in respect to coverage of a dataset.

[18] Extensiveness combines two typical dimensions found in DQFs: Completeness of data coverage and Coverage. They are here combined as they both relate to the amount of data available.
19 There is a fundamental distinction between missing data that are known to exist (e.g., the date of birth of a patient), or missing data whose existence is unknown (e.g., the presence of co-pathologies). Quite often in a data model the definition of a variable as "required "implies that the relative values are known to exist, and therefore when such data is missing it is a "missing known". In general, in the absence of explicit negation, it may not be possible to distinguish" missing known" from" missing unknowns". When some data point is expected to be captured, the inability to distinguish" missing knowns" from" missing unknowns" is an issue of Reliability (one is unable to assesses if data corresponds to the reality it is meant to represent).

6.2.2. Considerations for Extensiveness

The Extensiveness of the information collected depends on the specification of the data collection process. However, when combining different datasets for secondary use, there is no guarantee about the completeness of the overall dataset. On an intrinsic level, one can resort to metrics to assess the level of completeness of data. Metrics that assess how much data are present in a dataset in respect to what could be present in a given data model are fairly simple to compute. Metrics that assess how complete the data are with respect to the population they intend to measure, are more complex and may involve the engagement of gold standards. Completeness with respect to a schema is easily definable, while Coverage depends on some assumptions that can be defined only with respect to a question. Thresholds used as acceptance criteria can also be defined with respect to a question (e.g., 80% complete).

6.2.3. Examples of metrics for Extensiveness

Table 2 – Example of metrics for Extensiveness

Sub-dimension	Metric group	Abstract rule	Reference	Example
Completeness	Missing required values	Missing values with respect to a local schema – over time	Verification	Breed or sex of the animal should not be NULL.
		Missing values with respect to a local schema – single time	Verification	The encounter ID variable has missing values.
	Estimated missing values	Missing values with respect to common expectations	Verification	Sudden drop of diagnosis codes due to a defective feed from a claim clearing house vendor.
		Relative assessment of missing values with respect to a trusted source of knowledge	Validation	The current encounter ID variable is missing twice as many values as the institutionally validated database. A drop in ICD-9CM codes upon implementation of ICD-10-CM.
Coverage		Coverage of a population	Verification	The percentage of a target population present in a database.

6.3. Coherence

Coherence (also referred to as Consistency[20]) is defined as the dimension that expresses how different parts of an overall dataset are consistent in their representation and meaning.

The Coherence dimension answers the questions: is the dataset analysable as a "whole" or are additional steps needed like linkage of multiple datasets? Is the format of values (e.g., dates) the same across the dataset? Is the precision of values the same (e.g., age always approximated to

[20] Consistency and Coherence can be considered largely synonymous, with the caveat that detection of inconsistencies is often a way to measure the reliability of data.

years)? Are references to entities consistent so that information about the same entity is properly "linked" across parts of the dataset?

When considering the "fit for purpose" definition of quality, coherence relates to the analysability of data.

6.3.1. Sub-dimensions of Coherence

Coherence is a complex and nuanced dimension that includes the following sub-dimensions:

- **Format Coherence**: whether data are expressed in the same way throughout a dataset (e.g., a dataset mixing dates represented as DD-MM-YYYY and MM-DD-YYYY will not be suitable for an integrated analysis).
- **Structural or Relational Coherence**[21]: whether the same entities are identified in the same way throughout a dataset. A sub-aspect of Structural Coherence is that references are resolved to the correct entities (e.g., a sample annotation table with refer to the correct value in a result table).
- **Semantic Coherence**: whether the same value mean the same thing throughout a dataset. For instance, whether "anuria" means a condition of total cessation of urine production or the measurement of the amount of urine, or whether the same notion of a measure is intended to have the same precision throughout a dataset.
- **Uniqueness**: Uniqueness is the property that the same information[22] is not duplicated but appears in the dataset once. This problem is typical for data aggregated from different sources. Note that data with some redundancy will score lower in the Uniqueness dimension, but those extra records could help improving other dimensions, such as Reliability.

- Other DQ concepts related to Coherence is **Conformance** when this is defined with respect to a specific reference or data model. Conformance may practically be the best way to assess Coherence, and it also specialised as format, Structural and Semantic Conformance. As an example, conformance would assess if the representation of data is coherent by assessing if it is the same as an overall target standard (e.g.: DD-MM-YYYY)[23].**Validity**[24] is a narrower case of Conformance that is defined when the reference model is specific to the dataset being assessed. As an example, if a file is associated to a schema specifying that all dates in the D.O.B. filed should be in DD-MM-YYYY format, the file could be directly assessed as valid or not.

6.3.2. Considerations for Coherence

Coherence of data at source largely depends on foundational determinants such as the synchronisation of processes and systems across an organisation generating data, or when multiple data are aggregated on the commitment of such organisation(s) to the use of internal or external data standards. By extension, Coherence for data aggregated and repurposed for secondary usage depends on the availability of shared standards and reference data. The intrinsic aspects of Coherence of a dataset can be improved, largely within a data standardisation processing step. However, improving Coherence involves approximating or clarifying the meaning of data. Access to the source system and

[21] Structural and relational Coherence as synonyms here. It may be the case that these two concepts are distinct or non-tabular data. This distinction will be addressed if the need arises, in extensions of this framework to specific data types.
[22] It is worth noting that" information" is distinct from data. Two distinct measurements resulting in the same data would not constitute duplicate information (and such measurements would most likely differ in value, when metadata is included). Whereas the same measurement reported two times would amount to a duplication.
[23] A file could be coherent, but not conformant, if all values are coherent (e.g.: MM-DD-YYYY) while an overall target standard proposed to assess conformance requires DD-MM-YYYY.
[24] As noted in 5.4, this is a different meaning (in common use) then what is defined for" Validation".

processes is often required for clarifications as an example. Some aspects of semantic Coherence may be difficult to assess with a metric and hence can only be assessed with respect to a specific question and analysis strategy.

6.3.3. Examples of metrics for Coherence

Table 3 – Example of metrics for coherence

Sub-dimension	Metric group	Abstract rule	Reference	Example
Format coherence (conformance)	Syntactic constraints	Data Values conform to internal formatting constraints	Verification	Sex is only one ASCII character.
	Allowed values	Data values conform to allowable values or ranges	Verification	Sex for the animal only has values "M", "F". or "U".
		Data values conform to the representational constraints based on external standards	Validation	Values for primary language conform to ISO standards.
Relational coherence (conformance)	Reference Coherence	Data values conform to relational constraints	Verification	Patient medical record number links to other tables as expected.
		Unique (key) data values are not duplicated	Verification	A medical record number is assigned to a single patient.
		Data values conform to relational constraints based on external standards	Validation	Data values conform to all not-NULL requirements in a common multi-institutional data exchange format.
	Schema Coherence	Changes to the data model or data model versioning	Verification	Version 1 data does not include medical discharge hour.

Sub-dimension	Metric group	Abstract rule	Reference	Example
	Computational Coherence	Computed values conform to programming specifications	Verification	Database calculated and hand calculated BMI (body mass index) values are identical.
		Computed results based on published algorithms yield values that match validation values provided by external sources	Validation	Computed BMI percentiles yield identical values compared to test results and values provided by EMA.
Semantic coherence (conformance)	Precision Coherence	The precision of values is fitting a target standard	Verification	E.g., two decimal digits are used and generally not zero.
	Semantic Coherence	Use of code lists is consistent across data	Verification	E.g., the level of a MedDRA coding for an indication doesn't vary across the dataset.
Uniqueness		Same subject is represented with the same identity	Verification	William Smith is also represented as Bill Smith with the same DOB.
		Same subject is represented with multiple identities	Verification	William Smith and William Smith appear as separate individuals instead of the same individual.
		The data records of individuals are matched using unique keys	Validation	William Smith's DOB ID matches with Bill Smith's DOB and ID.

6.4. Timeliness

Timeliness is defined as the availability of data at the right time for regulatory decision making, that in turns entails that data are collected and made available within an acceptable time[25].

The Timeliness dimension answers the question: are the data reflecting the intended reality at the point of time of its use?

When considering the "fit for purpose" definition of quality, Timeliness covers how closely the data reflect the intended reality, at the time in which it is used.

6.4.1. Sub-dimensions of Timeliness

- **Currency** is a specific aspect of Timelines that considers how fresh the data are (e.g., current, and immediately useful)[26].

6.4.1.1. Other DQ concepts related to Timeliness

In the context of this Framework **Lateness**, intended as the aspect of data being captured later than asserted, falls in the dimension of Reliability (does the data correspond to reality, at the time it intended to measure?).

6.4.2. Considerations for Timeliness

Timeliness is determined by the systems and processes used to collect and make data available.

6.4.3. Examples of metrics for Timeliness

Table 4 – Example of metrics for timeliness

Sub-dimension	Metric group	Abstract rule	Reference
Currency	N/A	The average time of updates in a database (or timestamp)[27]	Verification
		The last update of a database (or timestamp)	Verification

6.5. Relevance

For the purpose of Data Quality assessment, 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 and more data-focused than the more commonly understood meaning of "relevance"[28] (i.e.: relevance of a

[25] While Timeliness is not further distinguished in this version of this framework, the definition highlights two different aspects of Timeliness: respect to the time data is measured (e.g.: delay between measurements of body temperature respect to the onset of fever), and respect to the time data is collected (e.g.: made available in a database).
[26] Note that the lack of currency doesn't imply a lack of timeliness: historic data may lack currency, but still be timely for retrospective studies,
[27] Measures of Currency focus on a narrower aspect of Timeliness and are generally based on the time data are actually recorded in a database (rather than the time of data collection).
[28] Relevance as a common term is defined as the degree to which something is related or useful to what is happening, discussed about or for a given objective.

data source to generate valid evidence informing a specific research question based on the study design).

To distinguish these two meanings, this text explicitly makes use of the terms "Relevance DQ Dimension" (relevance as here defined), and "Relevance to a question" (the more generic meaning).

The Relevance DQ dimension answers the question: does the dataset present the values (or data elements) that are needed to address a specific question, using a specific method?[29] .

When considering the "fit for purpose" definition of quality, the Relevance DQ dimension describes how the data cover the aspects of reality that are intended to be measured.

In this framework, relevance to a question is captured by "question specific determinants" that apply to all dimensions.

The dimension previously introduced partition DQ aspects on the basis for some driving questions (is data truly representing reality? How much data is there? Is data analysable as a whole? Is data available at the right time?). A missing question is about what type of data is there, and this is what the "Relevance DQ dimension" is covering.

Given the context described, Relevance can only be characterised in relation to a research question and a data analysis strategy[30]. However, in some cases, it is possible to identify a set of frequently required research questions that can be characterised from the Relevance point of view, in the context of a specific type of data source. This is referred to as **Relevance for a domain**, where 'domain' is a shorthand for a 'research questions domain'.

6.5.1. Examples of metrics for Relevance[31]

Table 5 – Example of metrics for relevance

Sub-dimension	Metric group	Abstract rule	Reference	Example
N/A	N/A	The number of variables (columns) available in a given dataset vs the number of required variables.	Verification	N/A

7. General recommendations and maturity models

Selecting datasets to use in regulatory decision making ultimately requires knowledge of the degree to which such data satisfy the Reliability, Extensiveness, Coherence, Timeliness and Relevance criteria. Such quality dimensions build up along an overall life cycle from generation through processing to aggregation and ultimately analysis, and in such process, data originally gathered for other usages can be repurposed when ethical or legal requirements are met [19].

The choice of quality measures and checks varies broadly depending on data types and their intended use. However, it is possible to organise such measures and checks following a coherent structure that helps achieve homogeneity and identify gaps.

The following tables exemplify how determinants of quality (Foundational, Intrinsic or Question-Specific) affect the different quality dimensions for both data and metadata. These tables provide a

[29] The distinction between Extensiveness and Relevance can be clarified by the two distinct questions that these dimensions are answering: how much data do we have? (Extensiveness) vs what data to we have? (Relevance).
[30] By data analysis strategy, the definition of assumptions, decisions, and methods to address a specific question is intended.
[31] This metric is provided as an example to clarify what pertains to the dimension of Relevance. Not all variables are equivalent and actual metrics will need to be specified for specific use cases and/or data types.

guidance for what metrics and actions apply at which stage of the data life cycle. For example, the dimension of Extensiveness is determined exclusively by Foundational determinants at production time. Further in the data life cycle, data intrinsic measures can only partially assess the degree of Reliability (plausibility metrics).

These tables also form the basis for the development of maturity models for the characterisation of DQ for regulatory purposes. The maturity models provide guidance as to how determinants can be characterised in successive levels of maturity. Higher maturity levels support the strongest possible evidence in the most efficient way.

Three distinct maturity models are provided, corresponding to the three determinants, to depict how maturity evolves with respect to process characterisation, intrinsic aspects (metrics) and the definition of target questions. These models are meant to apply to the different steps and actions that compose an overall evidence-generation framework.

It should be noted that the maturity models provided are abstract in the sense that they provide the classes or recommendations that need to be complemented with implementation detail for specific data types and use cases.

It takes time to characterise and implement processes to achieve higher maturity levels both for data source holders, but also for regulatory assessors to understand the impact of a higher maturity model. This is also context dependent e.g., disease area, disease frequency, health system etc. The DQF will be updated in the upcoming years with further deep dives in regulatory use cases of particular interest to guide clinical assessment for medicines regulation.

Table 6 – Characterisation of the effect of determinants on data quality dimensions

Determinant/ Dimension	Reliability	Extensiveness	Coherence	Timeliness	Relevance
Foundational		**Primary** Data collected following established protocols can be sufficient to address regulatory questions.			**Primary** Normally guaranteed by the design of the data collection process.
	Primary and secondary Data reliability (in all its aspects) results from systems and processes in place for data generation or collection. Reliability is affected by data processing and transformations at later stages e.g., standardisation to a CDM.	**Primary and secondary** The data collection protocol determines what data are collected.	**Primary and secondary** Dependent on the orchestration of processes originating data and on the commitment to internal or external data standards.	**Primary and secondary** Solely determined by systems and processes.	
	Secondary Precision may decrease during data transformation and harmonisation processes.	**Secondary** There is no guarantee on the completeness of an integrated dataset or its coverage for a different use case, and this can only be assessed or controlled.	**Secondary** Relies on shared standards and reference data. Documentation on data generation processes may be needed to enhance coherence.		**Secondary** Normally assessed for a specific use or a class of usages when datasets are selected.
Intrinsic	**Primary and secondary** Plausibility measures can be used to detect a (limited) class of reliability issues.	**Primary and secondary** Completeness measures based on a data model are easy to implement.	**Primary and secondary** Coherence can be measured exclusively based on data (with eventual access to	**Primary and secondary** Some aspects of timeliness may be observed in the datasets (e.g., event	**Primary and secondary** Relevance of data is not dependent on a dataset itself.

Determinant/ Dimension	Reliability	Extensiveness	Coherence	Timeliness	Relevance
	Direct measures of accuracy require access to the source of data.		datasets-independent reference data).	dates to determine currency). A dataset itself cannot in general reveal how current its information is.	
		Secondary Coverage measures are more complex and may require confrontation to a golden standard.	**Secondary** Coherence can be largely improved based solely on a dataset and data-independent elements (e.g., mapping to a common standard). A full resolution of coherence may require access to additional information on processes. Coherence needs to be assessed every time a new data source is "integrated".		
Question specific	**Primary** Processes and systems to collect data are usually designed to answer a specific question and to meet the required targets, across DQ dimensions, that such target entails.				
	Secondary Threshold for acceptable reliability can be defined only respect to a specific question and method.	**Secondary** Coverage and completeness depend on a question: metrics can be defined only respect to a specific question and method, or for a domain. For completeness, typically a question would determine a set of acceptance thresholds and general metrics.	**Secondary** Some assessment of semantic coherence (data distribution coherence or abstraction coherence) may only be measured respect to a specific question and method.	**Secondary** Acceptable timeliness depends on the question and its broader regulatory usage (e.g., approval vs monitoring).	**Secondary** Relevance can only be determined in relation to one or more questions.

Table 7 – Characterisation of the effect of determinants on Metadata quality dimensions

Determinant/Dimension	Reliability	Extensiveness	Coherence	Timeliness	Relevance
Foundational		**Primary** For primary data, the extensiveness of metadata can be characterised at source.	**Primary** Metadata coherence relies on the presence of common standards and terminologies.		**Primary** Normally guaranteed by the design of data collection process.
	Primary and secondary Reliability of Metadata relies on the processes to collect it, along the whole data processing chain. One key aspect to ensure reliability is to capture metadata as close to the source as possible.			**Primary and secondary** Timeliness of Metadata are purely dependent on the processes supporting its collection.	
			Secondary For secondary data, coherence relies on the presence on widely agreed standards and shared resources such as ontologies or reference data services.	**Secondary** When data are repurposed and used in different systems, timeliness of metadata should be enforced by design (metadata should be in synch with the data).	**Secondary** Relevant metadata can be required and controlled by a downstream system but cannot be guaranteed at source.
Intrinsic	**Primary and secondary** Some metadata (e.g., summary statistics) can be generated from a dataset. When data and metadata are considered as whole, traceability can also be assessed by intrinsic measures.	**Primary and secondary** Intrinsic measures for meta DQ mimic the ones for data (e.g., completeness and missing fields). Unlike data, metadata assessment may not require references to	**Primary and secondary** Metadata coherence solely depends on a specific metadata and data-independent elements (e.g., shared reference data).	**Primary and secondary** The assessment of timelines aspect of data typically depends on metadata (e.g., timestamps).	**Primary and secondary** Relevance of metadata does not depend on a dataset itself.

		golden standards (e.g., missing metadata values is not related to sampling of a population).			
Question specific	**Primary** Metadata requirements are designed for a specific question and are normally sufficient to address it.				
	Primary and secondary Metadata should be in general reliable independently of a specific question (not all metadata collected may be relevant for all questions).		**Primary and secondary** The coherence of metadata is independent from a specific question.	**Primary and secondary** Timeliness of metadata are independent from a specific question.	
		Secondary The characterisation of what metadata are necessary is ultimately dependent on a question (or set of typical questions)			**Secondary** Relevance of metadata is purely dependent on a question (or range of questions).

7.1. Foundational determinants: Recommendation and maturity levels

A characterisation of the systems and processes underpinning data generation and manipulation (foundational determinants) is necessary to manage DQ. Below is a set of defined maturity levels, each providing a progressive hierarchy of recommendations for the characterisation of foundational determinants, with the intention to chart a direction of improvement towards adequate and efficient characterisation of these aspects of DQ (see figure 5). It is recommended that FAIR principles [16] for data and metadata be implemented as early as possible, or partially, along maturity models.

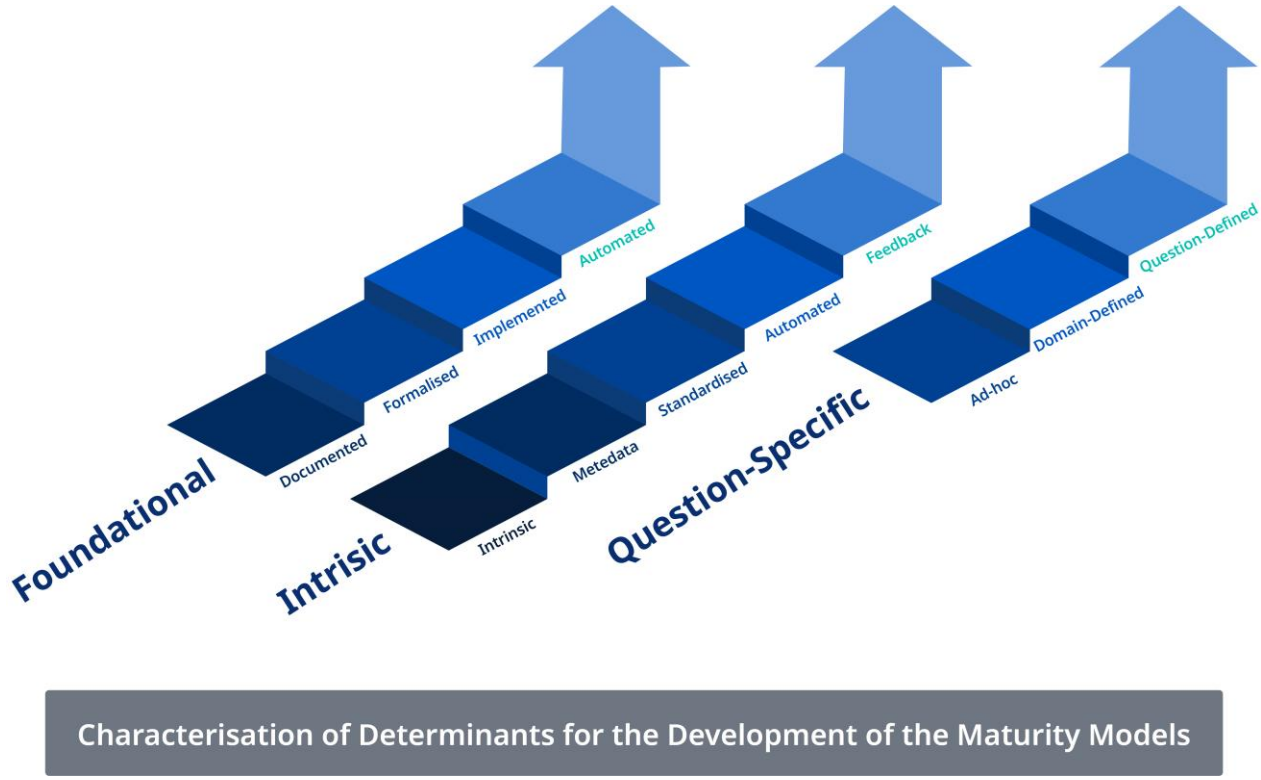

Figure 5 - Maturity model for data quality determinants

7.1.1. Level 1: documented

For data to be adequate for decision making, at a minimum, the processes that pertain to data generation and manipulation should be documented, true, verifiable (when relevant, this may extend to training procedures) and versioned. This is fundamental and ensures the reliability of any derived information. The documentation should cover determinants for Reliability (Precision), Extensiveness, Coherence and, when relevant, Timeliness. While some of these determinants depend on a specific question, data collection processes and systems will generally be designed with some generic questions in mind. The provision of documentation for data processing and transformation are also essential to guarantee that Reliability is preserved and should be provided for all such processing by different actors along the data life cycle.

From a metadata perspective, this means metadata (in some form) should always accompany a dataset it refers to.

To guarantee the quality of data, audit procedures or other controls should be in place.

When a system is designed for continuous data collection (as opposed to a one-off capture), additional processes of performance monitoring and improvement should be in place.

7.1.2. Level 2: formalised

The second level of the maturity model includes and extends the first level, by requiring that, whenever possible, documentation and metadata should be following an industry standard framework or QMS[32] . Level 2 should be considered the minimal level of acceptable maturity, though exceptions may arise for novel data types. The recommendation to use standards extends to metadata.

7.1.3. Level 3: implemented

Systems are in place that implement industry standard DQ processes systematically and by design. Infrastructure should be in place to support data management, including support for standardisation (e.g., reference data management or MDM). By reducing the potential for human error, such an implementation can generally improve Reliability and Coherence. Such an implementation may also be necessary to guarantee Timeliness and it should ensure that metadata are collected by design, and as close to the data generation or collection events as possible.

7.1.4. Level 4: automated

The operations and output of the above systems and infrastructure should be machine readable as to unify data and DQ elements for direct downstream consumption. All data and metadata should be represented following FAIR principles [15] to allow complete automatic processing of quality parameters This is intended to be an aspirational level.

7.2. Intrinsic determinants: Recommendations and maturity levels

Beyond documented evidence of how data were collected or generated, measures of intrinsic aspects of DQ can be applied. These can be directly derived from the dataset, but their computation could also rely on some external body of knowledge.

7.2.1. Level 0: intrinsic

There are no hard minimal requirements for quality, as any piece of data can be assessed before being used to generate evidence[33]. Nevertheless, the propagation of data without an associated quality assessment should be discouraged.

7.2.2. Level 1: metadata

Data are provided with a set of quality metrics as metadata. Some of these data can be directly derived from the dataset while other derive from the overall data collection process (e.g., sampling).

[32] What industry standards or frameworks applies depend on specific data types and use cases, and as such can be defined only in specialisations of this framework. Some initial references are however provided in the ”implementation notes” session.
[33] This initial level is assigned "0" to clarify that it corresponds to data "as is" irrespective of their intended use for regulatory decision making.

Metadata should also cover the description of data elements that are necessary for its interpretation (e.g., data dictionaries).

7.2.3. Level 2: standardised

Data are provided with a standardised set of quality metrics, which can be compared across datasets. When applicable or possible, standards should extend to cover reference knowledge that can be used to assess a dataset in respect to what it is meant to represent (e.g., typical population distributions to assess biases). Metadata makes use of shared definitions, which also enable comparability and integration across datasets.

7.2.4. Level 3: automated

Quality assessment is automated (at least for a large extent of metrics). In general, this is feasible only when data are represented in standard ways (e.g., in a CDM), so that a standard library of tests can be run on incoming data. Data and metadata should follow FAIR principles [16][34].

7.2.5. Level 4: feedback

There is a data ecosystem in place so that quality assessment by data consumers can provide feedback to improve the data collection and production process, thus allowing a continuous monitoring and improvement of DQ.

(Note that the order of maturity of level 2 and 3 may change for particular data types.)

7.3. Question-specific determinants: Recommendations and maturity levels

In general, it is not possible to assess the Relevance of a dataset, or aspects of Extensiveness and Precision, without a target question and a defined analysis strategy. However, when considering the adoption of a large body of data for regulatory decision making and its possible use beyond primary use cases, it becomes important to articulate to what degree DQ, including Relevance, can be assessed a-priori.

7.3.1. Level 1: ad-hoc

All dimensions that are question specific are assessed only at "question time" on an ad-hoc basis.

7.3.2. Level 2: domain-defined

A range of common questions is identified, from which metrics and thresholds can be derived that can be used to guarantee acceptable levels of quality. Data published in data catalogues should make use of such metrics.

7.3.3. Level 3: question-defined

The requirements for a specific question are precisely codified and can be mapped to metrics and thresholds in a way that could automatically assess the Relevance of a dataset for a specific question. At this level, the context under which data will be interpreted for decision making should be formalised

[34] As for foundational determinants, FAIR principles should be applied as early as possible, at least partially. Level 3 requires a full implementation of FAIR principles.

and shareable. This is the natural level for primary use cases, while for secondary use of data, this should be intended as an aspirational level.

8. Considerations for implementation of DQF

This section provides a set of observations and recommendations to guide the implementation of this DQF and its specialisations, to help achieving higher levels of maturity.

8.1. Quality at source

As a general guideline, in designing data collection and generation processes, aspects of DQ should be addressed as early as possible. For instance, assessment of quality done close to the moment of production can help in correcting a collection error. The further data travels from the original context, the harder it becomes to correct issues. This is particularly relevant for metadata as knowledge of the context of data generation is maximally present only at generation time.

8.2. The role of Master Data Management (MDM) and reference data

The availability of MDM and reference data has a direct impact on DQ. It is often a pre-requisite for data consistency, and it can even impact Reliability in some data production scenarios (e.g., materials data), as disconnected information can result in erroneous information. More broadly, MDM and reference data enable automation of a range of DQ checks and hence have an impact on Reliability as well.

Shared MDM and reference data can address aspects of Coherence beyond the primary use case that the data were generated for, as the use of standards guarantees some level of Semantic Coherence is maintained even beyond data aggregation steps.

8.3. The role of QMS and computerised systems

The implementation of a DQF at higher maturity levels requires the formalisation and implementation of systems and processes to support DQ.

A Quality Management System (QMS) [20] 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 [GCP], Good Laboratory Practices [GLP] or Good Manufacturing Practice [GMP]). 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. Life Science Industry specific QMSs should be considered depending on the nature of the data:

- Clinical trial data: ISO 14155 and EU Directive 2001/20/EC for GCP (clinical trial data)
- Data from medical devices or diagnostic products: ISO 13485 Quality System Regulation (QSR)
- Data from lab research: EU Directive 2004/9/EC and 2004/10/EC, GLP
- Data from clinical labs: ISO 15189 and ISO 17025

Whenever possible DQ processes should be framed in the context of standard QMSs.

Furthermore, in today's digital world, foundational DQ determinants are also impacted by computerised systems, that are used to create, modify, maintain, archive, retrieve, or transmit data. A software development life cycle including software quality assurance system ensures the appropriate design, development and testing of the software. This can be targeted through the ISO 250xx standard family,

named Systems and Software Quality Requirements and Evaluation (more specifically, data models can be addressed with ISO 25012). Computer system qualification and validation ensures the software is appropriately implemented, and necessary process controls are in place for using it according to its specifications, including documentation, access control, vendor management and audits. The EMA/226170/2021 Guideline on computerised systems and electronic data in clinical trials provides direction for GCP but can be adopted more broadly. The following decision tree provide guidance on how to consider QMSs for a DQF implementation (see figure 6).

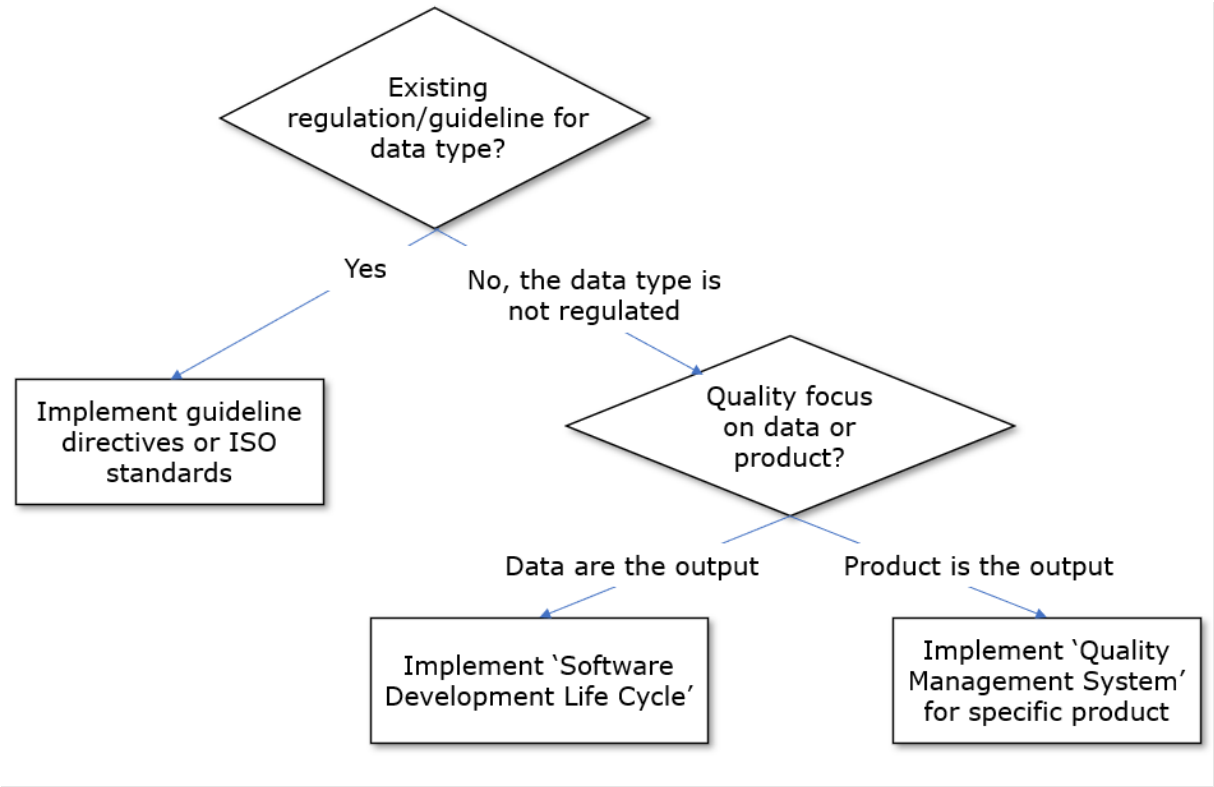

Figure 6 - Decision tree for QMS adoption in DQF implementation

8.4. The role of ISO and industry standards

The International Organisation for Standardisation (ISO) has produced standards providing frameworks for the implementation of various data management aspects, that are field tested and for which platforms, supporting services and certification bodies are established. These standards are often developed for implementation of industries where EMA's regulatory decision making does not apply.

ISO 9000: Describes the standards for quality management systems on all levels of an organisation. The adoption of this standard could be considered if no industry specific QMS applies.

ISO 8000: Describes the standards for the quality of Master Data and their exchange between systems[35]. The standard describes how Master Data conform to a set of specification expressed in a formal syntax and use specified identifiers to check against data requirements that point to a data dictionary. This standard also covers the methods for achieving data governance, data quality management, data quality assessment and rules for determining the quality of master data and industrial data. This includes the exchange of characteristics of data and identifiers and data

[35] To clarify, the ISO 8000 series does not establish a new management system. The series, instead, extends and clarifies ISO 9001 for the case where data are the product.

processing like creating, collecting, storing, maintaining, transferring, exploiting, and presenting data to deliver information. This standard is valuable Master Data play an integrate role in the process of generation of data for regulatory decision making.

ISO 25012: Defines a general data quality model for data retained in a structured format within a computer system, typical for data considered for regulatory decision making. It provides a framework for establishing data quality requirements, data quality measures, and a plan to perform data quality evaluations. It could be used across the entire life cycle from data collection or generation, management and processing, publishing, aggregation, and consumption and to evaluate the compliance of data with regulations. An example of implementation of this standard is done by Statistics Finland, which is also in line with the European Interoperability Framework, the FAIR principles and the Code of Practice for Statistics [13].

ISO 13485: Specifies requirements for a QMS for an organisation to design, build and obtain authorisation for medical devices that consistently meet customer and regulatory requirements. As with all QMS, this standard focusses on the quality of the product, and affects DQ as they are relevant for the design, development, production, and use of the device.

8.5. Notes on ALCOA+

ALCOA+ is a framework for data integrity used across the pharmaceutical industry in areas such as research, manufacturing, testing and supply chain. It postulates a set of principles that data and its documentation should comply to. In the specifics, data should be Attributable, Legible, Contemporaneous, Original, Accurate (ALCOA). The + refers to the following attributes: Consistent, Enduring, Available, Traceable. More information on these principles is available in the Guideline on computerised systems and electronic data [21] and in [22].

In relation to this Framework, ALCOA+ provides recommendations that focus on foundational determinants and that affect primarily the Reliability, but also the Extensiveness, Coherence and Timeliness dimensions.

When considering the ALCOA+ principles, they can be closely aligned to the dimensions of the present DQF, with the caveat that the ALCOA definitions are more focused and operational. For instance, the ALCOA definition of "Accurate" is expressing a set of characteristics (e.g., verifiable coding processes, validated data transfer) that should be in place so that "data should be an accurate representation of the observation made", that is how reliability (and more precisely Accuracy) is defined in this DQF. Other principles such as "Legible" and "Original" also falls under the Reliability dimension (as they answer the question "is data reflecting reality?", but they are not explicitly articulated in this Framework as they are a pre-condition to a regulatory submission.

For suitable use cases, ALCOA+ compliant specifications can enable level 2 (formalised) and above maturity levels for foundational determinants.

8.6. Notes on implementation of DQ controls

There are different possible implementations of data quality controls (i.e., testing).

If the **true facts** the data are representing are known and accessible, data can be tested using validation vs the source records containing these facts (see framework of reference session above). However, validation can be costly and time consuming, and often requires the use of adjudicators if the

facts are not available in machine-readable structured form. Alternatively, data can be tested via intrinsic plausibility metrics, and specifically by assessing the dataset respect to (See figure 7):

Other data in the same dataset: The test would detect logical or factual contradictions (e.g., embedding background knowledge on relations between entities and events). For example, the timing of a causal effect must occur after its causing intervention, or a female patients cannot have observations only occurring in males.

External reference ranges (or Gold Standards)**:** Some measured quantity cannot exceed a certain magnitude, such as a blood pressure of 1000/500 mmHg.

Plausible trends: certain data can be valid when observed individually, but the collective trend of all data of a kind should follow expected distributions or trends. For example, the incidence of a disease is unlikely to grow drastically from 2% to 80% in a population from one year to another, or the exposed cell line in an experiment cannot show less effect than the unexposed comparator. In this case data are assessed with respect to background knowledge on typical characteristics of data.

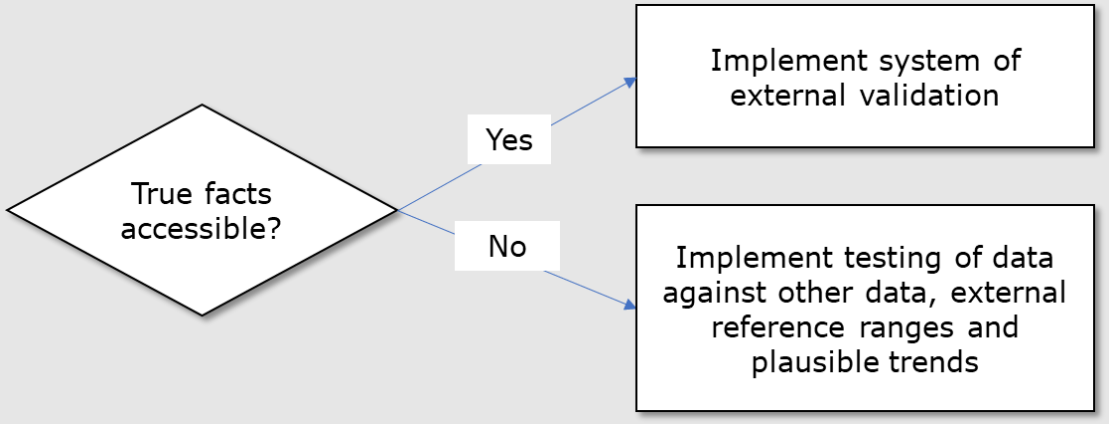

Figure 7 – Overview of external reference ranges

9. Glossary

This glossary addresses the main terms and definitions that have been used in this Data Quality Framework for regulatory decision making.

Definitions	Explanation
Data accessibility	The ability of data to be accessible for public use in terms of discoverability, exportability, and usability.
Data conciseness	The characteristic of data to be expressed in a compact representation. Sometimes also defined as the characteristic of data to include only essential, and not spurious, information.
Data immutability	Data immutability is the concept that data is never deleted or altered. Once some data is "stated" (e.g.: entered in a database), it can only be augmented (eventually with additional information meant to invalidate or supersede previous data) but never remoted. In other words, data that has been entered in a system (and on which some other data or actions may depend) cannot be changed without explicitly mentioning of a new state of the information and maintaining the knowledge of the previous state.
Data integrity	Data integrity refers to the maintenance and assurance of data reliability and consistency over time, encompassing the whole data life cycle. It is a broader concept than Data Quality, whose precise definition varies across contexts, extending from physical to logical aspects of data processing and storage.
Data quality metrics	DQ metrics can be defined as indicators that can be applied to a data source to derive assessments of one of more quality dimensions.
Data quality	Data quality is defined as fitness for purpose for users' needs in relation to health research, policy making, and regulation and that the data reflect the reality, which they aim to represent. Data quality is relative to the research question and does not address the question on what level is the quality measured e.g., variable, data source or institutional level. These aspects are addressed in the data quality determinants and dimensions of data quality.
Data quality determinants	What contributes to data quality or its characterisation. In this Framework determinants are classified into three categories: - Foundational determinants: what affects the quality of a dataset, being external to the dataset itself (e.g., systems and processes that generate data). - Intrinsic determinants: what can be derived in terms of quality for a dataset itself, without information on how the data came to or its intended usage. - Question specific determinants: what affects the assessment of a dataset quality, that strictly depends on the dataset intended or actual usage.

Definitions	Explanation
Data quality dimensions	Data quality aspects are partitioned into different group that answers different questions about data. Such partitions are called "dimensions". This Framework distinguishes five dimensions that can be divided further into sub-dimensions: 1) extensiveness, 2) coherence, 3) timeliness, 4) relevance, and 5) reliability.
Data quality framework	A Data Quality Framework provides a set of definitions, guidelines, and recommendation to assess and govern data quality. The framework here presented addresses a wide range of data sources for the purpose of characterising, assessing, and assuring data quality for regulatory decision making.
Entity	An entity is a collection of similar values that belong to a specific variable (e.g., weight). This is also referred to as row level.
Healthcare data	Medical data gathered from different settings containing various clinical measurements of specific populations. In most cases this is electronically stored data known as electronic health data.
Fit for purpose	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.
Foundational determinant	A data quality determinant that covers aspects related to the generation of data. It affects the quality of data, but it's not part of the data themselves e.g., software systems, training, audit processes. It can be seen as data generation specific.
Intrinsic determinant	A data quality determinant that covers aspects that are inherent to a given dataset e.g., the formatting of the data. This can be seen as a dataset specific determinant.
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 organization 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.
MDM	Master data management, a system that helps synthesise data from different systems and secure and clean it (eliminates duplications etc) to deal with the right information.
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

Definitions	Explanation
	the purpose of allowing data to be discoverable and checked for fitness for purpose, without revealing the data themselves.
Primary use of data	Primary use of (electronic) health data is the processing of personal health data for the provision of health services to assess, maintain or restore the state of health of the person it belongs it, including the prescription, dispensation and provision of medicinal products and medical devices, as well as for relevant social security, administrative or reimbursement services.
Secondary use of data	Secondary use of (electronic) health data is the processing of health data for other purposes rather than primary use such as national statistics, education/teaching, scientific research etc. The data used may include personal health data initially collected in the context of primary use, but also electronic health data collected for the purpose of secondary use.

10. References

1. European Medicines Regulatory Network Data Standardisation Strategy. December 16th, 2021. Available from: https://www.ema.europa.eu/en/documents/other/european-medicines-regulatory-network-data-standardisation-strategy_en.pdf.
2. European Health Data Space Data Quality Framework, Deliverable 6.1 of TEHDAS EU 3rd Health Program (GA: 101035467). May 18th, 2022, accessed at https://tehdas.eu/results/tehdas-develops-data-quality-recommendations
3. ISO 9001:2015 Quality Management System, accessed November 18th, 2022. https://www.iso.org/standard/62085.html.
4. Kahn, M.G., et al., A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC), 2016. **4**(1): p. 1244.
5. Healthcare Data Quality: A 4-Level Actionable Framework 2020 [September 5th, 2022]. Available from: https://www.healthcatalyst.com/insights/healthcare-data-quality-4-level-actionable-framework.
6. Schmidt, C.O., et al., Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med Res Methodol, 2021. **21**(1): p. 63.
7. Sentinel QA Program. Quality Assurance - Sentinel Version Control System. [February 14th, 2022]. Available from: https://dev.sentinelsystem.org/projects/QA/repos/qa_package/browse.
8. 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.
9. The National Patient-Centered Clinical Research Network. PCORnet - Data Quality Framework. [February 18th, 2022]. Available from: https://pcornet.org/data.
10. Duke-Margolis Center for Health Policy. Characterising RWD Quality and Relevance for Regulatory Purposes, 2018. Available from: https://healthpolicy.duke.edu/sites/default/files/2020-03/characterising_rwd.pdf.
11. Duke-Margolis Center for Health Policy. Determining Real-World Data's Fitness for Use and the Role of Reliability, 2019. Available from: https://healthpolicy.duke.edu/sites/default/files/2019-11/rwd_reliability.pdf.
12. Data Utility Framework. Available from: https://www.hdruk.ac.uk/wp-content/uploads/2020/11/201105-Updates-to-the-Data-Utility-Framework-v2.pdf.
13. Statistics Finland, Data Quality Framework, National data quality criteria and indicators. Available from: https://stat.fi/org/tiedon-laatukehikko/tiedon-laatukriteerit_en.html.
14. Cave, A., X. Kurz, and P. Arlett, Real-World Data for Regulatory Decision Making: Challenges and Possible Solutions for Europe. Clin Pharmacol Ther, 2019. **106**(1): p. 36-39.
15. Big Data Steering Group, Big Data Workplan 2022-2025. Available from: https://www.ema.europa.eu/en/news/big-data-use-public-health-publication-big-data-steering-group-workplan-2022-25.
16. FAIR. Available from: https://www.go-fair.org/fair-principles/.
17. Good Practice Guide for the use of the Metadata Catalogue of Real-World Data Sources, EMA/787647/2022. Available from: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/good-practice-guide-use-metadata-catalogue-real-world-data-sources_en.pdf.
18. Wang, S.V. and S. Schneeweiss, A Framework for Visualizing Study Designs and Data Observability in Electronic Health Record Data. Clin Epidemiol, 2022. **14**: p. 601-608.
19. European Commission. Can we use data for another purpose? [Internet, cited 27 Sept 2022] 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.
20. American Society for Quality. What is a Quality Management System (QMS)? | ASQ. Available from: https://asq.org/quality-resources/quality-management-system.
21. Guideline on computerised systems and electronic data in clinical trials. Available from: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/draft-guideline-computerised-systems-electronic-data-clinical-trials_en.pdf.
22. Data Integrity and Compliance with CGMP, guidance for Industry. Available from: https://www.fda.gov/files/drugs/published/Data-Integrity-and-Compliance-With-Current-Good-Manufacturing-Practice-Guidance-for-Industry.pdf.